ESC European Society of Cardiology

European Heart Journal (2022) **43**, 2921–2930
https://doi.org/10.1093/eurheartj/ehac238

STATE OF THE ART REVIEW
Digital health and innovation

Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease

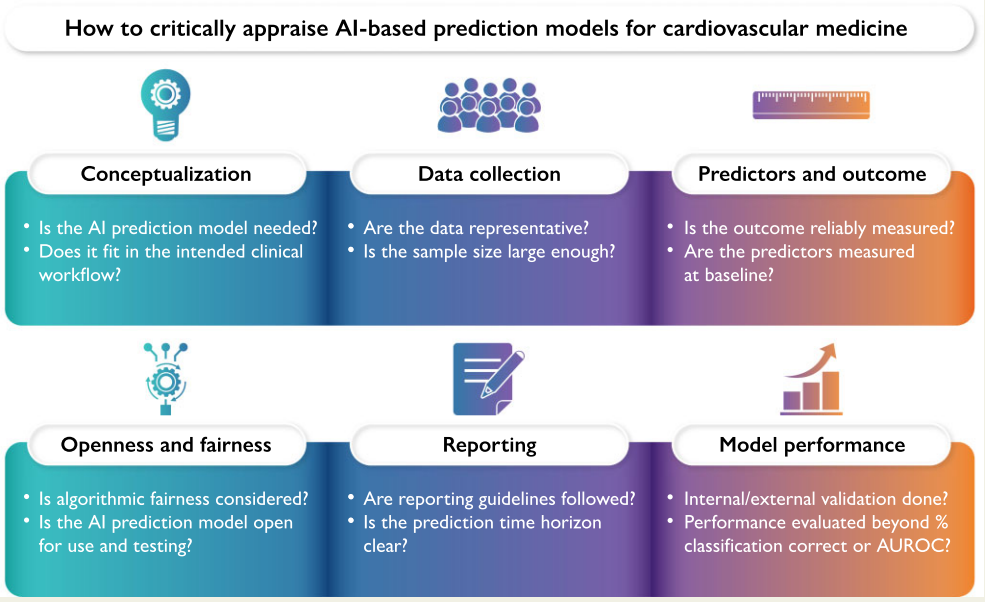
Maarten van Smeden [1]*, Georg Heinze [2], Ben Van Calster [3,4,5], Folkert W. Asselbergs[6,7,8], Panos E. Vardas [9,10], Nico Bruining [11], Peter de Jaegere[11], Jason H. Moore[12], Spiros Denaxas [8,13], Anne-Laure Boulesteix[14], and Karel G.M. Moons[1]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands; [2]Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria; [3]Department of Development and Regeneration, KU Leuven, Leuven, Belgium; [4]EPI Centre, KU Leuven, Leuven, Belgium; [5]Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, The Netherlands; [6]Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; [7]Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, UK; [8]Health Data Research UK and Institute of Health Informatics, University College London, London, UK; [9]Department of Cardiology, Heraklion University Hospital, Heraklion, Greece; [10]Heart Sector, Hygeia Hospitals Group, Athens, Greece; [11]Department of Cardiology, Erasmus MC, Thorax Center, Rotterdam, The Netherlands; [12]Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA; [13]The Alan Turing Institute, London, UK; and [14]Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Germany

Received 9 November 2021; revised 29 March 2022; accepted 26 April 2022; online publish-ahead-of-print 26 May 2022

Listen to the audio abstract of this contribution.

Graphical Abstract



Twelve critical questions to be asked by readers and reviewers when confronted with prediction models that are based on AI.

* Corresponding author. Tel: +31 648 931 109, Email: M.vanSmeden@umcutrecht.nl

Abstract

The medical field has seen a rapid increase in the development of artificial intelligence (AI)-based prediction models. With the introduction of such AI-based prediction model tools and software in cardiovascular patient care, the cardiovascular researcher and healthcare professional are challenged to understand the opportunities as well as the limitations of the AI-based predictions. In this article, we present 12 critical questions for cardiovascular health professionals to ask when confronted with an AI-based prediction model. We aim to support medical professionals to distinguish the AI-based prediction models that can add value to patient care from the AI that does not.

Keywords Digital health • Artificial intelligence • Machine learning • Diagnosis • Prognosis • Prediction

Introduction

Artificial intelligence (AI) and its subdiscipline machine learning are receiving increasing attention throughout medicine, including cardiovascular medicine.[1,2] Proponents promise AI will change the way medicine and healthcare is practiced, by making use of technological advancements that allow for collection of increasingly detailed and diverse data and the ever-increasing computational ability to analyse and combine such data. An important part of these promises is the development and implementation of more accurate clinical prediction models (algorithms, tools, or rules, from here onwards simply referred to as prediction models) to improve—or according to some advocates, even revolutionize—screening, diagnosis, and prognostication of diseases. Prediction models usually fall within one of two major categories: diagnostic prediction models that estimate an individual's probability of a specific health condition being currently present, and prognostic prediction models that estimate the probability of developing a specific health outcome over a specific time period.[3]

Indeed, technological developments in machine learning drive the ability to derive increasingly complex prediction models, from data sources that are *structured*, data from a sample of individuals that can simply be captured in a spreadsheet format, and *unstructured*, such as free text in electronic patient health records, medical images, and electrophysiology. Taking advantage of these technological developments, AI has now been introduced for a large variety of healthcare challenges within cardiovascular diseases, for instance, the automated detection of cardiac arrhythmias from electrocardiograms,[4] early detection of aortic stenosis,[5] and mortality prediction of patients undergoing cardiac resynchronization therapy.[6]

The increased interest in the development, testing, implementation, and impact of AI-based prediction models in cardiovascular patient care, also comes with new challenges. One of these challenges is that it requires the cardiovascular disease professional and researcher to familiarize themselves with the opportunities of AI prediction models, as well as their inherent limitations when developed for and applied in their own setting. This article aims to assist researchers and professionals, readers, and reviewers in appraising the development and testing (i.e. validation) of AI prediction models. We propose 12 critical questions for health professionals and researchers to consider (Graphical abstract).

Question 1: Is artificial intelligence needed to solve the targeted medical problem?

The development of a new AI prediction model should be clearly linked to a relevant medical problem it tries to solve.

The literature is populated with many prediction models to detect (diagnose) or prognosticate new onset cardiovascular diseases and predict future health in patients diagnosed with a cardiovascular disease. For instance, over 360 models for cardiovascular disease in the general population,[7] over 160 female-specific models for cardiovascular diseases,[8] and over 80 models for sudden cardiac arrest,[9] already exist. Prime examples of such prediction models in cardiovascular diseases are the Framingham risk score[10] and the recently updated SCORE2[11] for cardiovascular disease prediction in the general population, and the EuroSCORE[12] and revised cardiac risk index[13] for inpatient predictions. With such large numbers of prediction models already existing—and few models used in practice, before developing a new model one may question: is a new prediction model really needed?

While the potential of AI technology to improve predictions over the existing prediction models is beyond dispute when trained in the right way and on the right data, actual incremental value of AI over prevailing prediction models is not *per se* guaranteed for any healthcare application.[14] This was for instance shown in a recent systematic review where traditional regression techniques were not outperformed by modern AI-based prediction models.[15] Given the targeted medical problem at hand, claims about the incremental value of a more complex and possibly more difficult to implement AI prediction model over existing and often simpler prediction models should therefore be based on solid evidence coming, e.g. from careful model and prediction comparisons (see also question 8). For example, an AI-based prediction model aiming to predict 10-year risk of cardiovascular events might be compared with a canonical model like the Framingham risk score.

Question 2: How does the artificial intelligence prediction model fit in the existing clinical workflow?

Knowing the intended place of a prediction model within the existing clinical workflow is essential to identify early on the barriers towards implementation of the model in daily practice.

To understand and to be able to appraise the intended use of an AI prediction model, not only should it be clear what the model aims to predict, for whom it predicts (i.e. target population) and over what

time period (prediction horizon—see also question 4), but also where in the clinical workflow the prediction model is aimed to be implemented. A new AI prediction model may be developed as an add-on to improve the efficiency of the existing diagnostic testing process. Or it may be aimed at replacing an existing prognostic prediction model that is already part of the workflow. Such contextual information is of critical importance to identify relevant limitations and barriers of the AI-based prediction model early, well before implementation later-on in the daily clinical processes.

For example, the identification of cultural barriers, such as the trust of intended users in a complex AI model, and technical barriers, such as the mismatch between required technology platforms to execute and maintain the AI technology available in the daily clinical process of intended use, early in AI model development could, to some degree, be addressed during the AI prediction model development.[16] These particular barriers may be addressed by aiming for more transparent and simpler modelling strategies and by giving insight into which features contribute most to making the predictions (see question 12).

Operational challenges and perceived clinical utility may also play an important role in the adoption of an AI algorithm. For instance, a study of a diagnostic AI model for detecting diabetic retinopathy from retinal images found that the increasing workload of medical personnel associated with uploading images was an important barrier to implementation.[17] Another study identified a lack of perceived utility for decision-making of a prognostic model to predict post-operative nausea and vomiting as an important barrier.[18] Such studies of barriers that prevent implementation of prediction models are rare.

Question 3: Are the data for prediction model development and testing representative for the targeted patient population and intended use?

Representativeness of the data for the targeted population and intended use is important for development of well-calibrated prediction models and valid testing of AI predictions.

The predictive performance and clinical utility of any prediction model highly depend on the quality and representativeness of the data available for the model's development. When data of low quality are used to train a prediction model, for instance, data subject to large incompleteness, measurement and misclassification errors, or using data which are based on (partly) wrong and biased human decisions, important patterns may be missed that would otherwise be identified. This usually results in loss in the model's predictive performance.[19] Conversely, using high-quality data that are not representative for data quality in the targeted population may also result in disappointing predictive performance and even misleading predictions (a phenomenon known as predictor measurement heterogeneity[20,21]) once the model is tested or applied in the targeted setting with data that were not collected primarily for research, such as routine care data, or data collected in settings that differ greatly from the development setting. For example, the aforementioned AI algorithm for detecting diabetic retinopathy from retinal images performed poorly when employed under poor lighting conditions in eye clinics in Thailand.[17] Detailed information on the conditions under which data were collected and standardization of data collection where possible may reduce the chance of unexpected prediction failures due to measurement heterogeneity.

Representativeness of the data for the targeted population and context is also critical to ensure that individualized risks (i.e. predicted risks of the outcome given the individuals' feature values) are appropriately calibrated. Individualized risk estimates are often used to make medical decisions; inaccurate estimates of these risks (i.e. miscalibration) can thus lead to poor medical decisions. Data sets used for a prognostic model development with an incidence of the predicted outcome that is not representative for the incidence of the predicted outcome in the targeted clinical setting may require model recalibration.[22] Likewise, models can be poor performing and miscalibrated when developed on data in which certain groups are underrepresented (e.g. based on gender, ethnicity, and comorbidities), which may require model updating or even complete redevelopment.[22] Developments in transfer learning, in which an AI prediction model can be pre-trained on a data set that is not representative, may be used to alleviate some of the problems encountered with non-representative data.[23]

Representativeness of data for the targeted population and context is arguably even more important for testing than for developing any prediction model. First, this is required for a valid assessment of the model's calibration.[24] Second, it avoids artificial inflation of summary measures of predictive performance, e.g. by including healthy controls that are not part of the targeted population, or through overrepresentation of individuals with advanced diseases, in which prediction errors are less likely to occur.[25] Likewise, exclusion should be avoided of individuals with data that are incomplete (i.e. missing data) or of individuals for whom the outcome is more difficult to be determined, e.g. cases with an atypical presentation which are harder to predict. Excluding such individuals from testing may create a selection bias that results in unrealistic expectations of performance when the model is eventually applied in daily practice.

Question 4: Is the (time)point of prediction clear and aligned with the feature measurements?

The intended timing of prediction and the measurement of the feature data should be aligned, and the prediction model should not be developed or tested using measurements that are unavailable at the time of prediction.

In diagnostic prediction models, the goal is to determine whether the condition of interest is present or absent at the moment of prediction—the time a prediction is made. For instance, a disease may already be manifest but not yet assessed by a reference test.[26] Hence, to be aligned with the intended use of a diagnostic prediction model, the feature data (i.e. the data that serves as input in the AI model) should be measured before the true disease status is known.

For both diagnostic and prognostic prediction models, measurement and construction of features should generally be done without knowledge of the outcome to avoid artificial inflation of the associations between features and outcome.[27] Feature data should not include information that becomes available only after the intended moment of prediction. For example, an AI model that was developed to pre-operatively predict in-hospital mortality in patients undergoing transcatheter aortic valve replacement included features related

to post-operative complications,[28] such as acute kidney injury, sepsis, and cardiac arrest. As post-operative complications will be unknown pre-operatively, the intended point of prediction, such a model cannot be applied as intended.

Prognostic models in particular require specification of a prediction horizon—how far ahead in time the model aims to predict outcome occurrence by—and follow-up time to measure the outcome needs to be matched to that. Variation in follow-up times, e.g. because of administrative censoring or competing risks, can be accounted for using survival analysis techniques.[29]

Question 5: Is the outcome variable labelling procedure reliable, replicable, and independent?

Verification of the outcome status for each individual in the data set that is accurate and independent of the feature data is essential for the development and valid testing of the AI prediction model.

Like all other domains of medicine, there are many situations in cardiovascular disease research in which a perfectly accurate gold standard to diagnose a cardiovascular disease or condition is not available. This is, for instance, applicable to the diagnosis of heart failure with preserved ejection fraction,[30] but can also be relevant when interest is in cause-specific mortality or myocardial infarctions registered in, for instance, a routine healthcare database.[31] When developing an AI prediction model for an outcome for which no perfect reference standard is available, misclassification of the outcome status becomes probable. This can severely hamper the performance of the prediction model developed on the misclassified outcome data.[32] The AI prediction model may then be able to adequately predict the imperfect reference standard but not the true condition of interest.

To increase reliability and completeness of the verification of the outcome status, it may therefore be desired to rely on the judgement of individual patients by an expert, or a group of experts, or even independent outcome adjudication committees as commonly used in randomized therapeutic intervention trials. In image recognition applications of AI, such a process is known as labelling, often requiring large numbers of images to be scrutinized and annotated, which is a burdensome task that itself carries a risk of error.[33] To ensure verification of the outcome status can be appraised and replicated, detailed information must be provided regarding the experts involved, such as the education, expertise, years of experience of experts, and the setting, such as the number of experts per case, available information per case, time constraints and how discrepancies or disagreements between experts were resolved. Earlier studies into inter-observer variability within cardiovascular testing can serve as good examples.[34–37] Recent innovations for semi-automated labelling[38] may also be a promising area of development to overcome some of the mentioned limitations of case-by-case labelling.

While verification of the outcome status should be done as accurately as possible, it should in general be done without knowledge of the patients' characteristics that are used as candidate features in the development of the AI prediction model. This is important in situations where the outcome status can be influenced by knowledge from patient characteristics (e.g. not likely when non-cause-specific mortality is the outcome), which in turn can create artificial inflation of the associations between features and outcome.[27]

Question 6: Was the sample size sufficient for artificial intelligence prediction model development and testing?

It is essential to ensure the sample size available for developing and testing of the AI prediction model suffices to allow for reliable predictions in new individuals.

The sample size of the data set for development of the AI prediction model must be large enough to develop a model that is reliable when applied to new individuals in the target population and context. For regression-based prediction modelling, there is guidance and easy-to-use software to calculate the minimally required sample size.[39,40]

The minimally required sample size for a prediction model increases (i) the further away the incidence or prevalence in the target population is from 0.5, (ii) the lower the predictive value in the features (i.e. the extent to which the features are able to explain the variance in the outcome), and (iii) the more features and complexities are considered during modelling. Hence, to be able to make full use of the potential of AI prediction models, often with much higher complexity than default regression-based modelling, a much larger sample size than for traditional regression-based approaches is usually needed.[39] In particular, with rare outcomes such as inherited cardiomyopathy occurring only in 1/250 to 1/5000,[41] the minimally required sample size for traditional regression-based approaches may already be very high.[40]

While the three minimally required sample size drivers mentioned above can be used as a starting point for AI prediction modelling, currently no formal sample size criteria exist for alternatives to regression-based prediction models, such as random forest and neural networks including deep learning, let alone for settings in which the number of candidate features is much larger than the number of available individuals (i.e. high-dimensional data analyses), which limits the possibilities to perform *a priori* sample size calculations for such applications.

However, *a posteriori* sample size calculation can also be performed to justify the sample size of the development data and to ensure it suffices for developing the AI prediction model. A flexible *a posteriori* approach that can be used in retrospective and prospective model development studies is the learning curve approach.[42] A recent review of sample size determination methodologies in medical imaging research shows such an *a posteriori* sample size calculation is still rarely applied.[43]

The sample size for prediction model test data sets should be large enough to ensure the predictive performance measures (see question 8) can be estimated with sufficient precision (i.e. small confidence intervals). For reasonably precise model testing results, usually data are required for several hundreds of individuals. Recent formulas for minimally required sample size for prediction model testing or validation based on various predictive performance criteria have been published.[44] Unlike with model development, sample size formulas for prediction modelling testing or validation are applicable regardless of the method used to

develop the AI prediction model and can thus be calculated and justified *a priori*.

Question 7: Is optimism of predictive performance of the artificial intelligence prediction model avoided?

Testing of AI prediction is done through internal and external validation approaches that avoid reporting of optimistic model performance.

The AI prediction models should foremost be evaluated for their performance in making valid predictions. Estimates of predictive performance, such as the area under the receiver-operating characteristic curve (AUROC), should not be obtained directly from the same data used to develop the AI prediction model, as this will lead to estimates of performance that are too optimistic, for instance too high estimates of the AUROC.[22,45] Instead, performance of AI prediction models must be tested through rigorous internal and external prediction model validation procedures, to provide reliable estimates of their predictive performance.

Internal validation procedures use only the original model development study data to get estimates of predictive performance and include methods such as bootstrapping or cross-validation.[46] These methods do not prevent model overfitting but can provide insight in the extent of performance overfitting and aim to get an unbiased assessment of the model's performance. All steps taken for development should be integrated in the internal validation, i.e. considered as part of the AI modelling process and, in case of bootstrap or cross-validation, repeated in each bootstrap or cross-validation iteration. This includes steps for pre-filtering and selection of features and tuning and selection of the models, to avoid so-called *incomplete validation*.[47] Strictly speaking, such procedures test the *AI modelling process* rather than the *final model* itself.

Another common approach to test the AI prediction models is on a single test set after a train–test split of the data (or sometimes training–validation–test split, where validation here confusingly refers to data used for model tuning and selection). While train–test splits are common in AI prediction model development studies, this is typically referred to as an internal validation approach and reduces the effective sample size available for developing the model as compared with bootstrap and cross-validation procedures and is commonly mistaken for actual external validation of an AI prediction model.

To perform external validation, data may come from the same setting as used for development of the AI prediction model, collected in a different time period but often by the same researchers (*narrow external validation*[46]), or by other researchers in another geographical area (*broad external validation*[46]), or from even other types of patients, deviating from the original intended use. In a recent example of a narrow external validation, Al-Farra *et al.*[48] performed a temporal validation of updated prediction models of early mortality after transcatheter aortic valve implantation. Routine healthcare data from 13 Dutch heart centres between 2013 and 2016 were used for updating prediction models, while data from the same hospitals collected in 2017 were used for a narrow external validation of the updated models. For an overview of broad external validations of cardiovascular disease models, see Damen *et al.* [7]

Performance of any prediction model is expected to vary over time and place,[49,50] and therefore an AI prediction model is never truly *validated*, in the sense that it can be proven to work adequately across time and place. An external validation of an AI prediction model should therefore not be viewed as a method to generate a definitive verdict on the adequacy of performance of the AI prediction model, but a snapshot that can generate knowledge about the performance and, importantly, variation in performance over time and place, and clues to the need for replacing, updating, and tailoring AI prediction models to specific settings.

Question 8: Was the artificial intelligence model's performance evaluated beyond simple classification statistics?

The AUROC and classification accuracy statistics do not provide the full picture of the performance and utility of an AI prediction model. A broader view on performance is necessary.

While AI prediction models often have a binary outcome, usually current (diagnosis) or future (prognosis) presence or absence of a certain target status, other outcomes such as multi-category, continuous, and time-to-event outcomes are possible, and their evaluation require different performance parameters to be evaluated (beyond this article). However, for the more common binary outcomes to be predicted, there is also a large variation in measures that can quantify performance.[51] In general, measures that are sensitive to the relative frequency of the outcome, such as the percentage of correctly predicted individuals, should be interpreted with caution especially when the outcome is rare (e.g. when the relative frequency of the outcome is only 1%, a naive model that predicts everyone in the majority outcome class already has a percentage correctly predicted of $100-1\% = 99\%$).

For AI prediction models that quantify the probability of (current or future) presence of the target status for individuals (i.e. risk prediction models), the calibration of the model is important and often the weakest link.[24] Calibration of the model describes the degree to which the estimated risks agree with the observed risks. However, calibration and other traditional predictive performance measures, such as the AUROC, do not describe clinical consequences of the use of a prediction model. For this, decision curves[52] could be useful to determine the relation between a chosen risk threshold—for instance, a threshold above which treatment might be started—and the relative value of false-positive and false-negative predictions. This is to evaluate the net benefit of using the model at that specific risk threshold.[53]

Another aspect that has received increasing interest is the comparison of the performance of AI models to that of clinical experts, notably for diagnostic tasks. In 2019, a systematic review that compared the performance of deep learning algorithms to that of health professional assessment in diagnosis of various diseases from medical images found that only 17% of the studies compared performance with that of health professionals in other data than the data used to train the model.[54] Such comparative studies come with additional challenges,[55] such as the need for creating realistic settings where physicians work under realistic practical time constraints and have access to all regular patient information (possibly including existing prediction model results), where performance of model and physicians are evaluated on the same scale, and where optimism about

performance is prevented.[56] For a broader discussion of human vs. machine, we refer to Mayer-Schönberger and Cukier.[57]

Question 9: Were the relevant reporting guidelines for artificial intelligence prediction model studies followed?

Detailed and transparent reporting of AI prediction model development and testing are essential to ensure reproducible, replicable, and critically appraisable results.

Replicability (i.e. re-development and evaluation of the AI prediction model on different data with similar results) and reproducibility (i.e. re-development and evaluation of the AI prediction model on the original data with exactly the same results) should be the core principles for development and testing of AI prediction models. This requires detailed planning, conduct, documentation, and transparent reporting of all steps of the prediction modelling, including all data preparation steps (e.g. feature engineering, initial data analysis[58]), all model selection, tuning, recalibration, and testing steps, and the results from internal and/or external validation approaches.

Recent reviews of AI prediction models showed that the reporting of AI prediction modelling in academic journals is often poor.[59–62] Not only does incomplete and inaccurate reporting prevent adequate reproducibility and replicability of the study findings, it also prevents reviewers, readers, and users appreciating the appropriateness of the used methodology for model development, tuning, and testing, compromising their ability to critically appraise the results. Such problems can easily be avoided by following established reporting guidelines.

For prediction model development, validation, and updating studies, TRIPOD[46,63] has been the widely accepted reporting guideline. An extension specifically for AI prediction models is underway and soon anticipated.[64,65] A full overview of existing reporting guidelines with a specific focus on reporting guidelines of prediction models can be found on https://www.tripod-statement.org/ and on all other type of reporting guidelines on the Equator Network website: https://www.equator-network.org/.

Question 10: Is algorithmic (un)fairness considered and appropriately addressed?

AI prediction models can be a source of unfairness due to, among other reasons, choices in methodology or the data used for model development.

Data driven approaches, including AI prediction models, to inform healthcare professionals about the likely diagnosis and prognosis of patients are often considered to provide objective sources of diagnostic and prognostic information. Applying such prediction models in daily practice can, however, in some cases do more harm than good; some degree of prediction error is inescapable when applying AI prediction models in medical practice. Such errors—and thus potential harm—may be more likely to occur in particular groups of patients. For instance, a comprehensive study of a commercial AI algorithm to manage health in the USA showed Black patients were on average sicker than White patients with the same level of risk.[66] This was attributed to the model using healthcare costs as a proxy for healthcare needs; less money is spent on Black patients with the same level of healthcare needs.[66] There is growing concern about the potential of AI prediction models to increase such racial, gender, and minority group disparities via either choice in model development or existing inequalities encoded in the data used.[67]

When developing or testing prediction models, regardless of the used modelling technique, the algorithmic biases that create potential algorithmic unfairness should be acknowledged and investigated where possible.

Question 11: Is the developed artificial intelligence prediction model open for further testing, critical appraisal, updating, and use in daily practice?

Proprietary algorithms, complex algorithms, and algorithms that received regulatory approval may be more limited in openness, testing, updating, and less welcoming to critical appraisal than non-proprietary algorithms.

The AI prediction model development, testing, and deployment are increasingly the domain of commercial developers. These developers may choose not to disclose their algorithm and ask for a fee per patient for model use.[68] For example, a biomarker-based model to diagnose ovarian cancer has a cost of $897 per patient, which in order to test this model through an external validation approach may require more than $400K investment on model use costs alone.[68] Hence, the ability to test proprietary models may be severely hampered because of financial constraints of the user or tester. That commercially available prediction models are also not guaranteed to perform well was recently illustrated in the context of a widely implemented commercial model to predict sepsis, showing very poor calibration and discrimination in a broad external validation.[69]

Even in the absence of fees for use or testing, complex and proprietary algorithms often largely remain a 'black box' for testers or users, with limited ability for critical appraisal and updating of the AI prediction model. The AI prediction models rarely come with easily applicable model coefficients (see also question 12) that can easily be updated. However, the model may be encoded in closed software. Arguably, such closed software AI prediction models require extra scrutiny of their output through thorough testing with special attention to potential algorithmic unfairness.

Under the current regulatory standards, commercial and non-commercial AI prediction models that have already received regulatory approval—for instance, via a Conformitè Europëenne mark for a medical device—are limited in their opportunities to be updated. If a model is updated, for instance, to become better calibrated in a new setting (e.g. hospital), the updated AI prediction model may require additional regulatory approval before it can be used in practice.

Question 12: Are presented relations between individual features and the outcome not overinterpreted?

Approaches to identify which features are most important in making the predictions can increase interpretability of a black-box AI prediction model, but come with the risk of overinterpretation, including incorrectly inferring that the strongest associations between features and outcome indicate causal relations.

Table 1 Twelve critical questions about artificial intelligence-based prediction of cardiovascular disease

Question	Key considerations
Is AI needed to solve the targeted medical problem?	• Many prediction models already exist, few of them are used • Value of a new complex model over existing simpler model is not guaranteed
How does the AI prediction model fit in the existing clinical workflow?	• Knowing the place of a model in the clinical workflow is essential to identify and address cultural and technical barriers early on
Are the data for prediction model development and testing representative for the targeted patient population and intended use?	• Representative data at development is essential for model calibration • Excluding individuals with atypical presentation or missing data can create bias in predictive performance measures
Is the (time)point of prediction clear and aligned with the feature measurements?	• Feature data should not include information that becomes available only after the intended moment of prediction • Prognostic models require specification of a clear prediction horizon
Is the outcome variable labelling procedure reliable, replicable, and independent?	• Verification of the outcome status should be done accurately • Inaccurate verification may bias predictions and estimates of predictive performance
Was the sample size sufficient for AI prediction model development and testing?	• A priori or a posteriori sample size calculations can be used to justify the sample size
Is optimism of predictive performance of the AI prediction model avoided?	• Performance of AI prediction models must be tested through rigorous internal and external validation procedures
Was the AI model's performance evaluated beyond simple classification statistics?	• There is a large variety of statistics to quantify predictive performance • Traditional performance statistics do not describe clinical consequences of using the AI prediction model
Were the relevant reporting guidelines for AI prediction model studies followed?	• Reporting of prediction modelling studies is often poor • TRIPOD can be used to guide reporting for model development and testing
Is algorithmic (un)fairness considered and appropriately addressed?	• Prediction models have the potential to do harm when applied • Choices in model development and existing inequalities encoded in the data can create prediction models that are harmful to particular groups
Is the developed AI prediction model open for use, further testing, critical appraisal, and updating and use in daily practice?	• Proprietary AI prediction models can be difficult or expensive to test and critical appraisal • Regulatory standards can hamper the opportunities to update existing models that already received regulatory approval
Are presented relations between individual features and the outcome not overinterpreted?	• Explainable AI methodology can be helpful to identify which features contribute most to making predictions • Conclusions about cause and effect purely based on prediction modelling results are rarely justified

The AI prediction models are often criticized by healthcare workers, patients, lawmakers, and scientists for their black-box nature.[70] Unlike regression models, which are usually presented as equations with regression coefficients representing the strength of the multivariable relation between individual features and the outcome, outside of the regression framework (e.g. random forest, neural networks) the strength of multivariable feature—outcome relations may not be directly represented in the software output. Indeed, when only the output of a black-box model is presented to the user (i.e. the predictions), whereas the associations between the individual features and the outcome remain hidden, the predictions are difficult to scrutinize and to question, which in turn may hamper trust of the user in the AI prediction model.

Several approaches exist that aim at opening the black box after the AI prediction model has been developed, to 'explain' which features contribute most to making the prediction. Common examples of these so-called explainable AI approaches[71] are Local Interpretable Model-agnostic Explanations (LIME) and Additive exPlanations (SHAP)[72] values. Analogous to regression coefficients, SHAP and LIME values express both the direction of a feature–outcome association as well as the magnitude of these associations. For an application of SHAP in the context of obstructive coronary artery disease,, we refer the reader to Al'Aref *et al.*[73]

Despite the increasing popularity of approaches that aim to increase interpretability of AI prediction models, several authors have warned against overinterpretation of their results.[74,75] Such approaches do not generally allow for statements about *causal* relations between the selected features and the outcome. This is because causal inference inherently requires information that cannot directly be derived from data but must be provided by the analyst as explicit sets of assumptions.[76] Conclusions about cause and effect purely based on prediction model feature–outcome associations are rarely justified, even (or also) when the modelling is done using AI techniques. For discussions on AI that are specifically designed to explore cause and effect, we refer the reader to Blakely *et al.*[77]

Conclusion

In this article, we proposed 12 critical questions to be asked by readers and reviewers when they are confronted with prediction models that are based on AI. Many of these questions may also have relevance for prediction models that are not based on AI (*Table 1*).

With the increasing use of AI in medicine in general and, in particular, the diagnosis, prognostication, and treatment of cardiovascular diseases, it is important to keep asking critical questions before these prediction models are implemented in practice. Before implementation, many steps need to be taken including steps for data preparation, training of the model and software, as well as the evaluation of performance and impact of the model on clinical decision-making. For an overview of contemporary detailed guidance for each of these steps, we refer to a recent scoping review.[78] For an overview of ethical guidelines related to AI, see Hagendorff.[79]

Acknowledgement

The authors thank Dr Karin R. Jongsma for her valuable comments on an earlier version of this manuscript. Our acknowledgement does not imply endorsement by Dr Jongsma of this article.

Funding

F.W.A. and S.D. are supported by UCL Hospitals, NIHR Biomedical Research Centre and received support from the BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 joint undertaking under grant agreement no. 116074. J.H.M. is supported by National Institutes of Health (USA) grant LM010098. A.L.B. is supported by the German Research Foundation (grant BO3139/4-3) and the German Federal Ministry of Education and Research (grant 01IS18036A).

Conflict of interest: The authors declare no conflict of interest.

Glossary

AI prediction model	An algorithm, tool, software, or rule developed using AI methodology that provides information about the risk (often a probability) for an individual (often a patient) to have a certain disease (diagnostic) or experience a certain health state over time (prognostic)
Features/predictors	The AI model inputs used to make predictions. This may be predictors from structured data (e.g. demographics, patient history, biomarker values) or measurable properties or characteristics of unstructured data (e.g. text, CT images, electrophysiology)
Validation/testing	The evaluation of the predictive performance and/or clinical utility of the AI prediction model. This may be done using internal validation (evaluation on data from same population data as used for the prediction model development, such as by cross-validation) and/or external validation (evaluation on data from other population data as used for the model development). Both types of validation are here referred to as testing
Discrimination	The ability of the AI prediction model to discriminate between individuals with the outcome (e.g. having the disease) and individuals without the outcome (e.g. not having the disease). This is often quantified using the concordance index or area under the ROC curve (AUROC)
Calibration	The ability of the AI prediction model to accurately estimate the risk of the outcome (i.e. the probability of the event of interest). This is often quantified by the calibration in the large, calibration slope, and/or a calibration plot that depicts the estimated risks from the AI model vs. the observed outcome proportions
Model tuning	Tuning is the process of optimizing the so-called hyperparameters of a model, which control the flexibility of the model but also guard against overfit. The optimal values of such parameters are often found by cross-validation procedures
Sample size	The number of individuals from whom data obtained to develop and/or test the AI prediction model

References

1. Friedrich S, Groß S, König IR, Engelhardt S, Bahls M, Heinz J, *et al.* Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: a systematic review with recommendations. *Eur Heart J Digit Health* 2021;**2**:424–436.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;**25**:44–56.
3. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. *J Clin Epidemiol* 2021;**132**:142–145.
4. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 2019;**25**:65–69.
5. Cohen-Shelly M, Attia ZI, Friedman PA, Ito S, Essayagh BA, Ko WY, *et al.* Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *Eur Heart J* 2021;**42**:2885–2896.
6. Tokodi M, Schwertner WR, Kovács A, Tősér Z, Staub L, Sárkány A, *et al.* Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: the SEMMELWEIS-CRT score. *Eur Heart J* 2020;**41**:1747–1756.

7. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;**353**:i2416.
8. Baart SJ, Dam V, Scheres LJJ, Damen JAAG, Spijker R, Schuit E, *et al.* Cardiovascular risk prediction models for women in the general population: a systematic review. *PLoS One* 2019;**14**:e0210329.
9. Carrick RT, Park JG, McGinnes HL, Lundquist C, Brown KD, Janes WA, *et al.* Clinical predictive models of sudden cardiac arrest: a survey of the current science and analysis of model performances. *J Am Heart Assoc* 2020;**9**:e017625.
10. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, *et al.* General cardiovascular risk profile for use in primary care: the Framingham heart study. *Circulation* 2008;**117**:743–753.
11. SCORE2 Working Group and ESC Cardiovascular Risk Collaboration, Hageman S, Pennells L, Ojeda F, Kaptoge S, Kuulasmaa K, *et al.* SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J* 2021; **42**:2439–2454.
12. Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. *Eur Heart J* 2003;**24**:881–882.
13. Lee TH, Marcantonio ER, Mangione CM, Thomas EJ, Polanczyk CA, Cook EF, *et al.* Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation* 1999;**100**:1043–1049.
14. Hand DJ. Classifier technology and the illusion of progress. *Statist Sci* 2006;**21**:1–15.
15. 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.
16. Watson J, Hutyra CA, Clancy SM, Chandiramani A, Bedoya A, Ilangovan K, *et al.* Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? *JAMIA Open* 2020;**3**:167–172.
17. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, *et al.* A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, USA. 2020: pp. 1–12. https://doi.org/10.1145/3313831.3376718.
18. Kappen TH, van Loon K, Kappen MAM, van Wolfswinkel L, Vergouwe Y, van Klei WA, *et al.* Barriers and facilitators perceived by physicians when using prediction models in practice. *J Clin Epidemiol* 2016;**70**:136–145.
19. Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *J Clin Epidemiol* 2019;**105**:136–141.
20. Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. *Stat Med* 2019;**38**:3444–3459.
21. Luijken K, Wynants L, van Smeden M, Van Calster B, Steyerberg EW, Groenwold RHH, *et al.* Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J Clin Epidemiol* 2020;**119**:7–18.
22. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;**98**:691–698.
23. Kouw WM, Loog M. An introduction to domain adaptation and transfer learning. *Arxiv* 2018. https://arxiv.org/abs/1812.11806.
24. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;**17**:230.
25. Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KGM. Advantages of the nested case-control design in diagnostic research. *BMC Med Res Methodol* 2008;**8**:48.
26. Moons KG, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. *Clin Chem* 2004;**50**:473–476.
27. Moons KGM, Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic studies? *J Clin Epidemiol* 2002;**55**:633–636.
28. Hernandez-Suarez DF, Kim Y, Villablanca P, Gupta T, Wiley J, Nieves-Rodriguez BG, *et al.* Machine learning prediction models for In-hospital mortality after transcatheter aortic valve replacement. *JACC Cardiovasc Interv* 2019;**12**:1328–1338.
29. Wolbers M, Koller MT, Stel VS, Schaer B, Jager KJ, Leffondré K, *et al.* Competing risks analyses: objectives and approaches. *Eur Heart J* 2014;**35**:2936–2941.
30. Myhre PL, Vaduganathan M, Greene SJ. Diagnosing heart failure with preserved ejection fraction in 2019: the search for a gold standard. *Eur J Heart Fail* 2020;**22**: 422–424.
31. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, *et al.* Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ* 2013;**346**:f2350.
32. Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PMM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 2007;**11**:iii, ix–51.
33. Bertens LCM, Broekhuizen BDL, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, *et al.* Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS Med* 2013; **10**:e1001531.
34. Zir LM, Miller SW, Dinsmore RE, Gilbert JP, Harthorne JW. Interobserver variability in coronary angiography. *Circulation* 1976;**53**:627–632.
35. Bunting KV, Steeds RP, Slater LT, Rogers JK, Gkoutos GV, Kotecha D. A practical guide to assess the reproducibility of echocardiographic measurements. *Joe Am Soc Echocardiogr* 2019;**32**:1505–1515.
36. Koivumäki JK, Nikus KC, Huhtala H, Ryödi E, Leivo J, Zhou SH, *et al.* Agreement between cardiologists and fellows in interpretation of ischemic electrocardiographic changes in acute myocardial infarction. *J Electrocardiol* 2015;**48**:213–217.
37. Nagueh SF, Abraham TP, Aurigemma GP, Bax JJ, Beladan C, Browning A, *et al.* Interobserver variability in applying American Society of Echocardiography/ European Association of Cardiovascular Imaging 2016 Guidelines for Estimation of Left Ventricular Filling Pressure. *Circ Cardiovasc Imaging* 2019;**12**:e008122.
38. Desmond M, Duesterwald E, Brimijoin K, Brachman M, Pan Q. Semi-automated data labeling. *J Mach Learn Res* 2021;**133**:156–169.
39. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;**368**:m441.
40. van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, *et al.* Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res* 2019;**28**:2455–2474.
41. McKenna WJ, Judge DP. Epidemiology of the inherited cardiomyopathies. *Nat Rev Cardiol* 2021;**18**:22–36.
42. Christodoulou E, van Smeden M, Edlinger M, Timmerman D, Wanitschek M, Steyerberg EW, *et al.* Adaptive sample size determination for the development of clinical prediction models. *Diagn Progn Res* 2021;**5**:6.
43. Balki I, Amirabadi A, Levman J, Martel AL, Emersic Z, Meden B, *et al.* Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Can Assoc Radiol J* 2019;**70**:344–353.
44. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, Smeden M, *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med* 2021;**40**:4230–4251.
45. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, *et al.* Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012;**98**:683–690.
46. Moons KGM, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;**162**: W1–W73.
47. Hornung R, Bernau C, Truntzer C, Wilson R, Stadler T, Boulesteix AL. A measure of the impact of CV incompleteness on prediction error estimation with application to PCA and normalization. *BMC Med Res Methodol* 2015;**15**:95.
48. Al-Farra H, de Mol BAJM, Ravelli ACJ, ter Burg WJPP, Houterman S, Henriques JPS, *et al.* Update and, internal and temporal-validation of the FRANCE-2 and ACC-TAVI early-mortality prediction models for Transcatheter aortic Valve Implantation (TAVI) using data from the Netherlands heart registration (NHR). *Int J Cardiol Heart Vasc* 2021;**32**:100716.
49. Hickey GL, Grant SW, Caiado C, Kendall S, Dunning J, Poullis M, *et al.* Dynamic prediction modeling approaches for cardiac surgery. *Circ Cardiovasc Qual Outcomes* 2013;**6**:649–658.
50. Wessler BS, Nelson J, Park JG, McGinnes H, Gulati G, Brazil R, *et al.* External validations of cardiovascular clinical prediction models: a large-scale review of the literature. *Circ Cardiovasc Qual Outcomes* 2021;**14**:e007858.
51. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;**21**:128–138.
52. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making* 2015;**35**:162–169.
53. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;**3**:18.
54. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;**1**: e271–e297.
55. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009; **338**:b606.
56. van Smeden M, Van Calster B, Groenwold RHH. Machine learning compared with pathologist assessment. *JAMA* 2018;**319**:1725–1726.
57. Mayer-Schönberger V, Cukier K. *Big Data: A Revolution that will Transform how We Live, Work, and Think. First Mariner Books edition*. Boston: Mariner Books, Houghton Mifflin Harcourt; 2014.

58. Huebner M, Vach W, le Cessie S. A systematic approach to initial data analysis is good research practice. *J Thorac Cardiovasc Surg* 2016;**151**:25–27.
59. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;**368**:m689.
60. Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M, *et al.* Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020;**369**:m1328.
61. Navarro CLA, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, *et al.* Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021;**375**:n2281.
62. Dhiman P, Ma J, Navarro CA, Speich B, Bullock G, Damen JA, *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.
63. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;**162**:55–63.
64. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;**393**:1577–1579.
65. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, *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.
66. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;**366**:447–453.
67. Chouldechova A, Roth A. The frontiers of fairness in machine learning. *Arxiv*. 2018. https://doi.org/10.48550/arXiv.1810.08810.
68. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? *J Am Med Inform Assoc* 2019; **26**:1651–1654.
69. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, *et al.* External validation of a widely implemented proprietary sepsis prediction Model in hospitalized patients. *JAMA Intern Med* 2021;**181**:1065.
70. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Ann Intern Med* 2020;**172**:59.
71. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, *et al.* Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 2020;**58**:82–115.
72. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: ACM. 2020: pp. 180–186.
73. Al'Aref SJ, Maliakal G, Singh G, van Rosendael AR, Ma X, Xu Z, *et al.* Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. *Eur Heart J* 2020;**41**:359–367.
74. Lipton ZC. The Mythos of Model Interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* 2018;**16**:31–57.
75. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digital Health* 2021;**3**: e745–e750.
76. Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, *et al.* Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit Health* 2020;**2**:e677–e680.
77. Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. *Int J Epidemiol* 2021;**49**:2058–2064.
78. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med* 2022;**5**:2.
79. Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. *Minds Mach* 2020; **30**:99–120.