

# 2024 Revision of the level of evidence grading system for ESC clinical practice guideline recommendations II: diagnostic tests and prediction models

Emanuele Di Angelantonio <sup>1,2,3,4,5,6,\*†</sup>, Lisa Pennells <sup>1,2,\*†</sup>,  
Magdy Abdelhamid <sup>7</sup>, Victor Aboyans <sup>8</sup>, Riccardo Asteggiano <sup>9</sup>,  
Jelena Čelutkienė <sup>10</sup>, Diederick E. Grobbee <sup>11,12</sup>, Bernard lung<sup>13</sup>, Peter Jüni <sup>14</sup>,  
John William McEvoy <sup>15,16</sup>, Amina Rakisheva <sup>17</sup>, Xavier Rossello <sup>18,19</sup>,  
Frank L. J. Visseren <sup>20</sup>, Colin Baigent <sup>14</sup>, and Eva B. Prescott <sup>21</sup>

<sup>1</sup>British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Papworth Road, Cambridge Biomedical Campus, Cambridge CB2 0BB, UK; <sup>2</sup>Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Papworth Road, Cambridge Biomedical Campus, Cambridge CB2 0BB, UK; <sup>3</sup>NIHR Blood and Transplant Research Unit in Donor Health and Behaviour, University of Cambridge, Cambridge, UK; <sup>4</sup>BHF Centre of Research Excellence, School of Clinical Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK; <sup>5</sup>Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK; <sup>6</sup>Health Data Science Centre, Human Technopole, V.le Rita Levi-Montalcini 1, 20157 Milan, Italy; <sup>7</sup>Department of Cardiovascular Medicine at the Faculty of Medicine, Kasr El Ainy, Cairo University, Cairo, Egypt; <sup>8</sup>Department of Cardiology at the Dupuytren University Hospital, Limoges, France; <sup>9</sup>Faculty of Medicine, Insubria University, Varese, and Laboratorio Analisi e Ricerca Clinica, Turin, Italy; <sup>10</sup>Clinic of Cardiac and Vascular Diseases, Vilnius University, Vilnius, Lithuania; <sup>11</sup>University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands; <sup>12</sup>The Julius Center for Health Sciences and Primary Care, Utrecht, Netherlands; <sup>13</sup>Hospital Bichat-Claude Bernard, Université de Paris, Paris, France; <sup>14</sup>Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK; <sup>15</sup>Department of Cardiology, University of Galway, School of Medicine, Galway, Ireland; <sup>16</sup>National Institute for Prevention and Cardiovascular Health, Galway, Ireland; <sup>17</sup>Cardiology Department, City Cardiology Center, Almaty, Kazakhstan; <sup>18</sup>Cardiology Department, Hospital Universitari Son Espases, Health Research Institute of the Balearic Islands (IdiSBA), Palma de Mallorca, Spain; <sup>19</sup>Facultad de Medicina, Universitat Illes Balears (UIB), Palma de Mallorca, Spain; <sup>20</sup>Department of Vascular Medicine, University Medical Center Utrecht, Utrecht, Netherlands; and <sup>21</sup>Department of Cardiology, Bispebjerg University Hospital, Copenhagen, Denmark

Received 12 September 2024; revised 20 November 2024; accepted 8 January 2025; online publish-ahead-of-print 21 March 2025

## Abstract

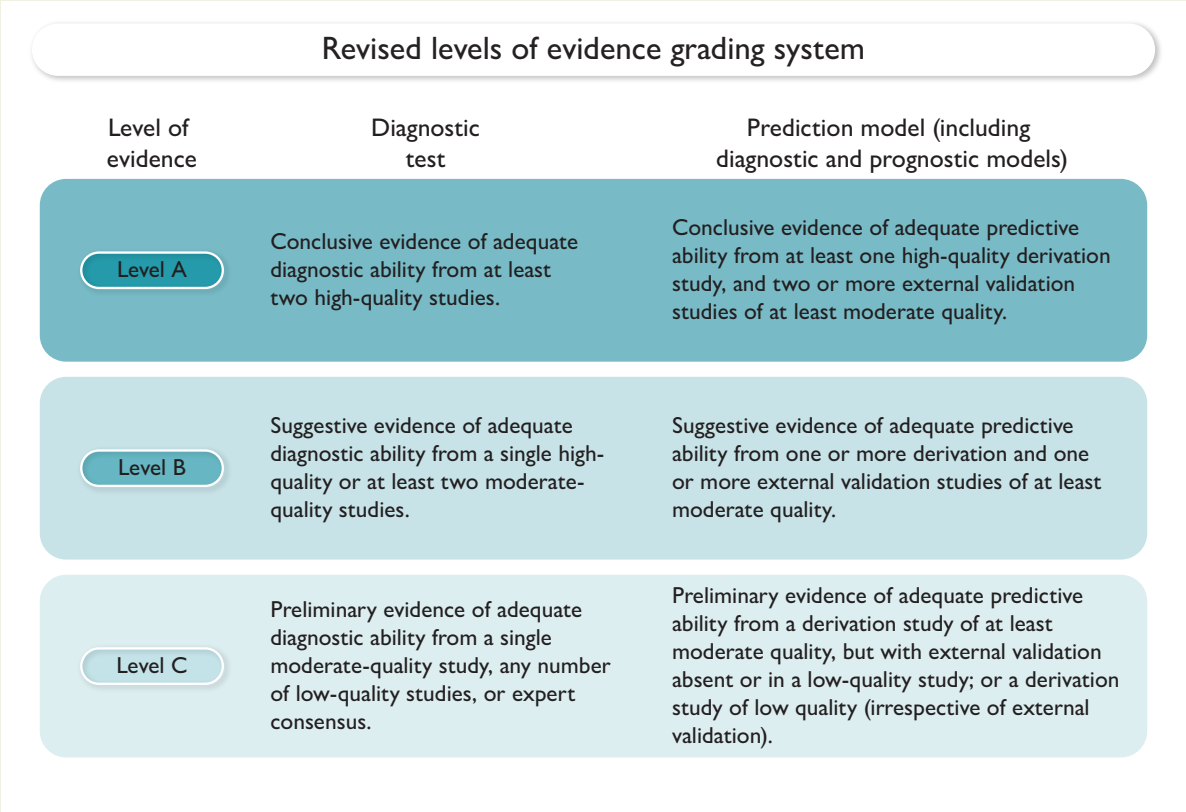
The level of evidence (LOE) grading system for European Society of Cardiology (ESC) Clinical Practice Guidelines (CPG) classifies the quality of the evidence supporting a recommendation. However, the current taxonomy does not fully consider the optimal study design necessary to establish evidence for different types of recommendations in ESC guidelines. Therefore, two separate task forces of clinical and methodological experts were appointed by the CPG Committee, with the first tasked with updating the LOE grading system for therapy and prevention and the second responsible for developing a LOE grading system for diagnosis and prediction. This report from the second of these Task Forces develops a new system for diagnostic tests and prediction models which maintains the three-level grading structure to classify the quality of the evidence but introduces new definitions specific for diagnosis and prediction. For diagnostic tests, LOE A represents conclusive evidence of adequate diagnostic ability from at least two high-quality studies. Level of evidence B represents suggestive evidence from one high-quality or at least two moderate-quality studies. Level of evidence C represents preliminary evidence not classified as A or B, including evidence from less than two moderate-quality studies, or from expert consensus. For prediction models, LOE A represents conclusive evidence of adequate predictive ability from at least one high-quality derivation and two or more external validation studies of at least moderate quality. Level of evidence B represents suggestive evidence in one or more derivation studies and one or more external validation studies of at least moderate quality. Level of evidence C represents preliminary evidence not classified as A or B, including evidence from a derivation study of at least moderate quality, but with low quality or no external validation, or a derivation study of low quality.

\* Corresponding authors. Email: [ed303@medschl.cam.ac.uk](mailto:ed303@medschl.cam.ac.uk) (E.D.A.); Tel: +44(0)1223 748498, [lamp2@medschl.cam.ac.uk](mailto:lamp2@medschl.cam.ac.uk) (L.P.)

† The first two authors contributed equally to the study.

© The European Society of Cardiology 2025. All rights reserved. For permissions, please email: [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

Graphical Abstract



The revised level of evidence grading system.

**Keywords** Clinical Practice Guidelines • Cardiovascular disease • European Society of Cardiology • Evidence grading • Diagnostic tests • Prediction models • Study quality • Diagnostic ability • Predictive ability

Introduction

The European Society of Cardiology (ESC) is a professional medical society that has published Clinical Practice Guidelines (CPG) on the prevention, diagnosis, and treatment of cardiovascular diseases since 1994.<sup>1</sup> These cover a wide range of topics, including the management of stable and acute coronary syndromes, heart failure, arrhythmias, hypertension, valvular heart disease, congenital heart disease, and cardiovascular prevention. Each guideline is developed by a task force composed of experts in the relevant area, including as necessary cardiologists, surgeons, allied healthcare professionals, patients, clinical trialists, epidemiologists, and methodologists. Each guideline undergoes a rigorous review process involving experts and representatives from up to 57 National Cardiac Societies of the ESC. A strict declaration of interest policy is enforced for both task force members and reviewers. All recommendations must receive at least 75% of the task force members' votes to be approved.

The ESC uses a three-tiered level of evidence (LOE) grading system, which classifies the quality of the evidence supporting each recommendation in a guideline. The current criteria for each LOE was first published in 2000 as part of the ESC guidelines on the management of acute coronary syndromes without persistent ST segment elevation<sup>2</sup> based on written recommendations of the ESC Committee for Scientific and Clinical Initiatives which were made available on the ESC website in the same year.<sup>1</sup> Since

then, the LOE system has been widely adopted and used in nearly all subsequent ESC guidelines without major adaptations.<sup>3</sup>

A limitation of the current LOE system is that it is uniformly applied to recommendations covering therapy, prevention, diagnosis, and prediction.<sup>4</sup> However, the optimal study designs necessary to establish evidence for therapy and prevention vary from those necessary to establish evidence for diagnosis or prediction, where randomized controlled trials are not central. For example, in relation to diagnostic tests and prediction models the current LOE does not specify the appropriate study design, methodology, and the quality needed to classify the evidence for such recommendations. As a result, with the current system, LOE A is rarely achieved for diagnostic tests and prediction models, while LOE B might be achieved by underpowered studies of suboptimal quality.

In view of these limitations, the ESC CPG Committee appointed a methodology task force and requested its members revise the LOE grading system as it relates to therapy and prevention (Part I) and to diagnostic tests and prediction models (Part II). The objective of the second phase of the task force's work was to develop a new LOE concept to inform guideline recommendations for diagnostic tests and prediction models. The concept was collaboratively developed through a process that involved six collective meetings of the task force members over a span of eight months. The drafting of the text was undertaken by a writing group (E.D.A., L.P., C.B., and E.B.P.), and subsequently reviewed by all members of the task force. The new LOE system was

**Table 1** New levels of evidence grading system for diagnostic tests and prediction models

| Diagnostic test     |                                                                                                                                                  | Prediction model<br>(including diagnostic and prognostic models)                                                                                                                                                                                    |
|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Level of evidence A | Conclusive evidence of adequate diagnostic ability from at least two high-quality studies <sup>a</sup>                                           | Conclusive evidence of adequate predictive ability from at least one high-quality derivation study, and two or more external validation studies of at least moderate quality <sup>a</sup>                                                           |
| Level of evidence B | Suggestive evidence of adequate diagnostic ability from a single high-quality or at least two moderate-quality studies <sup>a</sup>              | Suggestive evidence of adequate predictive ability from one or more derivation and one or more external validation studies of at least moderate quality <sup>a</sup>                                                                                |
| Level of evidence C | Preliminary evidence of adequate diagnostic ability from a single moderate-quality study, any number of low-quality studies, or expert consensus | Preliminary evidence of adequate predictive ability from a derivation study of at least moderate quality, but with external validation absent or in a low-quality study; or a derivation study of low quality (irrespective of external validation) |

A study is defined as a single dataset from a distinct study population.

Level of evidence C includes all forms of evidence that are not classed as A or B.

The given criteria can also be applied to evidence comparing diagnostic tests or prediction models to assess evidence for preferential use of one test or model over another.

<sup>a</sup>Criteria for defining quality of studies and adequate diagnostic or predictive ability are described in the 'Explanatory notes'.

subjected to an impact assessment to evaluate its influence on LOE in previous guidelines. The document underwent a rigorous peer review process involving expert methodologists and the CPG Committee, and the final document was approved by the ESC Board.

## Definitions of diagnostic tests and prediction models

### Diagnostic tests

A diagnostic test is a measurement or a combination of measurements which is used to help make a diagnosis of a given disease or condition. Diagnostic test studies commonly focus on evaluating the accuracy and reliability of the test in making a disease diagnosis, i.e. the focus is on assessing the performance of the test where the 'true' disease status is determined using a reference standard approach.

### Prediction models

Prediction models include any statistical model that uses single or multiple measurements to give a probability estimate or other probability-related quantity and includes both diagnostic and prognostic models. A diagnostic prediction model estimates the probability that a target disease or condition is already present in an individual (e.g. the Wells criteria for pulmonary embolism<sup>5</sup>), whereas a prognostic model estimates the probability a target disease will occur in the future (e.g. SCORE2<sup>6</sup> or CHA<sub>2</sub>DS<sub>2</sub>-VASc<sup>7</sup>). The statistical modelling commonly involves regression approaches, but may only provide a 'scoring' or classification rule to guide diagnosis or quantify future risk. As previously described,<sup>8,9</sup> diagnostic and prognostic prediction models have similar characteristics, with both aiming to estimate the probability of a clinical condition, either being currently present or potentially occurring in the future, and similar considerations are necessary when evaluating the evidence for their adequate performance. In contrast to diagnostic tests, assessing the evidence for prediction model quality involves two-fold consideration of (i) the quality of the data and methods used to create the model (derivation) and (ii) the quality of the data and methods used in the independent assessment of predictive ability (validation).

The current guidance relates to the assessment of evidence relating only to the ability of a diagnostic test or prediction model to *identify* the current or future presence of a clinical condition. It does not relate to the ability of the test or model usage to improve or prevent the clinical outcome through resulting intervention. Such a connection between tests/models and clinical outcomes typically operates indirectly, mediated by these treatment choices, and unlike therapeutic interventions (i.e. treatments), diagnostic tests and prediction models are seldom subjected to randomized controlled trials where outcome prevention is assessed.<sup>10</sup> Therefore, the appraisal of the evidence for tests or models requires a different approach than the assessment of the evidence for treatments, which are mostly tested using randomized clinical trials. However, if a diagnostic test or prediction model is an integral part of an intervention strategy, directly influencing treatment decisions and risk of outcomes, evidence supporting its use in preventing or improving outcomes should be assessed according to the same standards as for other therapeutic interventions. This aligns with the principles outlined in Part I of the updated ESC LOE grading system (insert ref when available).

## Overview of the new level of evidence grading system

The LOE grading system for diagnostic tests and prediction models is presented in [Table 1](#). It provides a standardized approach for evaluating the strength of evidence in ESC guidelines relating to the use of diagnostic tests or prediction models. Accompanying criteria for defining quality of studies and diagnostic or predictive ability are described in the 'Explanatory notes'.

## Description of levels of evidence for diagnostic tests

### Level of evidence A

Level of evidence A applies to conclusive evidence of the diagnostic ability of a test, which is unlikely to be refuted in a future study of the same test in a similar population of individuals that are representative of the target

clinical population. A conclusion of adequate diagnostic ability from at least two independent high-quality studies is required to merit LOE A. Each study should be assessed for quality, to confirm there is no important source of bias resulting from study design and participant selection, practical execution of study methods, or analysis of data. What constitutes adequate diagnostic ability, high-quality data (including data size and statistical power), and lack of bias is context specific, and criteria for defining these requirements are provided in the 'Explanatory notes'.

## Level of evidence B

Level of evidence B applies to evidence that is suggestive but not conclusive, and further research, if conducted, could conceivably result in changes to the recommendation. Here, the demonstration of adequate diagnostic ability is limited to either a single high-quality study (and therefore generalizability of findings is not ensured) or two or more studies of at least moderate quality, in which each has a non-negligible potential for bias but when combined the replication of adequate diagnostic ability reduces the chance of clinically important bias.

## Level of evidence C

Level of evidence C is the lowest LOE and represents preliminary findings on diagnostic ability from studies and scenarios that do not meet the criteria for LOE A or B, including findings based on expert consensus. Either bias cannot be ruled out or generalizability to the clinically relevant population is not ensured, or both. Further studies are needed to determine true diagnostic ability.

## Description of levels of evidence for prediction models

### Level of evidence A

Level of evidence A applies to conclusive evidence of the predictive ability of a prognostic model, which is unlikely to be refuted in a future study of the same model in a population of individuals that are representative of the same target clinical population. Derivation of the model should be from at least one high-quality study. In addition, to ensure evidence is reliable, demonstrating adequate predictive ability in external validation studies is also necessary. External studies should be from data sources collected independently from those used for model development, and a withheld portion of data (split sample) from the original development dataset is not classed as an external study. A conclusion of adequate predictive ability from two or more external studies of at least moderate quality is required. All derivation and external validation studies should be assessed to confirm there is no important source of bias resulting from study design and participant selection, practical execution of study methods, or analysis of data. What constitutes adequate predictive ability, high-quality data (including data size), and lack of bias is context specific, and criteria for defining these requirements are provided in the 'Explanatory notes'.

### Level of evidence B

Level of evidence B applies to evidence that is suggestive but not conclusive, and further research, if conducted, could conceivably result in changes of the recommendation. Here, adequate predictive ability is demonstrated in (i) one or more risk model derivation studies of at least moderate quality and (ii) one or more external validation studies of at least moderate quality. The presence of only one external validation study reduces confidence in the generalizability of findings to other external clinically relevant populations. Nevertheless, in this situation

predictive ability will have been demonstrated in both derivation and external validation studies, thereby providing some evidence of generalizability and consistency across two independent studies. Likewise, for each individual study of moderate rather than high quality the potential for bias may be non-negligible, but replication of adequate predictive ability in two settings reduces the chance of clinically important bias.

## Level of evidence C

Level of evidence C is the lowest LOE and represents preliminary findings on predictive ability from studies and scenarios that do not meet the criteria for LOE A or B. These may include findings of adequate predictive ability in a single model derivation study only, with external validation either absent or in a low-quality study, or any situation where the derivation study is of low quality (irrespective of external validation). In this case, optimism and bias cannot be ruled out, and generalizability to the clinically relevant population is not ensured. Further studies are needed to determine true predictive ability.

For each LOE, the commonality across diagnostic tests and prediction models is the number of external validation studies which demonstrate adequate diagnostic or predictive ability. Level of evidence A requires two external validation studies and LOE B only one. Validation across at least two external validation studies provides evidence of generalizability, which is particularly important when a test or model is recommended for use in multiple populations or countries. However, the required quality of these studies differs between the diagnostic test and prediction model categories. Lower quality is accepted for external validation of prediction models, since adequate performance must also be demonstrated in the model derivation study; the demonstration of predictive ability in both derivation and external validation data sources increases generalizability, provides evidence against overfitting, and lowers the chance of clinically relevant bias, therefore offsetting the need for the more stringent study quality criteria used for diagnostic tests. It should also be noted that the term study implies analysis of a single data source from a distinct population. Some single publications may report findings on diagnostic or predictive ability from multiple distinct data sources, each of which should be classed as a separate study for the purposes of determining the LOE. For example, the publication describing development of the SCORE2 10-year risk prediction algorithms includes evidence of predictive ability from several high-quality derivation and external validation data sources<sup>6</sup> and so is sufficient to conclude LOE A.

## Explanatory notes

Guidance is provided in the following sections on criteria relevant for assessing data quality, and the diagnostic ability of tests and predictive ability of models. [Tables 2](#) and [3](#) summarize key criteria and include separate sections on (i) data quality and (ii) diagnostic or predictive ability. Each section contains a list of criteria that are classified as either essential or desirable. Criteria can be used to determine adequate diagnostic/predictive performance and data quality in the following way.

### Determination of data quality

A study dataset can be defined as:

- (1) High quality: if all essential and most desirable data quality criteria are met.
- (2) Moderate quality: if all essential but only some desirable criteria are met.
- (3) Low quality: if not all essential criteria are met.

**Table 2** Summary of criteria for assessing data quality and diagnostic ability of diagnostic tests

| Domain                                                                             |                                | Criteria                                                                                                                                                                                                           | Classification: Essential (E); Desirable (D) |
|------------------------------------------------------------------------------------|--------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------|
| Data quality                                                                       | Patient selection/study design | • Cross-sectional study—consecutive or randomly selected individuals with appropriate inclusion/exclusion criteria                                                                                                 | E                                            |
|                                                                                    |                                | • Sufficient sample size <sup>a</sup>                                                                                                                                                                              | E                                            |
|                                                                                    |                                | • Representative of the target population in terms of disease or condition prevalence                                                                                                                              | E                                            |
|                                                                                    |                                | • Representative of target population in terms of absolute values and variance of measurements recorded by the diagnostic test                                                                                     | E                                            |
|                                                                                    |                                | • Representative of the target population in terms of characteristics not included in the test measurement (e.g. age, comorbidities, socio-demographic characteristics, recent calendar period of data collection) | E                                            |
|                                                                                    | Index test                     | • Blinding of 'true' disease status/diagnosis when applying the test <sup>b</sup>                                                                                                                                  | D                                            |
|                                                                                    |                                | • Pre-specification of test threshold—rather than determination according to test ability (sensitivity/specificity)                                                                                                | D                                            |
|                                                                                    |                                | • Test characteristics (e.g. technology used, consistency, execution, interpretation) consistent with that which will be used in practice                                                                          | D                                            |
|                                                                                    | Reference standard             | • Reference standard test used to assess the research test under study is sufficiently accurate in defining/identifying true presence of the condition/outcome                                                     | E                                            |
|                                                                                    |                                | • Reference standard results blinded to index test result (i.e. reference standard test applied similarly, regardless of index test result) <sup>b</sup>                                                           | D                                            |
|                                                                                    |                                | • Condition identified by reference standard is aligned to that which will occur in practice, in terms of severity and other relevant condition/disease-specific characteristics                                   | E                                            |
|                                                                                    | Flow and timing                | • Reference standard consistently applied to define true status of study participants in terms of timing and standardized application methods                                                                      | D                                            |
|                                                                                    |                                | • All patients included in the analysis with minimal risk of informative drop out                                                                                                                                  | D                                            |
|                                                                                    | Diagnostic ability             | • Metric of diagnostic ability is quantified and appears sufficient <sup>c</sup> given the clinical context                                                                                                        | E                                            |
| Measure of diagnostic accuracy reported: sensitivity, specificity, PPV, NPV, AUROC |                                |                                                                                                                                                                                                                    |                                              |

Criteria marked as E are considered essential, whereas those considered desirable are marked D. Some desirable criteria may be considered more important or essential depending on clinical context.

<sup>a</sup>Sufficient sample size can be judged by justification in the published work or by calculation using published tools.<sup>11</sup>

<sup>b</sup>These criteria may be considered essential for tests where subjective judgment is required.

<sup>c</sup>Which statistics are reported and the required value is specific to the clinical context.

## Determination of diagnostic or predictive ability

Adequate diagnostic or predictive ability can be concluded only if all essential criteria are met.

## Diagnostic tests: guidance on assessment of diagnostic ability

Metrics to assess the diagnostic ability of a test are primarily focused on measures of accuracy such as sensitivity, specificity, positive predictive

value (PPV), and negative predictive value (NPV). Measures of reliability may also be reported when repeat testing is available (these include, for example, Cohen's kappa and coefficient of variation). Where a test is used to classify patients into categories (e.g. positive, negative, and uncertain), accuracy statistics for more than one threshold can be considered. The combination of sensitivity and specificity as an area under the receiver operating characteristic curve may also add confidence to conclusions since this can demonstrate diagnostic ability over the full range of test values. Accuracy metrics will commonly be calculated based on the ability of the test to detect the presence of a condition that is defined in the study using a measured reference standard. The alignment

**Table 3** Summary of criteria for assessing data quality and predictive ability of risk prediction models

| Domain             |                            | Criteria                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | Classification:<br>essential (E);<br>desirable (D) |
|--------------------|----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|
| Data Quality       | Participant Selection      | <ul style="list-style-type: none"> <li>Appropriate study design:<br/>For diagnostic prediction models: cross-sectional cohort, nested case-cohort, or nested case-control study or RCT with appropriate inclusion/exclusion criteria<br/>For prognostic prediction models: cohort, nested case-cohort, or nested case-control study or RCT with appropriate inclusion/exclusion criteria. With RCT, any treatment effects must be allowed for, either by excluding participants from the model derivation or adjustment for treatment at the analysis stage</li> </ul> | E                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>Individuals in study representative of the target population considering:               <ul style="list-style-type: none"> <li>(i) prevalence (diagnostic model) or incidence (prognostic model) of the condition</li> <li>(ii) value and variance of measurements used in the model</li> <li>(iii) other characteristics not included in the model (e.g. geographical location of residence)</li> </ul> </li> </ul>                                                                                                            | E                                                  |
|                    |                            | Sufficient sample size for model derivation or validation <sup>a</sup> [see sections 'Sample size considerations for prediction model derivation (both diagnostic and prognostic models)' and 'Sample size considerations for model validation (both diagnostic and prognostic models)']                                                                                                                                                                                                                                                                               | E                                                  |
|                    | Predictors                 | <ul style="list-style-type: none"> <li>Predictors defined and assessed in a similar way for all participants</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                | D                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>Blinding of the outcome applied when measuring the predictors</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                        | D                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>All predictors available to measure in the target population</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                         | E                                                  |
|                    | Outcome/reference standard | <ul style="list-style-type: none"> <li>Predictor characteristics (e.g. definition, technology used to measure, consistency, timing of measurement, and interpretation) are consistent with that which will be used in clinical practice</li> </ul>                                                                                                                                                                                                                                                                                                                     | E                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>Outcome is determined appropriately according to current clinical practice</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                           | E                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>Reference standard used for outcome ascertainment is sufficiently accurate in identifying true presence of the condition/outcome</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                     | E                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>The reference standard/outcome definition is pre-specified</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                           | E                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>The predictors are excluded from the outcome definition (i.e. are not part of outcome diagnosis)</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                     | D                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>The reference standard/outcome is defined and determined in a consistent way for all participants</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                    | E                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>The reference standard/outcome results are blinded to predictor measurement</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                          | D                                                  |
| Predictive ability | Analysis                   | <ul style="list-style-type: none"> <li>The time interval between predictor measurement and reference standard/outcome assessment is appropriately aligned with the time period over which the model predicts</li> </ul>                                                                                                                                                                                                                                                                                                                                                | D                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>The condition identified by reference standard/predicted in the study is comparable to that which will occur in practice, in terms of severity and other relevant condition/disease-specific characteristics</li> </ul>                                                                                                                                                                                                                                                                                                         | D                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>Continuous predictors maintained without unnecessary categorization, with incorporation of non-linear effects where relevant through use of splines or fractional polynomials. If categorization is applied, this should be predefined rather than chosen to optimize model fit</li> </ul>                                                                                                                                                                                                                                      | D                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>Predictors defined consistently in derivation and validation datasets</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                | D                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>All eligible participants included in the analysis (e.g. not excluded due to drop out during follow-up)</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                              | D                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>Missing data appropriately handled (see section 'Missing data')</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                      | D                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>Initial selection of candidate predictors not based on univariable association with outcome. Rather, prior evidence should be considered regarding clinical relevance, as well as reliability, consistency, applicability, availability, and cost of measurement. This should be followed by multivariable modelling</li> </ul>                                                                                                                                                                                                 | D                                                  |
|                    |                            | <ul style="list-style-type: none"> <li>Appropriate handling of data complexity in terms of censoring, competing risks, and sampling design</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                  | E                                                  |

Continued

**Table 3 Continued**

| Domain | Criteria                                                                                                                                                                                                                                                      | Classification:<br>essential (E);<br>desirable (D) |
|--------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|
|        | <ul style="list-style-type: none"> <li>Calibration<sup>b</sup> and discrimination assessed using methods appropriate to study design and data characteristics (e.g. allowing for time to event, censoring, sampling design, competing risks, etc.)</li> </ul> | E                                                  |
|        | <ul style="list-style-type: none"> <li>Optimism allowed for or assessed (model derivation/internal validation only)<sup>c</sup></li> </ul>                                                                                                                    | E                                                  |
|        | <ul style="list-style-type: none"> <li>Predictor effects/weights fully specified and aligned with final model presented in analysis (e.g. not crudely weighted or simplified without re-fitting)</li> </ul>                                                   | D                                                  |

Criteria marked as E are considered essential, whereas those considered desirable are marked as D. Some desirable criteria may be considered more important or essential depending on clinical context.

<sup>a</sup>Sufficient sample size can be judged by justification in the published work or by calculation using published tools<sup>12,13</sup>.

<sup>b</sup>Calibration is only relevant in external validation if the dataset is truly representative of the contemporary target population.

<sup>c</sup>Optimism assessment may not be necessary for very large studies, or for models developed with simultaneous presentation of results from an external validation in a separate data source.

of the reference standard with presence of the condition of interest is crucial to consider. Any mismatch between the two should reduce confidence in the values of accuracy metrics obtained. Some reference standards will be improved over time, and this should be kept in mind when assessing diagnostic ability of historical studies.

The diagnostic threshold level and the reliability of these tools should be based on consideration of how the test is to be used in the clinical context. For example, sometimes it is more important to rule out a particular disease with high risk of a severe adverse clinical outcome (e.g. D-dimer for pulmonary embolism). In this case, a high sensitivity and NPV may be most important. Tests that are more invasive have higher risk of harms, which must be balanced against the harm associated with misdiagnosis due to alternative test options that are less invasive but also less accurate. These include both the harms of a 'false-negative' and 'false-positive' test result. We acknowledge that this description, while illustrative, may be simplistic relative to the complexities of clinical practice, where tests are most often used as part of multi-component diagnostic and treatment algorithms and do not stand alone.

## Diagnostic tests: guidance on determining data quality for assessment of diagnostic tests

Studies for diagnostic tests can be cross-sectional or prospective/longitudinal (where follow-up may be necessary to determine whether a condition was present at the time of the test). While in theory nested case-control or case-cohort studies can be used if the sampling fraction is known and accounted for in the analysis, such study designs may lead to optimistic results.<sup>14</sup> High-quality studies are those which are:

- (1) Sufficiently large in size,
- (2) Representative of the target population in terms of the prevalence of the condition that the test aims to diagnose (this is particularly important for PPV and NPV),
- (3) Representative of the target population in terms of the absolute values and variance of measurements recorded by the diagnostic test,
- (4) Representative of the target population in terms of other characteristics not included in the test measurement (e.g. geographical location of residence, socio-demographic and ethnic representation, recent calendar period of data collection).

Sufficient sample size can be judged based on detail given in the study-specific publication as well as published tools.<sup>11</sup> Careful consideration of points (2), (3), and (4) will be needed in the context of what is known about the target population, which is likely to include multiple countries. Where possible, reported statistics on disease prevalence and test values should be sought in order to judge whether study data are representative, although this may not always be available and clinical judgement may be needed. A more detailed assessment of potential bias and applicability/generalizability can be made for each data source, and the way in which it has been utilized, using additional signalling questions from the QUADAS-2 tool.<sup>14</sup> This tool outlines approaches to assess potential bias and applicability within four different domains: (i) appropriate study design and study participant selection, (ii) quality and application of the diagnostic test in question (the index test), (iii) appropriate definition and use of a 'reference standard' test to define 'true' positives, and (iv) appropriate and consistent application of study methods to all (flow and timing).

A summary of all the key criteria to be considered in each domain, when determining quality of data and diagnostic test ability, is given in [Table 2](#), which includes but is not limited to the QUADAS-2 criteria. Further detail if needed should be sort from the original QUADAS-2 publication.<sup>14</sup> [Supplementary data online, Appendix](#) contains example applications of these criteria to assess the LOE for diagnostic tests.

## Prediction models: guidance on assessment of adequate predictive ability

Three key criteria should be considered regarding quantitative assessment of the performance of prediction model: (i) association of predictors with the clinical outcome of interest; (ii) adequate discrimination; and (iii) adequate calibration.

## Predictors importantly associated with the clinical outcome of interest

All included predictors should be individually associated with the outcome being predicted [also see discussion of artificial intelligence (AI)-derived models below]. However, the strength and significance of the associations is context specific and should not be used in isolation to determine the predictive ability of a model. If important non-linear

associations (i.e. stronger associations at certain levels of the predictor) and effect modifications (e.g. different associations for men and women) are relevant, these should be included.

## Adequate discrimination

Discrimination is the ability of the model to distinguish between individuals who do or do not have the disease outcome of interest. In the context of prognostic prediction models, discrimination can include assessment of the ability of the model to correctly predict the time order of disease outcomes (i.e. to put an individual who has the outcome first at higher risk than an individual who has the outcome later). Discrimination should be assessed in the context of both (i) internal validation in the data used to derive the model and (ii) external validation in an independent study.<sup>15</sup> Allowance for optimism in discrimination should be made in the context of internal validation (see section 'Optimism correction for discrimination or calibration'). Demonstration of adequate discrimination using a C-statistic is appropriate (taking into account the time to event component and censoring for survival data with Harrell's C-index or a time-dependent C-statistic). Sensitivity and specificity can be reported if there is a predefined clinically relevant time frame and established risk threshold for risk estimates.<sup>16,17</sup> Values of C-statistics that constitute adequate predictive ability are difficult to define as these will vary with clinical context, the consequences of the risk estimate, the disease severity, and potential for harm with inaccurate identification of high- or low-risk individuals. Also of note, differences in C-statistics between models can be statistically significant without necessarily being clinically significant (see section 'Assessing evidence for comparison of alternative tests or models, or measurement of additional risk predictors'). C-statistic values will also vary with the characteristics of the data sample used in the way that relative risk estimates do not. For example, among study samples that are more homogeneous than those expected in clinical practice, discrimination may be lower than might be expected in the target population. Also the severity of the observed disease outcome (e.g. mortality vs. hospitalization) and timeframe of follow-up may impact on observed discrimination (the latter, particularly if risk predictors have time-varying effects).<sup>18</sup> Testing in a sample representative of the target population is therefore ideal.<sup>17</sup> A C-statistic value with confidence limits overlapping of 0.5 implies that decisions based on risk estimation may be no better than chance alone. Values above 0.7 are typically considered to constitute good evidence for discriminative ability; values between 0.5 and 0.7 must be interpreted in the context of clinical scenario and representativeness of the sample.

## Adequate calibration

Calibration refers to the agreement between the proportion of predicted and observed outcomes. As with discrimination, calibration should be checked both in the model derivation study (internal validation) and in an independent external validation data source.<sup>15,17</sup> The primary aim of internal validation of calibration is to check for model mis-specification (e.g. insufficient allowance for non-linear predictor effects or change of effects with age or gender). Allowance for optimism in calibration should be made in the context of internal validation (see section 'Optimism correction for discrimination or calibration'). External validation of calibration should check for transferability of the model to a clinically representative population. External validation checks of calibration may result in re-calibrating the model to be more appropriate for the target population, in which case this could be considered evidence of good calibration for the recalibrated version.

Calibration should be shown graphically, ideally as a smoothed version over the possible range of risk predictions, and with confidence intervals to indicate precision. This should ideally be accompanied by simple metrics to indicate absolute degree of calibration, e.g. the calibration slope, the ratio of predicted to observed risks, or a summary of absolute differences. A single statistical test (*P*-value) quantifying evidence for concordance between observed and predicted risk is less helpful since this summarizes statistical significance and not magnitude of miscalibration, nor which risk groups have less accurate predictions (low vs. high risk). Any miscalibration identified should be considered in terms of clinical relevance and potential impact on diagnostic or treatment decisions. For example, if miscalibration is only observed in those at the extreme upper end of the risk scale, beyond any clinically relevant threshold, this may not impact on the conclusion or resulting action. Likewise, many datasets will not represent the entire contemporary target population if they represent past periods of time or selected subsets of the population (e.g. due to eligibility criteria, or healthy volunteer bias) and calibration should be interpreted with this in mind.

## Optimism correction for discrimination or calibration

Prediction models will often perform better in the data source used for their derivation, a phenomenon known as optimism,<sup>16</sup> commonly due to overfitting (overly tailoring the model to the specific-observed dataset). Therefore, in the context of internal validation optimism correction should be applied as appropriate to study size and design; bootstrapping or resampling approaches to internal validation are preferred over crude data splitting which can provide inefficient use of data and induce model instability.<sup>19</sup> Very large studies may not need cross validation if evidence of minimal optimism is provided (e.g. a shrinkage factor). Likewise, if, during model derivation, predictive ability is also tested in an external validation dataset (independent data source), this can provide sufficient evidence that the model performance is not optimistic. The section on sample size provides further information on determination of sufficient data size to avoid overfitting.

## Additional desirable assessments of predictive ability

The following additional metrics may reinforce the finding of adequate predictive ability.

### Measure of overall fit

A measure of  $R^2$  indicates the amount of variation in the outcome that can be explained by the predictors in the model. Various versions have been proposed for binary or survival data; the Cox–Snell version may have advantages since it can be used for sample size calculations.<sup>17</sup> A measure of the mean squared error in prediction may also be useful, e.g. the Brier score, which summarizes the absolute difference between estimated risks/event probabilities and real observed status (0 or 1) across all individuals.<sup>17</sup>

### Measures of clinical utility

Net benefit curves (sometimes known as decision curves) can be helpful to assess benefit of the model for clinical decision-making over a range of thresholds.<sup>17</sup> These provide the balance of true positives, minus the false positives across a range of risk thresholds, with threshold-dependent weighting applied. The inherent assumption with the weighting is that the choice for threshold indicates how much more important it is to identify a case than to incorrectly identify a control.

## Prediction models: guidance on determining data quality for assessment of prediction models

Various aspects should be considered when determining the data quality of studies including: (i) the study design; (ii) how representative the study participants are of the target population; and (iii) the sample size.

### Study design: diagnostic prediction models

Studies for diagnostic prediction models can be cross-sectional or prospective (where follow-up may be necessary to determine whether a condition was present at the time of the test) and should allow accurate estimation of the outcome prevalence. Case-control studies are generally unsuitable since prevalence and incidence are impacted by design and biases may be introduced by measurement of predictors after case ascertainment. Nested case-cohorts or nested case-control studies can be used if the sampling fraction is known and accounted for in the analysis.

### Study design: prognostic prediction models

Studies for prognostic models should be designed such that incidence of the relevant outcome can be accurately observed: prospective/longitudinal cohorts, or nested case-cohorts where the sampling fraction is known and incorporated appropriately into derivation and validation analyses. Clinical trial data can also be used provided that any intervention as part of the trial either does not impact on the outcome or is appropriately accounted for (either by exclusion of active trial arms or appropriate allowance in analysis). However, clinical trial data might not be representative of the population to which the prediction model is applied, particularly due to inclusion/exclusion criteria.

## Participants representative of the target population (both diagnostic and prognostic prediction models)

Aside from having appropriate study design, high-quality studies are those which are:

- (1) Representative of the target population in terms of the incidence of the disease outcome which the model aims to predict;
- (2) Representative of the target population in terms of other characteristics not included in the model (e.g. geographic location, comorbidity, socio-demographic characteristics, recent calendar period of data collection);
- (3) Representative of the target population in terms of the absolute values and variance of the input variables used in the model (particularly important for validation since measures of predictive ability are very sensitive to the distribution of input variables).

How representative studies are should be considered in the context of what is known about the target population, which is likely to include multiple countries. Where possible, reported statistics on disease prevalence and test values should be sought in order to judge whether study data are representative, although this may not always be available and clinical judgement may be needed. For model derivation, criteria 1), 2), and 3) can be relaxed if appropriate steps are taken to ensure translation/recalibration of the model to the target population, e.g. a model derived in data from one geographical area can be recalibrated to ensure relevance to the target geographical area(s) using summary statistics from the target population. This does not replace the need for external validation in a relevant population.

## Sample size considerations for prediction model derivation (both diagnostic and prognostic models)

Defining a sufficiently large sample size required to derive a risk prediction model is context specific and depends on a combination of factors: (i) the number of outcomes relative to the number of parameters included in the prediction model, (ii) the total number of study participants included in the model, (iii) the outcome proportion (incidence), and (iv) the expected predictive ability of the model. To determine whether a study sample size was sufficient for model derivation, calculations can be made using published tools with accompanying software to guide this judgement.<sup>12,13,20</sup> Research has shown that the crude rule of thumb that 10 events per predictor are needed in the dataset is often an underestimation of event numbers required to avoid overfitting the model and optimism in apparent performance.<sup>12,13,21</sup> Very large sample sizes will therefore be required when the outcome prevalence or incidence is low; this is further exacerbated if included risk predictors have low predictive ability. For machine learning approaches to model derivation, over 200 events per predictor may be required.

## Sample size considerations for model validation (both diagnostic and prognostic models)

Previous rules of thumb that >100 events and >100 non-events are needed for assessment of discrimination and >200 events and >200 non-events are needed for calibration assessment are a helpful starting point but may be simplistic, and published tools to estimate the necessary sample size for a given clinical circumstance are available and ideally preferred.<sup>12</sup> To assess calibration of a prognostic model in particular, additional consideration of the number of people followed up for the time frame of the prediction estimate is important. Observed risk can vary widely if few individuals remain in the study at the time point of interest. Examination of confidence intervals for observed risks in calibration assessments across the range of risk predictions can also reveal precision in estimation and whether sufficient data have been included.

A more detailed assessment of potential bias and applicability/generalizability of the studies for prediction model assessment can be made using signalling questions from the PROBAST tool.<sup>16</sup> This tool can be applied after careful definition of the intended use of the prediction test, the target population, the predictors that can be measured in the target population, and the disease/condition outcome that is to be diagnosed or predicted and the reference standard approach to outcome determination. A summary of criteria that are helpful to assess when determining the data quality and predictive ability of a risk model, including the PROBAST criteria and additional points from other relevant published guidance,<sup>12,15–17</sup> is given in [Table 3](#). Further detail if needed should be sought from the original PROBAST publication.<sup>16</sup> Supplementary data online, [Appendix](#) contains example applications of these criteria to assess the LOE for risk prediction models.

## Additional considerations

### Artificial intelligence methods for diagnostic tests and risk prediction

Recent years have seen substantial increases in the number of AI-based methods proposed for disease diagnosis and prediction of

medical outcomes (e.g. AI assessment of medical images for disease diagnosis). Such approaches are based on computerized learning of trends in the data that are related to the outcome of interest and iterative minimization of the differences in predicted and observed outcomes.

Methodology for AI-based diagnostic testing and prediction model development and validation is a developing area of research with wide-reaching areas of application. Formal guidance for the use and reporting of AI approaches for medical diagnoses and risk predictions are under development (STARD-AI, TRIPOD-AI, QUADAS-AI, and PROBAST-AI) and should be consulted when available.<sup>21–23</sup> The current report is primarily focused on guidance relevant to non-AI methods for diagnostic tests and prediction modelling. Much of the guidance is transferable to AI-based approaches, although some context-specific adaptation and additional information may be needed. In particular, the sample size needed for model derivation with AI approaches may be substantially larger than that needed for traditional statistical models to avoid optimism in apparent performance and ensure generalizability to target clinical populations.<sup>19</sup> Likewise, internal validation approaches utilized in the model derivation stage may be different for AI prediction schemes. Transparent reporting and justification of the chosen approaches and sample sizes are important considerations when assessing the quality of published evidence. Transparent reporting is also important to facilitate the practical implementation of AI-based models and improve trust and uptake of these somewhat ‘black box’ style calculations. The initial selection of potential risk predictors may be determined in a more agnostic way for AI models, and prior knowledge of association with the outcome and clinical rationale for inclusion may be less relevant. This is particularly true for large panels of predictors (e.g. metabolite assays) or imaging features. In such circumstances, demonstration of model stability (consistency in the derived model over multiple samples) is particularly pertinent and can be shown using bootstrapping or resampling techniques.<sup>19</sup>

## Missing data

Incomplete data concerning risk predictors and diagnostic features within datasets used to either derive or evaluate risk prediction models should be appropriately addressed, as relevant to the particular context. Missing data in a dataset used to derive a prediction model can lead to bias in estimated model parameters and inaccurate prediction model derivation. Similarly, missing data in datasets used to validate a model can yield biased estimates of model performance. Therefore, strategies for managing missing data are imperative. Various methods to impute data are possible, with multiple imputation using chained equations (MICE) known to be a robust approach when relevant assumptions about the ‘distribution of missingness’ are met.<sup>24</sup> This assessment is context specific. Other imputation approaches can be preferred given the specific context, source of data, and characteristics of the missing predictors. For all approaches, reporting should be transparent with appropriate justification and relevant testing/exploration of assumptions whenever possible, particularly when the amount of missing data is substantial. There may be situations where data are likely to be missing in patients from the target population when the model is applied in clinical practice. It may therefore be relevant for assessments of predictive ability to apply the same approach to dealing with missing data as will be used in clinical practice. For example, if a risk predictor will be set to a standard value when missing in clinical practice, a similar approach applied when validating the model is reasonable, assuming the degree and cause of missingness in the validation dataset is similar to that expected in practice.

## Assessing evidence for comparison of alternative tests or models, or measurement of additional risk predictors

There may be circumstances where a diagnostic test or prediction model should be assessed against alternative tests/models, e.g. adding a risk predictor to an existing prediction model. This may be relevant for only a subset of the population (e.g. those that are assessed to be at intermediate risk using the initial risk assessment model). In such instances, the evidence should be evaluated using the same process and criteria outlined for diagnostic tests and prediction models, with emphasis on the evaluation data being representative of the specific target population. To quantify improvement in predictive ability, clinically as well as statistically significant improvement in measures of discrimination (e.g. C-statistics, sensitivity, specificity) should be demonstrated. These metrics can potentially be accompanied by improvements in net benefit (e.g. comparison of net benefit curves),<sup>19</sup> in absolute risk prediction (e.g. using the integrated discrimination improvement), and in risk classification [e.g. using the net reclassification index (NRI)].<sup>25</sup> Such additional assessments should not be used in isolation and are relevant only where the dataset is truly representative of the target population. Use of the NRI should be limited to situations where there are existing clinical risk thresholds already proposed or in use.<sup>26</sup> What constitutes a clinically relevant improvement in discrimination is often difficult to define, and modelling to translate potential statistically significant improvements into tangible benefits to case identification may be necessary. Benefits to clinical performance should ideally be weighed against feasibility, acceptability, infrastructure needed for, and cost of risk assessment.

## Consideration of competing risks

In the context of prognostic models, the presence of competing events in the derivation dataset may influence the risk estimates provided by the fitted model.<sup>27</sup> A competing event is that which precludes the occurrence of the event of interest.<sup>28</sup> For example, a non-cardiovascular death would be considered a competing event if it occurs before, and therefore precludes the observation of a cardiovascular disease event. There may be circumstances where the incidence of competing events is substantial enough to importantly inflate the risk estimates provided by the model, and correction for this in the statistical modelling may be appropriate. Such a correction allows risk estimates to be interpreted as ‘the probability of the relevant outcome occurring in a certain period of time given that the competing event may happen first’. Without a competing risk adjustment, estimates from the model can be interpreted as ‘the probability of the relevant outcome occurring in the theoretical situation in which experiencing the competing event is impossible’. The choice of model may be context specific. If a true event probability is desired and competing event occurrence is substantial, then competing risk adjustment should be applied. Conversely, a more accurate picture of an individual’s underlying disease-specific health status may be captured by a model that is not adjusted for competing risks. Likewise, if the incidence of competing risks is small, then models without competing risk adjustment are sufficient. If a prediction model has been derived using approaches that adjust for competing risks, then any accompanying validation approaches (i.e. discrimination and calibration assessment) should account for competing risks in a similar manner.

## Discussion

In this paper, we present a new system for assessing and grading evidence relating to diagnostic tests and prediction models to be applied

in future ESC guidelines. The task force acknowledged and valued the widespread acceptance of ESC three-tiered system, which has been used in ESC guidelines for over 20 years, and the proposed system therefore retains the hierarchical three LOEs ranging from A to C. The new system provides specific criteria for each LOE, considering factors such as adequate study design, methodology, quality, and the absence of major bias (*Graphical Abstract*). The proposed system aims to improve the reliability and accuracy of guideline recommendations in cardiovascular medicine. 'Explanatory notes' provide additional information to enable guideline task force members to understand the rationale of the grading system and use it appropriately.

The new LOE grading system for diagnostic tests and prognostic models could lead to upgrading of the evidence for some recommendations from LOE B to A, as well as to downgrading of the evidence for some recommendations from LOE B to C. These changes in LOE could potentially reflect the uncertainty derived from applying a grading system that was originally designed for evidence related to interventions to evidence related to diagnosis and prediction.

While the upgrading of the evidence for diagnosis and prediction from LOE B to A reflects the quality of the underlying evidence and the recognition of the specific methods and study design necessary to establish evidence for diagnosis or prediction, the downgrading for some recommendations from LOE B to C could be perceived as a weakness. However, it is essential to understand that this is not a shortcoming of the current guidelines themselves but rather a reflection of the limitations of the underlying evidence and tools used for evaluation of evidence that become more evident when applying the revised LOE grading system. Therefore, no change to previously published guidelines is needed; the upgraded LOE will be implemented in future guidelines only, starting with those published in 2026.

## Limitations of the revised level of evidence grading system

The focus of the revised LOE does not address in detail new analytical methods for diagnosis and prediction, such as those based on AI. As more experience is gained on this type of methodology, the grading system and the accompanying guidance might need to be adapted. Furthermore, the task force emphasized stringent criteria, recognizing that this approach might sometimes result in assigning a lower LOE than that which could be inferred from the overall body of evidence. The task force preferred to err on the side of caution rather than assigning an inappropriately high LOE to a recommendation. These priorities were closely linked to the need for a grading system that can be used relatively easily by clinical experts who contribute to the development of ESC CPG.

It is important to note that the LOE provided here may not offer a clear classification in every instance. There will be areas open to interpretation, which is an inherent challenge. While the class of recommendation should generally align with the LOE, other considerations can and should be factored in. Each guideline's task force will need to offer recommendations and guidance based on their best assessment of the evidence and their understanding of clinical realities. The ESC guideline development process, which involves the careful selection of experts for each task force and a rigorous review process, ensures that the final recommendations and corresponding LOE are robust and well-informed.

## Conclusions

The ESC has recognized the limitations of the current LOE grading system and developed a revised, standardized approach to evaluate the

strength of evidence in their guidelines. The proposed system for the first time provides specific criteria for assessing evidence of diagnostic tests and prognostic models. The proposed system considers quality of studies, external validation, and presence of bias and is accompanied by guidance on how to apply these criteria for future ESC guidelines task forces.

## Supplementary data

Supplementary data are available at *European Heart Journal* online.

## Declarations

### Disclosure of Interest

All authors declare no disclosure of interest for this contribution.

## Data Availability

No data were generated or analysed for this manuscript.

## Funding

E.D.A. and L.P. were supported by funding from the British Heart Foundation (RG/18/13/33946: RG/F/23/110103), NIHR Cambridge Biomedical Research Centre (NIHR203312) [\*], BHF Chair Award (CH/12/2/29428), Cambridge BHF Centre of Research Excellence (RE/24/130011, RE/18/1/34212) and by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and the Wellcome Trust. \*The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

## References

1. Klein WW. Current and future relevance of guidelines. *Heart* 2002;**87**:497–500. <https://doi.org/10.1136/heart.87.6.497>
2. Bertrand ME, Simoons ML, Fox KA et al. Management of acute coronary syndromes: acute coronary syndromes without persistent ST segment elevation; recommendations of the Task Force of the European Society of Cardiology. *Eur Heart J* 2000;**21**:1406–32. <https://doi.org/10.1053/euhj.2000.2301>
3. Tantawy M, Marwan M, Hussien S, Tamara A, Mosaad S. The scale of scientific evidence behind the current ESC clinical guidelines. *Int J Cardiol Heart Vasc* 2023;**45**:101175. <https://doi.org/10.1016/j.ijcha.2023.101175>
4. van Dijk VWB, Grobbee DE, de Vries MC, Groenwold RHH, van der Graaf R, Schuit E. A systematic breakdown of the levels of evidence supporting the European Society of Cardiology guidelines. *Eur J Prev Cardiol* 2019;**26**:1944–52. <https://doi.org/10.1177/2047487319868540>
5. Wells PS, Anderson DR, Rodger M et al. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. *Thromb Haemost* 2000;**83**:416–20. <https://doi.org/10.1055/s-0037-1613830>
6. SCORE2 working group and ESC Cardiovascular Risk Collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J* 2021;**42**:2439–54. <https://doi.org/10.1093/eurheartj/ehab309>
7. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest* 2010;**137**:263–72. <https://doi.org/10.1378/chest.09-1584>
8. Moons KG, Altman DG, Reitsma JB et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;**162**:W1–73. <https://doi.org/10.7326/M14-0698>

9. 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–W33. <https://doi.org/10.7326/M18-1377>
10. Gale CP, Stocken DD, Aktaa S et al. Effectiveness of GRACE risk score in patients admitted to hospital with non-ST elevation acute coronary syndrome (UKGRIS): parallel group cluster randomised controlled trial. *BMJ* 2023;**381**:e073843. <https://doi.org/10.1136/bmj-2022-073843>
11. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol* 2005;**58**:859–62. <https://doi.org/10.1016/j.jclinepi.2004.12.009>
12. Riley RD, Snell KIE, Archer L et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *BMJ* 2024;**384**:e074821. <https://doi.org/10.1136/bmj-2023-074821>
13. Riley RD, Ensor J, Snell KIE et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;**368**:m441. <https://doi.org/10.1136/bmj.m441>
14. Whiting PF, Rutjes AW, Westwood ME et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;**155**:529–36. <https://doi.org/10.7326/0003-4819-155-8-201110180-00009>
15. Collins GS, Dhiman P, Ma J et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 2024;**384**:e074819. <https://doi.org/10.1136/bmj-2023-074819>
16. Wolff RF, Moons KGM, Riley RD et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;**170**:51–8. <https://doi.org/10.7326/M18-1376>
17. 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. <https://doi.org/10.1136/bmj-2023-074820>
18. Rossello X, Bueno H, Gil V et al. MEESSI-AHF risk score performance to predict multiple post-index event and post-discharge short-term outcomes. *Eur Heart J Acute Cardiovasc Care* 2021;**10**:142–52. <https://doi.org/10.1177/2048872620934318>
19. Riley RD, Pate A, Dhiman P, Archer L, Martin GP, Collins GS. Clinical prediction models and the multiverse of madness. *BMC Med* 2023;**21**:502. <https://doi.org/10.1186/s12916-023-03212-y>
20. Riley RD, Collins GS, Ensor J et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Stat Med* 2022;**41**:1280–95. <https://doi.org/10.1002/sim.9275>
21. 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. <https://doi.org/10.1136/bmjopen-2020-048008>
22. Sounderajah V, Ashrafian H, Golub RM et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;**11**:e047709. <https://doi.org/10.1136/bmjopen-2020-047709>
23. Sounderajah V, Ashrafian H, Rose S et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med* 2021;**27**:1663–5. <https://doi.org/10.1038/s41591-021-01517-0>
24. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011;**30**:377–99. <https://doi.org/10.1002/sim.4067>
25. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2007;**27**:157–72. <https://doi.org/10.1002/sim.2929>
26. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarker. *Stat Med* 2011;**30**:11–21. <https://doi.org/10.1002/sim.4085>
27. Hageman SHJ, Dorresteijn JAN, Pennells L et al. The relevance of competing risk adjustment in cardiovascular risk prediction models for clinical practice. *Eur J Prev Cardiol* 2023;**30**:1741–7. <https://doi.org/10.1093/eurjpc/zwad202>
28. Rossello X, Gonzalez-Del-Hoyo M. Survival analyses in cardiovascular research, part II: statistical methods in challenging situations. *Rev Esp Cardiol (Engl Ed)* 2022;**75**:77–85. <https://doi.org/10.1016/j.recsep.2021.06.027>