PROBAST: A tool to assess the risk of bias and applicability of prediction model studies

Robert F. Wolff[1,#], Karel G. M. Moons[2,3,#], Richard D. Riley[4], Penny F. Whiting[5,6], Marie Westwood[1], Gary S. Collins[7], Johannes B. Reitsma[2,3], Jos Kleijnen[1,8], Susan Mallett[9] on behalf of the PROBAST group[*]

[1] Kleijnen Systematic Reviews Ltd, York, United Kingdom

[2] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

[3] Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands

[4] Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Keele, United Kingdom

[5] Bristol Medical School, University of Bristol, Bristol, United Kingdom

[6] NIIHR CLAHRC West, University Hospitals Bristol NHS Foundation Trust, Bristol, United Kingdom

[7] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Diseases, University of Oxford, Oxford, United Kingdom

[8] School for Public Health and Primary Care (CAPHRI) Maastricht University, Maastricht, The Netherlands

[9] Institute of Applied Health Research, NIHR Birmingham Biomedical Research Centre, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom

[#] Both authors contributed equally

[*] All members of the PROBAST group are listed in the appendix

Corresponding author:
Dr Robert Wolff
Kleijnen Systematic Reviews Ltd
Unit 6
Escrick Business Park
Riccall Road
Escrick
York YO19 6FD
United Kingdom
Tel. +44 (0)1904 727987
Fax. +44 (0)1904 720429
Email. robert@systematic-reviews.com

Short title: PROBAST
Word count: 3,008 words (Introduction; Methods; Results; Discussion)

Keywords: Bias (Epidemiology); Diagnosis, Evidence-Based Medicine; Multivariable Analysis; Prediction; Prognosis; Reproducibility of Results

Abstract

(309 words)

Background: Clinical prediction models combine several predictors (risk or prognostic factors) to estimate the risk whether a particular condition is present (diagnostic model) or whether a certain event will occur in the future (prognostic model). Large numbers of diagnostic and prognostic prediction model studies are published each year and a tool facilitating their quality assessment is needed, e.g. to support systematic reviews and evidence syntheses.

Objective: To introduce and describe the development of PROBAST, a tool for assessing the risk of bias and applicability of prediction model studies.

Methods: Web-based Delphi procedure (involving 40 experts in the field of prediction model research) and refinement of the tool through piloting. The scope of PROBAST was determined with consideration of existing risk of bias tools and reporting guidelines, such as CHARMS, QUADAS, QUIPS, and TRIPOD.

Results: After seven Delphi rounds, a final tool was developed which utilises a domain-based structure supported by signalling questions. PROBAST assesses the risk of bias of prediction model studies and any concerns for their applicability. Studies that PROBAST can be used for include those developing, validating, and extending a prediction model. We define risk of bias to occur when shortcomings in the study design, conduct or analysis lead to systematically distorted estimates of model predictive performance or to an inadequate model to address the research question. The predictive performance is typically evaluated using calibration and discrimination, and sometimes (notably in diagnostic model studies) classification measures. Applicability refers to the extent to which the prediction model study matches the systematic review question in terms of the target population, predictors, or outcomes of interest.
PROBAST comprises 20 signalling questions grouped into four domains: participant selection, predictors, outcome, and analysis.

Conclusions: PROBAST can be used to assess the risk of bias and any concerns for applicability of studies developing, validating or extending (adjusting) prediction, both diagnostic and prognostic, models.

Introduction

(415 words)

Prediction relates to determining the probability of something currently unknown. In the context of medical research, prediction typically relates to either diagnosis (probability of a certain condition being present but not yet detected) or prognosis (probability of developing a future outcome).(1-3) Prognosis does not only pertain to sick individuals or with an established diagnosis, but also to, for example, prognosis of pregnant women at risk of developing diabetes(4) or of individuals in the general population at risk of developing osteoporotic fractures(5).

Prediction research includes predictor finding studies, prediction model development, validation and adjusting or updating studies, and prediction model impact studies.(1) Predictor finding studies (also known as risk factor or prognostic factor studies) aim to identify which predictors independently contribute to the prediction of a diagnostic or prognostic outcome.(1, 6) Prediction model studies typically aim to develop, validate or adjust (e.g. extend) a multivariable prediction model. In a prediction model, multiple predictors are used in combination for estimating individual probabilities to inform and often guide individual care.(2, 7, 8) These models can either predict an individual's probability of currently having a particular outcome or disease (diagnostic prediction model) or experiencing a particular outcome in the future (prognostic prediction model). Well known examples are the Wells rule for diagnosing deep venous thrombosis,(9) the Ottawa ankle rules for detecting fractures,(10, 11) QRISK2 for predicting cardiovascular risk,(12) and the PCPT risk calculator for prostate biopsies.(13)

Prediction models, both diagnostic and prognostic, are widely used for a variety of medical domains and settings,(14-16) evidenced by the large number of models developed, especially in cancer,(17, 18) neurology,(19, 20) and cardiovascular disease domains.(21) Prediction models are sometimes described as risk prediction models, predictive models, prediction indices or rules, or risk scores.(2, 8) Prediction model impact studies evaluate the effect of using a model to guide patient care compared to not using such a model, and focus on the effect of its use on clinical decision making, patient outcomes, or costs of care, using a comparative design.(1)

Systematic reviews have a key role in evidence based medicine and in the development of clinical guidelines.(22-24) They are considered to provide the most reliable form of evidence for the effects of an intervention or diagnostic test.(25, 26) Review and synthesis of prediction model studies is a relatively new and evolving area. Systematic reviews of prediction models are increasingly undertaken to appraise and summarise evidence on the performance of prediction models.(1, 7, 27) They typically aim to systematically identify, appraise and summarise primary studies reporting the development or validation of one or more prediction models.(27)

Quality assessment of included studies is a crucial step in any systematic review.(25, 26) The QUIPS tool has been developed to assess the risk of bias in predictor finding (prognostic factor) studies.(28) The methodological quality of studies investigating the impact of a prediction model using a comparative randomised design can be assessed using the revised Cochrane risk of bias tool (ROB 2.0)(29) or ROBINS-I for non-randomised comparative designs.(30) With the increased numbers of systematic reviews for prediction model studies, a tool facilitating quality assessment for individual prognostic and diagnostic prediction model studies is urgently needed.

We present PROBAST (Prediction model Risk Of Bias ASsessment Tool), a tool to appraise the quality of prediction model studies. The tool allows the assessment of risk of bias and concerns for the applicability of diagnostic and prognostic prediction model studies. PROBAST can be used to assess both model development and model validation studies, including those adjusting (e.g. extending) a prediction model (Box 1). We explicitly refer to the accompanying Explanation and Elaboration (E&E) paper for detailed explanations on how to use the PROBAST tool and how to make risk of bias and applicability judgements.[REF E&E paper] To the best of our knowledge, PROBAST is the first tool which has been rigorously developed for this purpose.

Methods – Development of PROBAST

(840 words)

Development of PROBAST was based on a four-stage approach for developing health research reporting guidelines: define the scope, review the evidence base, web-based Delphi procedure, and refine the tool through piloting.(31) Guidelines explicitly aimed at the development of quality assessment tools were not available at the time.(32)

Development stage 1: Define the scope

A steering group of nine experts in the area of prediction model studies and quality assessment tool development was established. This group agreed on key features of the desired scope of PROBAST through regular teleconferences and face-to-face meetings. The scope was further refined during the web-based Delphi procedure with a panel of 40 experts.

It was agreed that PROBAST should not cover all multivariable diagnostic or prognostic studies but only primary studies that developed, validated or adjusted (e.g. extended) one or more multivariable prediction models for diagnosis or prognosis. A multivariable prediction model is defined as any combination or equation of two or more predictors for estimating the probability or risk of a diagnostic or prognostic outcome for an individual.(7, 8, 33-35) Hence, a relevant primary prediction model study was one that included a model development, model validation or model adjustment (or a combination of these) for the purpose of making individualised predictions of a diagnostic or prognostic outcome (Box 1). Diagnostic and prognostic model studies often use different terms for the predictors and outcomes (Box 2). Studies using multivariable modelling techniques to identify predictors (e.g. risk or prognostic factors) associated with an outcome but not attempting to develop, validate or adjust (e.g. extend) a model for making individualised predictions are not covered by PROBAST.(6) Hence PROBAST is not intended for predictor finding studies and prediction model impact studies.

PROBAST was designed to assess primary studies included in a systematic review. The group agreed that PROBAST would assess both the *risk of bias and concerns for applicability* of a study that evaluates (develops, validates or extends) a multivariable diagnostic or prognostic prediction model to be used for individualised predictions. A domain-based structure was adopted similar to that used in other risk of bias tools such as the revised Cochrane Risk of Bias tool,(29) QUADAS-2,(36) ROBINS-I(30) and ROBIS.(37)

Development stage 2: Review the evidence

Three different approaches were used to provide an evidence base to inform the development of PROBAST: (1) identification of relevant methodological reviews in the area of prediction model research, (2) asking members of the steering group to identify relevant methodological studies, and (3) use of the Delphi procedure to ask members of the wider group to identify additional evidence.

Identified literature was used to guide the scope and produce an initial list of signalling questions for consideration for inclusion in PROBAST.(1, 2, 6-8, 34, 35, 38-44) Signalling questions were grouped into common themes in order to identify possible domains.

Development stage 3: Web-based Delphi procedure

A modified Delphi process was used to gain feedback and agreement on the scope, structure and content of PROBAST. Web-based surveys were developed to gather structured feedback for each round. The Delphi group included 40 members comprising methodological experts in the areas of prediction model research and quality assessment tool development, experienced systematic reviewers, commissioners, and representatives of reimbursements agencies. Different potential stakeholders were included to ensure that the views of end-users, methodological experts and decision makers were represented.

The Delphi process consisted of seven rounds. Round 1 asked about the scope of the tool and it was agreed to focus on prediction model studies only and to follow a domain-based structure. Round 2 aimed at identifying and finding a consensus regarding the relevant domains to be included. The signalling questions for domains were refined in rounds 3 to 5. Respondents were asked to rate each proposed signalling question for inclusion using a 1 to 5 Likert scale. They were also given the opportunity to provide suggested rephrasing, provide any supporting evidence (e.g. references to relevant studies) and suggest any missing signalling questions. Round 6 refined the domains and introduced further optional guidance for the use of PROBAST. In the last round, participants were sent the agreed draft version of PROBAST and given the opportunity to provide any final feedback.

Development stage 4: Piloting and refining of the tool

Five workshops on PROBAST were held at consecutive annual Cochrane Colloquia (Quebec 2013, Hyderabad 2014, Vienna 2015, Seoul 2016, Cape Town 2017) and numerous consecutive workshops with MSc and PhD students (e.g. MSc Epidemiology program of Utrecht University, The Netherlands, and Evidence Based Health Care program of Oxford University, UK). In these, we piloted the then current version of the PROBAST tool to gather feedback on the practical issues associated with using the tool so we could further refine and subsequently validate the tool. Finally, over thirty review groups have already piloted PROBAST versions, included the final version, in their reviews. Topics included cancer, cardiology, endocrinology, pulmonology and orthopaedics.

All feedback received from these initiatives was used to further inform the content and structure of the PROBAST tool, wording of the signalling questions, and content of the guidance documents.[REF E&E paper]

Results – The PROBAST tool

(1,508 words)

What does PROBAST assess?

PROBAST assesses both the *risk of bias* and *concerns for applicability* of primary studies that developed or validated one or more multivariable prediction models for diagnosis or prognosis (Boxes 1 and 2). Development of a prediction model can also include adding new predictors to established predictors, i.e. the extension of an existing prediction model. Similarly, validation of an existing model can be accompanied by adjusting (updating) and also extending of the model, i.e. the development of a new model. PROBAST is also applicable to these two situations (Box 1). A multivariable prediction model is defined as any combination or equation of two or more predictors for estimating the probability or risk for an individual.(7, 8, 33-35)

Target users

PROBAST was developed with authors of systematic reviews in mind and is therefore written from that perspective. Other potential users of PROBAST include organisations supporting decision making (e.g. NICE, IQWiG), researchers and clinicians with an interest in evidence-based medicine or involved in guideline development, journal editors and manuscript reviewers.

Definition of risk of bias and applicability

Bias is usually defined as presence of systematic error within a study leading to distorted or flawed study results, hampering the internal validity of that study. In prediction model development and validation, there are known features which make a study at risk of bias, although there is limited *empirical* evidence to demonstrate the most important sources of bias. We define risk of bias to occur when shortcomings in the study design, conduct or analysis lead to systematically distorted estimates of model predictive performance or to an inadequate model to address the research question. Model predictive performance is typically evaluated using calibration and discrimination, and sometimes (notably in diagnostic model studies) classification measures.(8) To understand bias in study estimates of model predictive performance, it helps to think about how a hypothetical methodologically robust prediction model (development or validation) study would have been designed, conducted and analysed. Many sources of bias identified in other medical research areas are also relevant to prediction model studies, such as blinding of assessors of study outcomes to other features of the study, and the use of consistent definitions and measurements for predictors and outcomes within the study.

Concerns for the applicability of primary studies to the review question can arise when the study population, predictors or outcomes of a primary study differ from those specified in the review question. Applicability concerns may arise when participants in the prediction model study are from a different medical setting than the population defined in the review question. For example, participants in a primary prediction model study may be enrolled from a hospital setting but the review question specifically relates to participants in primary care. The reported prediction model discrimination and calibration may not be applicable, as patients in hospital settings typically have more severe disease than patients in primary care.(9, 45)

For systematic reviews where eligibility criteria, predictors and outcomes of the primary studies, directly match the review question, there will be no concerns for applicability of a primary study for the review. However, typically systematic reviews have inclusion criteria that are broader than the focus of the review question. The broader inclusion criteria allow for variation in the searching of the primary studies and thus require careful assessment of applicability of each primary study to the actual review question.(8)[REF E&E paper]

Types of prediction model study

A primary study identified as relevant for the review may include the development, validation or update of one or more prediction models. For each study, a PROBAST assessment should be completed for each distinct model that is developed, validated, or adjusted (e.g. extended) for making individualised predictions, relevant to the systematic review question.

PROBAST includes four steps. We stress the importance of the accompanying Explanation and Elaboration paper which provides detailed explanations and guidance carrying out each step.[REF E&E paper]

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

Step 1: Specify your systematic review question

Assessors are first asked to report their systematic review question based on the guidance given in the CHARMS checklist.(41)

Step 2: Classify the type of prediction model evaluation

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

Step 3: Assess risk of bias and applicability

Step 3 aims to identify areas where bias may be introduced into the prediction model study or where there may be concerns for applicability. It involves the assessment of four domains to cover key aspects of prediction model studies: (1) participant selection, (2) predictors, (3) outcome, and (4) analysis. The risk of bias component of each domain comprises four sections: information used to support the judgment, 20 signalling questions (2 to 9 per domain), judgment of risk of bias, and rationale regarding the judgment (Table 1).

The support for judgement box provides space to record the information used to answer the signalling questions. Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). Risk of bias is judged as "low", "high", or "unclear".' All signalling questions are phrased so that "yes" indicates absence of bias. Any signalling question rated as "no" or "probably no" flags the potential for bias; assessors will need to use their own judgment to determine whether the domain should be rated as "high", "low" or "unclear" risk of bias. A "no" rating does not automatically result in a "high" risk of bias rating. The "no information" category should be used only when insufficient information is reported to permit a judgment. By recording the rationale for the risk of bias rating, the rating will be transparent and, where necessary, facilitate discussion among review authors completing assessments independently.

The first three domains are also rated for concerns for applicability (low / high / unclear) to the review question defined above. Concerns for applicability are rated in a similar way to risk of bias but there are no signalling questions.

All domains should be completed separately for each evaluation of a distinct model in each study. The team completing a PROBAST assessment is likely to need both subject content and methodological expertise to complete an assessment. For further details on how to score risk of bias and applicability concerns we refer to the E&E paper and www.probast.org.[REF E&E paper]

- **Domain 1 (Participant selection)** covers potential sources of bias and applicability concerns related to how participants were selected for enrolment into the study and the data sources (e.g. study designs) used. Three signalling questions support the assessment of risk of bias.
- **Domain 2 (Predictors)** covers potential sources of bias and applicability concerns related to the definition and measurement of the predictors evaluated for inclusion in the prediction model. Four signalling questions support the assessment of risk of bias.
- **Domain 3 (Outcome)** covers potential sources of bias and applicability concerns related to the definition and measurement of the outcome that is predicted by the model. Seven signalling questions support the assessment of risk of bias.
- **Domain 4 (Analysis)** covers potential sources of bias regarding the statistical analysis methods. It assesses aspects related to the choice of analysis method and whether key statistical considerations (e.g. in regards to missing data) were correctly addressed. Nine signalling questions support the assessment of risk of bias.

Table 1 presents an overview of step 3. Detailed examples how to rate signalling questions and judge domains can be found in the E&E publication and on www.probast.org.[REF E&E paper]

Step 4: Overall judgement

Based on the risk of bias classifications for each domain in step 3, an *overall* judgement about risk of bias of the prediction model should be made. An overall rating of either low, high, or unclear risk of bias should be used. We recommend rating the prediction model to be of a low risk of bias if no relevant shortcomings were identified in the risk of bias assessment, i.e. all domains were rated as "low risk of bias". If at least one domain was judged to be of high risk of bias, an overall judgement of high risk of bias should be used. Similarly, unclear risk of bias should be assigned once an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.

However, if a prediction model was developed without any (external) validation and even all four domains were rated as low risk of bias, downgrading to high risk of bias should still be considered unless the model development was based on a very large data set and included some form of internal validation. For details we refer to the E&E paper.[REF E&E paper]

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

Detailed explanation and examples on how to judge the overall risk of bias and concerns for applicability can be found in the E&E publication and on www.probast.org.[REF E&E paper] Table 2 suggests a way to present the results of the PROBAST assessments.

Discussion

[245 words]

Assessment of the quality of included studies is an essential component of all systematic reviews and evidence syntheses. Systematic reviews of prediction model studies are a rapidly evolving area.(27) PROBAST is the first rigorously developed tool designed specifically to assess the quality of prediction model studies for development, validation or updating models for both diagnostic and prognostic models.

Explicit guidance and explanation about how to use PROBAST is provided in the E&E paper.[REF E&E paper] To understand and use the PROBAST tool, we stress that this E&E paper should be read in conjunction with the current paper. A multidisciplinary team, combining both subject content and methodological expertise, should ideally be used when assessing prediction model studies.

Detailed examples and future revisions of the tool will be made available via the website; please check this website to ensure that you have the current version of the tool. We welcome feedback from users via www.probast.org.

We adopted a domain based structure similar to that used in other recently developed tools such as the revised Cochrane risk of bias tool (ROB 2.0),(29) QUADAS-2 for diagnostic accuracy studies,(36) ROBINS-I for non-randomised studies,(30) and ROBIS for systematic reviews.(37) All stages of PROBAST development included a wide range of stakeholders with piloting starting with early versions of the tool allowing feedback from direct reviewer experience to be incorporated into the final tool. We feel that these two features have resulted in a tool that is both methodologically sound and user-friendly.

Acknowledgements

The authors would like to thank the members of the Delphi panel (see below) for their valuable input. Furthermore, the authors would like to thank all testers, especially C. Braun, J.G. Damen, P. Heus, L. Hooft, and R. Pajouheshnia, for providing feedback on PROBAST. The authors are grateful to Janine Ross and Steven Duffy for their support in managing the references.

KGM Moons and JB Reitsma gratefully acknowledges financial contribution by the Netherlands Organisation for Scientific Research (ZONMW 918.10.615 and 91208004). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

R Riley is a member of the Evidence Synthesis Working Group funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR) [ProjectNumber 390]. The views expressed are those of the author(s) and not necessarily those of the NIHR, the NHS or the Department of Health.

PF Whiting (time) was supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) West at University Hospitals Bristol NHS Foundation Trust.

GS Collins was supported by the NIHR Biomedical Research Centre, Oxford.

S Mallett is supported by NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham.

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

Potential Conflicts of Interest

Robert F. Wolff: None to declare

Karel G. M. Moons: None to declare

Richard D. Riley: None to declare

Penny F. Whiting: None to declare

Marie Westwood: None to declare

Gary S. Collins: None to declare

Johannes B. Reitsma: None to declare

Jos Kleijnen: None to declare

Susan Mallett: None to declare

Author Contributions

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

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

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

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

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

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

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

Administrative, technical, or logistic support: R.F. Wolff, K.G.M. Moons

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

References

1. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Medicine. 2012;9(5):1-12.
2. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Medicine. 2013;10(2):e1001381.
3. Knottnerus JA. Diagnostic prediction rules: principles, requirements and pitfalls. Prim Care. 1995;22(2):341-63.
4. Lamain-de Ruiter M, Kwee A, Naaktgeboren CA, de Groot I, Evers IM, Groenendaal F, et al. External validation of prognostic models to predict risk of gestational diabetes mellitus in one Dutch cohort: prospective multicentre cohort study. BMJ. 2016;354:i4338.
5. 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.
6. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Medicine. 2013;10(2):e1001380.
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. BMJ. 2015;350:g7594.
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. Annals of Internal Medicine. 2015;162(1):W1-73.
9. 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(2):100-7.
10. Bachmann LM, Kolb E, Koller MT, Steurer J, ter Riet G. Accuracy of Ottawa ankle rules to exclude fractures of the ankle and mid-foot: systematic review. BMJ. 2003;326(7386):417.
11. Dowling S, Spooner CH, Liang Y, Dryden DM, Friesen C, Klassen TP, et al. Accuracy of Ottawa Ankle Rules to exclude fractures of the ankle and midfoot in children: a meta-analysis. Acad Emerg Med. 2009;16(4):277-87.
12. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008;336(7659):1475-82.
13. Ankerst DP, Boeck A, Freedland SJ, Thompson IM, Cronin AM, Roobol MJ, et al. Evaluating the PCPT risk calculator in ten international biopsy cohorts: results from the Prostate Biopsy Collaborative Group. World J Urol. 2012;30(2):181-7.
14. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine. 2011;9:103.
15. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA. 2011;306(15):1688-98.
16. 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(5):e19994.
17. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. Cancer Investigation. 2009;27(3):235-43.
18. Shariat SF, Karakiewicz PI, Suardi N, Kattan MW. Comparison of nomograms with other methods for predicting outcomes in prostate cancer: a critical analysis of the literature. Clinical Cancer Research. 2008;14(14):4400-7.

19. Counsell C, Dennis M. Systematic review of prognostic models in patients with acute stroke. Cerebrovascular Diseases. 2001;12(3):159-70.

20. 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.

21. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416.

22. Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, eds. Clinical practice guidelines we can trust. Washington, DC: National Academies Press; 2011.

23. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014;129(25 Suppl 2):S49-73.

24. Rabar S, Lau R, O'Flynn N, Li L, Barry P. Risk assessment of fragility fractures: summary of NICE guidance. BMJ. 2012;345:e3698.

25. Centre for Reviews and Dissemination. Systematic Reviews: CRD's guidance for undertaking reviews in health care [Internet]. York: University of York; 2009 [accessed 27.10.17].

26. Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions [Internet]. Version 5.1.0 [updated March 2011]: The Cochrane Collaboration; 2011 [accessed 27.10.17].

27. 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.

28. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of Internal Medicine. 2013;158(4):280-6.

29. Higgins JPT, Sterne JAC, Savović J, Page MJ, Hróbjartsson A, Boutron I, et al. A revised tool for assessing risk of bias in randomized trials. In: Chandler J, McKenzie J, Boutron I, Welch V, eds. Cochrane Methods; 2016.

30. Sterne JA, Hernan 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.

31. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Medicine. 2010;7(2):e1000217.

32. Whiting P, Wolff R, Mallett S, Simera I, Savović J. A proposed framework for developing quality assessment tools. Systematic Reviews. 2017;6(1):204.

33. Canet J, Gallart L, Gomar C, Paluzie G, Valles J, Castillo J, et al. Prediction of postoperative pulmonary complications in a population-based surgical cohort. Anesthesiology. 2010;113(6):1338-50.

34. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. Journal of Clinical Epidemiology. 2013;66(3):268-77.

35. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375.

36. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine. 2011;155(8):529-36.

37. Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. Journal of Clinical Epidemiology. 2016;69:225-34.

38. Harrell FE. Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.

39. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:e5595.
40. Mallett S, Royston P, Dutton S, Waters R, Altman DG**.** Reporting methods in studies developing prognostic models in cancer: a review. BMC Medicine. 2010;8:20.
41. 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 Medicine. 2014;11(10):e1001744.
42. Altman DG, Vergouwe Y, Royston P, Moons KG**.** Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.
43. 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.
44. Royston P, Moons KG, Altman DG, Vergouwe Y**.** Prognosis and prognostic research: developing a prognostic model. BMJ. 2009;338:b604.
45. Knottnerus JA**.** Between iatrotropic stimulus and interiatric referral: the domain of primary care research. J Clin Epidemiol. 2002;55(12):1201-6.

Table 1. Overview of step 3 (Assessment of risk of bias and concerns for applicability)

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

Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). Risk of bias and concerns for applicability are rated as low, high, or unclear.
D = Development studies only; ROB = Risk of bias; V = Validation studies only

Table 2. Suggested Tabular Presentation for PROBAST Results

Study	Risk of bias				Applicability			Overall	
	Participant selection	Predictors	Outcome	Analysis	Participant selection	Predictors	Outcome	Risk of bias	Applicability
Study 1	+	-	?	+	+	+	+	-	+
Study 2	+	+	+	+	+	+	+	+	+
Study 3	+	+	+	?	-	+	+	?	-
Study 4	-	?	?	-	+	+	-	-	-
Study 5	+	+	+	+	+	?	+	+	?
Study 6	+	+	+	+	?	+	?	+	?
Study 7	?	?	+	?	+	+	+	?	+
Study 8	+	+	+	+	+	+	+	+	+

Boxes

Box 1. Types of diagnostic and prognostic modelling studies or reports addressed by PROBAST (adopted from the TRIPOD and CHARMS guidance(8, 41))

Prediction model development without (external) validation
These studies aim to develop one or more 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 individualised predictions, and quantify the predictive performance of that model in the development set. Sometimes, model development studies may also focus on adding one or more new predictors to established predictors. In any prediction model study, overfitting may occur, particularly in small data sets. Hence, development studies should ideally include some form of resampling or "internal validation" (internal because no data other than the development sample are used), 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
Studies that have the same aim as the previous type, but the development of the model is followed by quantifying the model predictive performance in data *external* to the development sample. This may be data 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. model developed in secondary care and tested in similar participants from primary care); or even in other types of participants (e.g. 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 two so created data sets only differ by chance and sample size of model development is reduced.

Prediction model (external) validation
These studies aim to assess the predictive performance of one or more existing prediction models by using other participant data that were not used (i.e. external to) in the development of the model. 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 in fact a new model is being developed after the external validation of the existing model.

Box 2. Differences between diagnostic and prognostic prediction model studies

Diagnostic prediction models aim to estimate the probability that a target condition measured using a reference standard (referred to as 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 although sometimes follow-up is used as part of the reference test to determine the target condition presence at the moment of prediction.

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

Despite the different timing of the predicted outcome, there are many similarities between diagnostic and prognostic prediction models, including:

- Type of outcome is often binary (target condition or disease presence (yes/no) or future occurrence of an outcome event (yes/no).
- 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.
- Same challenges when developing or validating a multivariable prediction model. The same measures for assessing predictive performance of the model can be used although diagnostic models more commonly 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:

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 presence/absence of target condition	Event (future occurrence yes or no) Event measurement
Missing outcome assessment	
Partial verification, lost to follow-up	Lost to follow-up and censoring

Appendix

To ensure the use of the latest version download from the website www.probast.org.

PROBAST

(Prediction model study Risk Of Bias Assessment Tool)

What does PROBAST assess?
PROBAST assesses both the *risk of bias* and *concerns for applicability* of a study that evaluates (develops, validates or adjusts) a multivariable diagnostic or prognostic prediction model. It is designed to assess primary studies included in a systematic review.

Bias occurs if systematic flaws or limitations in the design, conduct or analysis of a primary study distort the results. For the purpose of prediction modelling studies, we define risk of bias to occur when shortcomings in the study design, conduct or analysis lead to systematically distorted estimates of model predictive performance or to an inadequate model to address the research question. Model predictive performance is typically evaluated using calibration and discrimination, and sometimes (notably in diagnostic model studies) classification measures, and these are likely inaccurately estimated in studies with high risk of bias. *Applicability* refers to the extent to which the prediction model from the primary study matches your systematic review question, for example in terms of the participants, predictors or outcome of interest.

A primary study may include the development, validation or adjustment of more than one prediction model. A PROBAST assessment should be completed for each distinct model that is developed, validated or adjusted in a study, so there may be more than one PROBAST assessment for a primary study. Assessors are advised to focus only on the prediction models included in a study that are of interest for the systematic review question. Where a publication assesses multiple prediction models, only complete a PROBAST assessment for those models that meet the inclusion criteria for your systematic review. Please note that subsequent use of the term “model” includes derivatives of models, such as simplified risk scores, nomograms, or recalibrations of models.

PROBAST can be used to assess any type of diagnostic or prognostic prediction model examining individualised predictions, regardless of the predictors used, outcomes being predicted, or method to develop, validate or adjust the model.

PROBAST includes four steps.

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

If this is your first time using PROBAST, we strongly recommend reading the detailed explanation and elaboration (E&E) paper and to check the examples on www.probast.org.

Step 1: Specify your systematic review question

State your systematic review question to facilitate the assessment of the applicability of the evaluated models to your question. *The following table should be completed once per systematic review.*

Criteria	**Specify your systematic review question**
Intended use of model:	
***Participants** including selection criteria and setting:*	
***Predictors (used in modelling)** including (1) types of predictors (e.g. history, clinical examination, biochemical markers, imaging tests), (2) time of measurement, (3) specific measurement issues (e.g. any requirements/ prohibitions for specialised equipment):*	
***Outcome** to be predicted:*	

Step 2: Classify the type of prediction model evaluation

Use the following table to classify the evaluation as model development, model validation, or combination. Different signalling questions apply for different types of prediction model evaluation.
When a publication focuses on adding one or more new predictors to established predictors then use "development only". When a publication focuses on validation of an existing model in other data though followed by adjusting (updating) or extending of the model such that in fact a new model is being developed, then use "development and validation in the same publication".
If the evaluation does not fit one of these classifications then PROBAST should not be used.

Classify the evaluation based on its aim		
Type of model evaluation	**Tick as appropriate**	**PROBAST classification**
Prediction model development without testing its predictive performance in other individuals, i.e. no external validation. Model development should ideally include internal validation, such as bootstrapping or cross-validation.		Development (Dev) only
Prediction model development as well as testing of predictive performance in other individuals (external validation), both in the same publication.		Development (Dev) and external validation (Val)
Evaluating the predictive performance of a previously developed prediction model in other individuals (external validation).		External validation (Val) only

This table should be completed once for each publication being assessed and for each relevant outcome in your review.

Publication reference	
Models of interest	
Outcome of interest	

Step 3: Assess risk of bias and applicability

PROBAST is structured as four key domains. Each domain is judged for risk of bias (low, high or unclear) and includes signalling questions to help make judgements. Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). All signalling questions are phrased so that "yes" indicates absence of bias. Any signalling question rated as "no" or "probably no" flags the potential for bias; you will need to use your judgement to determine whether the domain should be rated as "high", "low" or "unclear" risk of bias. The guidance document contains further instructions and examples on rating signalling questions and risk of bias for each domain.
The first three domains are also rated for concerns for applicability (low/ high/ unclear) to your review question defined above.
Complete all domains separately for each evaluation of a distinct model. Shaded boxes indicate where signalling questions do not apply and should not be answered.

DOMAIN 1: Participant selection			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection:			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?			
1.2 Were all inclusions and exclusions of participants appropriate?			
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*		
Rationale of bias rating:			
B. Applicability			
Describe included participants, setting and dates:			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*		
Rationale of applicability rating:			

DOMAIN 2: Predictors			
A. Risk of Bias			
List and describe predictors included in the final model, e.g. definition and timing of assessment:			
		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?			
2.2 Were predictor assessments made without knowledge of outcome data?			
2.3 Are all predictors available at the time the model is intended to be used?			
Risk of bias introduced by predictors or their assessment	**RISK:** *(low/ high/ unclear)*		
Rationale of bias rating:			
B. Applicability			
Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** *(low/ high/ unclear)*		
Rationale of applicability rating:			

DOMAIN 3: Outcome			
A. Risk of Bias			
Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:			
		Dev	Val
3.1 Was the outcome determined appropriately?			
3.2 Was a pre-specified or standard outcome definition used?			
3.3 Were predictors excluded from the outcome definition?			
3.4 Was the outcome defined and determined in a similar way for all participants?			
3.5 Was the outcome determined without knowledge of predictor information?			
3.6 Was the time interval between predictor assessment and outcome determination appropriate?			
Risk of bias introduced by the outcome or its determination	**RISK:** *(low/ high/ unclear)*		
Rationale of bias rating:			
B. Applicability			
At what time point was the outcome determined: *If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:*			
Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** *(low/ high/ unclear)*		
Rationale of applicability rating:			

<table>
<tr><td colspan="4"><b>DOMAIN 4: Analysis</b></td></tr>
<tr><td colspan="4"><b>Risk of Bias</b></td></tr>
<tr><td colspan="4"><i>Describe numbers of participants, number of candidate predictors (for DEV only), outcome events and events per candidate predictor (for DEV only):</i></td></tr>
<tr><td colspan="4"><i>Describe how the model was developed (predictor selection, optimism, risk groups, model performance):</i></td></tr>
<tr><td colspan="4"><i>Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants):</i></td></tr>
<tr><td colspan="4"><i>Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit:</i></td></tr>
<tr><td colspan="4"><i>Describe any participants who were excluded from the analysis:</i></td></tr>
<tr><td colspan="4"><i>Describe missing data on predictors and outcomes as well as methods used for missing data:</i></td></tr>
<tr><td colspan="2"></td><td>Dev</td><td>Val</td></tr>
<tr><td colspan="2">4.1 Were there a reasonable number of participants with the outcome?</td><td></td><td></td></tr>
<tr><td colspan="2">4.2 Were continuous and categorical predictors handled appropriately?</td><td></td><td></td></tr>
<tr><td colspan="2">4.3 Were all enrolled participants included in the analysis?</td><td></td><td></td></tr>
<tr><td colspan="2">4.4 Were participants with missing data handled appropriately?</td><td></td><td></td></tr>
<tr><td colspan="2">4.5 Was selection of predictors based on univariable analysis avoided?</td><td></td><td>▨</td></tr>
<tr><td colspan="2">4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?</td><td></td><td></td></tr>
<tr><td colspan="2">4.7 Were relevant model performance measures evaluated appropriately?</td><td></td><td></td></tr>
<tr><td colspan="2">4.8 Was model overfitting and optimism in model performance accounted for?</td><td></td><td>▨</td></tr>
<tr><td colspan="2">4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?</td><td></td><td>▨</td></tr>
<tr><td><b>Risk of bias introduced by the analysis</b></td><td><b>RISK:</b>
<i>(low/ high/ unclear)</i></td><td></td><td></td></tr>
<tr><td colspan="4"><i>Rationale of bias rating:</i></td></tr>
</table>

Step 4: Overall assessment

Use the following tables to reach overall judgements about risk of bias and concerns for applicability of the prediction model evaluation (development and/or validation) across all assessed domains.
Complete for each evaluation of a distinct model.

Reaching an overall judgement about 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 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 at least one domain is judged to be at **high risk of bias**.
Unclear risk of bias	If an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.

Reaching an overall judgement about 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 at least one domain, the prediction model evaluation is judged to have **high concerns for applicability**.
Unclear concerns for applicability	If unclear concerns (but no "high concern") for applicability for at least one domain, the prediction model evaluation is judged to have **unclear concerns for applicability** overall.

Overall judgement about risk of bias and applicability of the prediction model evaluation		
Overall judgement of risk of bias	**RISK:** *(low/ high/ unclear)*	
Summary of sources of potential bias:		
Overall judgement of applicability	**CONCERN:** *(low/ high/ unclear)*	
Summary of applicability concerns:		

Members of PROBAST Delphi group

Doug Altman, University of Oxford, United Kingdom
Patrick Bossuyt, University of Amsterdam, The Netherlands
Gary S. Collins*, University of Oxford, United Kingdom
Nancy R. Cook, Harvard University, United States of America
Gennaro D´Amico, Ospedale V Cervello, Italy
Thomas P. A. Debray, Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Jon Deeks, University of Birmingham, United Kingdom
Joris de Groot, University of Utrecht, The Netherlands
Emanuele di Angelantonio, University of Cambridge, United Kingdom
Tom Fahey, Royal College of Surgeons in Ireland, Ireland
Paul Glasziou, Bond University, Australia
Frank Harrell, Vanderbilt University, United States of America
Jill A. Hayden, Dalhousie University, Canada
Martijn W. Heymans, Department of Epidemiology and Biostatistics, Amsterdam Public Health research institute, VU University Medical Center, Amsterdam, The Netherlands
Lotty Hooft, Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Chris Hyde, Peninsula Technology Assessment Group, United Kingdom
John Ioannidis, Stanford University, United States of America
Alfonso Iorio, McMaster University, Canada
Stephen Kaptoge, University of Cambridge, United Kingdom
Jos Kleijnen*, Kleijnen Systematic Reviews, United Kingdom
André Knottnerus, Maastricht University, The Netherlands
Mariska Leeflang, University of Amsterdam, The Netherlands
Susan Mallett*, University of Birmingham, United Kingdom
Carl Moons*, Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Frances Nixon, National Institute for Health and Care Excellence (NICE), United Kingdom
Michael Pencina, University of Boston, United States of America
Pablo Perel, London School of Hygiene and Tropical Medicine, United Kingdom
Bob Phillips, Centre for Reviews and Dissemination (CRD), York, United Kingdom
Heike Raatz, Kleijnen Systematic Reviews, United Kingdom
Johannes B. Reitsma*, University of Utrecht, The Netherlands
Rob Riemsma, Kleijnen Systematic Reviews, United Kingdom
Richard Riley*, Keele University, United Kingdom
Maroeska Rovers, University of Utrecht, The Netherlands
Anne W. S. Rutjes, Institute for Social and Preventive Medicine (ISPM) and Institute of Primary Health Care (BIHAM), University of Bern, Switzerland
Willi Sauerbrei, University of Freiburg, Germany
Stefan Sauerland, Institute for Quality and Efficiency in Healthcare (IQWiG), Germany
Fülöp Scheibler, University Medical Center Schleswig-Holstein, Germany
Rob Scholten, Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Ewoud Schuit, Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Ewout Steyerberg, Erasmus University Medical Center Rotterdam and Leiden University Medical Center, The Netherlands
Toni Tan, National Institute for Health and Care Excellence (NICE), United Kingdom
Gerben ter Riet, Department of General Practice, University of Amsterdam, The Netherlands
Danielle van der Windt, Keele University, United Kingdom
Yvonne Vergouwe, Erasmus University Medical Center, Rotterdam, The Netherlands
Andrew Vickers, Memorial Sloan-Kettering Cancer Center, United States of America
Marie Westwood*, Kleijnen Systematic Reviews, United Kingdom
Penny Whiting*, University of Bristol, United Kingdom
Robert Wolff*, Kleijnen Systematic Reviews, United Kingdom
Angela M. Wood, University of Cambridge, United Kingdom
** denotes members of the PROBAST steering group*