PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods

Karel G M Moons,[1] Johanna A A Damen,[1,2] Tabea Kaul,[1] Lotty Hooft,[1,2] Constanza Andaur Navarro,[1] Paula Dhiman,[3] Andrew L Beam,[4] Ben Van Calster,[5,6] Leo Anthony Celi,[7,8] Spiros Denaxas,[9,10] Alastair K Denniston,[11] Marzyeh Ghassemi,[12] Georg Heinze,[13] André Pascal Kengne,[14] Lena Maier-Hein,[15,16] Xiaoxuan Liu,[11,17,18,19] Patricia Logullo,[3] Melissa D McCradden,[20] Nan Liu,[21] Lauren Oakden-Rayner,[22] Karandeep Singh,[23] Daniel S Ting,[21,24] Laure Wynants,[5,25] Bada Yang,[1,2] Johannes B Reitsma,[1] Richard D Riley,[18,19] Gary S Collins,[3] Maarten van Smeden[1]

For numbered affiliations see end of the article

Correspondence to:
K G M Moons
k.g.m.moons@umcutrecht.nl
(or @carlmoons on X;
ORCID 0000-0003-2118-004X)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2025;388:e082505
http://dx.doi.org/10.1136/bmj-2024-082505

Accepted: 16 January 2025

The Prediction model Risk Of Bias ASsessment Tool (PROBAST) is used to assess the quality, risk of bias, and applicability of prediction models or algorithms and of prediction model/algorithm studies. Since PROBAST's introduction in 2019, much progress has been made in the methodology for prediction modelling and in the use of artificial intelligence, including machine learning, techniques. An update to PROBAST-2019 is thus needed. This article describes the development of PROBAST+AI. PROBAST+AI consists of two distinctive parts: model development and model evaluation. For model development, PROBAST+AI users assess quality and applicability using 16 targeted signalling questions. For model evaluation, PROBAST+AI users assess the risk of bias and applicability using 18 targeted signalling questions. Both parts contain four domains: participants and data sources, predictors, outcome, and analysis. Applicability of the prediction model is rated for the participants and data sources, predictors, and outcome domains. PROBAST+AI may replace the original PROBAST tool and allows all key stakeholders (eg, model developers, AI companies, researchers, editors, reviewers, healthcare professionals, guideline developers, and policy organisations) to examine the quality, risk of bias, and applicability of any type of prediction model in the healthcare sector, irrespective of whether regression modelling or AI techniques are used.

SUMMARY POINTS

PROBAST (Prediction model Risk Of Bias ASsessment Tool), launched in 2019, assesses the risk of bias and applicability of prediction models and prediction model studies

In response to feedback from multiple users, advances in prediction modelling and artificial intelligence (AI)/machine learning methods, and numerous recent reviews indicating poor quality of AI/machine learning based prediction model studies, an update of PROBAST-2019 was necessary

The update was also needed to better address all novel and necessary methodological considerations for a broader set of modelling approaches than only prevailing statistical techniques

PROBAST+AI extends and replaces PROBAST-2019

The updated tool allows all key stakeholders (eg, prediction model developers, readers, editors, healthcare professionals, and health policy organisations) to examine the quality, risk of bias, and applicability of any type of prediction model study in the healthcare sector, regardless of data analytical (prevailing statistical or AI/machine learning) techniques used

The original PROBAST Explanation and Elaboration document still provides a comprehensive background for PROBAST+AI

In healthcare, prediction models or algorithms (hereafter referred to as prediction models) estimate the probability of a health outcome for individuals. In the diagnostic setting—including screening and monitoring—the model typically aims to predict or classify the presence of a particular outcome, such as a disease or disorder. In the prognostic setting the model aims to predict a future outcome—typically health related—in patients with a diagnosis of a particular disease or disorder, or in the general population. The primary use of a prediction model in healthcare is to support individual healthcare counselling and shared decision making on, for example, subsequent medical

testing, referral to another healthcare professional or facility, treatment, discharge from hospital, or lifestyle changes. For example, the tool QR4 predicts the probability of developing a cardiovascular event within the next 10 years and informs whether individuals should undergo changes to their lifestyle or be prescribed drugs.[1] Prediction models are developed for, and used, in all healthcare settings and for all medical conditions to predict all types of outcomes. Thousands of models are published annually in the healthcare domain to predict the same health condition or the same types of outcomes, often for the same target population.[2-5] For example, within the first 15 months of the covid-19 pandemic, 381 prognostic prediction models for the disease were published.[5]

For decades, traditional statistical modelling approaches, in particular regression modelling, have been the prevailing approaches when developing prediction models. In more recent years, however, interest has increased in other analytical approaches, such as artificial intelligence (AI), including machine learning, techniques. Popular examples of such AI/machine learning methods are support vector machines, tree based learning (eg, random forests), and neural networks (including deep learning).[6] As software and high computational power, such as through cloud computing, has become increasingly accessible, the development of prediction models in the healthcare domain using AI/machine learning methods has become even more overwhelming. The ease with which prediction models can now be developed has contributed to their vast numbers mentioned in the biomedical literature. Also, many changes have occurred in the infrastructure for healthcare data, such as the increasing use of electronic health records and advances in natural language processing to make use of unstructured data from these records. This all has resulted in large amounts of data further facilitating the development and training or evaluation of prediction models, with both prevailing statistical and AI/machine learning techniques.

Despite the abundance of prediction models in the healthcare literature and guidelines, numerous reviews in the past two decades have shown that most of the published models, including those based on AI/machine learning methods, are of poor quality, reported predictive performances are at high risk of bias,[2-5 7 8] and fairness related issues affect the predictive performance of models related to certain groups.[9 10] Prediction model studies including AI/machine learning based prediction model studies, also systematically are beset by overinterpretation (otherwise known as spin) of the applicability, validity, and usefulness of the resulting models.[11 12] Furthermore, poor science practices and inefficient translation of poorly performing models lead to research waste.[13] All these issues are compounded by lack of scrutiny and oversight because the use of prediction models is largely unregulated and non-standardised. Accordingly, all these issues cast doubt on the validity and accuracy and thus the safety and applicability of prediction models in medical guidelines or healthcare practice, and they potentially create or further widen existing inequities in healthcare.[14 15]

In response to these developments, the Cochrane Prognosis Methods group was established in 2008, with a focus on systematic reviews of prognosis, including prediction model, studies.[16-18] To facilitate the appraisal of prediction models , the Prediction model Risk Of Bias ASsessment Tool (PROBAST; www.probast.org) for the appraisal of model development and evaluation (validation) studies in the healthcare domain, was published in 2019.[19 20] Risk of bias refers to the potential for a systematic error (bias) in the estimators of the model's predictive performance for the target population or populations of interest. Bias can act in either direction, with potential for overestimation or underestimation of the true model performance. Applicability refers to whether a prediction model or its study is relevant to the assessor's review question or to the assessor's intended use of the prediction model, including the target population and setting. Several tools are available for assessing methodological quality and applicability of diagnostic and prognostic tests and models. These were recently summarised and accompanied by a decision tree to determine which quality assessment tool to use for which context, purpose, and situation.[21]

PROBAST assesses the risk of bias and applicability using 20 signalling questions across four domains: participants, predictors, outcome, and analysis.[19 20] The tool enables a focused and transparent approach to assessing risk of bias and applicability of studies that develop, update, or evaluate (validate) the performance of a prediction model. PROBAST was accompanied by a detailed explanation and elaboration document providing the rationale behind each domain and signalling question, examples of how to use the tool, and a discussion of issues causing concerns about risk of bias and applicability in prediction model studies.[19] Additional guidance is available in other methodological papers.[18 22-25]

Advances in AI/machine learning methods and extensive feedback from numerous PROBAST users plus evaluation of its use among hundreds of users, necessitated an update of PROBAST to allow additional data and appraisal of prediction models' quality, and to better address the methodological considerations for a broader set of modelling approaches given the uptake in AI/machine learning based prediction models.[26] For example, inherently different approaches to handling predictors in tree based learning and neural networks, or the often wrongly used methods to address the so called imbalance between classes in a dataset, were not dealt with in the original PROBAST (PROBAST-2019). Moreover, in recent years, important methodological advances have taken place, including guidance on appropriate sample size for developing[27-31] and evaluating[32-34] prediction models using either regression based or AI/machine learning techniques. Finally, the introduction of AI/machine learning based prediction models has been accompanied by

Box 1: Glossary of terms*

Algorithmic bias
When the predictions or classifications by the algorithm (model) benefit or disadvantage certain groups of individuals, without a justified reason for such unequal impacts.

Apparent performance
A type of model performance evaluation. In apparent performance, model (prediction or classification) performance is estimated using the same data as used for model development.

Applicability
Whether the study in question is relevant (applicable) to the assessor's review question or the assessor's intended use of a model, including target population and setting.

Artificial intelligence
Many definitions of AI exist, some of which are extensive and complex (eg, European AI Act[40]). In the context of prediction in healthcare, the term AI is commonly used for statistical learning approaches that do not fall under the family of generalised linear models (eg, logistic regression) or survival modelling (eg, Cox regression). For example, analytical models are commonly referred to as AI if they are based on support vector machines, tree based learning (eg, random forests), and neural networks (eg, deep learning). In this context, machine learning is often used synonymously. Any strict distinction between statistical versus AI/machine learning, however, quickly becomes a false dichotomy.[39]

Calibration
The agreement between the model's estimated probabilities and the observed outcome probabilities. Calibration is typically assessed graphically using a plot of the observed outcome values on the y axis and the estimated outcome values on the x axis, with a calibration curve for individual data.

Data leakage
When data from the model development phase are somehow inadvertently included during the model evaluation (testing) phase, typically leading to overly optimistic performance estimates or inaccurate predictions or classifications of the model. This occurs because the evaluated model has learnt from information in the leaked data.

Development or training data
The data used to build or fit (referred to as develop or train) the prediction model.

Discrimination
How well the estimated outcome values from the model differ from those with observed outcome values. Discrimination is typically quantified by the C index for time-to-event outcomes and the C statistic (sometimes referred to as the area under the receiver operating characteristic curve) for binary outcomes.

Evaluation or test data
The data used to estimate the prediction or classification performance of a prediction model, sometimes referred to as test data or validation data. Evaluation data should ideally be different from the data used to train the model, do model selection, or tune hyperparameters, such that participants do not overlap between the training and evaluation data (see also data leakage).

External validation
A type of model performance evaluation. In external validation, model performance is estimated using participant data that were not used for development (including internal validation) of the model.

Fairness
Property of prediction models that do not disadvantage groups of people based on characteristics such as age, sex or gender, race or ethnicity, or socioeconomic status.

Feature
Measurable property that is used as input for a prediction model. In this paper consistently referred to as predictor.

Hyperparameters
Values that control the model development or learning process.

Hyperparameter tuning
Finding the optimal settings of hyperparameters and parameters in the building strategy for the prediction model.

Imprecision
When a model's performance estimate is based on a small evaluation sample, leading to wide confidence intervals of the performance estimates.

Internal validation or evaluation
A type of model performance evaluation. The process of assessing a prediction model's performance using some form of splitting, resampling, or cross validation technique on the development dataset.

Machine learning
A subspecialty of AI that focuses on developing models that are capable of learning and making predictions or decisions from data, without being explicitly programmed. In the context of prediction models in health, machine learning is often used synonymously with AI.

(Continued)

Box 1: (Contonued)

Model evaluation
Evaluating the predictive performance of a model by estimating, for example, its overall predictive accuracy (eg, Brier score), model discrimination (eg, C statistic), model calibration (eg, calibration plot, calibration slope), and clinical usefulness (eg, decision curve analysis). Evaluation types can include the assessment of the model's apparent performance, internal validation performance, and external validation performance.

Outcome
The diagnostic or prognostic health state or value, or their probabilities that are being predicted. In machine learning, this is often referred to as the target value or response variable.

Predictor
A characteristic that can be measured or attributed at an individual level (such as age, sex, systolic blood pressure, disease stage), or group level (eg, country). A predictor is often referred to as a feature, input, independent variable, or covariate.

Risk of bias
The potential for a systematic error (bias) in the estimators of the model's predictive performance for the target population. Bias can act in either direction, such that overestimation or underestimation of the true model performance might occur.

Data preprocessing
Typical preparatory step for predictors before data analysis. For example, transforming a continuous predictor or outcome, categorising or recategorising a predictor or outcome, or collapsing rare predictor or outcome categories.

Validation data
Validation data can have different meanings. Typically, in the medical literature these refer to data that are not used for model development (or training) but are only used to evaluate (validate) a model's (predictive or classification) performance, often referred to as external validation. The differences between internal and external validation are explained above and in the main text. In the computer science literature, validation data typically refer to data that have been held back and used after the model development phase for parameter or hyperparameter tuning of the model, that will then go forward for a model's predictive performance evaluation. To avoid any ambiguity and harmonise terminology, in this paper validation data refers to any data used to evaluate a model's performance.

AI=artificial intelligence; PROBAST=Prediction model Risk Of Bias ASsessment Tool; TRIPOD=Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

*Definitions relate to the specific context of, and use of these terms in, PROBAST+AI and therefore are not necessarily applicable to other areas of research. Developed based on TRIPOD+AI statement by Collins et al.[38]

an important recognition of the concerns about the models' fairness/unfairness, discrimination,[6 10-12 35 36] and reproducibility.[37] An update of PROBAST-2019 was therefore considered necessary to reflect these latest developments and to capture the potential consequences for the quality, risks of bias, and applicability assessment of prediction models in the healthcare domain, regardless of the data analytical method that was used for the prediction modelling (ie, prevailing statistical or AI/machine learning techniques).

This paper describes the process for updating PROBAST-2019 to PROBAST+AI, presents the PROBAST+AI tool, and provides guidance on how to use the tool. PROBAST+AI harmonises the landscape of quality assessment of any type of prediction or classification model and algorithm in healthcare. Consistent with the TRIPOD+AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; www.tripod-statement.org) reporting guideline,[38] the + sign indicates that PROBAST+AI provides consolidated recommendations regardless of whether regression models or AI methods have been used.[39] This nomenclature circumvents false dichotomies between statistical and AI (including machine learning) techniques. We also use the suffix +AI to the PROBAST acronym to be consistent with existing guidelines for studies broadly labelled as "involving AI," with machine learning as the most prominent class of AI for prediction modelling in the healthcare domain (see box 1). PROBAST+AI is not only useful for assessing the applicability, quality, and risk of bias of prediction model studies when conducting a systematic review. It can also be used generally, by all key stakeholders (eg, model developers, AI companies, researchers, editors, reviewers, healthcare professionals, guideline developers, health policy organisations, and ethical review boards) in their critical appraisal, use, implementation, and uptake of prediction models in healthcare, without conducting an explicit systematic review (see table 1). Box 1 provides a glossary of terms for key concepts used in the specialty of prediction modelling.

Development of PROBAST+AI

A working group with extensive experience in prediction model research (using statistical or AI/machine learning methods), systematic reviews, and application of PROBAST-2019 was formed (KGMM, JAAD, TK, CAN, PD, LH, JBR, RDR, GSC, and MvS) to oversee the developmental process for PROBAST+AI. The protocol for updating PROBAST-2019 has been published[41] and is also available on the Open Science Framework.[42] On the PROBAST website (www.probast.org) in August 2019, we announced a large international project comprising a series of systematic

Table 1 | Users/stakeholders, actions, and potential benefits of PROBAST+AI

Users/stakeholders	Proposed actions	Potential benefits
Academic institutions	Promote or require adherence to PROBAST+AI by investigators developing, evaluating, assessing, or implementing prediction models Provide training to early career researchers on the importance of methodological quality assessment, including requiring doctoral students to adhere to the quality criteria underlying PROBAST+AI	Enhances transparency in the design, analysis, and reporting of prediction model research
		Improves quality, accountability, reproducibility, replicability, fairness, and usefulness of produced research
		Avoids research waste
Researchers	Adhere to quality criteria underlying PROBAST+AI when developing or evaluating prediction models	Improves methodological quality
		Increases knowledge of the minimal quality criteria required and expected when performing a prediction model study
		Improves accountability, reproducibility, replicability, fairness and usefulness of produced research
		Avoids research waste
Systematic reviewers and meta-researchers	Use PROBAST+AI to assess quality, risk of bias, and applicability of prediction models	Can be used to assess study quality (eg, design, methods) and applicability of prediction models
		Increases trust in research findings
		Improves quality, accountability, reproducibility, replicability, fairness, and usefulness of published research
Journal editors	Recommend or mandate authors to use PROBAST+AI and submit a completed methodological quality assessment as part of a systematic review/meta-analysis Recommend peer reviewers to use PROBAST+AI to assess the methodological quality of prediction models for studies developing/validating a model	Improves understanding of journal requirements and expectations for prediction model publications
		Increases efficiency of peer review resulting from improved author understanding of journal requirements for prediction model publications
		Improves quality, accountability, reproducibility, replicability, fairness, and usefulness of published research
Peer reviewers	Use PROBAST+AI to assess the methodological quality of prediction models	Improves efficiency of peer reviews
		Facilitates and directs specific feedback to authors on where important details are missing
Commercial manufacturers of prediction models	Adhere to quality criteria underlying PROBAST+AI when developing or evaluating a particular prediction model	Improves methodological quality
		Increases awareness of the minimal quality criteria required and expected when developing a prediction model
		Improves accountability, reproducibility, replicability, fairness, and usefulness of produced models
Funders	Recommend or mandate use of quality criteria established in PROBAST+AI by investigators when reviewing a grant for prediction model research	Increases usefulness and fairness of research findings
		Reduces avoidable research waste due to inadequate methodological quality
		Ensures that funded research can be used by others
Policy makers	Use or promote PROBAST+AI to ensure research is methodologically sound	Ensures decisions to implement a prediction model are based on adequate methodology
		Adds integrity for evidence based policy recommendations
Regulators	Clinical reviewers use PROBAST+AI to assess adequate methodological quality for "software as medical device" regulatory submissions when the operating principle of the product is a prediction model	Aligns reported intended use with regulatory intended purpose
		Aligns medical device regulatory review with pivotal investigational reporting
		Encourages manufacturers to publish clinical investigation reports by encouraging one common standard
Healthcare professionals	Verify whether a prediction model meets methodological standards and whether the model is applicable before purchasing or using a model to support clinical use	Improves understanding of the target population of a model and the clinical decision for which it is intended
		Improves understanding of model predictions and awareness of limitations
		Improves trust in research findings
Institutional ethical review boards	Verify whether a proposal for a prediction model (study) meets the required methodological standards	Improves quality of prediction model development and evaluation studies
Patients, public, study participants	Understand and advocate use of PROBAST+AI by authors, peer reviewers, journals, and funders	Improves trust in research findings
		Improves understanding of prediction model research
		Promotes health equity considerations in research

Developed based on PROBAST-2019[19 20] and the TRIPOD+AI statement by Collins et al.[38]
AI=artificial intelligence; PROBAST=Prediction model Risk Of Bias ASsessment Tool; TRIPOD=Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

reviews on the methodological (including risk of bias) and reporting quality of prediction models, including AI/machine learning based prediction models, published in the specialty of cancer as well as in the generic healthcare literature. Furthermore, after previous research on the interrater agreement with PROBAST-2019,[43 44] we conducted a comprehensive study on the application of the tool.[26] All these findings informed the update of PROBAST-2019.[6 11 12 19-21 26 45 46]

These reviews consistently showed that most studies on AI based prediction models were poorly reported and poorly conducted, with both the development and the evaluation of such models rated at high risk of bias.[6 11 12 45 46] Many factors contributed to this high risk of bias, including small sample sizes, poor handling of missing data, failure to deal with model overfitting, and lack of adequate assessment of predictive performance. The conclusions from these reviews were supported by

many other prediction model reviews that have since been published.[32 47-60]

Generation of candidate domain and signalling questions

PROBAST-2019 was used as the starting point.[19 20] It included 20 signalling questions across four domains: participants, predictors, outcome, and analysis. We anticipated that these domains and signalling questions would already largely apply to AI/machine learning based prediction model studies as well as regression based approaches.

In October 2020, we conducted a survey among more than 50 authors who had participated in an earlier living review on prediction models for covid-19.[5] Each of these authors had applied PROBAST-2019 on multiple occasions, including on papers with AI/machine learning based prediction models. The survey asked about their experiences of PROBAST-2019, its applicability to AI/machine learning based prediction model studies, suggestions for improvement or changes in wording of existing domains and signalling questions, and whether domains or signalling questions needed to be deleted or added. All suggestions and recommendations were harmonised into an initial list of 26 candidate signalling questions distributed over four risk of bias domains, then labelled as participants and data sources, predictors, outcomes, and analysis. The domains of participants and data sources, predictors, and outcomes each included an applicability subdomain. This initial list served as the basis for the first round of the online large scale Delphi survey.

Recruitment of Delphi participants

The PROBAST+AI working group identified participants from authors of relevant publications, through social media (eg, X, formerly Twitter), and based on personal recommendations of Delphi participants (snowballing recruitment). Participants were recruited covering all key stakeholder groups (table 1) from a range of settings (eg, university, primary care, hospital, biomedical journal, patient and non-profit organisations, and for-profit organisations), and bearing in mind geographical and other aspects of diversity.

Delphi process

We designed and shared the Delphi surveys electronically using the REDCap online platform (www.projectredcap.org), and later the Castor online platform (www.castoredc.com). Owing to a change within University Medical Centre Utrecht's system, we needed to switch data capturing programs after the first round.

In the first round of the Delphi process, we asked for each of the initial 26 candidate items to be rated. For each signalling question, we asked participants whether they strongly disagreed, disagreed, neither agreed nor disagreed, agreed, or strongly agreed to its inclusion in the development of PROBAST+AI. Participants were also invited to comment on any domain or signalling question, and to suggest new items. For an item to be included, a level of agreement (response to strongly agree or agree) of ≥80% had to be achieved. For items that did not surpass the 80% threshold but were identified as essential, we checked whether a reformulation of the signalling question—that is, based on received comments and expert opinion—led to a higher result in the next round. CAN, JAAD, and TK analysed the narrative responses, with agreement from KGMM, MvS, JBR, LH, GSC, RDR, and PD. After each round, we presented the aggregated quantitative results for agreement to the participants of the next round. Responses were anonymous.

Round 1

The first round was opened from 12 July to 12 September 2021. Of 201 people invited to participate, 95 completed the survey (see supplementary figure 1). Panellists were based in various countries and represented six continents.

Round 2

The second round was opened from 20 January to 10 March 2023. All participants who completed the first round were invited to the second round. Those who did not respond to the first round were reinvited. To improve diversity of the participants due to expertise (eg, experts in AI, algorithmic bias, or fairness) or geographical location, additional participants who were identified or recommended after the first round were also invited. These additional participants were identified by their participation in the development of TRIPOD+AI[38] or had contacted us after some PROBAST+AI coauthors had advertised the survey through X (formerly Twitter). Of 294 people invited to participate in the second round, 144 responded to the survey, including 12 who provided partial responses (see supplementary information, figure 1 and table 1).

In the second round, participants were given a summary of and link to the aggregated ratings from the first round (available at doi:10.17605/OSF.IO/W3CFE). The second round included four domains, like those in PROBAST-2019 and the first round, although slightly reworded: participants and data sources, predictors, outcome, and analysis. Major changes to the previous round were the more explicit distinction between model development and evaluation of model performance, and the more distinguished focus on quality assessment (for model development) and risk of bias assessment (for evaluation of model performance) (see box 1). The number and type of signalling questions differed between model development (21 items) and model evaluation (22 items) (see supplementary table 2).

Round 3

The third and final round was opened from 11 May to 7 July 2023. All participants who completed the first or second round were invited to take part in the third round. Those who did not respond to the previous

Box 2: Noteworthy changes and additions to PROBAST+AI

- Updated tool for assessing quality, risk of bias, and applicability that covers prediction model studies regardless of the modelling approach applied (eg, regression or AI/machine learning methods.
- A more explicit consideration of model development and evaluation of model performance as separate phases, with updated signalling questions.
- Distinguishes three types of model performance evaluation: apparent performance, internal validation, and external validation (see box 1).
- Particular emphasis on fairness and on algorithmic bias, to assess whether specific methods were used to ensure fairness and deal with algorithmic bias (see box 1). Aspects of fairness and algorithmic bias are embedded throughout the signalling questions of the four domains.
- Harmonisation of nomenclature between different specialties of expertise (eg, statistics, AI/machine learning, data science, epidemiology) (see box 1).
- Useful not only for assessing prediction model studies when the aim is to conduct a systematic review of prediction models, but also for appraising the applicability, quality, and risk of bias of one or more specific prediction models (eg, when developing healthcare guidelines, policy, or healthcare recommendations, or to make a decision on whether or not to use or implement a prediction algorithm into daily practice).

AI=artificial intelligence; PROBAST=Prediction model Risk Of Bias ASsessment Tool.

rounds were reinvited, as well as participants who were identified or recommended after these rounds. All those invited received the aggregated responses of the first two rounds. Of 299 people who received an invitation to participate in the third round, 131 responded to the survey (see supplementary information, figure 1 and table 1).

The list of signalling questions for the third round for the model development phase included the same four domains as for the second round, with 18 items for model development and 19 items for model evaluation (see supplementary table 2).

Consensus meeting

A hybrid consensus meeting chaired by KGMM, LH, JAAD, PD, RDR, and MvS was held on 17 October 2023. Of 29 participants invited to the consensus meeting, 26 attended (see supplementary figure 1). Participants were identified to ensure a balanced representation of the key stakeholder groups and geographical and other aspects of diversity. In preparation for the final consensus meeting, KGMM, JAAD, TK, CAN, PD, LH, JBR, RDR, GSC, and MvS developed a pre-final PROBAST+AI based on input from the Delphi survey rounds in several hybrid meetings. Ten days before the consensus meeting, participants were emailed a document containing a brief overview of PROBAST+AI, the format of and instructions for the consensus meeting, a summary of the aggregated responses from the last Delphi survey round, and the draft PROBAST+AI. The tool shared with the consensus meeting participants included four domains and 16 signalling questions for quality assessment of model development, and the same four domains and 18 signalling questions for risk of bias assessment of the model evaluation. Given the high endorsement for many items in the third round of the Delphi survey, we selected only a subset of 10 signalling questions for plenary discussion and voting during the consensus meeting. The 10 items had either undergone rewording after the third round or were new items introduced after that round.

PROBAST+AI

After the consensus meeting, the working group developed the final PROBAST+AI tool (see supplementary tables 3 and 4). The most noteworthy change in PROBAST+AI compared with the PROBAST-2019 tool[20] was the more explicit distinction between signalling questions to assess the methodological quality of the process of model development versus assessment of the risk of bias in the evaluation of the model performance (see next section for a detailed rationale for this distinction). Box 2 summarises noteworthy changes and additions to PROBAST-2019, such as this more explicit distinction between quality of model development and risk of bias in the model performance evaluation, as well as more explicit attention for algorithmic bias and fairness throughout the tool. Supplementary table 5 provides a more detailed comparison between PROBAST+AI and PROBAST-2019.

In summary (table 2 and supplementary tables 3 and 4), PROBAST+AI includes four domains, 34 signalling questions (16 for model development, 18 for model evaluation), and six applicability items (three each for model development and for model evaluation). The first three domains share the same signalling questions for assessing either the quality of model development or the risk of bias in model evaluation. All domains focus on concerns of quality and applicability (for model development) and on concerns of risk of bias and applicability (for model evaluation). Domain 1 (participants and data sources) covers issues related to the participants and data sources used for model development or model evaluation. Domain 2 (predictors) covers the definition or measurement of predictors included in the development or evaluation of the prediction model, whereas domain 3 (outcome) covers the same aspects regarding definition and measurement of the outcome predicted. Domain 4 (analysis) deals with data analysis methods and assesses aspects related to the choice of analysis method and whether key statistical considerations (eg, handling of missing data) were dealt with correctly. Domain 4 has five signalling questions to support the quality assessment for the model development, and seven to support the risk of bias assessment for model evaluation. Detailed information on all items is available in the Explanation and Elaboration Light document (see supplementary table 4). Supplementary table 5 shows the differences between PROBAST-2019 and PROBAST+AI.

Assessment of quality (model development) versus risk of bias (model evaluation)

In PROBAST-2019,[19 20] signalling questions for both model development and model evaluation could

Table 2 | Summary of step 3 (assessment of quality, risk of bias, and concerns about applicability) of PROBAST+AI

Participants and data sources	Predictors	Outcomes	Analyses
Model development			
Signalling questions*:			
1.1 Were appropriate data sources used?	2.1 Were predictors defined and assessed in a similar way for all participants?	3.1 Were outcomes defined and assessed appropriately?	4.1 Was there evidence that the sample size was reasonable?
1.2 Was an appropriate study design used?	2.2 Was any preprocessing of predictors similar for all participants?	3.2 Were outcomes defined and assessed in a similar way for all participants?	4.2 Were continuous and categorical predictors handled appropriately?
1.3 Did the inclusions and exclusions of study participants result in a representative dataset?	2.3 Were predictor assessments made without knowledge of outcome data?	3.3 Were outcome assessments made without use or knowledge of predictor data?	4.3 Were participants with missing or censored data handled appropriately in the analysis?
	2.4 Were the predictors included in the model available at the time the model was intended to be used?	3.4 Was the time interval between predictor assessment and outcome assessment appropriate?	4.4 If methods to address class imbalance were used, was the model or the model predictions recalibrated?
			4.5 Were methods used to address potential model overfitting?
Quality†:			
Concern regarding quality of selection of participants and data sources	Concern regarding the quality of the predictors or their assessment	Concern regarding quality of the outcome or its determination	Concern regarding quality of the analysis
Applicability†:			
Concern that the data of the included participants do not match the review question or the assessor's intended use of the prediction model	Concern that the definition, preprocessing, assessment, or timing of assessment of the predictors in the model do not match the review question or the assessor's intended use	Concern that the outcome, its definition, assessment, or timing of assessment do not match the review question or the assessor's intended use	
Model evaluation			
Signalling questions*:			
1.1 Were appropriate data sources used?	2.1 Were predictors defined and assessed in a similar way for all participants?	3.1 Were outcomes defined and assessed appropriately?	4.1 Was model evaluation based on only apparent performance avoided?
1.2 Was an appropriate study design used?	2.2 Was any preprocessing of predictors similar for all participants?	3.2 Were outcomes defined and assessed in a similar way for all participants?	4.2 Was there evidence that the sample size was reasonable?
1.3 Did the inclusions and exclusions of study participants result in a representative dataset?	2.3 Were predictor assessments made without knowledge of outcome data?	3.3 Were outcome assessments made without use or knowledge of predictor data?	4.3 Were participants with missing or censored data handled appropriately in the analysis?
	2.4 Were the predictors included in the model available at the time the model was intended to be used?	3.4 Was the time interval between predictor assessment and outcome assessment appropriate?	4.4 If methods to address class imbalance were used, was the evaluation done in a dataset without correction for imbalance?
			4.5 If data splitting was done to create training and test datasets, was there evidence that data leakage was avoided?
			4.6 If resampling methods were used to evaluate model performance, were all model development steps replicated in the resampling process?
			4.7 Was the predictive performance of the model evaluated appropriately—for example, calibration, discrimination, and net benefit?
Risk of bias†:			
Risk of bias introduced by the selection of participants and data sources	Risk of bias introduced by the predictors or their assessment	Risk of bias introduced by the outcome or its determination	Risk of bias introduced by the analysis
Applicability†:			
Concern that the data of the included participants do not match the review question or the assessor's intended use of the prediction model	Concern that the definition, preprocessing, assessment, or timing of assessment of the predictors in the model do not match the review question or the assessor's intended use	Concern that the outcome, its definition, assessment, or timing of assessment do not match the review question or the assessor's intended use	

Developed based on PROBAST-2019.[19 20] For further details see the PROBAST+AI Explanation and Elaboration Light in supplementary table 4 and the PROBAST-2019 Explanation and Elaboration paper.[19]
AI=artificial intelligence; PROBAST=Prediction model Risk Of Bias ASsessment Tool.
*Answered as yes, probably yes, probably no, no, no information, or not applicable.
†Rated as low, high, or unclear.

be used to assess risk of bias—that is, systematic error in the estimate of the model's true predictive performance. With PROBAST+AI, we clarified that assessments of model development rather address quality, whereas assessments of model evaluation address bias. The former assesses the quality of how a prediction model is developed (model development), whereas the latter addresses the risk of bias in the predictive or classification performance of a developed model (model evaluation).

Model development is the actual process of constructing, producing, or manufacturing a prediction model, from data collection and study design to fitting the model on data and producing or fitting the final prediction model or algorithm. Each model is developed only once; it can be compared with the manufacturing or production of a medical test, device, or drug. When methodological weaknesses or shortcomings are present in the design, conduct, and analysis of a model development process, these might lead to a prediction model with less reliable or accurate predictions and weak predictive or classification performance when evaluated or applied to data from individuals other than those used for the model development.[19 20] With PROBAST+AI, we introduced the concept of methodological quality (or simply quality) of the actual model development or production process: concerns about a lower quality of the model development process, as indicated by the signalling questions of the first part of PROBAST+AI should thus be seen as a red flag or raise concern about a poorly developed (manufactured) model.

Model evaluation is the process of estimating the model's predictive performance—for example, in terms of calibration, discrimination, or net benefit. Although a prediction model is developed or manufactured only once, it can and ideally should be evaluated more than once on its predictive performance, in participant data not used for model development. In other words, a particular model has only one model development process (or study), but it can have multiple external evaluations or model evaluation studies. This process can also be compared to a medical test, device, or drug that is manufactured only once but can be evaluated on its accuracy or effectiveness multiple times. When methodological weaknesses or shortcomings in the design, conduct, and analysis of a prediction model evaluation are present, reported model performance estimates may systematically differ from the true model performance.[19 20] Estimating model performance can be done in various ways (see box 1)[19 24 38]: using exactly the same participant data as that used for model development (ie, apparent performance); using some form of splitting, resampling, or cross validation technique on the data of the development set (ie, internal validation); or using different participant data from the development dataset (ie, external validation). Thus, the second part of PROBAST+AI assesses the risk of bias in the quantification or evaluation of the performance estimates of a prediction model (for each of the apparent, internal and external validation components, as relevant) by assessing the study's design, conduct, and analysis using a series of signalling questions.

In the context of evaluating the performance of a prediction model, bias thus refers to systematic error in the estimates of the model's true predictive performance.[19 20] Bias can act in either direction, potentially leading to systematic overestimation or underestimation of the true prediction model performance. In the context of developing or manufacturing a prediction model, we cannot speak of the true performance of that model, and thus it is more appropriate to speak of quality than of bias. A poorly developed prediction model (ie, the development study was judged as having low quality) may likely have weak predictive performance—for example, small sample sizes used to develop models tend to result in lower discrimination performance when the model is evaluated or applied in data from new individuals. Models of lower quality may also be more susceptible to concerns about bias in the predictive performance of the model when evaluated in or applied to new participant data, but these need to be examined using the evaluation component of PROBAST+AI.

Finally, bias must not be confused with imprecision, which arises when a model's performance estimate is based on a small evaluation sample, leading to wide confidence intervals of the performance estimates.[19]

Potential users and the utility of PROBAST+AI

PROBAST+AI includes a formal tool for quality appraisal of the model development process, provides a formal tool to assess risk of bias of a model's predictive performance, and enables an assessment of the applicability of a prediction model to the intended purposes of PROBAST+AI users. We emphasise that PROBAST+AI is not only useful for researchers, authors, and reviewers of prediction model (development or evaluation) studies but for anyone who wants to appraise the applicability, quality, and risk of bias of prediction models themselves (see table 1). PROBAST+AI is thus also useful for researchers or medical technology or device manufacturers aiming to develop or evaluate a prediction model with or without accompanying software; healthcare professionals determining whether or not to implement a prediction model in their daily healthcare practice; health policy regulators and guideline organisations appraising prediction models for their clinical guidance, such as the World Health Organization, US Food and Drug Administration, and UK National Institute for Health and Care Excellence; journal editors, reviewers, and ethical review boards aiming to critically appraise prediction model studies; or others who want to judge the applicability, quality, and risk of bias of a prediction model for their specific context, situation, or purposes. Table 1 outlines potential users of PROBAST+AI, the different purposes for which the tool can be used, and their potential benefits.

Steps for using PROBAST+AI

PROBAST+AI can be used regardless of the modelling approach, prevailing statistical methods, or AI/machine learning techniques used for model development (see supplementary table 3). PROBAST+AI therefore supersedes PROBAST-2019.[19 20] The PROBAST-2019 Explanation and Elaboration document[19] remains the comprehensive background document of PROBAST+AI and serves as an important pedagogical document to provide rationale and examples for most of the PROBAST+AI items. Moreover, supplementary table 4

provides an additional bullet point structure for each signalling question, including a brief explanation and elaboration (ie, Explanation and Elaboration Light) to facilitate implementation of PROBAST+AI. Differences in item scoring between traditional regression based models and models based on AI/machine learning techniques are, when needed, also described in the Explanation and Elaboration Light.

PROBAST+AI uses the same four steps as PROBAST-2019[19 20] (see supplementary table 3 for explanations).

Step 1: Specify the intended purpose of the prediction model assessment or prediction model systematic review

In accordance with PROBAST-2019,[19 20] when using PROBAST+AI we advise specifying the purpose of the assessed prediction model. For this we recommend defining the PICOTS (Population, Index model, Comparator model, Outcome, Timing, Setting, and intended use of the prediction model) criteria as provided by the guidance of the Cochrane Prognosis Methods group (https://methods.cochrane.org/prognosis/) and described in CHARMS (checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies[61]). Defining the PICOTS directly indicates the aim of the assessment or review of the prediction models.

Step 2: Classify the type of prediction model study

Prediction model studies can include model development or model evaluation, or both.[19 20 38] PROBAST+AI includes different signalling questions depending on the type of prediction model study. Therefore, we recommend that assessors and reviewers state whether they address model development only or model evaluation only, or both. Furthermore, model evaluation distinguishes between estimation of the model's apparent performance, the internal validation performance, and the external validation performance (see box 1 for descriptions). When a publication focuses on updating a previously developed model, such as adding one or more new predictors, the model development part of PROBAST+AI should (also) be used. When a publication focuses on evaluating the performance of an existing model in other (external) participant data, only the model evaluation part should be used.

Step 3: Assess quality, risk of bias, and applicability of the prediction model for each domain

This step aims to identify areas where concerns about quality and risk of bias might be introduced in the prediction model study, or where concerns about applicability might exist.[19 20] For each domain the quality (for model development) and risk of bias (for model evaluation) assessment comprises four sections (see table 2, also see supplementary tables 3 and 4 for detailed guidance on use): Section 1—general information from the study or model to support answering the signalling questions of that domain; section 2—answering the signalling questions; section 3—a judgment of concerns about quality (for model development) or risk of bias (for model evaluation) per domain; and section 4—rationale for the overall quality judgment (separately for the development) or risk of bias judgment (separately for the evaluation) of the prediction model.

As with PROBAST-2019,[19 20] assessors can record any additional information used to answer the signalling questions in the box related to rationale for any judgment. Signalling questions are answered as yes, probably yes, no, probably no, no information, or, when appropriate, not applicable. Quality concerns (for model development) are judged as low, high, or unclear, and risk of bias (for model evaluation) is judged as low, high, or unclear. All signalling questions are phrased such that yes answers or probably yes answers indicate low concern for quality (ie, high quality) or low risk of bias. Any signalling question answered as no or probably no flags the potential for quality concerns or bias. Subsequently, assessors need to use their judgment to determine based on these answers whether the entire domain should be rated as low, high, or unclear quality concern (for the model development) or low, high, or unclear risk of bias (for the model evaluation). If a signalling question is answered with no, it does not automatically result in a high concern for quality or a risk of bias rating of the entire domain. The no information category should be used only when reported information is insufficient to permit a judgment. The not applicable category may be available for items that are not applicable for certain types of prediction models or situations. When the rationale for the overall domain judgment is recorded separately for the model development and for the model evaluation, the domain's quality or risk of bias rating will be more transparent. This can also facilitate discussion among different reviewers or assessors who complete assessments independently.

The first three domains, in accordance with PROBAST-2019,[19 20] also include assessment of concerns about the applicability of the prediction model (low, high, unclear) to the review question or to the assessors' intended use of the assessed model. Applicability is defined as any concern that the included data of the participants and setting (domain 1); or the definition, preprocessing, assessment, or timing of assessment of the predictors (domain 2); or the outcome definition, assessment, or timing of assessment (domain 3), do not match the prediction model review question or the assessor's intended use of the model. Accordingly, applicability refers to either applicability of a study to the question of the reviewer (for example, when one conducts a systematic review of prediction model studies) or whether a particular developed or evaluated model is indeed useful for the intended use or purpose of the assessor. For example, a model can have a low concern for quality if data and participant selection were appropriate for the modeller's intended purpose, but a high concern for

applicability if either does not match how the reviewer intends to use that model.

Step 4: Overall quality, risk of bias, and applicability judgment

The final step of the PROBAST+AI tool is similar to that in PROBAST-2019,[19 20] in which the four domain ratings are combined into an overall judgment on the quality and applicability of the model for model development and separately on the risk of bias and applicability for model evaluation. This overall judgment is scored as either low, high, or unclear concern of the quality; low, high, or unclear risk of bias; and low, high, or unclear concern of the applicability. Step 4 in supplementary table 3 provides guidance on how to make an overall judgment on quality, risk of bias, and applicability, as well as the original PROBAST-2019 guidance.[19 20]

For example, a high overall concern of applicability (for both model development and model evaluation) indicates a limited or poor applicability of the scored model for the review question or the assessor's intended use of the model, whereas a low overall concern of applicability indicates a good applicability of the scored model.[19 20] Similarly, for the quality judgment of model development, an overall high concern indicates a low quality of the model development (production) process, whereas an overall low concern indicates a high quality of the model development process. And similarly for the risk of bias judgment of the model evaluation, a low risk of bias indicates that the reported estimates for model performance are valid (unbiased), whereas a high risk of bias indicates the performance estimates might systematically differ from the true model performance. For all three (ie, judgment of applicability, quality, or risk of bias), an unclear overall judgment indicates that reported information was insufficient to make an adequate judgment.

These overall judgments may sometimes involve changing or reclassifying an overall high concern to low concern (for quality of the model development process) or a high risk of bias to low risk of bias (for model performance estimates).[19 20] For example, for the model evaluation part, if a model performance was evaluated without any external performance evaluation, domain 4 might have been scored as high risk of bias. If the model was (typically in that same study) developed (ie, fitted) on a large dataset and evaluated with some form of internal model performance validation, however, this high risk of bias might be changed to an overall low risk of bias rating, provided that the other three domains had low concern about quality and risk of bias.

Multiple PROBAST+AI assessments and extending answers from model development to model evaluation

When a study reports the development and evaluation of more than one prediction model, all domains should be completed for each distinct prediction model.[19 20] The same publication may even address the model development process, its evaluation with apparent performance estimates, its evaluation with performance estimates after internal validation, and its performance evaluation with some form of external validation (see box 1 for explanations of terms).

Also, the same report may describe the development and evaluation of a specific model combined with the evaluation of multiple other models. We recommend that a separate PROBAST+AI assessment is done for each model. If all this was done on the same dataset and using the same predictor and outcome definitions and measurements, however, the responses to the signalling questions in (notably) the domains of participants and data sources, predictors, and outcome would be the same for the model development and the model performance evaluations: the answers across the three domains can then easily be copied and pasted. However, if a prediction model was developed and external data sources were used for evaluation of its predictive performance, the responses for the first three domains could differ between the model development and model performance evaluations.

Notably, for studies only developing models, both parts of PROBAST+AI must usually still be completed since typically the apparent performance evaluation is also estimated—and (ideally) the internal validation performance as well. In these instances, the responses to the signalling questions (mainly in the domains of participants and data sources, predictors, and outcome) will be again the same for the development, apparent performance evaluation, and interval validation: accordingly, the answers across the three can be copied and pasted.

If in a study in which a prediction model was developed and external data sources were also used for evaluation of its performance, however, the responses for the first three domains (ie, participants and data sources, predictors, outcome) may differ.

Finally, a model evaluation study may only describe the evaluation of one or more prediction models that have been developed in other previously published development studies.[19 20 38] In these instances, only the second part of PROBAST+AI needs to be filled in, although separately for each evaluated (validated) model, where often the responses to many signalling questions will likely be the same for each assessment and can thus be copied.

The signalling questions in PROBAST+AI are in a natural order similar to PROBAST-2019,[19 20] as is roughly encountered when reviewing a prediction model report or study, although this may depend on journal formatting policies. The items and issues addressed by the domains and signalling questions of PROBAST+AI are the minimal and most essential items and issues to be assessed. Users can always assess additional aspects of a study to obtain a better view of the quality, risk of bias, or applicability of a prediction model.

Discussion

PROBAST+AI has been developed through a comprehensive, phased, and international consensus process with multi-stakeholders. It provides explicit

criteria for assessing the methodological quality, risk of bias, and applicability of studies or reports describing the development or evaluation of prediction models using any data analytical (ie, prevailing statistical or AI/machine learning) method. Notable changes in PROBAST+AI (see box 2) are a clear distinction between the concepts of quality in the model development process and risk of bias in the estimates of the model's performance, and a more explicit emphasis on fairness and algorithmic bias throughout the tool, as described in the PROBAST+AI Explanation and Elaboration Light (see supplementary tables 3 and 4). PROBAST+AI is a direct extension and update of PROBAST-2019[19 20] (see supplementary table 5), and because of these specific changes and additions, PROBAST+AI may replace PROBAST-2019.

We illustrated that methodological quality of a model development or manufacturing process differs from a risk of bias assessment in the evaluation or quantification of the model's performance. Assessors might have high confidence in the performance estimates (low risk of bias) from a well conducted evaluation of an initially poorly developed model (low quality). It is also possible that assessors are sceptical (high risk of bias) about the performance estimates in a poor evaluation study of a previously well developed model (high quality). We have emphasised that a single study might include a model development and several types of performance evaluations of that same model—that is, either using the same participant data for development and performance evaluation or using different (external) data for the performance evaluation.[19 20 38] PROBAST+AI is to be used to assess both the methodological quality of a model development process and the risks of bias in the evaluated predictive performance estimates.

We further stress that with improved data infrastructures and thereby increasing availability of data that were not collected primarily for research purposes (eg, patient data from administrative registries, electronic health records, or other real world contexts), it has become more important to assess not only the methods used to develop or evaluate a prediction model but also the quality of the source of data and the inclusiveness or fairness of the relevant individuals in the dataset. Fairness in prediction model research is particularly important in healthcare, also or perhaps certainly when AI/machine learning methods are used to develop or evaluate the models.[38 62] Fairness (see box 1) means that prediction models should be designed and used to avoid adverse discrimination against any group of individuals and not to perpetuate any inequities in healthcare provision and outcomes for patients or the general population.[62] One important aspect of fairness is ensuring that the data used to develop or evaluate prediction models are diverse and representative. The STANdards for data Diversity, INclusivity and Generalisability (STANDING) Together initiative has developed standards for data diversity, inclusivity, and generalisability.[36] This means that data sources should include information from individuals representing a diversity of characteristics, such as age, sex or gender, and race or ethnicity, as well as individuals with different health conditions or comorbidities and potentially from different geographical locations that are representative of the target population for which the prediction model is intended. If data used to develop the model are not diverse and representative, the resulting model may not be effective or fair and thus not applicable to those individuals for which the model is intended (by the assessor of the model). Furthermore, if data used to evaluate a model are not representative of the assessor's target population, the estimates of predictive performance in particular subgroups could be misleading. PROBAST+AI has therefore stressed more explicitly aspects on fairness of the model development and evaluation throughout the four domains, to ensure due consideration is given during the appraisal of prediction models and prediction model studies. Although algorithmic bias and fairness assessments for the model development and model evaluation procedure are crucial, they should always be approached with caution. Algorithmic bias and fairness related issues may not be identified by exploratory data analysis alone. The ultimate assessment of algorithmic bias and fairness occurs when the model is deployed in daily healthcare practice.[10]

Conclusion

We anticipate that PROBAST+AI will help all stakeholders (eg, prediction model developers and companies, researchers, editors, reviewers, healthcare professionals, patients, ethical review boards, guideline developers, and health policy organisations) who encounter prediction models in the healthcare sector to understand and appraise the quality, risk of bias, and applicability of prediction models and prediction model studies. Using PROBAST+AI to guide the design and analysis of a prediction model development study or evaluation (validation) study, or both, should help to reduce research waste while improving the accuracy, effectiveness, generalisability, and appropriate use and fairness of prediction models in any healthcare setting or domain where prediction or classification plays a role, and regardless of the data analytical modelling (ie, prevailing statistical or AI/machine learning) technique used.

AUTHOR AFFILIATIONS

[1]Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, 3508 GA Utrecht, Netherlands

[2]Cochrane Netherlands, University Medical Centre Utrecht, Utrecht University, Utrecht, Netherlands

[3]Centre for Statistics in Medicine, UK EQUATOR Centre, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

[4]Department of Epidemiology, Harvard T H Chan School of Public Health, Boston, MA, USA

[5]Department of Development and Regeneration, KU Leuven, Leuven, Belgium

[6]Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium

[7]Department of Biostatistics, Harvard T H Chan School of Public Health, Boston, MA, USA
[8]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA
[9]Institute of Health Informatics, University College London, London, UK
[10]British Heart Foundation Data Science Centre, Health Data Research Centre UK, London, United Kingdom
[11]College of Medicine and Health, University of Birmingham, Birmingham, UK
[12]Department of Electrical Engineering and Computer Science, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[13]Institute of Clinical Biometrics, Centre for Medical Data Science, Medical University of Vienna, Vienna, Austria
[14]Department of Medicine, University of Cape Town, Cape Town, South Africa
[15]Division of Intelligent Medical Systems, German Cancer Research Centre (DKFZ), Heidelberg, Germany
[16]National Centre for Tumour Diseases (NCT) Heidelberg, Heidelberg, Germany
[17]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
[18]School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, UK
[19]NIHR Birmingham Biomedical Research Centre, Birmingham, UK
[20]Department of Bioethics, The Hospital for Sick Children, Toronto, ON, Canada
[21]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore
[22]Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA, Australia
[23]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, USA
[24]AI Office, Singapore Health Service, Duke-NUS Medical School, Singapore, Singapore
[25]Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands

The PROBAST+AI authors are as follows: Karel Moons (UMC Utrecht, Netherlands), Maarten van Smeden (UMC Utrecht, Netherlands), Richard Riley (University of Birmingham, UK), Gary Collins (University of Oxford, UK), Paula Dhiman (University of Oxford, UK), Johannes Reitsma (UMC Utrecht, Netherlands), Johanna Damen (UMC Utrecht, Netherlands), Tabea Kaul (UMC Utrecht, Netherlands), Lotty Hooft (Cochrane Netherlands, Netherlands), Constanza Andaur Navarro (UMC Utrecht, Netherlands), Bada Yang (UMC Utrecht, Netherlands), Andrew Beam (Harvard School of Public Health, USA), Ben Van Calster (KU Leuven, Belgium), Leo Celi (Massachusetts Institute of Technology, USA), Spiros Denaxas (University College London, UK), Alastair Denniston (University of Birmingham, UK), Marzyeh Ghassemi (Massachusetts Institute of Technology, USA), Georg Heinze (Medical University of Vienna, Austria), André Pascal Kengne (University of Cape Town, South Africa), Xiaoxuan Liu (University of Birmingham, UK), Patricia Logullo (University of Oxford, UK), Lena Maier-Hein (German Cancer Research Centre, Germany), Melissa McCradden (The Hospital for Sick Children, Canada), Nan Liu (Duke-NUS Medical School, Singapore), Lauren Oaken-Rayner (University of Adelaide, Australia), Karandeep Singh (University of Michigan, USA), Daniel Ting (Stanford University, US), Laure Wynants (KU Leuven, Belgium).

We thank members of the PROBAST+AI Delphi panel for their time and valuable contribution in helping to develop the PROBAST+AI statement. We also gratefully acknowledge all contributors to the original PROBAST guidance.

The full list of Delphi survey participants and others who provided feedback on PROBAST+AI are as follows: Jose A Calvache, Elie Akl, Elena Albu, Lucy Archer, Sarah Barman, Valentina Bellini, Laura Bonnett, Patrick Bossuyt, Anne-Laure Boulesteix, Randy Boyes, Peter-Bram 't Hoen, Danilo Bzdok, Jennifer Camaradou, Guido Camps, Jonathan Chen, Evangelia Christodolou, Jeremie Cohen, Darren Dahly, Maarten De Vos, Thomas Debray, Jon Deeks, Andre Dekker, Jac Dinnes, Edgar Efrén Lozada Hernández, Joie Ensor, Ari Ercole, Andre Esteva, Ji Eun Park, Lavinia Ferrante di Ruffano, Alan Fraser, Shan Gao, Geert-Jan Geersing, Bart Geerts, Robert Golub, Benjamin Gravesteijn, Olivier Groot, Saskia Haitjema, Michael Harhay, Frank Harrell, Ulrike Held, Tina Hernandez-Boussard, Alejandro Hernández-Arango, Pauline Heus, Bethany Hillier, Michael Hoffman, Jeroen Hoogland, Mohammed Hudda, Merel Huisman, Ivana Isgum, Jan Jaap Baalbergen, Patricia Jaspers, David Jenkins, Kevin Jenniskens, Charles Kahn, Vineet Kamal, Michael Kammer, Evangelos Kanoulas, Ilse Kant, Teus Kappen, Christopher Kelly, Nina Kreuzberger, Jethro Kwong, Joanna Lane, Linda Lapp, Artuur Leeuwenberg, Tim Leiner, Brooke Levis, Qui Li, Christopher Lovejoy, Kim Luijken, Pat Lyons, Stephen M Borstelmann, Jie Ma, Dennis Makau, Sue Mallett, Konstantinos Margetis, Iain Marshall, Glen Martin, Bilal Mateen, Michael Matheny, Matthew McDermott, David McLernon, Jamie Miles, Antonio Moura, Leila Mureebe, Myura Nagendran, Charlie Nederpelt, Daan Nieboer, Wiro Niessen, Steven Nijman, Quentin Noirhomme, Daniel Oberski, Johan Ordish, Almilaji Orouba, Ravi Parikh, Seong Park, Andre Pascal Kengne, Niels Peek, Bas Penning de Vries, Daniel Pinto dos Santos, Robert Platt, Frank Rademakers, Erik Ranschaert, Kelly Reeve, Samuel Relton, Dimitris Rizopolous, Sherri Rose, Laura Rosella, Jan Roth, Alicja Rudnicka, Rupa Sakar, Pui San Tan, Katie Scandrett, Michael Schlussel, Ewoud Schuit, Martijn Schut, Mark Sendak, Jamie Sergeant, Chunhu Shi, Nicole Skoetz, Kym Snell, Adrian Soto-Mota, Viknesh Sounderajah, Matthew Sperrin, Benjamin Spivak, Ewout Steyerberg, Tom Stocker, Matthew Strother, Herdiantri Sufriyana, Xin Sun, Tom Syer, Toshihiko Takada, Halil Tanboga, Cristian Tebé, Paul Tiffin, Jim Tol, Eric Topol, Darren Treanor, Ioanna Tzoulaki, Wouter Veldhuis, Kamal Vineet, Christine Wallish, Junfeng Wang, Peter Watkinson, Wim Weber, Gary Weissman, Penny Whiting, Rebecca Whittle, Jack Wilkinson, Tyler Williamson, Marie Westwood, Robert Wolff, Aaron Y Lee, Christopher Yau, Valentijn de Jong, Annemarie van 't Veen, Wouter van Amsterdam, Bas van Bussel, Peter van der Heijden, Iwan van der Horst, Florien van Royen, Kim van der Braak.

Contributors: KGMM, JBR, RDR, GSC, and MvS conceived this paper. KGMM, JAAD, TK, CAN, PD, LH, JBR, RDR, GSC, and MvS designed the PROBAST+AI tool (signalling questions for the quality, evaluation, and applicability components) and subsequently designed the surveys carried out to inform the guideline content. TK, CAN, and JAAD analysed the survey results and free text comments from the survey. TK designed the materials for the consensus meeting with input from KGMM. TK took consolidated notes from the consensus meeting. KGMM and MvS led the drafting of the manuscript, with initial edits from JAAD, TK, CAN, PD, LH, JBR, RDR, and GSC. All authors were involved in revising the article critically for important intellectual content. All authors approved the final version of the article. KGMM is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: The research on PROBAST-AI is unfunded. RDR, GSC, and PD are supported by an EPSRC grant entitled "Artificial intelligence innovation to accelerate health research" (No EP/Y018516/1). RDR and GSC are supported by a National Institute for Health and Care Research (NIHR) Medical Research Council grant entitled "Better methods better research" (MR/V038168/1). RDR is supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. RDR and GSC are senior investigators for the NIHR. GSC and PL are supported by Cancer Research UK (programme grant C49297/A27294). PD is supported by Cancer Research UK (project grant PRCPJT-Nov21\100021). The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or Department of Health and Social Care.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work. KGMM is director of Health Innovation Netherlands (HI-NL), editor in chief of *BMC Diagnostic and Prognostic Research*, and principal investigator and author of the "Guidance for high quality AI in healthcare" (https://guideline-ai-healthcare.com). GSC is director of the UK EQUATOR Centre, editor in chief of *BMC Diagnostic and Prognostic Research*, and a statistical editor for *The BMJ*. PL is a meta-researcher with the UK EQUATOR Centre. RR is a statistical editor for *The BMJ* and receives royalties for two textbooks: *Prognosis Research in Healthcare* and *Individual Participant Data Meta-Analysis*.

Ethical approval: This project qualified as non-medical research involving human subjects (non-WMO) according to the Dutch Central Committee on Research Involving Human Subjects (CCMO) and formal ethical approval was waived. The modified Delphi survey procedure was approved by the data management board of the Julius Centre of

the University Medical Centre Utrecht (Netherlands). The modified Delphi process was subject to regular monitoring. Delphi survey participants provided electronic informed consent before completing the survey. All participant data were pseudonymised and stored securely on a server of the University Medical Centre Utrecht.[63]

Data sharing: Aggregated Delphi survey responses are available on the Open Science Framework "PROBAST+AI" repository (doi:10.17605/OSF.IO/W3CFE).

Transparency: The guarantor (KGMM) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Dissemination to participants and related patient and public communities: The published paper will be shared via email with all Delphi participants. Results will not be sent to patient and public communities as this is methodological rather than applied research.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Hippisley-Cox J, Coupland CAC, Bafadhel M, et al. Development and validation of a new algorithm for improved cardiovascular risk prediction. *Nat Med* 2024;30:1440-7. doi:10.1038/s41591-024-02905-y
2 Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416. doi:10.1136/bmj.i2416
3 Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. *BMJ* 2019;367:l5358. doi:10.1136/bmj.l5358
4 de Munter L, Polinder S, Lansink KW, Cnossen MC, Steyerberg EW, de Jongh MA. Mortality prediction models in the general trauma population: A systematic review. *Injury* 2017;48:221-9. doi:10.1016/j.injury.2016.12.009
5 Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020;369:m1328. doi:10.1136/bmj.m1328
6 Andaur Navarro CL, Damen JAA, van Smeden M, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *J Clin Epidemiol* 2023;154:8-22. doi:10.1016/j.jclinepi.2022.11.015
7 Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9:1-12. doi:10.1371/journal.pmed.1001221
8 White N, Parsons R, Collins G, Barnett A. Evidence of questionable research practices in clinical prediction models. *BMC Med* 2023;21:339. doi:10.1186/s12916-023-03048-6
9 Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med* 2021;27:2176-82. doi:10.1038/s41591-021-01595-0
10 Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. *Nat Med* 2024;30:2838-48. doi:10.1038/s41591-024-03113-4
11 Dhiman P, Ma J, Andaur Navarro CL, et al. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. *J Clin Epidemiol* 2023;157:120-33. doi:10.1016/j.jclinepi.2023.03.012
12 Andaur Navarro CL, Damen JAA, Takada T, et al. Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models. *J Clin Epidemiol* 2023;158:99-110. doi:10.1016/j.jclinepi.2023.03.024
13 London AJ, Kimmelman J. Against pandemic research exceptionalism. *Science* 2020;368:476-7. doi:10.1126/science.abc1731
14 van Royen FS, Moons KGM, Geersing G-J, van Smeden M. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. *Eur Respir J* 2022;60:2200250. doi:10.1183/13993003.00250-2022
15 Van Calster B, Wynants L, Riley RD, van Smeden M, Collins GS. Methodology over metrics: current scientific standards are a disservice to patients and society. *J Clin Epidemiol* 2021;138:219-26. doi:10.1016/j.jclinepi.2021.05.018
16 Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HC. Prognosis research: toward evidence-based results and a Cochrane methods group. *J Clin Epidemiol* 2007;60:863-5, author reply 865-6. doi:10.1016/j.jclinepi.2007.02.004
17 Moons KG, Hooft L, Williams K, Hayden JA, Damen JA, Riley RD. Implementing systematic reviews of prognosis studies in Cochrane. *Cochrane Database Syst Rev* 2018;10:ED000129. doi:10.1002/14651858.ED000129
18 Damen JAA, Moons KGM, van Smeden M, Hooft L. How to conduct a systematic review and meta-analysis of prognostic model studies. *Clin Microbiol Infect* 2023;29:434-40. doi:10.1016/j.cmi.2022.07.019
19 Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med* 2019;170:W1-33. doi:10.7326/M18-1377
20 Wolff RF, Moons KGM, Riley RD, 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
21 Kaul T, Kellerhuis BE, Damen JAA, et al. Methodological quality assessment tools for diagnosis and prognosis research: overview and guidance. *J Clin Epidemiol* 2025;177:111609. doi:10.1016/j.jclinepi.2024.111609
22 Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460. doi:10.1136/bmj.i6460
23 Debray TPA, de Jong VMT, Moons KGM, Riley RD. Evidence synthesis in prognosis research. *Diagn Progn Res* 2019;3:13. doi:10.1186/s41512-019-0059-4
24 Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 2024;384:e074819. doi:10.1136/bmj-2023-074819
25 Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ* 2024;384:e074820. doi:10.1136/bmj-2023-074820
26 Kaul T, Damen JA, Wynants L, et al. Assessing the quality of prediction models in healthcare using PROBAST: an evaluation of its use and practical application. *J Clin Epidemiol* 2025; doi:10.1016/j.jclinepi.2025.111732
27 Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Stat Med* 2019;38:1262-75. doi:10.1002/sim.7993
28 Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019;38:1276-96. doi:10.1002/sim.7992
29 Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441. doi:10.1136/bmj.m441
30 van Smeden M, de Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol* 2016;16:163. doi:10.1186/s12874-016-0267-3
31 van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Stat Methods Med Res* 2019;28:2455-74. doi:10.1177/0962280218784726
32 Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004
33 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. doi:10.1186/1471-2288-14-137
34 Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230. doi:10.1186/s12916-019-1466-7
35 Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform* 2021;28:e100289. doi:10.1136/bmjhci-2020-100289
36 Ganapathi S, Palmer J, Alderman JE, et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat Med* 2022;28:2232-3. doi:10.1038/s41591-022-01987-w
37 McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. *Sci Transl Med* 2021;13:13. doi:10.1126/scitranslmed.abb1655
38 Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378

39 Finlayson SG, Beam AL, van Smeden M. Machine Learning and Statistics in Clinical Research Articles-Moving Past the False Dichotomy. *JAMA Pediatr* 2023;177:448-50. doi:10.1001/jamapediatrics.2023.0034
40 Artificial Intelligence Act. (Regulation (EU) 2024/1689), Official Journal version of 13 June 2024. Interinstitutional File: 2021/0106(COD). https://eur-lex.europa.eu/eli/reg/2024/1689
41 Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008. doi:10.1136/bmjopen-2020-048008
42 Andaur Navarro CLD, Johanna AA, Hooft L, et al. Protocol for development of a risk of bias assessment tool for diagnostic and prognostic prediction studies based on artificial intelligence: PROBAST-AI 2022 [cited August 2024]. https://osf.io/w3cfe/
43 Langenhuijsen LFS, Janse RJ, Venema E, et al. Systematic metareview of prediction studies demonstrates stable trends in bias and low PROBAST inter-rater agreement. *J Clin Epidemiol* 2023;159:159-73. doi:10.1016/j.jclinepi.2023.04.012
44 Kaiser I, Pfahlberg AB, Mathes S, et al. Inter-Rater Agreement in Assessing Risk of Bias in Melanoma Prediction Studies Using the Prediction Model Risk of Bias Assessment Tool (PROBAST): Results from a Controlled Experiment on the Effect of Specific Rater Training. *J Clin Med* 2023;12:1976. doi:10.3390/jcm12051976
45 Dhiman P, Ma J, Navarro CA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol* 2021;138:60-72. doi:10.1016/j.jclinepi.2021.06.024
46 Andaur Navarro CL, Damen JAA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Med Res Methodol* 2022;22:12. doi:10.1186/s12874-021-01469-6
47 Rech MM, de Macedo Filho L, White AJ, et al. Machine Learning Models to Forecast Outcomes of Pituitary Surgery: A Systematic Review in Quality of Reporting and Current Evidence. *Brain Sci* 2023;13:495. doi:10.3390/brainsci13030495
48 Munguía-Realpozo P, Etchegaray-Morales I, Mendoza-Pinto C, et al. Current state and completeness of reporting clinical prediction models using machine learning in systemic lupus erythematosus: A systematic review. *Autoimmun Rev* 2023;22:103294. doi:10.1016/j.autrev.2023.103294
49 Kee OT, Harun H, Mustafa N, et al. Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. *Cardiovasc Diabetol* 2023;22:13. doi:10.1186/s12933-023-01741-7
50 Song Z, Yang Z, Hou M, Shi X. Machine learning in predicting cardiac surgery-associated acute kidney injury: A systemic review and meta-analysis. *Front Cardiovasc Med* 2022;9:951881. doi:10.3389/fcvm.2022.951881
51 Yang Q, Fan X, Cao X, et al. Reporting and risk of bias of prediction models based on machine learning methods in preterm birth: A systematic review. *Acta Obstet Gynecol Scand* 2023;102:7-14. doi:10.1111/aogs.14475
52 Groot OQ, Ogink PT, Lans A, et al. Machine learning prediction models in orthopedic surgery: A systematic review in transparent reporting. *J Orthop Res* 2022;40:475-83. doi:10.1002/jor.25036
53 Lans A, Kanbier LN, Bernstein DN, et al. Social determinants of health in prognostic machine learning models for orthopaedic outcomes: A systematic review. *J Eval Clin Pract* 2023;29:292-9. doi:10.1111/jep.13765
54 Li B, Feridooni T, Cuen-Ojeda C, et al. Machine learning in vascular surgery: a systematic review and critical appraisal. *NPJ Digit Med* 2022;5:7. doi:10.1038/s41746-021-00552-y
55 Groot OQ, Bindels BJJ, Ogink PT, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop* 2021;92:385-93. doi:10.1080/17453674.2021.1910448
56 Yusuf M, Atal I, Li J, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open* 2020;10:e034568. doi:10.1136/bmjopen-2019-034568
57 Araújo ALD, Moraes MC, Pérez-de-Oliveira ME, et al. Machine learning for the prediction of toxicities from head and neck cancer treatment: A systematic review with meta-analysis. *Oral Oncol* 2023;140:106386. doi:10.1016/j.oraloncology.2023.106386
58 Sheehy J, Rutledge H, Acharya UR, et al. Gynecological cancer prognosis using machine learning techniques: A systematic review of the last three decades (1990-2022). *Artif Intell Med* 2023;139:102536. doi:10.1016/j.artmed.2023.102536
59 Wang W, Kiik M, Peek N, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLoS One* 2020;15:e0234722. doi:10.1371/journal.pone.0234722
60 Miles J, Turner J, Jacques R, Williams J, Mason S. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. *Diagn Progn Res* 2020;4:16. doi:10.1186/s41512-020-00084-1
61 Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11:e1001744. doi:10.1371/journal.pmed.1001744
62 Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health* 2021;3:e260-5. doi:10.1016/S2589-7500(20)30317-4
63 Andaur Navarro CK, Tabea K. Data Management Plan: PROBAST+AI Delphi Survey. 2022. https://osf.io/w3cfe/

Supplementary information: Supplementary figure 1 and tables 1, 2, and 5
Supplementary information: Supplementary table 3
Supplementary information: Supplementary table 4