Annals of Internal Medicine RESEARCH AND REPORTING METHODS

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

Karel G.M. Moons, PhD*; Robert F. Wolff, MD*; 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

Prediction models in health care use predictors to estimate for an individual the probability that a condition or disease is already present (diagnostic model) or will occur in the future (prognostic model).

Publications on prediction models have become more common in recent years, and competing prediction models frequently exist for the same outcome or target population. Health care providers, guideline developers, and policymakers are often unsure which model to use or recommend, and in which persons or settings. Hence, systematic reviews of these studies are increasingly demanded, required, and performed.

A key part of a systematic review of prediction models is examination of risk of bias and applicability to the intended population and setting. To help reviewers with this process, the authors developed PROBAST (Prediction model Risk Of Bias ASsessment Tool) for studies developing, validating, or updating (for example, extending) prediction models, both diagnostic and prognostic.

PROBAST was developed through a consensus process involving a group of experts in the field. It includes 20 signaling questions across 4 domains (participants, predictors, outcome, and analysis). This explanation and elaboration document describes the rationale for including each domain and signaling question and guides researchers, reviewers, readers, and guideline developers in how to use them to assess risk of bias and applicability concerns. All concepts are illustrated with published examples across different topics. The latest version of the PROBAST checklist, accompanying documents, and filled-in examples can be downloaded from www.probast.org.

Ann Intern Med. 2019;170:W1-W33. doi:10.7326/M18-1377 **Annals.org**

For author affiliations, see end of text.

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

Prediction models in health care aim to predict for an individual whether a particular outcome, such as disease, is present (diagnostic models) or whether it will occur in the future (prognostic models) (1–6). Diagnostic models can be used to refer patients for further testing, to initiate treatment, or to inform patients. Prognostic models can be used to aid decisions about preventive lifestyle changes, therapeutic interventions, or monitoring strategies or to stratify risk in randomized trial design and analysis (7, 8). Potential users of prediction models include health care professionals, policymakers, guideline developers, patients, and the general public.

The medical literature contains thousands of studies developing and validating prediction models and often has numerous prediction models for the same target population and outcomes. For example, more than 60 models address breast cancer prognosis (9), more than 250 exist in obstetrics (10), and nearly 800 predict outcomes in patients with cardiovascular disease (11). This proliferation of prediction models will increase further with the growth of personalized or precision medicine.

Systematic reviews are considered the most reliable form of evidence when addressing randomized therapeutic studies and studies of diagnostic test accuracy (12). In the era of personalized and precision medicine, interest in systematic reviews of prediction model studies is rapidly growing, as exemplified by the formation of the Cochrane Prognosis Methods Group to support systematic reviews of prognosis, including prediction model studies (13, 14). Guidance to facilitate systematic reviews of prediction models has been developed (Table 1), including for search strategies (15, 41–43), formulation of the review question (16, 17), data extraction (16), and meta-analysis (17, 22–25, 40, 44, 45).

Assessment of risk of bias (ROB) is an essential step in any systematic review. Shortcomings in study design, conduct, and analysis can result in study estimates being "at ROB"—that is, at risk of results being flawed or distorted. When interpreting results from a systematic review, readers can draw stronger conclusions from a review based on primary studies at low ROB than from one based on studies at high or unclear ROB (46). Identifying the studies most relevant to the settings and populations targeted in the review (based on the applicability of primary studies to the review question) is also important. We therefore developed PROBAST (Prediction model Risk Of Bias ASsessment Tool) to address the lack of suitable tools designed specifically to assess ROB and applicability of primary prediction model studies.

PROBAST consists of 4 domains containing 20 signaling questions to facilitate ROB assessment (39). The structure and rating system are similar to those in tools designed to assess ROB in randomized trials (revised Cochrane ROB Tool [ROB 2.0]), diagnostic accuracy studies (QUADAS-2 [Quality Assessment of Diagnostic Accuracy Studies 2]), and systematic reviews (ROBIS) (37, 47, 48). Although PROBAST was designed for use in systematic reviews of prediction model studies, it can also be used as a general tool for critical appraisal of (primary) prediction model studies.

See also:

Related article 51

Table 1. Guidance on Conducting Systematic Reviews of Prediction Model Studies

Task	Guidance
Reporting of primary study	Transparent reporting of studies on prediction models for prognosis and diagnosis (TRIPOD) (7, 8)
Defining review question and developing criteria for including studies*	Guidance for defining review question and design of the review of prognosis studies (CHARMS) (16, 17), see also Table 4 guidance for protocol for DTA reviews (18, 19)
Searching for studies*	Search filters for prediction studies (15, 41–43) https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/filters-to-identify-studies-about-prognosis Search for DTA studies (20)
Selecting studies and extracting data*	Guidance and checklist for data extraction and critical appraisal of prediction model studies (CHARMS) (16) Guidance for DTA studies (19, 21)
Assessing risk of bias and applicability in included studies*	PROBAST (39)
Analyzing data and undertaking meta-analyses*	Meta-analysis of prediction model studies (17, 22-25, 40, 44, 45) Meta-analysis of DTA studies (26-33)
Interpreting results and drawing conclusions*	PROBAST (39) and guidance for interpretation of results of reviews of prediction model studies (17, 22-24, 40) Guidance for interpretation of DTA reviews (19)
Reporting of systematic reviews*	Transparent reporting of systematic reviews and meta-analysis (PRISMA and PRISMA-DTA) (34-36)
Assessing risk of bias of systematic reviews	ROBIS (37)

CHARMS = CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Model Studies; DTA = diagnostic test accuracy; PRISMA = Preferred Reporting Items for Systematic reviews and Meta-Analyses; PROBAST = Prediction model Risk Of Bias ASsessment Tool; ROBIS = Risk of Bias in Systematic Reviews; TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.
* Step in line with the general methods for Cochrane reviews (38).

Here, we describe the rationale behind the domains and signaling questions, how to use them, and how to reach domain-level and overall judgments about ROB and applicability of primary studies to a review question. At the Web site (www.probast.org), 5 filled-in examples from across the medical field illustrate these processes. Because this is an area of active research, the tool, examples, and accompanying guidance will be updated when needed, and the latest version of PROBAST should always be downloaded from the Web site.

FOCUS OF PROBAST

PROBAST is designed to assess primary studies that develop, validate, or update (for example, extend) multivariable prediction models for diagnosis or prognosis (**Boxes 1** and **2**). A multivariable prediction model is defined as any combination or equation of 2 or more predictors (such as age, sex, symptoms, signs, disease stage, or biomarkers) for estimating for an individual the probability or risk of having (diagnosis) or developing (prognosis) a particular outcome (1, 4, 6–8, 49, 50). Other names for prediction model include risk prediction model, predictive model, prediction index or rule, and risk score (1, 3–8, 49–51).

Diagnostic and Prognostic Models

Diagnostic prediction models estimate the probability that a certain outcome, the "target condition," is currently present. Diagnostic prediction model studies typically include individuals who are suspected—but not yet known—to have the target condition.

Prognostic prediction models estimate the probability that a future outcome or event will occur, such as death, disease recurrence, disease complication, or therapy response. The time period of prediction can vary from hours (for example, preoperatively predicting postoperative nausea and vomiting) to years (for example, predicting lifelong risk for a coronary event). Although many prognostic models enroll patients with an established diagnosis, this does not have to be the starting point, as seen in models for predicting development of diabetes in pregnant women (52) or of osteoporotic fractures in the general population (53). Consistent with the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement (7, 8), PROBAST thus broadly defines prognostic models as those predicting a future outcome in persons at risk for that outcome.

Diagnostic and prognostic model studies often use different terms for predictors and outcomes (**Box 2**). The cancer literature frequently distinguishes between prognostic and predictive models, such that predictive models identify individuals with differential treatment effects (54). These types of (predictive) models are outside the scope of this article.

Types of Predictors, Outcomes, and Modeling Techniques

PROBAST can be used to assess any type of diagnostic or prognostic prediction model aimed at individualized predictions regardless of the predictors used; outcomes being predicted; or methods used to develop, validate, or update (for example, extend) the model.

Predictors range from demographic characteristics, medical history, and physical examination results; to imaging results, electrophysiology, blood, urine, or tissue measurements, and disease stages or characteristics; to results from "omics" and other new biological measurements. Predictors are also called covariates, risk indicators, prognostic factors, determinants, index test results, or independent variables (4, 6–8, 49, 50, 55, 56, 57).

PROBAST distinguishes between candidate predictors and predictors included in the final model (57).

Candidate predictors are variables considered potentially predictive of the outcome presence (diagnosis) or occurrence (prognosis)–that is, all those evaluated in the study regardless of whether they are included in the final multivariable model.

PROBAST primarily addresses prediction models for binary and time-to-event outcomes because these are the most common in medicine. However, the tool can also be used to assess models predicting nonbinary outcomes, such as continuous scores (for example, pain scores or cholesterol levels) or categorical outcomes (for example, the Glasgow Coma Scale). Almost all PROBAST signaling questions apply equally to continuous and categorical outcomes, except questions addressing number of outcome events per predictor and certain measures of model performance (such as the c-statistic), which are not relevant to continuous outcomes.

Prediction models usually involve regression modeling techniques, such as logistic regression or survival models. Prediction models may also be developed or validated using nonregression techniques, such as neural networks, random forests, or support vector machines. As the use of routine care (and "big") data increases, additional modeling techniques are becoming more common, including machine and artificial learning models. The main differences between studies using regression and other types of prediction modeling include the methods of data analysis; nonregression development models can often have greater risks of overfitting when data are sparse, and the potential lack of transparency can affect the applicability and usability of the models. In the section on tailoring PROBAST with additional signaling questions, we provide guidance about how PROBAST can be adapted to address other types of outcomes and modeling techniques.

Types of Review Questions

PROBAST can be used to assess different types of systematic review questions. For some review questions, all prediction model studies are relevant (including both development and validation), but for other questions only validation studies are relevant. Box 3 gives examples of potential review questions for both prognostic and diagnostic prediction models where PROBAST is applicable. CHARMS (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) and Table 2 provide explicit guidance on how to frame a focused question for reviews of prediction model studies (16, 17).

Types of Prediction Model Studies

PROBAST addresses studies on multivariable models intended to make diagnostic and prognostic predictions in individuals–that is, *individualized predictions* (Box 1)–including studies on 1) developing new prediction models, 2) developing and validating the same prediction models, 3) validating existing prediction models, 4) developing new prediction models versus validating existing models, 5) updating (for example, adjusting model coefficients) or extending (for example, adding new predictors to) existing prediction models, and 6) combinations of these purposes.

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 (8, 16).

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

Diagnostic prediction models aim to estimate the probability that a target condition or disease 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 or 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.

PROBAST is not designed to assess predictor finding studies, where the aim of multivariable modeling is to identify predictors associated with the outcome rather than to develop a model for individualized predictions (16, 68, 69). The QUIPS (Quality in Prognosis Studies) tool has been developed for assessment of bias in these studies (70).

PROBAST is also not suitable for assessing comparative studies that quantify the impact on participants' health outcomes of using a prediction model (as part of a complex intervention) compared with not using a model or an alternative model. Such comparative model impact studies use either randomized or nonrandomized designs (8, 55, 71–74) and appropriate ROB tools for randomized studies (47) or nonrandomized studies (75).

Another ROB tool, QUADAS-2, has been developed for studies of diagnostic test accuracy (48). However, it should be noted that some diagnostic test accuracy studies include a diagnostic prediction model rather than a diagnostic test. In these cases, use of PROBAST should be considered where appropriate.

ROB AND APPLICABILITY

ROB

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 there is limited *empirical* evidence showing the most important sources of bias. We define ROB to occur when shortcomings in study design, conduct, or analysis could lead to systematically distorted estimates of a model's predictive performance. Model predictive performance is typically evaluated using measures of calibration and discrimination, and sometimes (notably in diagnostic model studies) of classification (Box 4) (8). 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.

Applicability

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. For example, such concerns may arise when participants in the prediction model study are from a different medical setting from the targeted population defined in the review question (Table 2). A prediction model developed in secondary care may have different discrimination and calibration in primary care because patients in hospital settings typically have more severe disease than those in primary care (71, 85).

When participants, predictors, and outcomes of the primary studies directly match a review question, small concerns about applicability of the studies will likely remain. However, the inclusion criteria for systematic reviews are typically broader than the precise focus of the review question.

Bias and applicability concerns should not be confused here with heterogeneity in predictive performance of a particular model across different validation studies, which may result, for example, from case mix or varying disease severity (17, 40, 44). Variation in the performance of a particular model across validations can be reported with relevant prediction intervals as part of the investigation of heterogeneity using meta-analysis methods (17, 40).

For example, in a review and meta-analysis of a specific single prediction model that includes all validation studies of that model, ROB and applicability assessments should be supplemented with an investigation of heterogeneity in the reported predictive performance of that model across the validation studies. The predictive performance of a specific model validated in other studies is expected to differ because of differences in (for example) participant characteristics, health care setting, geographic location, or calendar time period. This does not mean that there is ROB *within* the primary validation study or that there are concerns regarding applicability; it merely reflects expected variation in the predictive performance of a specific model across studies. Potential sources of heterogeneity between studies can be investigated using meta-analysis or presentation stratified by characteristics that differ across studies (17, 40, 44).

APPLYING PROBAST

PROBAST consists of 4 steps (Table 3). A PROBAST assessment should be completed for each distinct model that is relevant to the systematic review question. We use various examples to illustrate key issues relating to ROB and applicability (Table 4<(85)>). These examples address both diagnostic and prognostic models; focus on different medical areas, study designs, and predictor and outcome types; and include development and validation studies. Assessments of these examples are available at www.probast.org.

Step 1: Specify Your Systematic Review Questions

First, reviewers need to specify their review question in terms of the intended use of the prediction model, targeted participants, predictors used in modeling, and outcomes to be predicted. Structured reporting of these elements facilitates assessment of applicability. Existing guidance (the CHARMS checklist) can help reviewers define a clear and focused review question (16), as summarized in Table 2.

Step 1 is completed once per systematic review; Table 5 provides an example.

Step 2: Classify the Type of Prediction Model Evaluation

In Step 2, the type of prediction model evaluation is identified to link to the relevant signaling questions in PROBAST. When a single publication reports both development and validation (Box 1)—or both validation and adjustment or extension—of a particular model, each will be assessed separately. A model extension, where new predictors are added to an existing model, would be assessed as new model development.

Step 2 is completed once for each prediction model assessed for the review; Table 6 provides an example.

Step 3: Assess ROB and Applicability

Assessing ROB

PROBAST provides a structured approach to identify potential ROB, based on 4 domains with signaling questions. Signaling questions are *factual* and are answered as

Box 3. Examples of systematic review questions for which PROBAST is suitable.

A specific target population: **Review of all models developed or validated for predicting the risk of incident type 2 diabetes in the general population (58). Review of all prognostic models developed or validated for use in patients diagnosed with acute stroke (59).**

A specific outcome: **Review of all diagnostic models developed or validated for detecting venous thromboembolism regardless the type of patients (60). Review of all prognostic models developed or validated for predicting loss of daily activity, regardless the type of patients (61).**

A particular clinical field: **Review of all prognostic models developed or validated in reproductive medicine (62). Review of all prognostic models developed or validated in acute care of traumatic brain injury (63).**

A specific prediction model: **Review of the predictive performance of the EuroSCORE (a model to predict operative mortality following cardiac surgery) as found across all external validation studies of the EuroSCORE model (64). Review to compare the predictive performance of various prognostic models for developing cardiovascular disease in middle-aged individuals in the general populations, across all validation studies of these models (65).**

A specific predictor: **Meta-analysis of the added predictive value of C-reactive protein when added to the Framingham risk model (66). Meta-analysis of the added predictive value of carotid artery imaging to an existing cardiovascular risk prediction model (67).**

There are various different questions that systematic reviews of prediction models may address. The following are examples of different types of reviews in which PROBAST can be applied. EuroSCORE = European System for Cardiac Operative Risk Evaluation; PROBAST = Prediction model Risk Of Bias ASsessment Tool.

Table 2. PICOTS*

Item	Comments
1. Population	Define the target population in which the prediction model(s) under review will be used.
2. Index	Define the prediction model(s) under review.
3. Comparator	If applicable, define whether other prediction models are reviewed and compared with the index model.
4. Outcome(s)	Define the outcome(s) to be predicted by the model(s) under review.
5. Timing	Define at what moment or time point (e.g., in the patient work-up) the prediction model(s) under review are to be used, and over what time period the outcome(s) are predicted (the latter in case of prognostic models).
6. Setting	Define the intended clinical setting and intended use of the prediction model(s) under review.

PICOTS = population, index, comparator, outcomes, timing of prediction and of outcomes, and setting.
* Key items to guide the framing of the review aim as suggested in previous guidance (16, 17). PICOTS is a modification of the traditional PICO system used in systematic reviews of therapeutic intervention studies, by adding timing (the time point of using the prediction model and the time period of the prediction) and clinical setting (17).

yes (Y), probably yes (PY), no (N), probably no (PN), or no information (NI). All signaling questions are phrased so that "yes" indicates low ROB and "no" high ROB. The ratings PY and PN are included to allow judgments to be made when information is *not sufficient* to be confident in answering Y or N. To conform with other ROB tools, responses of Y are intended to have similar implications to responses of PY (and likewise for N and PN) but allow a distinction between something that is known and something that is likely to be the case (37, 47, 75). Assessors should use NI only when there is truly no information to answer a signaling question.

The answers to these signaling questions assist reviewers in judging the overall ROB for each domain. A domain where all signaling questions are answered as Y or PY should be judged as "low ROB." An answer of N or PN on 1 or more questions flags the potential for bias, whereas NI indicates insufficient information. This does not mean that bias is definitely present. For example, in a prognostic study where predictors were clearly determined before event occurrence and measurement but the report does not state whether predictor measurements were blinded to outcome occurrence, this question (see signaling question 2.3) is factually rated as NI. However, the assessor may still judge the overall ROB of this domain to be low, because it can be inferred that predictors were measured a long time before the outcome occurred. When judging ROB for a particular domain, reviewers thus need to use their judgment to determine whether issues identified by the signaling questions are likely to have introduced bias into the model development or validation.

Assessing Concerns Regarding Applicability

Applicability of a primary study to the review question is assessed for the first 3 domains using information reported in Table 5 (the review question) and Tables 7 to 9. The analysis domain relates to limitations with the data or how the analysis was performed, which are not related to the review question, so this domain has no applicability assessment. The degree of applicability is rated as "low," "high," or "unclear" concern. The "unclear" category should be used only when reported information is insufficient.

If the review question and primary study are a good match, concern regarding applicability is likely low. A review may address a focused question while study inclusion criteria are broader.

Support for Judgment and Rationale for Rating

To improve the transparency of the assessment process, PROBAST includes 2 text boxes for each domain. The first box allows reviewers to record support for judgment–that is, information that was used to answer the ROB signaling questions or inform the applicability assessment for that domain. Text can be either summarized or copied and pasted directly from the article being assessed. The second text box is "rationale of . . . rating," which allows reviewers to record the reason for judging the model to have high, low, or unclear ROB or high, low, or unclear concern regarding applicability. For example, if a domain is judged to be at high ROB, the reviewers can summarize which study features led to the rating. Or if a domain is rated at low ROB despite 1 or more signaling questions answered as N, PN, or NI, this box can be used to explain why issues identified by the questions are not likely to have introduced bias into the study.

Further guidance and examples are provided in the relevant domain sections and Tables 7 to 10. The latest updated versions of guidance can be downloaded from www.probast.org.

Domain 1: Participants

This domain covers potential sources of bias and applicability concerns related to the data sources used and how participants were selected for enrollment into the study. In the support for judgment box, reviewers should describe the data sources–for example, cohort study, randomized study, or routine care registry–and the criteria for participant selection in the primary study.

ROB. Two signaling questions facilitate an ROB judgment for this domain (Table 7).

1.1 Were appropriate data sources used, e.g., cohort, randomized controlled trial, or nested case–control study data?

Numerous data sources and study designs can be used in prediction model studies.

Prognostic model studies. Prognostic model studies are at low ROB when based on a prospective longitudinal cohort design, where methods tend to be defined and consistently applied for participant inclusion and exclusion criteria, predictor assessment, and outcome determination across a predefined follow-up (1). Using prespecified and consistent methods ensures that participant data are systematically and validly recorded.

Model development and validation studies have higher potential for ROB when participant data are from existing sources, such as existing cohort studies or routine care registries, because data are often col-

lected for a purpose other than development, validation, or updating of prediction models, and are also often without a protocol. In routine care registries, data relating to inclusion and exclusion criteria are often inconsistently measured and recorded (44, 90). For example, in reference to the Clinical Practice Research Datalink, Herrett and colleagues (90) state that "[t]he quality of primary care data is variable because data are entered by [general practitioners] during routine consultations, not for the purpose of research. Researchers must therefore undertake comprehensive data quality checks before undertaking a study."

Data from 1 or more groups of randomized intervention trials can also be used for prognostic model development, validation, or updating. However, the randomized treatments may need to be included as separate predictors to account for any treatment effects, because effective treatments are predictors of the outcome (91, 92). In addition, randomized trials usually have more restricted inclusion criteria, which typically lead to narrower predictor distributions ("smaller case mix"). Models developed or validated using data with narrower predictor distributions tend to show lower discriminative ability than those developed or validated using data sources with more broadly distributed predictors (93–96). This is because in the former, a model's range of predicted probabilities—and therefore its discriminative ability—is smaller.

In case–cohort or nested case–control studies, participants with the outcome (case patients) and those without the outcome (noncase or control patients) are sampled from preexisting, well-described cohorts or routine care registries of known size. These studies can be considered to be at low ROB as long as researchers appropriately adjust for the original cohort or registry outcome frequency in the analysis (see signaling question 4.6) (57, 97–100). If they do not, case–cohort and nested case–control studies are at high ROB for prediction model purposes. For example, in logistic prediction models, reweighting the control and case samples by the inverse sampling fraction (from the original cohort or registry) enables correct estimation of baseline risk, which allows researchers to obtain corrected absolute predicted probabilities and model calibration measures (97–100). Case–control studies in which case and control participants are not sampled from a prespeci-

Box 4. Prediction model performance measures.

Calibration **reflects the agreement between predictions from the model and observed outcomes. Calibration is preferably reported graphically, with observed risks plotted on the *y*-axis against predicted risks on the *x*-axis. This plot is commonly done by tenths of the predicted risk and is preferably augmented by a smoothed (lowess) line over the entire predicted probability range. This is possible both for prediction models developed by logistic regression (59, 76, 77) and by survival modeling (78, 79). The calibration plot displays the direction and magnitude of any model miscalibration across the entire predicted probability range, which can be combined with estimates of the calibration slope and intercept (79, 80). Calibration is frequently assessed by calculating the Hosmer–Lemeshow goodness-of-fit test; however, this test has limited suitability to evaluate poor calibration and is sensitive to the numbers of groups and sample size: The test is often nonsignificant for small data sets and nearly always significant for large data sets. Studies reporting only the Hosmer–Lemeshow test with no calibration plot or a table comparing the predicted versus observed outcome frequencies provide no useful information on the accuracy of the predicted risks (see signaling question 4.7).**

Discrimination **refers to the ability of a prediction model to distinguish between individuals who do or do not have (diagnosis) or develop (prognosis) the outcome. The most general and widely reported measure of discrimination, for both logistic and survival models, is the concordance index (c-index), which is equivalent to the area under the receiver-operating characteristic curve for logistic regression models.**

Calibration **and *discrimination*** **measures should take into account the type of outcome being predicted. For survival models, researchers should appropriately account for time-to-event and censoring using, e.g., Harrell's c-index or the D statistic (50, 81, 82).**

Many other model predictive performance measures are available, including measures to express model classification abilities (e.g., sensitivity and specificity) and reclassification parameters (e.g., the Net Reclassification Index) (80). These measures can be estimated after introducing 1 (or more) thresholds in the range of the model-predicted probabilities. Classification measures are frequently used in diagnostic test accuracy studies but less in prediction model studies. Categorization of the predicted probabilities for the estimation of a model classification measures leads to loss of information, since the entire range of predicted probabilities of the model is not fully utilized. Using thresholds can allow discrimination to be reported at potentially clinically relevant thresholds as opposed to across all potential thresholds that may not be clinically important. However, introducing probability thresholds implies that the chosen threshold is relevant to clinical practice, which often is not the case since these thresholds are often data driven, yielding biased classification parameters (83). Authors should rather assess these measures based on the general principles of prespecifying (probability) thresholds (see also signaling question 4.2) to avoid multiple testing of thresholds and potential selective reporting of thresholds based on the data itself.

There are many other measures of model predictive performance, including net benefit measures and decision curve analysis, but these are not commonly reported for prediction modeling studies (84). Many of these measures provide a link between probability thresholds and false-positive and false-negative results.

All the above model performance measures, when estimated on the development data, are often optimistic due to overfitting or choosing optimal thresholds and should therefore be estimated using bootstrapping or cross-validation methods (see signaling question 4.8).

Table 3. 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 prediction 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.

fied and well-defined cohort or registry are at high ROB because the definition and number of the selected case and control participants relative to the source population is unclear. Accordingly, baseline risks or hazards and absolute outcome probabilities cannot be correctly adjusted for (57).

Diagnostic model studies. Diagnostic models predict the presence or absence of an outcome (target disease) at the same time point as the index tests or predictors are measured (**Box 2**). Accordingly, the design with the lowest ROB for diagnostic model studies is a cross-sectional study where a group (cohort) of participants is selected on the basis of certain symptoms or signs that make them suspected of having the target condition of interest. Subsequently, the predictors (index tests) and outcome (disease presence or absence) according to the reference standard are measured in all participants (101–104). Diagnostic studies using a cross-sectional design in which the presence of disease cannot be determined in all patients by a reference standard at 1 time point (for example, some participants with a potential malignant mass have no lesion on imaging that can be biopsied) require additional follow-up to establish whether the target condition was present when the index tests were done.

As with prognostic models, a diagnostic model using a nested case–control design can be at low ROB only if researchers adjust the case and control samples by the inverse sampling fractions (see signaling question 4.6) to obtain a correct estimate of the outcome prevalence in the original cohort (105–109). Similarly, use of a nonnested case–control design where case patients with advanced conditions and healthy control participants are overrepresented will lead to incorrect estimates of disease prevalence and overestimated diagnostic model performance (105–109).

Example. In Perel and colleagues' study (88), data for the development of the prognostic model came from a randomized trial (CRASH-2 [Clinical Randomisation of an Antifibrinolytic in Significant Haemorrhage 2]) combining data from 2 treatment groups. Because the authors included the allocated treatment as a predictor in the model development, this signaling question should be answered as Y.

Example. Aslibekyan and colleagues (86) used a nonnested case–control study but did not adjust their analyses by weighting the case and control samples by the inverse of the sampling fractions. Accordingly, this signaling question should be answered as N.

1.2 Were all inclusions and exclusions of participants appropriate?

Studies that inappropriately include or exclude participants may produce biased estimates of model predictive performance, because the model is based on a selected subgroup of participants that may not represent the intended target population.

Inappropriate inclusion results from including participants already known to have the outcome at the time of predictor measurement. For example, in a study developing a model to predict development of type 2 diabetes, some participants may already have type 2 diabetes if inclusion criteria included absence of diabetes based on self-reported data. Including participants who already have the disease will most likely result in a model with overestimated predictive performance.

Similarly, for a diagnostic model that aims to detect the presence or absence of pulmonary embolism in symptomatic patients, exclusion of patients with preexisting lung disease could be considered inappropriate. Such patients may be harder to diagnose with pulmonary embolism than those without preexisting disease; diagnostic accuracy may be overestimated if a model, after excluding these patients, is developed for use in all patients suspected of pulmonary embolism. Authors should then explicitly state that the developed model is applicable only to suspected pulmonary embolism in patients without preexisting lung disease.

Note that this signaling question is not asking about loss to follow-up after inclusion in the primary study (that is, it is not about inappropriate exclusions during the study); this is dealt with in domain 4. This signaling question is about participants who were inappropriately included or excluded during enrollment. Further, it is important to distinguish between selection bias imposed on a study population by restrictions in inclusion criteria and a study population with characteristics that may limit the applicability of the study to the review question (see Applicability).

In summary, the key issue is whether any inclusion or exclusion criteria, or the recruitment strategy, could have made the included study participants unrepresentative of the intended target population. Some ROB tools (such as QUADAS-2) have a signaling question that asks whether the study recruited a consecutive or random sample of patients. Because this is rarely achievable for any study, we have not included this question in PROBAST.

Example. Aslibekyan and colleagues (86) excluded all participants with a fatal myocardial infarction (MI) because they used a case–control design. Participants who had died of MI were excluded because retrospective self-reported data could not be collected from them. The prediction model for nonfatal MI was thus based on selected healthier participants, including only those who survived their MI (case participants) or did not develop an MI (control participants). This likely in-

troduced bias because the study participants represented a selected lower-risk sample of the original population of persons at risk for (any) MI. Stating that the developed prediction model predicts only nonfatal MI does not solve the issue because at the moment of prediction, identifying participants who will develop fatal MI is not possible. This signaling question should be answered as PN.

Rating the ROB for domain 1. Table 7 shows how the signaling questions should be answered and an overall judgment for domain 1 reached.

Applicability Applicability for this domain considers the extent to which the population included in the primary study matches the participants specified in the systematic review question (step 1; Table 5). Consider a review that aims to identify all model development and validation studies to diagnose bacterial conjunctivitis in symptomatic children. The review could specify inclusion criteria such that prediction model studies with both adults and children were eligible. Studies that included only children would likely receive a rating of low concern regarding applicability, whereas those in adults and children may have high concern regarding applicability.

The generalizability and thus applicability of prediction model studies based on randomized trial data needs careful consideration. Randomized trials tend to apply strict inclusion and exclusion criteria and may measure fewer predictors and outcomes, thus reducing the applicability of a model developed or validated from their data. In contrast, study characteristics, predictors, and outcomes have a wider distribution in data from routine care or health care registries, and thus prediction model studies using such registries for model development or validation tend to have higher generalizability.

Identifying when certain issues relating to a primary study are likely to introduce ROB and whether these raise concerns regarding applicability is often challenging. Applicability assessment is entirely dependent on the systematic review question (Tables 2 and 5). Consider the hypothetical pulmonary embolism example in signaling question 1.2, where reviewers might restrict the target population of their review to patients suspected of having pulmonary embolism who have no preexisting lung disease. For this target population, inclusion of patients with preexisting lung disease in a primary study would constitute an applicability concern but not necessarily an ROB. Similarly, consider a diagnostic model development study that included patients with a broad age range (18 to 90 years). This may not have introduced any bias into the primary study, but it may limit the applicability of the model if the systematic review question focuses on young adults (aged 18 to 30 years).

Finally, primary studies sometimes validate a model in participant data that were (for the researchers) *intentionally* different from the specific population used in the development study. For example, cardiovascular prediction models developed using a healthy general population have been validated in patients diagnosed with type 2 diabetes mellitus (110), and a model to diagnose deep venous thrombosis (DVT) that was developed in an emergency secondary care setting was validated in a primary care setting (85). In both cases, heterogeneity in model performance between the development study and the validation studies should be expected (40).

Domain 2: Predictors

This domain covers potential sources of bias and applicability concerns related to the definition and measurement of the predictors. Predictors are variables evaluated for their association with the outcome of interest; they are ultimately combined to form the prediction model.

In the support for judgment box, reviewers may list and describe how the predictors were defined, the time point of their assessment, and whether other information was available when the predictors were assessed.

Note that for systematic reviews focusing on a specific prediction model, it is sufficient to list and describe only the predictors in the model being validated.

ROB Three signaling questions facilitate an ROB judgment for this domain (Table 8).

2.1 Were predictors defined and assessed in a similar way for all participants?

Predictors should be defined and assessed in the same way for all study participants to reduce ROB. If different definitions and measurements across participants are used for the same predictor, differences in its associations with the outcome can be expected. For example, active bleeding in the lower digestive tract may be included as a possible predictor in a diagnostic model developed to detect colorectal cancer. This predictor "blood in feces" could be assessed in some participants on the basis of visible blood in the stool and in others using fecal occult blood testing. However, if these methods (with different minimum detection levels) are used interchangeably as a single predictor, "blood in feces" could introduce bias, especially if the choice of measurement method was based on prior tests or symptoms.

The potential for this bias is higher for predictors that involve subjective judgment, such as imaging test results, which introduce risk for studying the predictive ability of the observer rather than that of the predictors (1, 111–114). Where special skill or training is required, specifying who assessed the predictor (for example, experienced consultant vs. inexperienced trainee) may also be important.

Example. Perel and colleagues (88) assessed the following predictors, all of which were recorded on the entry form for the CRASH-2 randomized trial: demographic characteristics (age and sex), characteristics of the injury (type and time since injury), and physiologic variables (Glasgow Coma Scale score, systolic blood pressure, heart rate, respiratory rate, and central capillary refill time). Because the data used to develop the prediction model came from a substudy of a randomized trial and predictors were taken from the study entry form, it is likely—although not specifically described in the paper—that all predictors were defined and assessed in the same way for all participants. This signaling question would therefore be answered as PY. If data were derived from multiple sources (such as in routine care data registries, where different versions of

Table 4. Example Papers

Study Author, Year (Reference)	Topic Area	Type of Prediction Model Study		Data Source	Study Population
		Development/ Validation	Diagnostic/ Prognostic		
Aslibekyan, 2011 (86)	MI	Development + validation	Prognostic	Nonnested case–control study, population of central valley in Costa Rica (1994–2004)	First nonfatal acute MI cases versus control patients without a nonfatal MI
Han, 2014 (87)	Severe TBI	Validation	Prognostic	Cohort study, 1 hospital in Singapore (February 2006–December 2009)	Patients with severe TBI (GCS ≤8)
Oudega, 2005 (85)	DVT	Validation	Diagnostic	Prospective cross-sectional study, 110 primary care practices in the Netherlands (January 2002–March 2003)	Patients with symptoms or signs of DVT
Perel, 2012 (88)	Traumatic bleeding	Development + validation	Prognostic	Development: Randomized trial, 274 hospitals in 40 countries (no dates reported) Validation: Registry, 60% of trauma hospitals in England and Wales (2000–2008)	Development: Patients with trauma and significant bleeding or risk of significant bleeding within 8 h Validation: Patients with trauma and estimated blood loss ≥20%
Rietveld, 2004 (89)	Infectious conjunctivitis	Development	Diagnostic	Cohort study, 25 primary care centers in the Netherlands (September 1999–December 2002)	Patients with signs of infectious conjunctivitis (defined as red eye + [muco-] purulent discharge or glued eyelid)

CT = computed tomography; DVT = deep venous thrombosis; GCS = Glasgow Coma Scale; MI = myocardial infarction; TBI = traumatic brain injury.

the Glasgow Coma Scale or different definitions of injury type were likely used), this signaling question would be answered as PN.

2.2 Were predictor assessments made without knowledge of outcome data?

Risk of bias is low when predictor assessments are made without knowledge of the outcome status, often called "blinding" or "masking." Blinding of predictor assessment to outcome data is particularly important for predictors that involve subjective interpretation or judgment, such as those based on imaging, histology, history, or physical examination. Lack of blinding increases risk for incorporating outcome information into predictor assessments, which likely increases their association and leads to biased, inflated estimates of model performance (1, 111–119).

Blinding of predictor assessors to outcome information occurs naturally in prognostic studies that use a prospective cohort design when prognostic predictors are assessed before the outcome occurs. This bias is more likely in studies that retrospectively record predictors (vulnerable to recall bias) or in cross-sectional studies, such as diagnostic model studies, where predictors and outcomes are assessed within a similar time frame (1, 111–120).

Most prediction model studies do not report information on blinding of predictor assessment to outcome data (121, 122). In prognostic studies, this signaling question should thus be answered as NI (**Table 8**). However, the domain can still be rated as low ROB in the overall ROB assessment because if predictors were measured and reported a long time before the outcome occurred, their measurement can be inferred to be "blinded to the outcome." Note that even in prognostic studies, predictors may sometimes be assessed after outcome information has been collected–for instance, when predictors are collected from reinterpretation of stored imaging information or when a retrospective follow-up design is used. An example is the reuse of frozen tissue or tumor samples to measure novel predictors (biomarkers); such samples will already be linked to participant follow-up information, so novel predictor measurement may happen after outcome occurrence and may not be blinded to outcome information.

Example. Oudega and colleagues (85) stated, "After informed consent was obtained, the primary care physician systematically documented information on the patient's history and physical examination by using a standard form on which the items and possible answers were specified. Patient history included sex, presence of previous DVT, family history of DVT, history of cancer (active cancer in the past 6 months), immobilization for more than 3 days, recent surgery (within the past 4 weeks), and duration of the 3 main symptoms (a painful, red, or swollen leg). Physical examination included the presence of tenderness along the deep venous system, distention of collateral superficial (nonvaricose) veins, pitting edema, swelling of the affected limb, and a difference between the circumference of the 2 calves. . . . After history taking and physical examination, all patients were referred to the hospital for D-dimer testing and leg ultrasonography" (85). Because the study reported that history and clinical information (that is, the predictors) for all participants were collected before D-dimer testing, and the assessments were therefore also blinded to the outcome, this signaling question should be answered as Y.

2.3 Are all predictors available at the time the model is intended to be used?

For a prediction model to be usable in a real-world setting, all included predictors need to be available at the time the model is intended to be applied (that is, at the moment of prediction) (**Table 2**). This sounds straightforward, but some models unfortunately include predictors or predictor information that could not be known at the time when the model would be used.

Table 4–Continued

Type of Predictors	Outcome	Total Study Sample Size (Participants With Outcome), *n*	Model Predictive Performance	
			Discrimination	Calibration
History taking, physical examination	First nonfatal MI	4547 (1984)	Yes	No
History taking, physical examination, laboratory parameters, CT	3 outcomes: Mortality at 14 d and at 6 mo and unfavorable events at 6 mo	300 (mortality at 14 d: 143; mortality at 6 mo: 162; unfavorable events at 6 mo: 213)	Yes	Yes
History taking, physical examination	DVT	1295 (289)	No	No
History taking, type of injury, physiological examination	Mortality within 28 d	Development: 20 127 (3076)	Yes	Yes
		Validation: 14 220 (1765)	Yes	Yes
History taking, physical examination	Positive bacterial culture	184 (57)	Yes	Yes

For example, a prognostic model to be used *preoperatively* to predict risk for nausea and vomiting within 24 hours after surgery should not include such predictors as *intraoperative* medication, unless this medication is preset and unchanged during surgery. Inappropriate inclusion of predictors not available at the time when the model would be applied makes a model unusable. It also inflates apparent model performance because such predictors are measured closer in time to the outcome assessment and are likely to be more strongly associated with the outcome. For predictors that are stable over time (such as sex and genetic factors), these aspects are not an issue.

Studies that aim to externally validate an existing prediction model are at high ROB when predictor data are missing at the time of validation and the researchers validate the model anyway by omitting these missing predictors. This is a common flaw in validation studies and effectively produces validation results for another model rather than for the intended model as originally developed. In such situations, this signaling question should be answered as N.

Example. Rietveld and colleagues (89) aimed to develop and validate a prediction model for the diagnosis of a bacterial origin of acute conjunctivitis in children presenting in primary care with symptoms of this disease, to guide decision making about the administration of antibiotics. All predictors should be available to the general practitioner during the initial consultation. The predictors in this study were indeed all obtained during history taking and physical examination, so this signaling question should be answered as Y. If the study had included laboratory testing (such as microscopy) among the predictors, the signaling question would probably be answered as N. Because obtaining microscopy results involves a delay, the general practitioner would be unlikely to have the results available during the initial consultation.

Rating the ROB for domain 2. Table 8 shows how the signaling questions should be answered and an overall judgment for domain 2 reached.

Applicability. A common reason for concerns regarding applicability in this domain is inconsistency between definition, assessment, or timing of predictors and the review question. Predictors should be measured using methods potentially applicable to the setting (Tables 2 and 5) addressed by the review. Primary studies that use specialized measurement techniques for predictors may yield optimistic predictions for the targeted setting of the review. For example, if a model should be used in a health setting with limited access to advanced imaging, a model development study that included results of positron emission tomography might not be applicable and so may be rated as high concern.

As in domain 1, a subtle distinction can exist between ROB and applicability assessment in this domain. Consider the example given in signaling question 2.1 of active bleeding in the lower digestive tract as a predictor for colorectal cancer presence. Such bleeding could be assessed on the basis of visible blood in the stool or fecal occult blood testing. Reviewers might focus their review on diagnostic models that used only the visual assessment as a predictor of colorectal cancer, meaning that a primary study using a fecal occult blood test would raise applicability concerns.

Similarly, as in domain 1, in reviews that aim to estimate the average predictive performance of a specific model, heterogeneity in the observed performance of that model across the development study and validation studies is expected due to differences in the definition and measurement of predictors (17, 40, 44). If different definitions or assessment methods are used, some validation studies might find different predictive performance from others and should be judged as a concern regarding applicability. Sometimes researchers intentionally apply different definitions or measurement methods—for example, using point-of-care rather than laboratory testing methods for certain blood values. Again, this might not be a problem if the explicit aim of the systematic review is to include all validations

Table 5. Example Step 1 Applied to the Perel Example Study*

Criteria	Specify Your Systematic Review Question
Intended use of model	Prognosis; at presentation at hospital accident and emergency
Participants, including selection criteria and setting	Trauma patients with or within 8 h at risk of significant bleeding, presenting at hospital accident and emergency department
Predictors (used in prediction modeling), including types of predictors (e.g., history, clinical examination, biochemical markers, imaging tests), time of measurement, specific measurement issues (e.g., any requirements/prohibitions for specialized equipment)	Patients' demographics; physiologic variables; injury characteristics; time from injury—all measured at presentation at the hospital accident and emergency department Imaging results available within 4 h of admission
Outcome to be predicted	Death within 28 d of injury

* Reference 90.

of a certain model regardless of the definition and measurement method of its predictors.

Domain 3: Outcome

This domain covers potential sources of bias and applicability concerns related to the definition and determination of the outcome. The ideal outcome determination would classify the outcome without error in all study participants.

In *diagnostic* model studies, the outcome is presence or absence of the target condition. Outcome determination or verification is measured using a reference standard (Box 2). In *prognostic* model studies, the predicted outcomes occur in the future, after the moment of prediction. For both types of model, the reference standard or outcome determination method may include a single test or procedure, a combination of tests (composite outcome), or a consensus by experts (for example, an outcome adjudication committee).

The support for judgment box enables reviewers to describe how and when the outcome was defined and determined and what information was available at the time of determination.

ROB. Six signaling questions facilitate an ROB judgment for this domain (Table 9).

3.1 Was the outcome determined appropriately?

This signaling question is intended to detect potential for bias due to outcome misclassification because suboptimal or inferior methods were used to determine the outcome. Errors in outcome classification can lead to biased regression coefficients, biased estimates of the intercept (logistic regression and parametric survival models models) or baseline hazard (Cox regression model), and thus biased performance measures of the prediction model.

When prediction model studies use data from routine care registries or existing studies originally designed and conducted to answer a different research question, assessors need to carefully appraise the appropriateness of outcome determination methods, sometimes using details from earlier publications about that study. In routine care registries, outcome data might not be recorded at all, or methods may have been suboptimal and may have missed or misclassified the outcome. In diagnostic studies, problems and bias due to misclassification of the target condition by suboptimal reference standard methods have been extensively studied (112, 116, 123–127).

As in measurement of predictors (signaling question 2.1), the potential for bias is higher for outcomes that involve subjective judgment, such as imaging, surgical, or even pathology results. Where special skill or training is required, specifying who determined the outcome (for example, experienced consultant vs. inexperienced trainee) may also be important.

Example. In Han and colleagues' study (87), "there were two defined outcomes for each of the models: one was mortality at 14 days, and the other was unfavorable outcome at 6 months," defined by the authors on the basis of the Glasgow Outcome Scale (GOS) as "severe disability, vegetative state, or death." Because outcomes, mortality, and the 3 categories based on the definition of GOS used well-established, appropriate measures for outcome determination, the signaling question should be answered as Y. Problems could arise if assessors who were not trained in determining this outcome had measured the GOS score. Despite the limited number of categories, misclassification is not uncommon for the GOS (128, 129). Use of inexperienced assessors could lead to a less favorable (PN or NI) answer for this signaling question.

3.2 Was a prespecified or standard outcome definition used?

This signaling question aims to detect potential ROB where model performance has been inflated by selection of an outcome definition that produces more favorable results, an example of selective outcome reporting (130).

Risk of bias is low when a prespecified or standard outcome definition is used and substantiated by a definition from clinical guidelines, previously published studies, or a published study protocol. Risk of bias is higher if an atypical threshold on a continuous scale has been used for defining an "outcome being present." Biased model performance can occur if authors test multiple thresholds to obtain the most favorable outcome definition to achieve the best estimate of model performance. For example, a biased assessment of model performance would result if authors used a continuous scale (like the GOS) ranging from 3 to 15 and chose thresholds for "good" and "poor" outcomes based on achieving the best model predictive performance.

Composite outcomes can also introduce ROB. For example, authors may introduce bias by adjusting a composite outcome definition to favor better model performance by excluding typical components or including atypical events.

Many outcomes have consensus-based definitions, including thresholds and preferred composite outcome definitions. The COMET (Core Outcome Mea-

sures in Effectiveness Trials) initiative (www.comet-initiative.org) was set up to facilitate development of agreed, standardized sets of outcomes. Determining whether standard or nonstandard definitions have been used may require specialist clinical knowledge.

Example. In Han and colleagues' study (87), "there were two defined outcomes for each of the models: one was mortality at 14 days, and the other was unfavorable outcome at 6 months," defined by the authors on the basis of the GOS as "severe disability, vegetative state, or death." Because mortality and the 3 categories based on the definition of GOS are well-established outcomes (that is, standard outcome definitions were used), the signaling question should be answered as Y. If instead of using a standard definition the authors had amended the categories of the GOS on the basis of their own clinical experience or internal hospital guidance, clinical judgment should be used to decide whether the altered GOS still constituted a standard outcome determination; if not, the signaling question should be answered as PN or N.

3.3 Were predictors excluded from the outcome definition?

Outcomes should ideally be determined without information about predictors (see signaling question 3.5), but in some cases it is not possible to avoid including predictors—for example, when outcomes require determination by a consensus panel using as much information as is available. If a predictor in the model forms part of the definition or assessment of the outcome that the model predicts, the association between predictor and outcome will likely be overestimated, and estimates of model performance will be optimistic; in diagnostic research, this problem is generally called incorporation bias (104, 111, 115, 117, 119, 131–134).

Where outcomes are difficult to determine by a single procedure (for example, a single reference test), determination of outcome presence or absence may be based on multiple components or tests (as in the World Health Organization criteria for diagnosis of MI), or even on all available information, including the predictors under study. The latter approach is known as consensus or expert panel outcome measurement and is also susceptible to incorporation bias (135).

Example. Aslibekyan and colleagues (86) aimed to develop a cardiovascular risk score based on the ability of predictors (such as dietary components, physical activity, smoking status, alcohol consumption, socioeconomic status, and measures of overweight and obesity) to predict nonfatal MI. The study reported that MI was defined according to World Health Organization criteria, including cardiac biomarkers, electrocardiography, imaging, or autopsy confirmation. Because the lifestyle and socioeconomic predictors Aslibekyan and colleagues used for modeling do not form any part of this definition of MI, the study would be rated as Y for this signaling question. If the study had included a cardiac biomarker (such as troponin T at initial hospital presentation) among the predictors assessed, this signaling question would likely be rated as N. This is because the initial troponin T measurement may have formed part of the information used to determine the outcome (MI).

3.4 Was the outcome defined and determined in a similar way for all participants?

The outcome should be defined and determined in the same way for all study participants, similar to predictors (signaling question 2.1).

Outcome definition and measurement should include the same thresholds and categories to define the presence of the outcome across participants. Where a composite outcome measure is used, the results of individual components should always be combined in the same way to establish outcome presence or absence. When a consensus or panel-based outcome committee is used, the same method for establishing the outcome (for example, majority vote) should be used (131, 135, 136).

Risk of bias can arise when outcome determination methods vary among participants—for example, because of variation between research sites in a multicenter study. Risk of bias is also higher in prediction model studies that are not based on predesigned studies but on data collected for a different purpose, such as routine care registry data, where inherently different outcome definitions and measurements are likely to be applied. In addition, when accuracy in determining the presence of an outcome varies among measurement methods (differential outcome verification) and the direction of bias is not easy to predict, ROB is higher. For example, in a *prognostic* model study aimed at predicting the future occurrence of diabetes in healthy adults, the presence of diabetes in an individual can be determined in various ways that may have different abilities to determine diabetes presence or absence, such as fasting glucose levels, oral glucose tolerance tests, or self-report. The potential for bias is again higher when outcomes require more subjective interpretation. Simi-

Table 6. Example Step 2 Applied to the Perel Example Study*

Type of Prediction Study	PROBAST Boxes to Complete	Tick as Appropriate	Definitions for Type of Prediction Model Study
Development only	Development		Prediction model development without external validation. These studies may include internal validation methods, such as bootstrapping and cross-validation techniques.
Development and validation	Development and validation	✓	Prediction model development combined with external validation in other participants in the same article.
Validation only	Validation		External validation of existing (previously developed) model in other participants.

PROBAST = Prediction model Risk Of Bias ASsessment Tool.
* Reference 90.

Table 7. Domain 1: Participants—Guidance Notes for Rating Risk of Bias and Applicability

Risk of bias assessment

Background

The overall aim for prediction models is to generate absolute risk predictions that are correct in new individuals. Certain data sources or designs are not suited to generate absolute probabilities. Problems may also arise if a study inappropriately includes or excludes participant groups from entering the study.

1.1 Were appropriate data sources used, e.g., cohort, RCT, or nested case–control study data?

Yes/probably yes: If a cohort design (including RCT or proper registry data) or a nested case–control or case–cohort design (with proper adjustment of the baseline risk/hazard in the analysis) has been used.

No/probably no: If a nonnested case–control design has been used.

No information: If the method of participant sampling is unclear.

1.2 Were all inclusions and exclusions of participants appropriate?

Yes/probably yes: If inclusion and exclusion of participants was appropriate, so participants correspond to unselected participants of interest.

No/probably no: If participants are included who would already have been identified as having the outcome and so are no longer participants at suspicion of disease (diagnostic studies) or at risk of developing outcome (prognostic studies),

or if specific subgroups are excluded that may have altered the performance of the prediction model for the intended target population.

No information: When there is no information on whether inappropriate inclusions or exclusions took place.

Risk of bias introduced by participants or data sources

Low risk of bias: If the answer to all signaling questions is "Yes" or "Probably yes," then risk of bias can be considered low. If ≥1 of the answers is "No" or "Probably no," the judgment could still be "Low risk of bias" but specific reasons should be provided why the risk of bias can be considered low.

High risk of bias: If the answer to any of the signaling questions is "No" or "Probably no," there is a potential for bias, except if defined at low risk of bias above.

Unclear risk of bias: If relevant information is missing for some of the signaling questions and none of the signaling questions is judged to put this domain at high risk of bias.

Applicability

Background

Included participants, the selection criteria used as well as the setting used in the primary prediction model study should be relevant to the review question.

Concern that included participants or the setting do not match the review question

Low concern for applicability: Included participants and clinical setting match the review question.

High concern for applicability: Included participants and clinical setting were different from the review question.

Unclear concern for applicability: If relevant information about the participants and clinical setting are not reported.

RCT = randomized controlled trial.

larly, outcomes measured on several occasions (such as clinic visits) are at ROB, particularly if the frequency of measurement is different between participants; more measurement occasions increase the likelihood of detecting the outcome.

In *diagnostic* studies, researchers sometimes explicitly do not or cannot apply the same outcome measure in each individual. For instance, in cancer detection studies, pathology results are likely to be available as a reference standard only for participants who have a positive result on a preceding index test, such as an imaging test. Two situations may then occur: *partial verification*, when outcome data are completely missing for the subset of participants with negative results on the index test and no reference standard result, or *differential verification*, when participants who were not referred to the preferred reference standard are assessed using an alternative reference standard of differing, usually lower, accuracy (106, 111, 117, 119, 131–134, 137). These differences in outcome determination affect the estimated associations between predictors and outcome and thus the predictive accuracy of the diagnostic models. Methods to account for partial and differential verification have been described (138–141).

Example. Han and colleagues (87) validated a model to predict an "unfavourable outcome after six months" in patients with severe traumatic brain injury. The outcome was determined using the GOS (levels 1 to 3 on the 5-point scale) for all patients included in this single-center study. This signaling question should be answered as Y. If a hospital in the study had used a different instrument to measure the outcome of interest, such as the Functional Status Examination, this would constitute a potential ROB because the tools are not directly comparable. This question would then be answered as PN or even N to highlight the potential ROB.

3.5 Was the outcome determined without knowledge of predictor information?

The outcome is ideally determined without information about predictors. This is comparable to randomized intervention trials, where the outcome is ideally determined without knowledge of treatment assignment. Knowledge of predictor results may influence outcome determination and lead to biased predictive accuracy of the model, usually due to overestimation of the association between predictors and outcome (111, 115, 117, 119, 132–134). This risk is lower for objective outcomes, such as death from any cause or whether childbirth was natural or by Cesarean section, but higher for outcome determinations requiring interpretation, such as death from a specific cause.

Some outcomes are inherently difficult to determine using a single measurement or test. As discussed in signaling question 3.3, sometimes diagnostic and prognostic research cannot avoid use of a consensus panel or end point committee, with which outcome determination includes knowledge of predictor information. If the explicit aim is to assess the incremental value of a particular predictor or compare the performance of competing models (for example, when validating >1 model on the same data set), the importance of blinded outcome determination increases to prevent overestimation of the incremental value of a particular predictor or to prevent biased preference for 1 model over another.

Review authors should carefully assess whether predictor information was available to those determining the outcome. If the information was present during outcome determination or if this is unclear, the potential consequences should be considered in the overall judgment of bias of this domain. This overall judgment should take into account the subjectivity of the outcome of interest and the underlying review question.

Example. In the diagnostic prediction model study of Rietveld and colleagues (89), the outcome of interest

was a bacterial infection of the eye established by culture as the reference standard procedure. Reading of the culture results was somewhat subjective. Therefore, the authors of the paper explicitly inform the reader about the degree of blinding in their study: "The general practitioners did not receive the culture results, and the microbiologist who analyzed the cultures had no knowledge of the results of the index tests [the candidate predictors of the study]" (89). This signaling question should therefore be answered as Y.

3.6 Was the time interval between predictor assessment and outcome determination appropriate?

This signaling question aims to detect situations where the time interval between predictor assessment and outcome determination is inappropriate (either too long or too short). Such judgment requires clinical knowledge to determine the appropriate time interval, and also depends on clinical context.

In *diagnostic* studies, where the model predicts whether the outcome (that is, target disease determined by a reference standard) is present at the moment of prediction (Box 2), the assessment of predictors (index tests) and outcome should ideally occur at the same point in time. In practice, an interval may elapse between the moments of assessment for predictors and outcome, in which the diagnostic outcome classification could improve or worsen. Sometimes determining the outcome presence requires clinical follow-up over time, so a delay between predictor and outcome assessment is built into the study design as a critical feature to reduce bias (as in Oudega and colleagues [85]).

A delay of a few days between predictor assessment and outcome determination may not be problematic for chronic conditions, whereas for acute infectious diseases even a short delay may be inappropriate. Conversely, when the reference standard involves follow-up, a minimum time may be required to capture the increase in symptoms or signs indicating that the disease was present at the moment when the predictors were assessed. Sometimes biological samples for predictor assessment and outcome determination are taken at the same time point, so the interval during which the disease status could change is effectively 0 even if the reference standard procedure on the sample is completed at a later time point.

In *prognostic* studies, the time interval between the moments of predictor assessment and outcome determination may also have been too short or too long to capture the clinically relevant outcome of interest.

For both *diagnostic and prognostic* models, bias can present in 2 ways. First, it can result if outcomes are determined too early, when relevant outcomes cannot be detected or the number of outcomes is unrepresentative. For example, in a model diagnosing the presence of metastases at the time of surgical removal of a colorectal cancer tumor, detection of metastases can be biased by the time point of follow-up used for the reference standard. Choice of a time point that is too early can introduce bias in the number of metastases detected due to limitations in current detection methods; at earlier follow-up times, metastases may not have grown large enough for detection. Second, the type of outcome may differ depending on the time interval. For example, metastases detected at earlier times might be mainly liver metastases, whereas at 1 year, more bone metastases may be detected. An ROB then occurs if the interval between predictor assessments and outcome determination results in a potentially unrepresentative number or type of outcomes (that is, metastatic locations).

Table 8. Domain 2: Predictors—Guidance Notes for Rating Risk of Bias and Applicability

Risk of bias assessment

Background

Bias in model performance can occur when the definition and measurement of predictors is flawed. Predictors are the variables evaluated for their association with the outcome of interest. Bias can occur, for example, when predictors are not defined in a similar way for all participants or knowledge of the outcome influences predictor assessments.

2.1 Were predictors defined and assessed in a similar way for all participants?

Yes/probably yes: If definitions of predictors and their assessment were similar for all participants.

No/probably no: If different definitions were used for the same predictor or if predictors requiring subjective interpretation were assessed by differently experienced assessors.

No information: If there is no information on how predictors were defined or assessed.

2.2 Were predictor assessments made without knowledge of outcome data?

Yes/probably yes: If outcome information was stated as not used during predictor assessment or was clearly not (yet) available to those assessing predictors.

No/probably no: If it is clear that outcome information was used when assessing predictors.

No information: No information on whether predictors were assessed without knowledge of outcome information.

2.3 Are all predictors available at the time the model is intended to be used?

Yes/probably yes: All included predictors would be available at the time the model is intended to be used for prediction.

No/probably no: Predictors would not be available at the time the model is intended to be used for prediction.

No information: No information on whether predictors would be available at the time the model is intended to be used for prediction.

Risk of bias introduced by predictors or their assessment

Low risk of bias: If the answer to all signaling questions is "Yes" or "Probably Yes," then risk of bias can be considered low.

If ≥1 of the answers is "No" or "Probably no," the judgment could still be "Low risk of bias" but specific reasons should be provided why the risk of bias can be considered low, e.g., use of objective predictors not requiring subjective interpretation.

High risk of bias: If the answer to any of the signaling questions is "No" or "Probably no," there is a potential for bias.

Unclear risk of bias: If relevant information is missing for some of the signaling questions and none of the signaling questions is judged to put the domain at high risk of bias.

Applicability

Background

The definition, assessment, and timing of predictors in the primary prediction model study should be relevant to the review question, e.g., predictors should be measured using methods potentially applicable to the daily practice that is addressed by the review.

Concern that the definition, assessment, or timing of predictors in the model do not match the review question.

Low concern for applicability: Definition, assessment, and timing of predictors match the review question.

High concern for applicability: Definition, assessment, or timing of predictors were different from the review question.

Unclear concern for applicability: If relevant information about the predictors is not reported.

Table 9. Domain 3: Outcome—Guidance Notes for Rating Risk of Bias and Applicability

Risk of bias assessment

Background

Bias in model performance can occur when methods used to determine outcomes incorrectly classify participants with or without the outcome. Bias in methods of outcome determination can result from use of suboptimal methods, tests, or criteria that lead to unacceptably high levels of errors in outcome determination, when methods are inconsistently applied across participants, or when knowledge of predictors influence outcome determination. Incorrect timing of outcome determination can also result in bias.

3.1 Was the outcome determined appropriately?

Yes/probably yes: If a method of outcome determination has been used which is considered optimal or acceptable by guidelines or previous publications on the topic.

Note: This is about level of measurement error within the method of determining the outcome (see concerns for applicability about whether the *definition* of the outcome method is appropriate).

No/probably no: If a clearly suboptimal method has been used that causes unacceptable error in determining outcome status in participants.

No information: No information on how outcome was determined.

3.2 Was a prespecified or standard outcome definition used?

Yes/probably yes: If the method of outcome determination is objective,

or if a standard outcome definition is used,

or if prespecified categories are used to group outcomes.

No/probably no: If the outcome definition was not standard and not prespecified.

No information: No information on whether the outcome definition was prespecified or standard.

3.3 Were predictors excluded from the outcome definition?

Yes/probably yes: If none of the predictors are included in the outcome definition.

No/probably no: If ≥1 of the predictors forms part of the outcome definition.

No information: No information on whether predictors are excluded from the outcome definition.

3.4 Was the outcome defined and determined in a similar way for all participants?

Yes/probably yes: If outcomes were defined and determined in a similar way for all participants.

No/probably no: If outcomes were clearly defined and determined in a different way for some participants.

No information: No information on whether outcomes were defined or determined in a similar way for all participants.

3.5 Was the outcome determined without knowledge of predictor information?

Yes/probably yes: If predictor information was not known when determining the outcome status,

or outcome status determination is clearly reported as determined without knowledge of predictor information.

No/probably no: If it is clear that predictor information was used when determining the outcome status.

No information: No information on whether outcome was determined without knowledge of predictor information.

3.6 Was the time interval between predictor assessment and outcome determination appropriate?

Yes/probably yes: If the time interval between predictor assessment and outcome determination was appropriate to enable the correct type and representative number of relevant outcomes to be recorded,

or if no information on the time interval is required to allow a representative number of the relevant outcome occur or if predictor assessment and outcome determination were from information taken within an appropriate time interval.

No/probably no: If the time interval between predictor assessment and outcome determination is too short or too long to enable the correct type and representative number of relevant outcomes to be recorded.

No information: If no information was provided on the time interval between predictor assessment and outcome determination.

***Table 9*—Continued**

Risk of bias introduced by predictors or their assessment

Low risk of bias: If the answer to all signaling questions is "Yes" or "Probably yes," then risk of bias can be considered low.

If ≥1 of the answers is "No" or "Probably no," the judgment could still be low risk of bias, but specific reasons should be provided why the risk of bias can be considered low, e.g., when the outcome was determined with knowledge of predictor information but the outcome assessment did not require much interpretation by the assessor (e.g., death regardless of cause).

High risk of bias: If the answer to any of the signaling questions is "No" or "Probably no," there is a potential for bias.

Unclear risk of bias: If relevant information about the outcome is missing for some of the signaling questions and none of the signaling questions is judged to put this domain at high risk of bias.

Applicability

Background

The definition of outcome in the primary study should be relevant for the outcome definition in the review question.

Concern that the outcome definition, timing, or determination do not match the review question

Low concern for applicability: Outcome definition, timing, and method of determination defines the outcome as intended by the review question.

High concern for applicability: Choice of outcome definition, timing, and method of outcome determination defines another outcome as intended by the review question.

Unclear concern for applicability: If relevant information about the outcome, timing, and method of determination is not reported.

Obviously, a review may specifically aim to determine either the short- or long-term prognosis of a certain condition, so the time interval between predictor assessment and outcome determination is also relevant to study applicability to the review question.

Example. Rietveld and colleagues (89) developed a diagnostic model to predict bacterial cause in conjunctivitis eye infection; ROB in the time interval is minimized because the same clinic visit was used to measure predictors from patient questionnaires and physical examination and to collect conjunctival samples for determination of the outcome of bacterial infection. Although the reference standard results require culture for more than 48 hours, this is not relevant to bias, because culture results reflect disease at the time of sample collection. This signaling question would be answered as Y, indicating a low potential for bias.

Example. In Aslibekyan and colleagues (86), where a model was developed to predict MI, this signaling question should be answered as NI due to lack of information on the time interval between predictor measurement and outcome determination. Different intervals could alter the number of MI events that would be detected.

Rating the ROB for domain 3. Table 9 shows how the signaling questions should be answered and an overall judgment for domain 3 reached.

Applicability. The applicability question for this domain considers the extent to which the outcome predicted in the developed or validated model matches the review question. If different definitions, timing, or determination methods are used, this should be judged an applicability concern. For example, a primary study might use a composite outcome that consists of components different from those included in the outcome definition of the review question (142).

Table 10. Domain 4: Analysis–Guidance Notes for Rating Risk of Bias

Risk of bias assessment

Background

Statistical analysis is a critical part of prediction model development and validation. The use of inappropriate statistical analysis methods increases the potential for bias in reported model performance measures. Model development studies include many steps where flawed methods can distort results. We recommend reviewers seek statistical advice when completing assessments of the analysis domain.

4.1 Were there a reasonable number of participants with the outcome?

Yes/probably yes: For model development studies, if the number of participants with the outcome relative to the number of candidate predictor parameters is ≥20 (EPV ≥20).*

For model validation studies, if the number of participants with the outcome is ≥100.

No/probably no: For model development studies, if the number of participants with the outcome relative to the number of candidate predictor parameters is <10 (EPV <10).*

For model validation studies, if the number of participants with the outcome is <100.

No information: For model development studies, no information on the number of candidate predictor parameters or number of participants with the outcome, such that the EPV cannot be calculated.

For model validation studies, no information on the number of participants with the outcome.

4.2 Were continuous and categorical predictors handled appropriately?

Yes/probably yes: If continuous predictors are not converted into ≥2 categories when included in the model (i.e., dichotomized or categorized),

or if continuous predictors are examined for nonlinearity using, for example, fractional polynomials or restricted cubic splines,

or if categorical predictor groups are defined using a prespecified method.

For model validation studies, if continuous predictors are included using the same definitions or transformations, and categorical variables are categorized using the same cut points, as compared with the development study.

No/probably no: If categorical predictor group definitions do not use a prespecified method.

For model development studies, if continuous predictors are converted into ≥2 categories when included in the model.

For model validation studies, if continuous predictors are included using different definitions or transformations, or categorical variables are categorized using different cut points, as compared with the development study.

No information: No information on whether continuous predictors are examined for nonlinearity and no information on how categorical predictor groups are defined.

For model validation studies, no information on whether the same definitions or transformations and the same cut points are used, as compared with the development study.

4.3 Were all enrolled participants included in the analysis?

Yes/probably yes: If all participants enrolled in the study are included in the data analysis.

No/probably no: If some or a subgroup of participants are inappropriately excluded from the analysis.

No information: No information on whether all enrolled participants are included in the analysis.

4.4 Were participants with missing data handled appropriately?

Yes/probably yes: If there are no missing values of predictors or outcomes *and* the study explicitly reports that participants are not excluded on the basis of missing data,

or if missing values are handled using multiple imputation.

No/probably no: If participants with missing data are omitted from the analysis,

or if the method of handling missing data is clearly flawed, e.g., missing indicator method or inappropriate use of last value carried forward,

or if the study had no explicit mention of methods to handle missing data.

No information: If there is insufficient information to determine if the method of handling missing data is appropriate.

4.5 Was selection of predictors based on univariable analysis avoided?†

Yes/probably yes: If the predictors are *not* selected on the basis of univariable analysis prior to multivariable modeling.

No/probably no: If the predictors are selected on the basis of univariable analysis prior to multivariable modeling.

No information: If there is no information to indicate that univariable selection is avoided.

***Table 10*–Continued**

4.6 Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?

Yes/probably yes: If any complexities in the data are accounted for appropriately,

or if it is clear that any potential data complexities have been identified appropriately as unimportant.

No/probably no: If complexities in the data that could affect model performance are ignored.

No information: No information is provided on whether complexities in the data are present or accounted for appropriately if present.

4.7 Were relevant model performance measures evaluated appropriately?

Yes/probably yes: If both calibration and discrimination are evaluated appropriately (including relevant measures tailored for models predicting survival outcomes).

No/probably no: If both calibration and discrimination are not evaluated,

or if only goodness-of-fit tests, such as the Hosmer–Lemeshow test, are used to evaluate calibration,

or if for models predicting survival outcomes performance measures accounting for censoring are not used,

or if classification measures (like sensitivity, specificity, or predictive values) were presented using predicted probability thresholds derived from the data set at hand.

No information: Either calibration or discrimination are not reported,

or no information is provided as to whether appropriate performance measures for survival outcomes are used (e.g., references to relevant literature or specific mention of methods, such as using Kaplan–Meier estimates),

or no information on thresholds for estimating classification measures is given.

4.8 Were model overfitting and optimism in model performance accounted for?†

Yes/probably yes: If internal validation techniques, such as bootstrapping and cross-validation including all model development procedures, have been used to account for any optimism in model fitting, and subsequent adjustment of the model performance estimates have been applied.

No/probably no: If no internal validation has been performed,

or if internal validation consists only of a single random split-sample of participant data,

or if the bootstrapping or cross-validation did not include all model development procedures including any variable selection.

No information: No information is provided on whether internal validation techniques, including all model development procedures, have been applied.

4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?†

Yes/probably yes: If the predictors and regression coefficients in the final model correspond to reported results from multivariable analysis.

No/probably no: If the predictors and regression coefficients in the final model do not correspond to reported results from multivariable analysis.

No information: If it is unclear whether the regression coefficients in the final model correspond to reported results from multivariable analysis.

Risk of bias introduced by the analysis

Low risk of bias: If the answer to all signaling questions is "Yes" or "Probably yes," then risk of bias can be considered low.

If ≥1 of the answers is "No" or "Probably no," the judgment could still be low risk of bias, but specific reasons should be provided why the risk of bias can be considered low.

High risk of bias: If the answer to any of the signaling questions is "No" or "Probably no," there is a potential for bias.

Unclear risk of bias: If relevant information about the analysis is missing for some of the signaling questions but none of the signaling question answers is judged to put the analysis at high risk of bias.

EPV = events per variable.

* For EPVs between 10 and 20, the item should be rated as either probably yes or probably no, depending on the outcome frequency, overall model performance, and distribution of the predictors in the model. For more guidance, see references 145 to 147.

† Development only.

As discussed for domains 1 and 2, in reviews that aim to estimate the average performance of a specific model across the included validation studies, heterogeneity in performance among validation studies is expected due to differences in definition and measurement of the outcome (17, 40, 44). Sometimes researchers intentionally apply different outcome definitions or measurement methods. This might not be a problem if the systematic review explicitly aimed to include all validations of the model regardless of outcome definition and measurement method.

Domain 4: Analysis

Use of inappropriate analysis methods or omission of important statistical considerations increases the potential for bias in the estimated predictive performance of a model. Domain 4 examines whether key statistical considerations were correctly addressed. Some of these aspects require specialist knowledge, and we recommend that this domain be assessed by at least 1 researcher with statistical expertise in prediction model studies. The support for judgment box should list and describe the important aspects needed to address this domain.

Nine signaling questions facilitate an ROB judgment for this domain (Table 10).

4.1 Were there a reasonable number of participants with the outcome?

As in all medical research, the larger the sample size, the better, because it leads to more precise results—that is, smaller standard errors and narrower CIs. In prediction model studies, overall sample size matters, but number of participants with the outcome is even more important. For a binary outcome, the effective sample size is the smaller of the 2 outcome frequencies, "with the outcome" and "without the outcome." For time-to-event outcomes, the key driver is the total number of participants with the event by the main time point of interest for prediction. More important, in prediction model studies, the number of participants with the outcome not only influences precision but also affects predictive performance—that is, is a potential source of bias. What is considered a reasonable number of participants with the outcome (yielding low ROB) differs between model development and validation studies.

Model development studies. The performance of any prediction model is to some extent overestimated when both model development and performance assessment use the same data set (49, 50, 146, 147). This overestimation is larger with smaller sample sizes and, notably, when fewer participants have the outcome. Concerns about optimistic performance are exacerbated when the predictors included in the final model are selected from many candidate predictors, relative to a low number of participants with the outcome, and when predictor selection was based on univariable analysis (see signaling question 4.5). Sample size considerations for model development studies have historically been based on the number of events per variable (EPV). More exactly, the number of events relative to the number of regression coefficients needs to be estimated for candidate predictors. For example, a candidate predictor with 6 categories will require 5 degrees of freedom (5 regression coefficients are estimated). Also, the word *candidate* is important: It indicates not the number of predictors included in the final model but rather the total number of predictors considered during any stage of the prediction model process.

Although an EPV of at least 10 has been widely adopted as a criterion to minimize overfitting (148–150), recent studies have shown that this threshold has no scientific basis (145), and various authors have suggested higher EPVs of at least 20 (145, 151, 152). In general, studies with EPVs lower than 10 are likely to have overfitting, whereas those with EPVs higher than 20 are less likely to have overfitting. However, the sample size needed to minimize overfitting is context-specific and depends on outcome prevalence, overall model performance (R^2), and predictor distributions (143–145). Therefore, deciding whether an appropriate sample size was used may be difficult, especially when EPV is between 10 and 20. Prediction models developed using machine-learning techniques often require substantially higher EPVs (often >200) to minimize overfitting (153).

Hence, the smaller the effective sample size and the lower the EPV, the higher the risk that the final prediction model has included spurious predictors ("overfitted" model) or failed to include important predictors ("underfitted" model). Overfitting and underfitting are likely to yield biased estimates of model apparent predictive performance (49–51, 146, 147, 154). With a small EPV, authors need to quantify the extent of misfitting of the developed prediction model (for example, by using internal validation techniques). Based on this internal validation, optimism-adjusted estimates of model performance can be produced and model parameters adjusted (that is, shrink regression coefficients) to decrease this bias (see signaling question 4.8).

Model validation studies. The aim of a validation study is to quantify the predictive performance of an existing model using a different data set from that used in model development (Box 1) (8, 49, 50, 155–157). The focus in a validation study is on accurate and precise estimation of model performance so that meaningful conclusions can be drawn. Validation studies are recommended to include at least 100 participants with the outcome, otherwise the risk for biased estimates of model performance becomes more likely (77, 78, 158).

Example. Aslibekyan and colleagues (86) developed 2 prognostic models (1 including only easy-to-obtain predictors and 1 extended with various dietary and blood markers) to predict risk for MI. Although the authors used a case–control design and many inclusion and exclusion criteria, their final sample had 839 case patients with an MI for developing prediction model 1 and 696 for model 2. The exact number of candidate predictors is not explicitly mentioned, but from the methods and supplementary tables 1 and 2 we can estimate that the authors likely used

20 to 30 predictors—or rather, degrees of freedom, because they categorized several continuous predictors into quintiles. This indicates that the EPV is between (taking the smallest number of events) 696/20 (that is, 35) and 696/30 (that is, 23). Because the EPV in either case is much larger than 10 and even larger than 20, this signaling question should be answered as Y, indicating low ROB.

Example. Oudega and colleagues (85) validated a diagnostic model for detecting the presence of DVT in patients who consulted their primary care physician about symptoms suggestive of DVT. The total sample size of their validation study was 1295 symptomatic patients, of whom 289 had DVT (as detected by D-dimer testing and leg ultrasonography). Because the study had more than the recommended 100 events for validation, the signaling question should be answered as Y, indicating low ROB. If this number had been lower (for example, 80 or 40 patients with DVT), the answer for this example would be PN or N, respectively.

4.2 Were continuous and categorical predictors handled appropriately?

Dichotomization of continuous predictors, such as age and blood pressure, should be avoided (159–161). Dichotomization requires choosing an often arbitrary cut point, for example, above which values are classified as high (or abnormal) and below which they are classified as low (or normal). The usual fallacious argument is that the approach aids clinical interpretation and maintains simplicity. However, it leads to loss of information, and a prediction model that includes dichotomized continuous predictors can have substantially reduced predictive ability (159–162).

For example, dichotomizing a variable at the median value has been shown to reduce power by about the same amount as discarding a third of the data (163). Also, the range of model-predicted risks across the spectrum of predictor values is lost: Persons just below the cut point are assumed to have a different risk from those just above the cut point, even though their predictor values barely differ. Conversely, 2 persons with very different values that are nonetheless both above (or both below) the cut point are assumed to have identical risks. Linear (or nonlinear) relationships between predictor and outcome risk are therefore lost. When a predictor is categorized using widely accepted cut points (that is, not based on the data at hand), although information has been lost, ROB is low because the cut point was predefined.

Model development studies. A developed model is at low ROB when included predictors are kept as continuous. The association between predictor and outcome risk should still be examined as linear or nonlinear by using, for example, restricted cubic splines or fractional polynomials (49, 50, 164).

A developed model is at high ROB when dichotomized continuous predictors are included, especially when cut points were chosen via data dredging on the same data set (for example, to identify the "optimal" cut point that maximizes predictor effects or minimizes associated *P* values) (159–162) and when a selection procedure was used to identify "significant thresholds" (49, 50).

Risk of bias is decreased when a model uses categorization of continuous predictors into 4 or more groups rather than dichotomization, especially when categories are based on widely accepted cut points (160, 162). However, for a model to be at low ROB, it should be clear that the number and placement of predictor cut points were chosen before data analysis—that is, prespecified. For similar reasons, as discussed for signaling question 4.1, an internal validation followed by optimism adjustment of model performance and prediction model parameters also decreases ROB (see also signaling question 4.8). For model development studies that have dichotomized continuous predictors after data analysis and did not adjust for this by applying internal validation and shrinkage techniques, this signaling question should be answered as N.

Model validation studies. In model validation studies, the model as originally fitted in the development data should be evaluated on its predictive accuracy in the validation data set. This means that the originally reported intercept (or baseline hazards) and regression coefficients are used for exactly the same format of the predictors. For example, if body mass index (BMI) is originally included as dichotomized in the model, validation studies should use BMI values dichotomized at the same cut point and not BMI as continuous or dichotomized using a different cut point. If predictors have different formats in validation and development models, the validation might be at high ROB because the predictor–outcome association (the regression coefficient) of BMI from the development study is effectively used in the validation study for a different version of the predictor.

Example. Oudega and colleagues (85) validated the Wells rule for identifying persons with DVT. However, the authors state, "The last item of the rule—presence of an alternative diagnosis—has never been unambiguously defined and often causes controversy among users of the rule. In our study, physicians were asked to give their own assessment of the patient's probability of having DVT by using a score of 1 to indicate high probability of DVT, no alternative diagnosis likely; 2 to indicate moderate probability of DVT, alternative diagnosis possible; or 3 to indicate low probability of DVT, alternative diagnosis certain. To tailor the judgment of the physician on this item, 7 common alternative diagnoses for patients with suspected DVT were provided on the study form. If a low or moderate probability was assigned to a patient, we subtracted 2 points from the Wells score in the analysis." Because this is not a true deviation from the original definitions, this signaling question should be answered as Y.

Example. Perel and colleagues (88) developed a prediction model (CRASH-2) for early death in patients with traumatic brain injury. During model development, they analyzed the 3-category variable "type of injury" (penetrating, blunt, or blunt and penetrating) as a 2-category variable (penetrating vs. a combined category of blunt and penetrating); the rationale for this

was not given. Nevertheless, continuous variables were analyzed as continuous in model development, so the collapse from 3 to 2 categories for this variable was probably due to few participants or events being in the "blunt" category. Further, the type of injury was not subsequently included in the final model, so it is unlikely that reduction in predictor categories was done to improve statistical significance for this predictor. Therefore, we should answer the signaling question as Y. When externally validating the CRASH-2 model, the authors "applied the coefficients of the model developed in CRASH-2" and have used the same predictors and scale as originally coded; thus, an answer of Y also is appropriate for the model validation assessment.

4.3 Were all enrolled participants included in the analysis?

As in all types of medical studies, all participants enrolled in a study should be included in the data analysis, or else ROB is possible (48, 111, 165, 166). This signaling question relates to exclusion of participants from the original study sample who met the inclusion criteria. It is not about inappropriate inclusion or exclusion criteria (which are addressed in signaling question 1.1) or the *handling* of missing data in predictors or outcomes (which is covered in signaling question 4.4).

Enrolled participants are often excluded because of uninterpretable (unclear) findings, outliers, or missing data in predictors or outcomes. Outlier, uninterpretable, or missing values occur in all types of medical research. Omitting enrolled participants from analysis can lead to biased predictor–outcome associations and biased predictive performance of the developed or validated model if the remaining analyzed individuals are not a completely random but rather a selective subsample. The relationship between predictors and outcomes is then different for analyzed versus excluded participants. For example, excluding participants whose predictor values (such as results on imaging or laboratory tests) were unclear likely yields a study sample with participants in the extremes of the predictor range. This in turn may result in biased, overestimated model discrimination (166). When only a low percentage of enrolled participants are not included in the analysis, ROB may be low. However, a minimal or acceptable percentage is hard to define because it depends on which participants were excluded and whether it was a selected subsample or not. The ROB increases with an increasing percentage of participants excluded.

Prediction model development or validation studies based on routine care databases or registries, where participants are not formally enrolled in some predesigned study and data are even collected for other reasons, are particularly susceptible to this form of bias. When such data sources are used for model development or validation, participant selection for the analysis should be based on clear criteria. In prediction model studies based on such routine care data sets, the extent of potential bias can be unclear because of insufficiently reported information on the applied eligibility criteria and on reasons for excluding study participants.

Example. In Han and colleagues' study (87), all 300 participants met eligibility criteria for validation of 3 versions of the IMPACT (International Mission for Prognosis and Analysis of Clinical Trials in Traumatic Brain Injury) models for traumatic brain injury, called core, extended, and laboratory IMPACT models. The investigators then excluded 36 participants (12%) from validation of the laboratory version of the model because of missing data on blood glucose levels; however, all participants could be included for the core and extended IMPACT models. For assessment of the core and extended models, the signaling question should be answered as Y because all participants were included in the analysis. For assessment of the laboratory model, the signaling question should be answered as either PN or PY, depending on the concern raised by exclusion of 36 of participants (12%) from the analysis. This would depend on clinical knowledge and judgment of whether the missing glucose measurements are likely to be associated with the severity of traumatic brain injury.

4.4 Were participants with missing data handled appropriately?

As noted in the previous item, simply excluding enrolled participants with any missing data from the analysis leads to biased predictor–outcome associations and biased model performance when the analyzed individuals are a selective rather than a completely random sample of the original full study sample (167–177). When a study report does not mention missing data, participants with any missing data have likely been omitted from analyses ("available-case" or "complete-case" analysis) because statistical packages automatically exclude persons with any missing value on any of the data analyzed unless prompted to handle otherwise. Numerous reviews show that available- or complete-case analysis is the most common way to handle missing data in prediction model studies (68, 178–186).

The most appropriate method for handling missing data is multiple imputation because it leads to the least biased results with correct standard errors and *P* values (167–173, 175–177). In prediction model studies, multiple imputation is superior in terms of bias and precision to other methods, in both model development (173, 176, 187) and validation studies (176, 188–190). In contrast to uninterpretable or outlier data, use of a separate category to capture missing data is not an appropriate method; this missing indicator method leads to biased results in prediction model studies, and the signaling question should then be answered as N (172, 177). The ROB due to missing data increases with increasing percentages of missing data, but a minimal acceptable percentage that can be used as a threshold for low ROB is hard to define (173). Judgment of possible ROB is facilitated when authors provide either the distributions (percentage, mean, or medians) of the predictors and outcomes between both groups (ex-

cluded vs. analyzed participants) or a comparison of the predictor–outcome associations and model predictive performance with and without inclusion of the participants with missing values. Similar results with and without such participants is a strong indication that the results of the analysis are less likely to be biased. If such a comparison is not presented and investigators have not used an imputation method, we recommend answering this signaling question as PN or N, especially if a relevant proportion of participants are excluded due to missing data.

Sometimes, when a model is validated in other data and a predictor of the model is systematically missing (for example, not measured), authors validate the original model (that is, the original predictor weights or regression coefficients) by simply omitting the predictor from the model. This leads to high ROB, and such studies should be rated as N for this question. If the model had originally been fitted without the omitted predictor, all of the remaining predictor coefficients would be different.

Example. Perel and colleagues (88) developed a prognostic model from a data set with very little missing data, and therefore they did a complete-case analysis. In the same article, the authors showed an external validation of this developed model where they applied multiple imputation. How few participants had missing data in the development study was unclear, and the completely observed and excluded sets of participants were not compared, making it hard to judge whether the model development had some ROB. In the validation study, the authors used multiple imputation, indicating that they know the procedure; if multiple imputation of missing data was needed in the development sample, they likely would have used it during model development as well. Accordingly, this signaling question should strictly be answered as NI for the development and Y for the validation part of the paper, although PY for the development part would also be possible.

Example. Aslibekyan and colleagues (86) stated that they used development complete-case analysis and excluded 10% of participants. No information was provided to confirm that complete-case analysis was a valid approach—that is, whether the included and excluded participants were similar such that the included participants approximated a completely random subset of the original study sample. Accordingly, this signaling question should be answered as N for development. For model validation, missing data and handling of missing data were not mentioned, so the answer for this signaling question for the model validation should strictly be NI, but perhaps even PN, given the reporting of their model development part and because all clinical studies tend to have some missing data.

4.5 Was selection of predictors based on univariable analysis avoided? (Model development studies only)

A data set will often have many features that could be used as candidate predictors, and in many studies researchers want to reduce the number of predictors during model development to produce a simpler model.

In a univariable analysis, individual predictors are tested for their association with the outcome. Researchers often select the predictors with a statistically significant univariable association (for example, $P < 0.05$) for inclusion in the development of a final prediction model. This method can lead to incorrect predictor selection because predictors are chosen on the basis of their statistical significance as a single predictor rather than in context with other predictors (49, 50, 191). Bias occurs when univariable modeling results in omission of variables from the model, because some predictors are important only after adjustment for other predictors, known from previous research to be important, did not reach statistical significance in the particular development set (for example, due to small sample size). Also, predictors may be selected on the basis of a spurious (accidental) association with the outcome in the development set.

A better approach to decide on omitting, combining, or including candidate predictors in multivariable modeling is to use nonstatistical methods—that is, methods without any statistical univariable pretesting of the associations between candidate predictors and outcome. Better methods include those based on existing knowledge of previously established predictors in combination with the reliability, consistency, applicability, availability, and costs of predictor measurement relevant to the targeted setting. Well-established predictors and those with clinical credibility should be included and retained in a prediction model regardless of any statistical significance (49, 50, 192). Alternatively, some statistical methods that are not based on prior statistical tests between predictor and outcome can be used to reduce the number of modeled predictors (for example, principal components analysis).

During modeling, predictor selection strategies may be used to omit predictors (for example, backwards selection procedures) and to fit a smaller, simpler final model (49, 50, 192). However, the effects of using such a multivariable predictor selection strategy on the potential overfitting of the prediction model to the development data at hand should be tested using internal validation and optimism adjustment strategies, which are discussed in signaling question 4.8.

When the model development correctly avoids univariable selection of candidate predictors and there is no evidence of univariable selection for predictors before the multivariable modeling, studies should be rated as Y or PY. When predictors are selected on the basis of univariable analysis before multivariable modeling, the signaling question for these studies should be answered as N.

Example. Before Perel and colleagues (88) developed their model, potential users of the model were consulted to identify candidate predictors and interactions based on known importance and convenience to the clinical settings of prehospital, battlefield, and emergency departments. The researchers then in-

cluded all so-defined candidate predictors in the multivariable analysis. Decisions about which predictors to retain in the final model were based on clinical reasoning, availability of predictor measurement at the time the model would be used, and practicalities of collecting predictors using equipment in the clinical settings. Although other predictors could have been considered important, the choice of predictors was not based on potentially biased univariable selection of predictors. The study would therefore be rated as Y for this signaling question.

Example. Rietveld and colleagues (89) used predictor selection based on univariable analysis ($P \leq 0.10$) to select predictors for the multivariable model. This signaling question would therefore be answered as N for this study. If all predictors had been entered into multivariable analysis without the prior univariable selection, an answer of Y would have been given.

4.6 Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?

The development and validation of prediction models must ensure that the statistical methods used and their underlying assumptions are appropriate for the study design and type of outcome data analyzed. Here, we draw attention to some key considerations related to complexities in the data that can lead to ROB in the estimated predictive performance of the model if not appropriately accounted for in the analyses.

As discussed under signaling question 1.1, if a case–cohort or nested case–control design is used for a prediction model, the analysis method must account for the sampling fractions (from the original cohort) to allow proper estimation of the absolute outcome probabilities (97, 99, 105, 109). For example, in a diagnostic prediction model (development or validation) study using a nested case–control design where a fraction of all control participants are sampled from the original cohort, a logistic regression in which the control participants are weighted by the inverse of their sampling fraction needs to be applied instead of a standard logistic regression, otherwise the predicted risks by the model will be biased. When such appropriate adjustments for sampling fractions are made, they alleviate the ROB concerns raised in signaling question 1.1. If they are not made, assessors should score an N only once to either signaling question 1.1 or this signaling question.

For prognostic models to predict long-term outcomes in which censoring occurs, a time-to-event analysis, such as a Cox regression, should be used to include censored participants up to the end of their follow-up. Use of logistic regression models that simply exclude censored participants with incomplete follow-up is inappropriate. Using a flawed logistic regression approach leads to a selected data set that includes fewer persons without the outcome, which biases predicted risks because those with the outcome are overrepresented. Time-to-event analysis correctly deals with these censored individuals.

When prominent competing risks are present, they should also be accounted for in the time-to-event analysis during development of a prognostic model. An example of competing risks would be in a model for occurrence of a second hip replacement, where death in elderly patients with a first hip replacement may occur before the second hip replacement. If competing risk is not correctly accounted for, absolute risk predictions will be overestimated and biased because patients with the competing event are simply censored (193).

Also, correct modeling methods are needed when each person can have more than 1 event, such as in a model of epilepsy seizure, where some persons have more than 2 seizures. Multilevel or random-effects (logistic or survival) modeling methods would be needed to avoid underestimation and bias in predictor effects (194–197).

Statistical expertise will be required to identify these and other issues in specific studies. The issues we have highlighted here will typically be the most important to be aware of in prediction modeling studies. If assessors deem that a study has ignored key statistical complexities, high ROB is indicated on this signaling question.

Example. Aslibekyan and colleagues (86) used a conditional logistic regression model to develop a prognostic prediction model for MI. Included participants provided data between 1994 and 2004; however, whether all individuals had predictor values recorded at the start of the period (vs. entering after 1994 and thus having a shorter follow-up) is unclear. If all individuals entered with predictor values in 1994, the model would predict risk for developing MI by 10 years (that is, by 2004) and be interpretable. However, if some individuals entered after 1994, then interpretation and bias of the logistic model is a concern because predictions are not specific to a particular time period and the length of follow-up is being ignored. If participants had different follow-up times, it would be better for a survival analysis model to be fitted to allow risk predictions over time and delayed entry of participants. Further, the prevalence of competing risk for death due to non-MI conditions was unclear, even though the population included persons up to age 86 years. Such issues may be a consequence of the case–control (rather than cohort) nature of the study. Thus, ROB was not avoided because of these statistical complexities and this signaling question should be rated as N or PN.

Example. In Rietveld and colleagues' study (89), the development of a diagnostic model using standard logistic regression was relatively straightforward because the developed model aimed to predict risk for bacterial conjunctivitis using a full cohort approach (without sampling) and therefore did not involve follow-up, censoring, or competing events. Here, this signaling question should be answered as Y.

4.7 Were relevant model performance measures evaluated appropriately?

Box 4 provides an overview of the various performance measures of a multivariable prediction model.

PROBAST is designed to assess studies on multivariable models that are developed or validated to make diagnostic or prognostic predictions in individuals–that is, *individualized predictions* (**Box 1**). Accordingly, to fully gauge the predictive performance of a model, reviewers must assess both model calibration and discrimination (such as the c-index) addressing the entire range of the model-predicted probabilities (7, 8). If calibration and discrimination are not assessed, the study is at ROB because the ability or performance of the model to provide accurate individual probabilities is not completely known (**Box 4**).

When calibration plots or tables are observed with small numbers of groups (for example, due to a small sample size with too few events), judgment of the plot is required to rate this signaling question properly. In the absence of a calibration plot or table comparing predicted versus observed outcome probabilities, studies reporting only a statistical test of calibration should be rated as N for this signaling question.

In addition, the methods used to assess model calibration and discrimination should be appropriate for the outcome the model is predicting. Such methods for models predicting a binary outcome developed using logistic regression will not be suitable for models using Cox regression to predict long-term outcome occurrences, such as 5-year mortality or survival, because censoring needs to be accounted for. Failure to account for censoring when assessing prognostic model calibration and discrimination–in either a development or a validation study–means that this signaling question should be answered as N or PN.

Some studies additionally provide classification measures, including sensitivity, specificity, predictive values, or reclassification measures (such as the net reclassification index), to indicate model predictive performance, sometimes without providing the model calibration and c-index (**Box 4**). Classification measures are most commonly provided in diagnostic model studies. Estimation of classification, as well as reclassification, parameters requires the introduction of 1 threshold (or more) in the range of the model-predicted probabilities. Use of thresholds allows the reporting of model predictive performance at probability thresholds that may be clinically relevant, as opposed to the entire range of the model-predicted probabilities. Nevertheless, use of probability thresholds typically leads to loss of information, because the entire range of predicted probabilities of the model is not fully utilized, and choice of thresholds can be data-driven rather than prespecified on clinical grounds (see also signaling question 4.2). This practice can cause substantial bias in the estimated classification (or reclassification) measures, especially when thresholds are chosen to maximize apparent performance (83, 198). When the choice of threshold is not prespecified, these methods are subject to ROB and this signaling question should be answered as N. The signaling question should also be answered as N when classification and reclassification measures are reported without model calibration. Before model-predicted probabilities are categorized, calibration is needed to understand whether the predicted probabilities are correct (**Box 4**).

Example. Rietveld and colleagues (89) assessed calibration using the Hosmer–Lemeshow test, which resulted in a *P* value of 0.117; they interpreted this as an indication that the model was well calibrated. If this were the only measure used to assess calibration of the model, this signaling question would be answered as N, because such a *P* value indicates neither the presence nor the magnitude of any miscalibration. However, in **Table 4**, the authors present the mean predicted probabilities with CIs across subgroups and the corresponding observed outcome frequencies. This calibration table gives an indication of the model calibration, such that the answer to the signaling question for this study would be PY.

Example. In the validation of their model for predicting early death in patients with traumatic bleeding, Perel and colleagues (88) evaluated calibration by presenting calibration plots of observed risks against predicted risks grouping by tenth of predicted risk. Presenting calibration in this format allows the reader to judge the accuracy of the model over the entire probability range. The plot could be enhanced by overlaying the figure with a nonparametric (lowess) smoother. The authors also reported a c-index, enabling readers to judge the discrimination ability of the model even without a 95% CI to indicate the uncertainty of the estimate. This study would be at low ROB and rated as Y for this signaling question.

4.8 Were model overfitting and optimism in model performance accounted for? (Model development studies only)

As discussed under signaling questions 4.1, 4.2, and 4.5, quantifying the predictive performance of a model on the same data from which the model was developed (apparent performance) tends to give optimistic estimates of performance due to overfitting–that is, the model is too much adapted to the development data set. This optimism is higher when any of the following are present: too few outcome events in total, too few outcome events relative to the number of candidate predictors (small EPV), dichotomization of continuous predictors, use of predictor selection strategies based on univariable analyses, or use of traditional stepwise predictor selection strategies (for example, forward or backward selection) in multivariable analysis in small data sets (small EPV) (49, 50).

Therefore, studies developing prediction models should always include some form of internal validation, such as bootstrapping and cross-validation. Internal validation is important to quantify overfitting of the developed model and optimism in its predictive performance, except when sample size and EPV are extremely large. Internal validation means that only the data of the original sample are used–that is, validation is based on the same participant data. If optimism is present, an important further step is to adjust or shrink the model predictive performance estimates (such as c-index) and predictor effects in the final model. Unfor-

tunately, this is not typically done. Use of regression coefficients that have not been shrunk or adjusted for optimism will lead to biased (commonly too extreme) predictions when the unshrunk model is used in other individuals. For example, a uniform (linear) shrinkage factor, as can be obtained from a bootstrap procedure, might be applied to all estimated predictor effects. Penalized regression approaches are also becoming popular, such as ridge regression and lasso regression, which allow each predictor effect to be shrunk differently and even allow exclusion of some predictors entirely (199). Some authors suggest that shrinkage methods do not differ much (200, 201), but others argue in favor of penalized approaches (49, 199).

When a prediction model is developed, the need to adjust for model overfitting and optimism is thus greater for studies with a small sample size and low EPV and those using stepwise predictor selection strategies. When internal validation and shrinkage techniques have been used, this signaling question should be answered as Y. Appropriate adjustments for overfitting alleviate ROB concerns due to the issues of low EPV (signaling question 4.1), dichotomization of continuous predictors (signaling question 4.2), and predictor selection procedures (signaling question 4.5). Studies that develop a prediction model but ignore or do not examine misfitted models should be rated as N for this signaling question, especially in the presence of small samples, low EPV, categorization of continuous predictors, or use of predictor selection strategies. An exception would be extremely large development studies with high EPV, where overfitting is of limited concern.

Some studies may use an inappropriate method to examine or adjust for optimism. Researchers often randomly split a data set at the participant level into 2 groups (1 for model development and 1 for internal validation), which has been shown to be an inadequate way to measure optimism (154, 202). Also, researchers often apply bootstrapping and cross-validation techniques to examine optimism but fail to replicate the exact model development procedure (for example, predictor selection procedures, in both univariable and multivariable analysis) and thus may underestimate the actual optimism for their model (203, 204). Such inappropriate methods would lead to an N for this signaling question.

Example. Perel and colleagues (88) examined the effect of overfitting in their model development by using bootstrapping. The authors state, "We drew 200 samples with replacement from the original data, with the same size as the original derivation data. In each bootstrap sample, we repeated the entire modelling process, including variable selection. We averaged the C statistics of those 200 models in the bootstrap samples. We then estimated the average C statistic when each of the 200 models was applied in the original sample. The difference between the two average C statistics indicated the 'optimism' of the C statistic in our prognostic model" (88). However, although the optimism in the c-statistic was examined, the optimism in absolute risk predictions was not considered, and thus no shrinkage factor was applied to the predictor coefficients. Nevertheless, the reported optimism in the c-statistic was very small (0.001), so the signaling question should be answered as PY or Y.

Example. Rietveld and colleagues' study (89) should be rated as PN or N because they did not use statistical methods to address overfitting. The authors used a predictor selection procedure based first on univariable *P* values and then on multivariable *P* values, and they also considered interactions between included predictors; thus, potential for overfitting is large. However, no examination of overfitting was made, and no attempt to shrink because of optimism was reported. The authors do report having used bootstrapping. However, this seems to be a check on the effect of outliers and estimating CIs rather than a way to examine overfitting and optimism in discrimination and calibration performance.

4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis? (Model development studies only)

Predictors and coefficients of the final developed model, including intercept or baseline components, should be fully reported to allow others to correctly apply the model to other individuals. Mismatch between the presented final model and the reported results from the multivariable analysis (such as the intercept and predictor coefficients) is frequent. A review of prediction models in cancer in 2010 found that only 13 of 38 final prediction model equations (34%) used the same predictors and coefficients as the final presented multivariable analyses, 8 used the same predictors but different coefficients, 11 used neither the same coefficients nor the same predictors, and 6 used an unclear method to derive the final prediction model from the presented results of the multivariable analysis (121).

Bias can arise when the presented final model and the results reported from the multivariable analysis do not match. One way this can occur is when nonsignificant predictors are dropped from a larger model to arrive at a final presented model but the predictor coefficients from the larger model are used to define the final model, which are no longer correct. When predictors are dropped from a larger model, it is important to reestimate all predictor coefficients of the smaller model because the latter has become the final model. These newly estimated predictor coefficients are likely different even if nonsignificant or irrelevant predictors from the larger model are dropped.

When a study reports a final model in which both predictors and regression coefficients correspond to the reported results of the multivariable regression analysis or model, this question should be answered as Y. If the final model is based only on a selection of predictors from the reported multivariable regression analysis without refitting the smaller model, it should be answered as N or PN. When no information is given

on the multivariable modeling from which predictors and regression coefficients are derived, it should be answered as NI.

This signaling question is not about detecting improper methods of selecting predictors for the final model; such methods are addressed in signaling question 4.5.

Example. Perel and colleagues (88) report the final model with odds ratios for each predictor and interaction term and the model formula with predictor coefficients. The full model would be rated as PY or Y because all predictors from the final multivariable analysis are included with coefficients derived from the multivariable analysis. Perel and colleagues also include a simplified model that was separately developed and validated, with the coefficient terms refitted in the simplified model. If instead the simplified model had not been refitted to correct coefficients for this simplified model with fewer predictors, the article would have been rated as N for this signaling question.

Example. Rietveld and colleagues (89) included all predictors in the final model in the simplified clinical score, but this score uses whole numbers to facilitate its usability. These rounded number scores are a derivative of the original weights of the predictors based on the final model: Each multivariably estimated regression coefficient was divided by the lowest regression coefficient (that is, the number 0.61, which was the regression coefficient for the predictor "itching") and then rounded to the nearest integer. However, for the predictor "two glued eyes," the coefficient of 2.707 was rounded to 5 rather than 4 (because 2.707/0.61 = 4.4). The signaling question should formally be answered as N because the assigned weights of the predictors do not completely correspond to the results in the final multivariable analysis.

Rating the ROB for domain 4. Table 10 shows how the signaling questions should be answered and an overall judgment for domain 4 reached.

Tailoring PROBAST With Additional Signaling Questions

We encourage researchers to also use PROBAST to appraise prediction model studies that consider outcome types other than binary or time-to-event outcomes (such as ordinal, nominal, or continuous outcomes) and for studies that use analysis methods other than regression-based techniques (such as tree-based or machine or artificial learning techniques). Reviewers may tailor PROBAST by adding additional signaling questions to address bias related to these other types of outcomes or modeling techniques. For example, when models for prediction of continuous outcomes are addressed, the signaling question about the number of events per studied predictor (domain 4) may be tailored to address the total number of study participants per studied predictor (49). When studies based on machine or artificial learning techniques are used, most if not all of the signaling questions will still apply. Additional questions may need to be added, because these techniques use different predictor selection strategies, predictor–outcome estimations, and methods to adjust for overfitting.

Also, when investigating studies on the added predictive value of a specific predictor to an existing model, users can add a signaling question that focuses on the methods used for quantifying added value (for example, net reclassification index or decision curve analysis) (84, 205). Similarly, when investigating studies that focus on recalibrating or updating an existing model to another setting, users can add a question on the method of recalibration or updating (for example, recalibrating the baseline risk or hazard, updating the original regression coefficients, or refitting the entire model).

Whenever reviewers tailor or add signaling questions, these need to be phrased such that the answer Y indicates low ROB and N high ROB to facilitate coherence with current signaling questions. Specific guid-

Table 11. Overall Assessment of Risk of Bias and Concerns for Applicability

Rating	Criteria
Reaching an overall judgment of risk of bias of the prediction model evaluation	
Low risk of bias	If all domains were rated low risk of bias. If a prediction model was developed without any external validation, and it was rated as low risk of bias for all domains, consider downgrading to high risk of bias. Such a model evaluation can only be considered as low risk of bias, if the development was based on a very large data set and included some form of internal validation.
High risk of bias	If ≥1 domain is judged to be at high risk of bias.
Unclear risk of bias	If an unclear risk of bias was noted in ≥1 domain and it was low risk for all other domains.
Reaching an overall judgment of concerns for applicability of the prediction model evaluation	
Low concerns for applicability	If low concerns for applicability for all domains, the prediction model evaluation is judged to have low concerns for applicability.
High concerns for applicability	If high concerns for applicability for ≥1 domain, the prediction model evaluation is judged to have high concerns for applicability.
Unclear concerns for applicability	If an unclear concern for applicability was noted in ≥1 domain and it was judged to have low concerns for applicability for all other domains.

Table 12. 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.

ance on how to assess each added signaling question should also be produced.

We do not recommend removing signaling questions from the tool unless they are clearly not relevant to a review question. If all studies would be answered as Y or N for a particular question, it is still helpful to leave the question in the tool. This shows whether a particular source of bias or concern regarding applicability is a potential problem for that review.

Step 4: Overall Judgment

Table 11 shows an overall judgment of the ROB and applicability of a prediction model evaluation. If a prediction model evaluation is judged as low on all domains relating to bias and applicability, it is appropriate to have an overall judgment of "low ROB" or "low concern regarding applicability." If an evaluation is judged as high for at least 1 domain, it should be judged as having "high ROB" or "high concern regarding applicability." If the prediction model evaluation is unclear in 1 or more domains and was rated as low in the remaining domains, it may be judged as having "unclear ROB" or "unclear concern regarding applicability."

PROBAST should not be used to generate a summary "quality score" for a study because of the well-known problems associated with such scores (206, 207). Rather than striving for a summary score, users should judge and discuss the effect of problems within each domain.

Presentation and Use of PROBAST Assessment in the Review

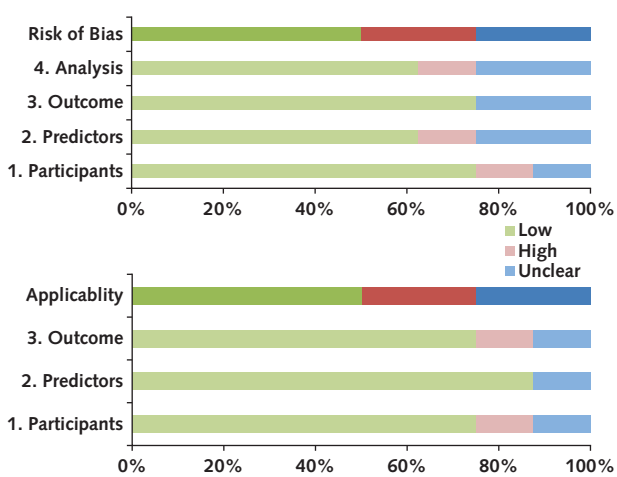
Presentation of the ROB and applicability assessment is an important aspect of communicating the strength of evidence in a review. All reviews should include a narrative summary of ROB and applicability concerns, linked to how this affects interpretation of findings and strength of inferences. In addition, a table showing the results of all assessments of ROB and applicability should be presented. Table 12 is an example to facilitate identification of key issues across all included prediction models and their studies. A graphical summary presenting the percentage of studies rated by level of concern, ROB, and applicability for each domain (Figure) can quickly sum up all studies. This is in line with item 22 of the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) statement (34, 35). These summaries are not sufficient on their own–that is, without an accompanying discussion of what any observed patterns mean for the evidence base in relation to the review question.

Further incorporation of ROB and concerns about applicability may be specified in the review planning stage or in the systematic review protocol. Users can include findings in the analysis by planning sensitivity analyses limited to studies with low concern for ROB or applicability either overall or for particular domains, or by investigating heterogeneity between studies using subgroups based on ratings of concern (17, 40, 44).

CONCLUDING REMARKS

To our knowledge, PROBAST is the first rigorously developed tool designed specifically to assess ROB and concerns regarding applicability of primary studies that develop, validate, or update (including extend) prediction models to be used for individualized predictions. PROBAST covers both diagnostic and prognostic

Figure. Suggested graphical presentation for PROBAST results.



PROBAST = Prediction model Risk Of Bias ASsessment Tool.

models, regardless of the medical domain, type of outcome, predictors, or statistical technique used.

This explanation and elaboration document provides explicit guidance on how to use PROBAST (39), including how to interpret each signaling question, grade the ROB per domain and overall, and present and incorporate PROBAST assessments into a systematic review, all accompanied with generic guidance on diagnostic and prognostic prediction model research. This detailed explanation and elaboration for PROBAST will enable a focused and transparent approach to assessing the ROB and applicability of studies developing, validating, or updating prediction models for individualized diagnostic or prognostic predictions. Five filled-in examples of PROBAST assessments, covering development studies, validation studies, and a combination of both and addressing both diagnostic and prognostic models, can be found at our Web site (www.probast.org). We also encourage and will make available translations of PROBAST.

Use of PROBAST requires the expertise and knowledge of prediction model researchers as well as clinicians. Guidance on methods for prediction model research is still at a relatively early stage compared with that for randomized intervention studies and studies of diagnostic test accuracy. We recognize that information currently necessary for assessment of bias and applicability is often not reported, and we hope that adherence of both journals and authors to the TRIPOD reporting guideline (7, 8) will reduce this problem.

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 always downloading the latest version of PROBAST and guidance from the Web site (www.probast.org).

From 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.); Kleijnen Systematic Reviews, York, United Kingdom (R.F.W., M.W.); Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Keele, United Kingdom (R.D.R.); Bristol 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 PROBAST Delphi panel (38) 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: Karel G.M. Moons, PhD, Julius Centre for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, PO Box 85500, 3508 GA Utrecht, the Netherlands; e-mail, K.G.M.Moons@umcutrecht.nl.

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

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

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

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

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

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

Final approval of the article: K.G.M. Moons, R.F. Wolff, 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: K.G.M. Moons, R.F. Wolff, J. Kleijnen, S. Mallett.

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

References

1. **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
2. **Harrell FE Jr, Lee KL, Pollock BG.** Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst. 1988;80:1198-202. [PMID: 3047407]
3. **Hlatky MA.** Evaluation of diagnostic tests. J Chronic Dis. 1986;39: 357-60. [PMID: 3700576]
4. **Laupacis A, Sekar N, Stiell IG.** Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA. 1997;277:488-94. [PMID: 9020274]
5. **Sox HC Jr.** Probability theory in the use of diagnostic tests. An introduction to critical study of the literature. Ann Intern Med. 1986; 104:60-6. [PMID: 3079637]
6. **Wasson JH, Sox HC, Neff RK, Goldman L.** Clinical prediction rules. Applications and methodological standards. N Engl J Med. 1985; 313:793-9. [PMID: 3897864]
7. **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
8. **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
9. **Altman DG.** Prognostic models: a methodological framework and review of models for breast cancer. In: Lyman GH, Burstein HJ, eds. Breast Cancer Translational Therapeutic Strategies. New York: Informa Healthcare; 2007:11-25.
10. **Kleinrouweler CE, Cheong-See FM, Collins GS, Kwee A, Thangaratinam S, Khan KS, et al.** Prognostic models in obstetrics: available, but far from applicable. Am J Obstet Gynecol. 2016;214:79-90. [PMID: 26070707] doi:10.1016/j.ajog.2015.06.013
11. **Wessler BS, Lai Yh L, Kramer W, Cangelosi M, Raman G, Lutz JS, et al.** Clinical prediction models for cardiovascular disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database. Circ Cardiovasc Qual Outcomes. 2015;8:368-75. [PMID: 26152680] doi:10.1161/CIRCOUTCOMES.115.001693
12. **Murad MH, Montori VM.** Synthesizing evidence: shifting the focus from individual studies to the body of evidence. JAMA. 2013; 309:2217-8. [PMID: 23736731] doi:10.1001/jama.2013.5616
13. **Hemingway H.** Prognosis research: why is Dr. Lydgate still waiting? J Clin Epidemiol. 2006;59:1229-38. [PMID: 17098565]
14. **Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HC.** Prognosis research: toward evidence-based results and a Cochrane methods group [Letter]. J Clin Epidemiol. 2007;60:863-5. [PMID: 17606185]
15. **Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG, et al.** Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One. 2012;7:e32844. [PMID: 22393453] doi:10.1371/journal.pone.0032844
16. **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
17. **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
18. **Deeks JJ, Wisniewski S, Davenport C.** Guide to the contents of a Cochrane diagnostic test accuracy protocol. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. London: The Cochrane Collaboration; 2013. Accessed at https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/Ch04_Sep2013.pdf on 25 November 2018.
19. **Bossuyt PM, Leeflang MM.** Developing criteria for including studies. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. London: The Cochrane Collaboration; 2008. Accessed at https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/Chapter06-Including-Studies%28September-2008%29.pdf on 25 November 2018.
20. **de Vet HCW, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D.** Searching for studies. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. London: The Cochrane Collaboration; 2008. Accessed at https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/Chapter07-Searching-%28September-2008%29.pdf on 25 November 2018.
21. **Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MM, Deeks JJ.** Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. London: The Cochrane Collaboration; 2009. Accessed at https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/ch09_Oct09.pdf on 25 November 2018.
22. **Ahmed I, Debray TP, Moons KG, Riley RD.** Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol. 2014;14:3. [PMID: 24397587] doi:10.1186/1471-2288-14-3
23. **Debray TP, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KG.** Meta-analysis and aggregation of multiple published prediction models. Stat Med. 2014;33:2341-62. [PMID: 24752993] doi:10.1002/sim.6080
24. **Debray TP, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW.** Aggregating published prediction models with individual participant data: a comparison of different approaches. Stat Med. 2012; 31:2697-712. [PMID: 22733546] doi:10.1002/sim.5412
25. **Snell KI, Ensor J, Debray TP, Moons KG, Riley RD.** Meta-analysis of prediction model performance across multiple studies: which scale helps ensure between-study normality for the C-statistic and

calibration measures? Stat Methods Med Res. 2018;27:3505-22. [PMID: 28480827] doi:10.1177/0962280217705678
26. **Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y.** Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. London: The Cochrane Collaboration; 2010. Accessed at https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/Chapter 10 - Version 1.0.pdf on 25 November 2018.
27. **Chu H, Guo H, Zhou Y.** Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. Med Decis Making. 2010;30:499-508. [PMID: 19959794] doi:10.1177/0272989X09353452
28. **Dendukuri N, Hadgu A, Wang L.** Modeling conditional dependence between diagnostic tests: a multiple latent variable model. Stat Med. 2009;28:441-61. [PMID: 19067379] doi:10.1002/sim.3470
29. **Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA.** A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8:239-51. [PMID: 16698768]
30. **Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH.** Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58:982-90. [PMID: 16168343]
31. **Rutter CM, Gatsonis CA.** A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20:2865-84. [PMID: 11568945]
32. **Takwoingi Y, Guo B, Riley RD, Deeks JJ.** Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. Stat Methods Med Res. 2017;26:1896-911. [PMID: 26116616] doi:10.1177/0962280215592269
33. **Takwoingi Y, Leeflang MM, Deeks JJ.** Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med. 2013;158:544-54. [PMID: 23546566] doi:10.7326/0003-4819-158-7-201304020-00006
34. **Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al.** The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med. 2009;151:W65-94. [PMID: 19622512]
35. **Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group.** Preferred Reporting Items for Systematic reviews and Meta-Analyses: the PRISMA statement. Ann Intern Med. 2009;151:264-9. [PMID: 19622511]
36. **McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, et al; the PRISMA-DTA Group.** Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA. 2018;319:388-96. [PMID: 29362800] doi:10.1001/jama.2017.19163
37. **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
38. **Higgins JPT, Green S, eds.** Cochrane Handbook for Systematic Reviews of Interventions. Chichester, United Kingdom: Wiley-Blackwell; 2011.
39. **Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins G, et al; PROBAST Group.** PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170:51-8. doi:10.7326/M18-1376
40. **Debray TP, Damen JA, Riley RD, Snell K, Reitsma JB, Hooft L, et al.** A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res. 2018:962280218785504. [PMID: 30032705] doi:10.1177/0962280218785504.30032705
41. **Ingui BJ, Rogers MA.** Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc. 2001;8:391-7. [PMID: 11418546]
42. **Keogh C, Wallace E, O'Brien KK, Murphy PJ, Teljeur C, McGrath B, et al.** Optimized retrieval of primary care clinical prediction rules from MEDLINE to establish a Web-based register. J Clin Epidemiol. 2011;64:848-60. [PMID: 21411285] doi:10.1016/j.jclinepi.2010.11.011
43. **Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R; Hedges Team.** Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. AMIA Annu Symp Proc. 2003:728-32. [PMID: 14728269]
44. **Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al.** External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140. [PMID: 27334381] doi:10.1136/bmj.i3140
45. **Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al.** Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. J Clin Epidemiol. 2016;69:40-50. [PMID: 26142114] doi:10.1016/j.jclinepi.2015.05.009
46. **Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al; Cochrane Bias Methods Group.** The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928. [PMID: 22008217] doi:10.1136/bmj.d5928
47. **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.
48. **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
49. **Harrell FE.** Regression Modeling Strategies, With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer; 2001.
50. **Steyerberg EW.** Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009.
51. **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
52. **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
53. **Hippisley-Cox J, Coupland C.** Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. BMJ. 2012;344:e3427. [PMID: 22619194] doi:10.1136/bmj.e3427
54. **Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al; PROGRESS Group.** Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ. 2013;346:e5793. [PMID: 23386361] doi:10.1136/bmj.e5793
55. **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
56. **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
57. **Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al.** Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98:683-90. [PMID: 22397945] doi:10.1136/heartjnl-2011-301246
58. **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

59. **Counsell C, Dennis M.** Systematic review of prognostic models in patients with acute stroke. Cerebrovasc Dis. 2001;12:159-70. [PMID: 11641579]
60. **Tamariz LJ, Eng J, Segal JB, Krishnan JA, Bolger DT, Streiff MB, et al.** Usefulness of clinical prediction rules for the diagnosis of venous thromboembolism: a systematic review. Am J Med. 2004;117:676-84. [PMID: 15501206]
61. **Veerbeek JM, Kwakkel G, van Wegen EE, Ket JC, Heymans MW.** Early prediction of outcome of activities of daily living after stroke: a systematic review. Stroke. 2011;42:1482-8. [PMID: 21474812] doi:10.1161/STROKEAHA.110.604090
62. **Leushuis E, van der Steeg JW, Steures P, Bossuyt PM, Eijkemans MJ, van der Veen F, et al.** Prediction models in reproductive medicine: a critical appraisal. Hum Reprod Update. 2009;15:537-52. [PMID: 19435779] doi:10.1093/humupd/dmp013
63. **Perel P, Edwards P, Wentz R, Roberts I.** Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak. 2006;6:38. [PMID: 17105661]
64. **Siregar S, Groenwold RH, de Heer F, Bots ML, van der Graaf Y, van Herwerden LA.** Performance of the original EuroSCORE. Eur J Cardiothorac Surg. 2012;41:746-54. [PMID: 22290922] doi:10.1093/ejcts/ezr285
65. **Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP.** Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ. 2012;344:e3318. [PMID: 22628003] doi:10.1136/bmj.e3318
66. **Tzoulaki I, Liberopoulos G, Ioannidis JP.** Assessment of claims of improved prediction beyond the Framingham risk score. JAMA. 2009;302:2345-52. [PMID: 19952321] doi:10.1001/jama.2009.1757
67. **Peters SA, den Ruijter HM, Bots ML, Moons KG.** Improvements in risk stratification for the occurrence of cardiovascular disease by imaging subclinical atherosclerosis: a systematic review. Heart. 2012;98:177-84. [PMID: 22095617] doi:10.1136/heartjnl-2011-300747
68. **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
69. **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
70. **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
71. **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
72. **Wallace E, Smith SM, Perera-Salazar R, Vaucher P, McCowan C, Collins G, et al; International Diagnostic and Prognosis Prediction (IDAPP) group.** Framework for the impact analysis and implementation of clinical prediction rules (CPRs). BMC Med Inform Decis Mak. 2011;11:62. [PMID: 21999201] doi:10.1186/1472-6947-11-62
73. **Reilly BM, Evans AT.** Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;144:201-9. [PMID: 16461965]
74. **Justice AC, Covinsky KE, Berlin JA.** Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130:515-24. [PMID: 10075620]
75. **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
76. **Austin PC, Steyerberg EW.** Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 2014;33:517-35. [PMID: 24002997] doi:10.1002/sim.5941
77. **Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW.** A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167-76. [PMID: 26772608] doi:10.1016/j.jclinepi.2015.12.005
78. **Collins GS, Ogundimu EO, Altman DG.** Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35:214-26. [PMID: 26553135] doi:10.1002/sim.6787
79. **Crowson CS, Atkinson EJ, Therneau TM.** Assessing calibration of prognostic risk scores. Stat Methods Med Res. 2016;25:1692-706. [PMID: 23907781] doi:10.1177/0962280213497434
80. **Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al.** Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128-38. [PMID: 20010215] doi:10.1097/EDE.0b013e3181c30fb2
81. **Grønnesby JK, Borgan O.** A method for checking regression models in survival analysis based on the risk score. Lifetime Data Anal. 1996;2:315-28. [PMID: 9384628]
82. **D'Agostino RB, Nam BH.** Evaluation of the performance of survival analysis models: discrimination and calibration measures. In: Balakrishnan N, Rao CR, eds. Handbook of Statistics, Survival Methods. Amsterdam: Elsevier; 2004:1-25.
83. **Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH.** Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clin Chem. 2008;54:729-37. [PMID: 18258670] doi:10.1373/clinchem.2007.096032
84. **Vickers AJ, Elkin EB.** Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-74. [PMID: 17099194]
85. **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]
86. **Aslibekyan S, Campos H, Loucks EB, Linkletter CD, Ordovas JM, Baylin A.** Development of a cardiovascular risk score for use in low- and middle-income countries. J Nutr. 2011;141:1375-80. [PMID: 21562240] doi:10.3945/jn.110.133140
87. **Han J, King NK, Neilson SJ, Gandhi MP, Ng I.** External validation of the CRASH and IMPACT prognostic models in severe traumatic brain injury. J Neurotrauma. 2014;31:1146-52. [PMID: 24568201] doi:10.1089/neu.2013.3003
88. **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
89. **Rietveld RP, ter Riet G, Bindels PJ, Sloos JH, van Weert HC.** Predicting bacterial cause in infectious conjunctivitis: cohort study on informativeness of combinations of signs and symptoms. BMJ. 2004;329:206-10. [PMID: 15201195]
90. **Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al.** Data resource profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44:827-36. [PMID: 26050254] doi:10.1093/ije/dyv098
91. **Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP, et al.** Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. J Clin Epidemiol. 2016;78:90-100. [PMID: 27045189] doi:10.1016/j.jclinepi.2016.03.017
92. **Schuit E, Groenwold RH, Harrell FE Jr, de Kort WL, Kwee A, Mol BW, et al.** Unexpected predictor-outcome associations in clinical prediction research: causes and solutions. CMAJ. 2013;185:E499-505. [PMID: 23339155] doi:10.1503/cmaj.120812
93. **Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG.** A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68:279-89. [PMID: 25179855] doi:10.1016/j.jclinepi.2014.06.018
94. **van Klaveren D, Gönen M, Steyerberg EW, Vergouwe Y.** A new concordance measure for risk prediction models in external validation settings. Stat Med. 2016;35:4136-52. [PMID: 27251001] doi:10.1002/sim.6997

95. **Kappen TH, Vergouwe Y, van Klei WA, van Wolfswinkel L, Kalkman CJ, Moons KG.** Adaptation of clinical prediction models for application in local settings. Med Decis Making. 2012;32:E1-10. [PMID: 22427369] doi:10.1177/0272989X12439755

96. **Vergouwe Y, Moons KG, Steyerberg EW.** External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172:971-80. [PMID: 20807737] doi:10.1093/aje/kwq223

97. **Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P, Ingelsson E.** Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. Am J Epidemiol. 2012;175:715-24. [PMID: 22396388] doi:10.1093/aje/kwr374

98. **Kengne AP, Beulens JW, Peelen LM, Moons KG, van der Schouw YT, Schulze MB, et al.** Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. Lancet Diabetes Endocrinol. 2014;2:19-29. [PMID: 24622666] doi:10.1016/S2213-8587(13)70103-7

99. **Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K.** Case-cohort design in practice—experiences from the MORGAM Project. Epidemiol Perspect Innov. 2007;4:15. [PMID: 18053196]

100. **Sanderson J, Thompson SG, White IR, Aspelund T, Pennells L.** Derivation and assessment of risk prediction models using case-cohort data. BMC Med Res Methodol. 2013;13:113. [PMID: 24034146] doi:10.1186/1471-2288-13-113

101. **Grobbee DE, Hoes AW.** Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research. London: Jones & Bartlett; 2009.

102. **Knottnerus JA.** The Evidence Base of Clinical Diagnosis. London: BMJ Books; 2002.

103. **Knottnerus JA, Muris JW.** Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol. 2003;56:1118-28. [PMID: 14615003]

104. **Sackett DL, Tugwell P, Guyatt GH.** Clinical Epidemiology: A Basic Science for Clinical Medicine. 2nd ed. Boston: Little, Brown; 1991.

105. **Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG.** Advantages of the nested case-control design in diagnostic research. BMC Med Res Methodol. 2008;8:48. [PMID: 18644127] doi:10.1186/1471-2288-8-48

106. **Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al.** Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6. [PMID: 10493205]

107. **Lumbreras B, Parker LA, Porta M, Pollán M, Ioannidis JP, Hernández-Aguado I.** Overinterpretation of clinical applicability in molecular diagnostic research. Clin Chem. 2009;55:786-94. [PMID: 19233907] doi:10.1373/clinchem.2008.121517

108. **Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM.** Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005;51:1335-41. [PMID: 15961549]

109. **van Zaane B, Vergouwe Y, Donders AR, Moons KG.** Comparison of approaches to estimate confidence intervals of post-test probabilities of diagnostic test results in a nested case-control study. BMC Med Res Methodol. 2012;12:166. [PMID: 23114025] doi:10.1186/1471-2288-12-166

110. **van der Leeuw J, van Dieren S, Beulens JW, Boeing H, Spijkerman AM, van der Graaf Y, et al.** The validation of cardiovascular risk scores for patients with type 2 diabetes mellitus. Heart. 2015;101:222-9. [PMID: 25256148] doi:10.1136/heartjnl-2014-306068

111. **Begg CB, McNeil BJ.** Assessment of radiologic tests: control of bias and other design considerations. Radiology. 1988;167:565-9. [PMID: 3357976]

112. **Begg CB.** Biases in the assessment of diagnostic tests. Stat Med. 1987;6:411-23. [PMID: 3114858]

113. **Elmore JG, Wells CK, Howard DH, Feinstein AR.** The impact of clinical history on mammographic interpretations. JAMA. 1997;277:49-52. [PMID: 8980210]

114. **Mackenzie R, Dixon AK.** Measuring the effects of imaging: an evaluative framework. Clin Radiol. 1995;50:513-8. [PMID: 7656516]

115. **Moons KG, Grobbee DE.** When should we remain blind and when should our eyes remain open in diagnostic studies? J Clin Epidemiol. 2002;55:633-6. [PMID: 12160909]

116. **Whiting PF, Rutjes AW, Westwood ME, Mallett S; QUADAS-2 Steering Group.** A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol. 2013;66:1093-104. [PMID: 23958378] doi:10.1016/j.jclinepi.2013.05.014

117. **Jaeschke R, Guyatt G, Sackett DL.** Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1994;271:389-91. [PMID: 8283589]

118. **Schwartz WB, Wolfe HJ, Pauker SG.** Pathology and probabilities: a new approach to interpreting and reporting biopsies. N Engl J Med. 1981;305:917-23. [PMID: 6268975]

119. **Swets JA.** Measuring the accuracy of diagnostic systems. Science. 1988;240:1285-93. [PMID: 3287615]

120. **Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM.** Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174:469-76. [PMID: 16477057]

121. **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

122. **Simon R, Altman DG.** Statistical aspects of prognostic factor studies in oncology [Editorial]. Br J Cancer. 1994;69:979-85. [PMID: 8198989]

123. **Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM.** A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009;62:797-806. [PMID: 19447581] doi:10.1016/j.jclinepi.2009.02.005

124. **van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KG, de Groot JA.** Latent class models in diagnostic studies when there is no reference standard—a systematic review. Am J Epidemiol. 2014;179:423-31. [PMID: 24272278] doi:10.1093/aje/kwt286

125. **Alonzo TA, Pepe MS.** Using a combination of reference tests to assess the accuracy of a new diagnostic test. Stat Med. 1999;18:2987-3003. [PMID: 10544302]

126. **Hui SL, Zhou XH.** Evaluation of diagnostic tests without gold standards. Stat Methods Med Res. 1998;7:354-70. [PMID: 9871952]

127. **Walter SD.** Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. Epidemiology. 1999;10:67-72. [PMID: 9888282]

128. **Lu J, Marmarou A, Lapane KL; IMPACT Investigators.** Impact of GOS misclassification on ordinal outcome analysis of traumatic brain injury clinical trials. J Neurotrauma. 2012;29:719-26. [PMID: 21815785] doi:10.1089/neu.2010.1746

129. **Lu J, Murray GD, Steyerberg EW, Butcher I, McHugh GS, Lingsma H, et al.** Effects of Glasgow Outcome Scale misclassification on traumatic brain injury clinical trials. J Neurotrauma. 2008;25:641-51. [PMID: 18578634] doi:10.1089/neu.2007.0510

130. **Ragland DR.** Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. Epidemiology. 1992;3:434-40. [PMID: 1391136]

131. **Naaktgeboren CA, Bertens LC, van Smeden M, de Groot JA, Moons KG, Reitsma JB.** Value of composite reference standards in diagnostic research. BMJ. 2013;347:f5605. [PMID: 24162938] doi:10.1136/bmj.f5605

132. **Sackett DL, Haynes RB, Tugwell P.** Clinical Epidemiology: A Basic Science for Clinical Medicine. Boston: Little, Brown; 1985.

133. **Feinstein AR.** Clinical Epidemiology: The Architecture of Clinical Research. Philadelphia: WB Saunders; 1985.

134. **Ransohoff DF, Feinstein AR.** Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299:926-30. [PMID: 692598]

135. **Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al.** Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med. 2013;10:e1001531. [PMID: 24143138] doi:10.1371/journal.pmed.1001531

136. **Naaktgeboren CA, de Groot JA, van Smeden M, Moons KG, Reitsma JB.** Evaluating diagnostic accuracy in the face of multiple reference standards. Ann Intern Med. 2013;159:195-202. [PMID: 23922065] doi:10.7326/0003-4819-159-3-201308060-00009
137. **de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N, Janssen KJ, et al.** Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ. 2011;343:d4770. [PMID: 21810869] doi:10.1136/bmj.d4770
138. **de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Brophy J, Joseph L, et al.** Adjusting for partial verification or workup bias in meta-analyses of diagnostic accuracy studies. Am J Epidemiol. 2012; 175:847-53. [PMID: 22422923] doi:10.1093/aje/kwr383
139. **Begg CB, Greenes RA.** Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983;39: 207-15. [PMID: 6871349]
140. **Harel O, Zhou XH.** Multiple imputation for correcting verification bias. Stat Med. 2006;25:3769-86. [PMID: 16435337]
141. **de Groot JA, Janssen KJ, Zwinderman AH, Moons KG, Reitsma JB.** Multiple imputation to correct for partial verification bias revisited. Stat Med. 2008;27:5880-9. [PMID: 18752256] doi:10.1002/sim .3410
142. **Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM.** Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess. 2007;11:iii. [PMID: 18021577]
143. **Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, et al.** Minimum sample size for developing a multivariable prediction model: Part I—continuous outcomes. Stat Med. 2018. [PMID: 30347470] doi:10.1002/sim.7993
144. **Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, et al.** Minimum sample size for developing a multivariable prediction model: PART II—binary and time-to-event outcomes. Stat Med. 2018. [PMID: 30357870] doi:10.1002/sim.7992
145. **van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, et al.** No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16:163. [PMID: 27881078]
146. **Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG.** Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003; 56:441-7. [PMID: 12812818]
147. **Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD.** Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making. 2001;21:45-56. [PMID: 11206946]
148. **Peduzzi P, Concato J, Feinstein AR, Holford TR.** Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48:1503-10. [PMID: 8543964]
149. **Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR.** A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49:1373-9. [PMID: 8970487]
150. **Vittinghoff E, McCulloch CE.** Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 2007;165: 710-8. [PMID: 17182981]
151. **Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV.** Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64:993-1000. [PMID: 21411281] doi:10.1016/j .jclinepi.2010.11.012
152. **Ogundimu EO, Altman DG, Collins GS.** Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol. 2016;76:175-82. [PMID: 26964707] doi: 10.1016/j.jclinepi.2016.02.031
153. **van der Ploeg T, Austin PC, Steyerberg EW.** Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14:137. [PMID: 25532820] doi:10.1186/1471-2288-14-137
154. **Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD.** Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774-81. [PMID: 11470385]
155. **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
156. **Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al.** Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98:691-8. [PMID: 22397946] doi:10.1136/heartjnl-2011-301247
157. **Altman DG, Royston P.** What do we mean by validating a prognostic model? Stat Med. 2000;19:453-73. [PMID: 10694730]
158. **Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD.** Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58:475-83. [PMID: 15845334]
159. **Altman DG, Lausen B, Sauerbrei W, Schumacher M.** Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994;86:829-35. [PMID: 8182763]
160. **Altman DG, Royston P.** The cost of dichotomising continuous variables. BMJ. 2006;332:1080. [PMID: 16675816]
161. **Royston P, Altman DG, Sauerbrei W.** Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25: 127-41. [PMID: 16217841]
162. **Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG.** Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med. 2016;35:4124-35. [PMID: 27193918] doi:10.1002/sim.6986
163. **MacCallum RC, Zhang S, Preacher KJ, Rucker DD.** On the practice of dichotomization of quantitative variables. Psychol Methods. 2002;7:19-40. [PMID: 11928888]
164. **Royston P, Sauerbrei W.** Multivariable Model-Building—A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables. Chichester, United Kingdom: Wiley; 2008.
165. **Begg CB, Greenes RA, Iglewicz B.** The influence of uninterpretability on the assessment of diagnostic tests. J Chronic Dis. 1986;39: 575-84. [PMID: 3090089]
166. **Shinkins B, Thompson M, Mallett S, Perera R.** Diagnostic accuracy studies: how to report and analyse inconclusive test results. BMJ. 2013;346:f2778. [PMID: 23682043] doi:10.1136/bmj.f2778
167. **Little RJA, Rubin DB.** Statistical Analysis With Missing Data. Hoboken, NJ: Wiley; 2002.
168. **Rubin DB.** Multiple Imputation for Nonresponse in Surveys. New York: J Wiley; 1987.
169. **Schafer JL.** Multiple imputation: a primer. Stat Methods Med Res. 1999;8:3-15. [PMID: 10347857]
170. **van Buuren S, Boshuizen HC, Knook DL.** Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18:681-94. [PMID: 10204197]
171. **White IR, Royston P, Wood AM.** Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30:377-99. [PMID: 21225900] doi:10.1002/sim.4067
172. **Donders AR, van der Heijden GJ, Stijnen T, Moons KG.** Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087-91. [PMID: 16980149]
173. **Janssen KJ, Donders AR, Harrell FE Jr, Vergouwe Y, Chen Q, Grobbee DE, et al.** Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol. 2010;63:721-7. [PMID: 20338724] doi:10.1016/j.jclinepi.2009.12.008
174. **Marshall A, Altman DG, Royston P, Holder RL.** Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Med Res Methodol. 2010;10:7. [PMID: 20085642] doi:10.1186/1471-2288-10-7
175. **Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al.** Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. [PMID: 19564179] doi:10.1136/bmj.b2393

176. **Vergouwe Y, Royston P, Moons KG, Altman DG.** Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol. 2010;63:205-14. [PMID: 19596181] doi:10.1016/j.jclinepi.2009.03.017
177. **Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG.** Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. CMAJ. 2012;184:1265-9. [PMID: 22371511] doi:10.1503/cmaj.110977
178. **Alba AC, Agoritsas T, Jankowski M, Courvoisier D, Walter SD, Guyatt GH, et al.** Risk prediction models for mortality in ambulatory patients with heart failure: a systematic review. Circ Heart Fail. 2013; 6:881-9. [PMID: 23888045] doi:10.1161/CIRCHEARTFAILURE.112 .000043
179. **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
180. **Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al.** External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40. [PMID: 24645774] doi:10.1186/1471 -2288-14-40
181. **Hussain A, Dunn KW.** Predicting length of stay in thermal burns: a systematic review of prognostic factors. Burns. 2013;39:1331-40. [PMID: 23768707] doi:10.1016/j.burns.2013.04.026
182. **Meads C, Ahmed I, Riley RD.** A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Res Treat. 2012;132:365-77. [PMID: 22037780] doi:10.1007/s10549-011-1818-2
183. **Medlock S, Ravelli AC, Tamminga P, Mol BW, Abu-Hanna A.** Prediction of mortality in very premature infants: a systematic review of prediction models. PLoS One. 2011;6:e23441. [PMID: 21931598] doi:10.1371/journal.pone.0023441
184. **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
185. **Dijk WD, Bemt Lv, Haak-Rongen Sv, Bischoff E, Weel Cv, Veen JC, et al.** Multidimensional prognostic indices for use in COPD patient care. A systematic review. Respir Res. 2011;12:151. [PMID: 22082049] doi:10.1186/1465-9921-12-151
186. **Vuong K, McGeechan K, Armstrong BK, Cust AE.** Risk prediction models for incident primary cutaneous melanoma: a systematic review. JAMA Dermatol. 2014;150:434-44. [PMID: 24522401] doi:10 .1001/jamadermatol.2013.8890
187. **Moons KG, Donders RA, Stijnen T, Harrell FE Jr.** Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59:1092-101. [PMID: 16980150]
188. **Marshall A, Altman DG, Holder RL, Royston P.** Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009;9:57. [PMID: 19638200] doi:10.1186/1471-2288-9-57
189. **Janssen KJ, Vergouwe Y, Donders AR, Harrell FE Jr, Chen Q, Grobbee DE, et al.** Dealing with missing predictor values when applying clinical prediction models. Clin Chem. 2009;55:994-1001. [PMID: 19282357] doi:10.1373/clinchem.2008.115345
190. **Jolani S, Debray TP, Koffijberg H, van Buuren S, Moons KG.** Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. Stat Med. 2015;34:1841-63. [PMID: 25663182] doi:10.1002/sim.6451
191. **Sun GW, Shook TL, Kay GL.** Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49:907-16. [PMID: 8699212]
192. **Harrell FE Jr, Lee KL, Mark DB.** Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87. [PMID: 8668867]
193. **Wolbers M, Koller MT, Witteman JC, Steyerberg EW.** Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology. 2009;20:555-61. [PMID: 19367167] doi:10.1097/EDE.0b013e3181a39056
194. **Crowther MJ, Look MP, Riley RD.** Multilevel mixed effects parametric survival models using adaptive Gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis. Stat Med. 2014;33:3844-58. [PMID: 24789760] doi:10 .1002/sim.6191
195. **Gail MH, Wieland S, Piantadosi S.** Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika. 1984;71:431e44.
196. **Greenland S, Robins MR, Pearl J.** Confounding and collapsibility in causal inference. Stat Sci. 1999;14:29-46.
197. **Wynants L, Vergouwe Y, Van Huffel S, Timmerman D, Van Calster B.** Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. Stat Methods Med Res. 2018;27:1723-36. [PMID: 27647815] doi:10.1177 /0962280216668555
198. **Ewald B.** Post hoc choice of cut points introduced bias to diagnostic research. J Clin Epidemiol. 2006;59:798-801. [PMID: 16828672]
199. **Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al.** How to develop a more accurate risk prediction model when there are few events. BMJ. 2015;351:h3868. [PMID: 26264962] doi:10.1136/bmj.h3868
200. **Janssen KJ, Siccama I, Vergouwe Y, Koffijberg H, Debray TP, Keijzer M, et al.** Development and validation of clinical prediction models: marginal differences between logistic regression, penalized maximum likelihood estimation, and genetic programming. J Clin Epidemiol. 2012;65:404-12. [PMID: 22214734] doi:10.1016/j.jclinepi .2011.08.011
201. **Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD.** Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000;19:1059-79. [PMID: 10790680]
202. **Austin PC, Steyerberg EW.** Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res. 2017;26:796-808. [PMID: 25411322] doi:10 .1177/0962280214558972
203. **Castaldi PJ, Dahabreh IJ, Ioannidis JP.** An empirical assessment of validation practices for molecular classifiers. Brief Bioinform. 2011; 12:189-202. [PMID: 21300697] doi:10.1093/bib/bbq073
204. **Varma S, Simon R.** Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. [PMID: 16504092]
205. **Pencina MJ, D'Agostino RB Sr, Steyerberg EW.** Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30:11-21. [PMID: 21204120] doi: 10.1002/sim.4085
206. **Jüni P, Witschi A, Bloch R, Egger M.** The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999;282:1054-60. [PMID: 10493204]
207. **Whiting P, Harbord R, Kleijnen J.** No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol. 2005;5:19. [PMID: 15918898]