Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance

Ben Van Calster, Gary S Collins, Andrew J Vickers, Laure Wynants, Kathleen F Kerr, Lasai Barreñada, Gael Varoquaux, Karandeep Singh, Karel GM Moons, Tina Hernandez-Boussard, Dirk Timmerman, David J McLernon, Maarten van Smeden, Ewout W Steyerberg, on behalf of Topic Group 6 of the STRATOS initiative

Numerous measures have been proposed to illustrate the performance of predictive artificial intelligence (AI) models. Selecting appropriate performance measures is essential for predictive AI models intended for use in medical practice. Poorly performing models are misleading and may lead to wrong clinical decisions that can be detrimental to patients and increase financial costs. In this Viewpoint, we assess the merits of classic and contemporary performance measures when validating predictive AI models for medical practice, focusing on models that estimate probabilities for a binary outcome. We discuss 32 performance measures covering five performance domains (discrimination, calibration, overall performance, classification, and clinical utility) along with corresponding graphical assessments. The first four domains address statistical performance, whereas the fifth domain covers decision–analytical performance. We discuss two key characteristics when selecting a performance measure and explain why these characteristics are important: (1) whether the measure's expected value is optimised when calculated using the correct probabilities (ie, whether it is a proper measure) and (2) whether the measure solely reflects statistical performance or decision–analytical performance by properly accounting for misclassification costs. 17 measures showed both characteristics, 14 showed one, and one (F1 score) showed neither. All classification measures were improper for clinically relevant decision thresholds other than when the threshold was 0·5 or equal to the true prevalence. We illustrate these measures and characteristics using the ADNEX model which predicts the probability of malignancy in women with an ovarian tumour. We recommend the following measures and plots as essential to report: area under the receiver operating characteristic curve, calibration plot, a clinical utility measure such as net benefit with decision curve analysis, and a plot showing probability distributions by outcome category.

Lancet Digit Health 2025

Published **Online**
https://doi.org/10.1016/j.landig.2025.100916

Department of Development and Regeneration (Prof B Van Calster PhD, L Wynants PhD, L Barreñada MSc, Prof D Timmerman PhD) **and Leuven Unit for Health Technology Assessment Research (LUHTAR)** (Prof B Van Calster, L Wynants, L Barreñada), **KU Leuven, Leuven, Belgium; Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands** (Prof B Van Calster, Prof E W Steyerberg PhD); **Department of Applied Health Sciences, School of Health Sciences, College and Medicine and Health, University of Birmingham, Birmingham, UK** (Prof G S Collins PhD); **Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA** (Prof A J Vickers PhD); **Department of Epidemiology, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, Netherlands** (L Wynants); **Department of Biostatistics, University of Washington School of Public Health, Seattle, WA, USA** (Prof K F Kerr PhD); **Parietal project team, INRIA Saclay-Île de France, Palaiseau, France** (Prof G Varoquaux PhD); **Division of Biomedical Informatics, Department of Medicine, University of California, San Diego, CA, USA** (K Singh PhD); **Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, Utrecht, Netherlands** (Prof B Van Calster, Prof K G M Moons PhD, M van Smeden PhD,

Introduction

The medical literature abounds with predictive artificial intelligence (AI) models that estimate the probability of individuals having (diagnostic) or developing (prognostic) a disease or health state of interest (the event), also known as clinical prediction models.[1,2] Although these models were traditionally developed using statistical methods such as regression analysis, the use of machine learning algorithms with improved flexibility is increasing. For instance, a traditional logistic regression model might aim to predict the risk of permanent stoma in patients undergoing resection of left-sided obstructive colon cancer using demographic, clinical, and laboratory measurements.[3] A more contemporary model built using deep learning might aim to predict the presence of atrial fibrillation based on sinus rhythm electrocardiograms.[4]

Regardless of the modelling approach, the performance of predictive AI models intended for medical practice should be properly evaluated. Consequently, selecting appropriate performance measures for predictive AI in health care is essential, since poorly performing models might lead to wrong clinical decisions that can be detrimental to patients and increase financial costs.[5] Although numerous such measures have been suggested, clarity is needed. There is occasional conflict regarding the measures that are recommended in the medical, statistical, and machine learning literature.[6–11]

Performance assessment is especially relevant for both external and internal validation studies. An external validation study evaluates model performance using a dataset that includes individual participant data from a target population in which the model might be used.[12–15] Unlike the training dataset, the external validation dataset includes data from individuals from different locations, time periods, or settings. In contrast, internal validation evaluates model performance using new individuals from the same population as the training dataset using methods such as cross-validation, bootstrapping, or (repeated) train–test splitting.[12,13] Thus, internal validation refers not to model selection but to an independent evaluation of the selected model.

In this Viewpoint, we assess classic and contemporary performance measures for model evaluation from the statistical and machine learning literature and provide recommendations for researchers, end users (ie, health care staff), and other stakeholders such as policy makers. We present a taxonomy of performance domains, describe key characteristics for performance measures, discuss these measures in combination with an illustrative case study, and formulate recommendations.

A taxonomy of five performance domains

We classify performance measures into five domains: discrimination, calibration, overall performance,

Prof E W Steyerberg); **Department of Medicine (Biomedical Informatics)** (Prof T Hernandez-Boussard PhD) **and Department of Biomedical Data Science** (Prof T Hernandez-Boussard), **Stanford University, Stanford, CA, USA; Department of Obstetrics and Gynecology, University Hospitals Leuven, Leuven, Belgium** (Prof D Timmerman)**; Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK** (D J McLernon PhD)

Correspondence to:
Dr Ewout W Steyerberg, Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, Utrecht 3508 GA, Netherlands
e.w.steyerberg@umcutrecht.nl

See **Online** for appendix

classification, and clinical utility. Among these domains, the first three evaluate performance based on probability estimates (Appendix p 2).

Discrimination focuses on the extent to which the model assigns higher probabilities of the event for individuals with the event than for those without. Discrimination reflects relative performance; ie, it does not matter how high or low the estimated probabilities are, only whether they allow to discriminate between individuals with versus without the event.

Calibration focuses on the extent to which the probability estimates correspond to observed event proportions. Calibration reflects absolute performance by evaluating whether estimates are too high or too low. Models can therefore have good discrimination but poor calibration, and vice versa.

Overall performance of a model combines discrimination and calibration by quantifying how closely the probability estimates approach the actual outcomes of 0 (no event) or 1 (event).[7,16,17]

The fourth and fifth performance domains require a threshold on the estimated risk of the event to classify individuals into two mutually exclusive groups: low-risk (estimated risk below the threshold) and high-risk (estimated risk equal to or above the threshold) groups. These groups are linked to a decision about an intervention (eg, surgery), which would be suggested for individuals at high risk but not for those at low risk. The threshold can therefore be referred to as the decision threshold. Although multiple decision thresholds can be used to separate individuals into three or more groups, we focus on the common single-threshold case.

The fourth performance domain, classification, focuses on the extent to which individuals are correctly classified as high or low risk. This domain is based on the contingency table or confusion matrix, a cross-tabulation of classifications (low *vs* high risk) and outcomes (event *vs* no event). Classification performance is perfect when all individuals with an event have a probability above the decision threshold and all individuals without an event have a probability below the threshold. Classification performance is influenced by discrimination and calibration performance.

The fifth domain, clinical utility, goes one step further by explicitly incorporating misclassification costs when evaluating classifications of individuals into low-risk and high-risk groups. Misclassification costs is an established term that refers broadly to the harms of misclassification of any kind, where misclassifications refer to false positives and false negatives.[18,19] In biomedical applications, the consequences of a false negative (for instance, not referring a woman with ovarian malignancy for advanced surgery) are almost always different from the consequences of a false positive (referring a woman with a benign tumour for advanced surgery). Clinical utility evaluates the quality of decisions based on the decision threshold, and whether using a model leads to better decisions than not using it or than a competing model. The decision threshold should therefore be clinically relevant and linked with misclassification costs (panel). Due to its focus on the quality of decisions, clinical utility is the most important performance domain.

We discuss 32 (three discrimination, six calibration, nine overall, 11 classification, and three clinical utility) performance measures (table 1), along with corresponding visual assessments.

Key characteristics of an informative performance measure

We define two key characteristics that a performance measure should meet: (1) the measure should be proper, and (2) it should have a clear focus on solely reflecting statistical or decision–analytical value by properly considering differential misclassification costs. Measures that do not possess the first characteristic cannot be trusted, whereas measures not possessing the second are equivocal. A third desirable characteristic is intuitive interpretation.[24,25] We do not discuss this characteristic further, as interpretability is subjective and influenced by background knowledge and familiarity.

Properness

A performance measure is called proper if its expected value is optimal when using the correct model, which is the model that gives the correct probabilities based on the predictors or features in the model.[26–30] Here, expected value refers to the average value obtained after repeating the validation study multiple times. In any given dataset, particularly when sample size is low, the correct model can be outperformed by an incorrect model due to random variation. The importance of properness is that a proper measure cannot be fooled: in expectation, the correct model cannot be outperformed by an incorrect one. A measure is strictly proper when its expected value is optimal only for the correct model. When the expected value is optimal for the correct model and for some incorrect models, a measure is called semi-proper. When an incorrect model can have a better expected value than the correct model, the measure is termed improper and cannot be trusted. The properness status for the 32 measures is listed in table 1, and an illustration is provided in the Appendix (pp 3–6).

Clear focus on statistical or decision–analytical evaluation

There is a clear distinction between statistical and decision–analytical performance evaluation of predictive AI models for medical practice. The first four performance domains (appendix p 2) focus on different aspects of statistical performance, whereas the clinical utility domain focuses on decision–analytical performance. Statistical performance measures are essential for model evaluation but cannot be used to find out whether a model should be used in practice: it is not appropriate to cite, for example, good discrimination and calibration or conclude that a model can be used to aid decisions about ovarian surgery. If a performance measure aims to go beyond measuring statistical value, it should incorporate misclassification

Panel: Defining a decision threshold

The primary aim of most predictive AI models in medical practice is to support subsequent decision making. Probability estimates may guide health professionals and patients to improve health outcomes by avoiding a burdensome intervention with limited expected benefit for those at low risk and facilitating intervention selection for those at high risk. Consequently, the decision threshold should be defined on medical rather than statistical grounds.[20]

Often, however, a threshold is chosen by optimising a statistical measure such as the Youden index (sensitivity + specificity – 1). When maximising the Youden index, sensitivity and specificity are considered equally important; this is rarely the case in medicine. Using statistical arguments to set a decision threshold is inconsistent with decision theory and detached from practical use by clinicians.

Instead, once the decision that the model intends to support is clearly defined, the four possible consequences of using the model to support that decision should be considered:

- True positives (individuals with the event and classified as high risk)
- True negatives (individuals without the event and classified as low risk)
- False negatives (individuals with the event and classified as low risk)
- False positives (individuals without the event and classified as high risk)[18,21,22]

The weight of these consequences might vary by the nature and effects of the intervention, the health-care system, or by clinician and patient.

The case study in this Viewpoint was when patients required surgical removal of an ovarian mass. The assessment of different neoplasias in the adnexa (ADNEX) model was used to decide whether advanced or conservative surgery was needed. A decision threshold of 0·1 (10%) for the probability of malignancy is often recommended.[23] Suggesting advanced surgery to patients who actually have a 10% risk of malignancy based on the ADNEX predictors implies performing advanced surgery in ten individuals per true positive (ie, performing advanced surgery in a patient with a malignant tumour).[18] In other words, we accept up to nine false positives (ie, performing advanced surgery in up to nine patients with benign tumour) per true positive. Hence, using this threshold assumes that the medical benefit of advanced surgery on a malignant tumour is nine times greater than the harm of unnecessary advanced surgery in individuals with a benign tumour.[18] In the section on Clinical utility, we describe how measures for clinical utility incorporate misclassification costs.

costs in accordance with decision–analytical principles (panel). If misclassification costs influence a performance measure in an implicit or ad hoc way, the measure neither assesses statistical performance nor adequately evaluates the quality of decisions for clinical practice.

Case study: external validation of a diagnostic model for ovarian cancer

As a case study, we consider prediction of malignancy in women with an ovarian tumour. The ADNEX model, developed by the International Ovarian Tumor Analysis (IOTA) consortium (authors BVC and DT are part of this consortium), preoperatively estimates the probability of malignancy in women with an ovarian tumour who are scheduled for surgery.[31] The model can inform decisions regarding the type of surgery (advanced *vs* conservative) for patients examined at an oncology centre or regarding referral to an oncology centre for patients examined elsewhere.[32] ADNEX was developed on the data from 5909 individuals recruited between 1999 and 2012 in 24 secondary and tertiary care centres from 10 countries (Italy, Belgium, Sweden, Czech Republic, Poland, France, United Kingdom, China, Spain, and Canada). The TransIOTA study externally validated ADNEX for its ability to distinguish benign from malignant tumours, using data from 894 women recruited between 2015 and 2019 in one secondary and five tertiary care centres in four countries (Belgium, Italy, Czech Republic, and United Kingdom). There were 434 women with a malignant tumour (prevalence 49%).[32] The retrospective use of data from the IOTA consortium was approved by the Research Ethics Committee from the University Hospitals Leuven, IOTA's leading ethics committee (S64709).

For didactic purposes, we use this dataset to calculate all discussed performance measures with 95% confidence intervals (95% CI) and show all discussed visualisations. Confidence intervals were obtained using the percentile bootstrap method on 1000 bootstrap samples. We evaluated the performance of ADNEX as is and after updating using logistic recalibration (table 1). To update ADNEX, we fitted a logistic regression model of the outcome on the logit of the estimated event probability (the linear predictor).[2,33] This method is similar to Platt scaling, a well known method in machine learning to improve the calibration of predictions.[34,35] Logistic recalibration essentially applies a linear transformation to the linear predictor. This method is thus a rank-preserving method, and the ranking of patients based on the estimated probability of malignancy remains the same before and after updating.

All R and Python scripts, as well as the estimated risk of malignancy and outcomes for the 894 participants, are available in the GitHub repository.

For more on the **R and Python scripts**, see https://github.com/benvancalster/PerfMeasuresOverview

Performance measures

We discuss the selected measures briefly in this section. A detailed description of the measures, including formulas, are presented in the Appendix (pp 7–23).

	Characteristics		ADNEX results	
	Properness*	Focus†	Before recalibration	After recalibration
Discrimination				
AUROC, AUC, or C-statistic	+	+	0·911 (0·894 to 0·927)	0·911 (0·894 to 0·927)
AUPRC or AP	+	–	0·895 (0·862 to 0·921)	0·895 (0·862 to 0·921)
pAUROC (sensitivity ≥0·8)	+	–	0·141 (0·130 to 0·151)	0·141 (0·130 to 0·151)
Calibration				
O:E ratio	+	+	1·228 (1·171 to 1·288)	1·000 (0·955 to 1·046)
Calibration intercept	+	+	0·810 (0·619 to 1·006)	0·000 (–0·180 to 0·184)
Calibration slope	+	+	0·934 (0·833 to 1·051)	1·000 (0·892 to 1·126)
ECI	+	+	0·105 (0·063 to 0·160)	0·002 (0·001 to 0·017)
ICI	+	+	0·094 (0·074 to 0·118)	0·014 (0·009 to 0·038)
ECE	+	+	0·091 (0·072 to 0·117)	0·017 (0·019 to 0·050)
Overall performance				
Loglikelihood	++	+	–370 (–407 to –334)	–337 (–368 to –307)
Logloss or cross entropy	++	+	370 (334 to 407)	377 (307 to 368)
Brier score	++	+	0·133 (0·118 to 0·147)	0·118 (0·106 to 0·131)
Scaled Brier, Brier skill score, or IPA	++‡	+	0·469 (0·412 to 0·527)	0·526 (0·475 to 0·576)
McFadden R^2	++‡	+	0·403 (0·343 to 0·461)	0·456 (0·405 to 0·504)
Cox–Snell R^2	++‡	+	0·427 (0·379 to 0·471)	0·469 (0·429 to 0·502)
Nagelkerke R^2	++‡	+	0·570 (0·505 to 0·629)	0·625 (0·573 to 0·670)
Coefficient of discrimination or discrimination slope	–	+	0·509 (0·478 to 0·540)	0·525 (0·495 to 0·556)
Mean absolute prediction error	–	+	0·243 (0·226 to 0·260)	0·237 (0·222 to 0·252)
Classification: summary measures (using *t*=0·1)				
Classification accuracy at *t*	–§	+	0·794 (0·768 to 0·819)	0·691 (0·661 to 0·723)
Balanced accuracy at *t*	–¶	+	0·799 (0·776 to 0·822)	0·700 (0·677 to 0·724)
Youden index at *t*	–¶	+	0·597 (0·551 to 0·643)	0·399 (0·353 to 0·448)
Diagnostic odds ratio at *t*	–	+	37·400 (24·600 to 68·500)	43·300 (23·600 to 119·000)
Kappa at *t*	–	+	0·592 (0·544 to 0·639)	0·392 (0·346 to 0·442)
F1 at *t*	–¶	–	0·818 (0·792 to 0·843)	0·756 (0·727 to 0·782)
MCC at *t*	–	+	0·625 (0·581 to 0·667)	0·480 (0·438 to 0·522)
Classification: partial measures (using *t*=0·1)				
Sensitivity or recall	–	+	0·954 (0·934 to 0·974)	0·984 (0·972 to 0·993)
Specificity	–	+	0·643 (0·603 to 0·686)	0·415 (0·370 to 0·463)
Positive predictive value or precision	–	+	0·716 (0·679 to 0·753)	0·614 (0·577 to 0·650)
Negative predictive value	–	+	0·937 (0·911 to 0·964)	0·965 (0·938 to 0·986)
Clinical utility‖ (net benefit: *t*=0·1; expected cost: costs 9:1)				
Net benefit	+	+	0·443 (0·411 to 0·475)	0·444 (0·411 to 0·478)
Standardised net benefit	+	+	0·912 (0·892 to 0·932)	0·915 (0·900 to 0·930)
Expected cost	+	+	0·355 (0·274 to 0·376)**	0·355 (0·274 to 0·376)**

*Properness: ++, strictly proper; +, semi-proper; –, improper. †Focus: +, measure focuses either on purely statistical or decision–analytical evaluation by properly addressing misclassification costs; –, measure mixes statistical and decision–analytical performance evaluation, and therefore, lacks a clear focus. ‡These measures are asymptotically strictly proper. §Semi-proper when *t*=0·5, which is rarely the clinically relevant threshold. ¶Semi-proper when *t* equals the true prevalence, which is rarely the clinically relevant threshold. ||For clinical utility in particular, the use of confidence intervals and p values for measures of clinical utility contradicts the principles of decision analysis. **Expected cost was minimised at a decision threshold of 0·06 for the original model and 0·15 for the recalibrated model. AP=average precision. AUC or AUROC=area under the receiver operating characteristic curve. AUPRC=area under the precision–recall curve. ECE=expected calibration error. ECI=estimated calibration index. ICI=integrated calibration index. IPA=index of prediction accuracy. MCC=Matthew's correlation coefficient. O:E ratio=observed over expected ratio. pAUROC=partial AUROC.

***Table 1:* Summary of performance measures discussed, assessment of two key characteristics, and results of ADNEX model in the case study before and after recalibration**

Discrimination

The definition of discrimination implies that discrimination measures should rely only on the ranks of the estimated probabilities in the dataset.[36] The key measure is the concordance probability or C-statistic. For binary outcomes, the C-statistic is equal to area under the receiver operating characteristic curve (AUROC).[37,38] Several researchers have advised against using AUROC when prevalence is far from 0·5 (class imbalance). AUROC has been described as misleading or overoptimistic when the event is rare because it ignores the difficulty of obtaining both acceptable positive predictive value (PPV or precision) and sensitivity (or recall) or does not consider misclassification costs.[39–44] The precision–recall (PR) curve and the area underneath (AUPRC) are often recommended as alternatives to the ROC curve and AUROC.[41,45,46] Another measure suggested instead of

AUROC is the partial AUROC (pAUROC), which focuses on the part of the ROC curve where specificity or sensitivity reach a specific minimum tolerable level.[42,47] AUROC, AUPRC, and pAUROC are semi-proper, because these rank-based measures are invariant to monotonic transformations of probability estimates.[48] Dividing all ADNEX probabilities by 100 does not change the value of these measures.

There are no grounds to label AUROC as misleading or overoptimistic.[36,49] Discrimination measures are not meant to reflect differential misclassification costs, and class imbalance should not be conflated with misclassification costs or medical relevance. Unlike AUROC, AUPRC and pAUROC do not have a clear focus (second key characteristic). AUPRC and pAUROC mix statistical performance with aspects of clinical utility without following decision–analytical principles. For AUPRC, the PR curve does not directly consider true negatives. Although true negatives might be highly irrelevant for selected non-medical applications, such errors are generally important in medical applications. For pAUROC, a statement such as “we need at least 90% sensitivity because we want to find at least 90% of the cancer cases” might appear reasonable at face value. However, depending on specificity and prevalence, this could require different decision thresholds to classify individuals as high risk.[50] Consequently, there is no decision–analytical basis for this approach.

Nevertheless, although discriminatory ability is essential for predictive AI, AUROC alone cannot be used to identify a model as good or useful.[51] Visualisation through ROC or PR curves is acceptable, but in our experience, these plots do not provide useful information beyond that provided by summary measures (eg, AUROC) or relevant clinical utility measures (eg, net benefit).[38,49]

Figure 1 shows the ROC and PR curves and presents pAUROC for the case study for the (unsupported) argument that a sensitivity less than 0·8 is unacceptably low. The ADNEX model had an AUROC of 0·91 (95% CI 0·89–0·93) and an AUPRC of 0·89 (95% CI 0·86–0·91). Ignoring sensitivity values below 0·8, pAUROC was 0·14 (95% CI 0·13–0·15).

Calibration

Several approaches have been suggested in the statistical and machine learning literature to address calibration. These approaches can be classified into three increasingly stringent levels, labelled as mean, weak, and moderate calibration.[52] The first two levels are mostly known from the statistical literature. Research on the quantification of a fourth level, strong calibration, is ongoing. Strong calibration is discussed in the appendix (p 9).

Mean calibration (or calibration-in-the-large) evaluates whether the model’s average estimated probability equals the observed prevalence in the dataset. Two measures for calibration-in-the-large are the observed over expected (O:E) ratio and the calibration intercept. ADNEX had an O:E ratio of 1·23 (95% CI 1·17–1·29), indicating that 23% more

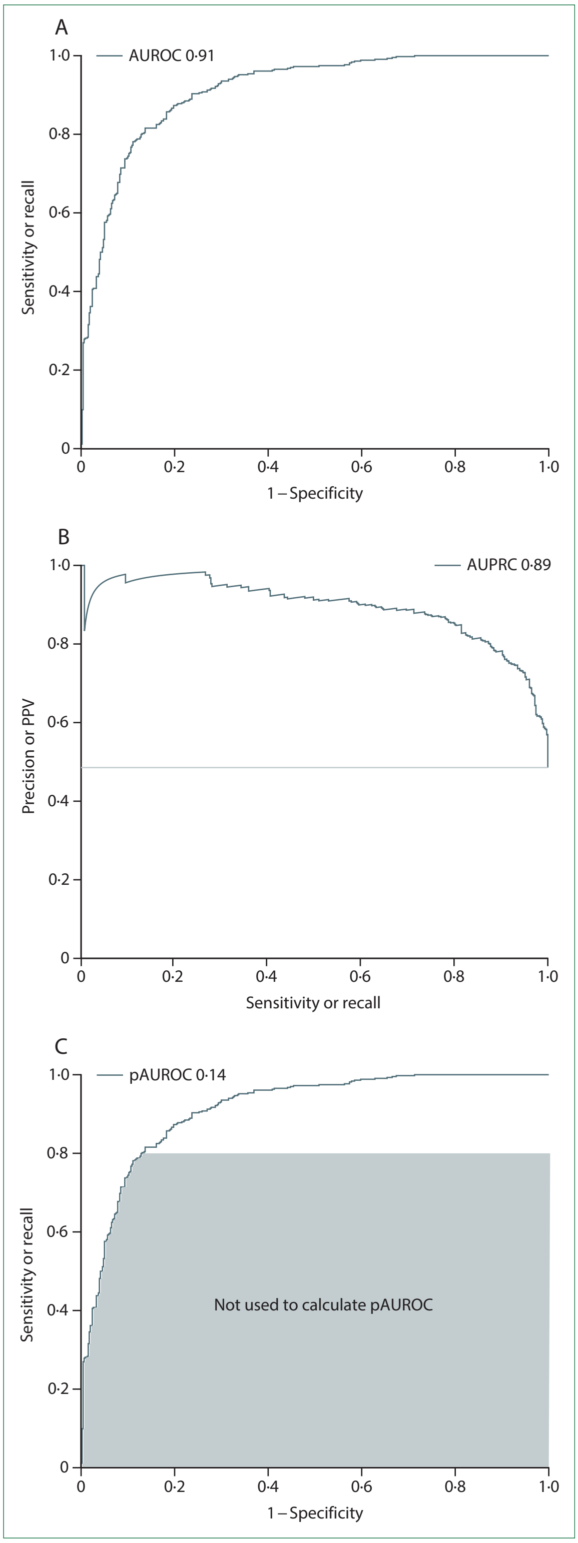

Figure 1: **ROC curve, PR curve, and pAUROC visualisation for the ADNEX model**
ROC (A) and PR curves (B) for ADNEX model. (C) Calculation of pAUROC from ROC curve; sensitivity less than 0·8 was considered to be unacceptably low. ROC=receiver operating characteristic. PR=precision–recall. AUROC=area under the ROC curve. pAUROC=partial AUROC. AUPRC=area under the PR curve. PPV=positive predictive value.

events were observed than expected based on the model (table 1). The calibration intercept of the ADNEX model was 0·81 (95% CI 0·62–1·01), suggesting underestimation of probabilities on average (intercept >0). The O:E ratio has a more intuitive interpretation than the calibration intercept.

A model has weak calibration if calibration-in-the-large is good and the estimated probabilities have, on average, neither too much nor too little spread, as quantified by the calibration slope.[33] Estimated probabilities with too much spread are on average too close to 0 and 1 (slope <1), whereas estimated probabilities with too little spread are on average too close to the prevalence (slope >1).[52,53] During internal validation, a calibration slope less than 1 can indicate potential overfitting.[54] In our case study, the ADNEX model had a calibration slope of 0·93 (95% CI 0·83–1·05), suggesting that the spread of the probabilities was adequate.

Moderate calibration means that, among people with an estimated probability of x, the observed proportion of the event also equals x. The most common way of assessing moderate calibration is by using a calibration plot, also referred to as a reliability diagram.[54–56] Calibration plots can be generated based on grouping individuals or smoothing.[56,57] Figure 2 presents the grouped (ten groups of equal size) and smoothed (using locally estimated scatterplot smoothing or loess) plots of the data used in the case study. The plots lie mostly above the diagonal, suggesting that the probabilities were underestimations across the whole range. A probable reason is that five of six participating centres in the validation study were tertiary care centres, resulting in a high prevalence of malignancy (49%). Grouped plots cannot be used to comprehensively address moderate calibration because individuals with very different estimated probabilities might still end up in the same group.

Several summary measures, such as the expected calibration error (ECE) for grouped plots and the estimated calibration index (ECI) and integrated calibration index (ICI) for smoothed plots, have been suggested for calibration plots.[58–60] Similar to statistical tests such as the Hosmer–Lemeshow test, the proposed summary measures cannot inform on the direction of miscalibration.[54,61] Further, ECE, ECI, and ICI depend on the grouping or smoothing methods used and have issues with statistical consistency.[62] Improved summary measures are being researched.[62] Therefore, calibration plots including confidence intervals are key tools for assessing calibration by visualising calibration performance conditional on estimated risk.

All discussed calibration measures are semi-proper,[52] with a focus on statistical performance (second key characteristic).

Overall performance

Basic measures of overall performance include likelihood-based measures, logloss (also known as cross-entropy or the negative loglikelihood) parameters, and Brier score.[63,64] Measures that express performance relative to a null model include scaled Brier (also known as Brier skill score or index of prediction accuracy) and R-squared measures of the proportion of explained variation, such as the McFadden, Cox–Snell, and Nagelkerke's R-squared.[65–69] Less common overall measures are the discrimination slope (also known as the coefficient of discrimination or the probabilistic AUROC) and the mean absolute prediction error.[16,17,70]

Loglikelihood, logloss, and Brier score are strictly proper, scaled Brier and R-squared measures are asymptotically strictly proper (ie, strictly proper when sample size is high, for instance, above 100), and discrimination slope and mean absolute prediction error are improper.[26,27,71] All discussed overall measures have a focus on statistical performance.

Overall performance measures used in our case study are presented in table 1. Plots for overall performance measures show the distribution of the estimated probabilities for events and non-events separately. Figure 3 shows violin plots for the ADNEX model, indicating that patients with benign tumour mostly had very low estimated probabilities of malignancy. Patients with malignancy mostly had moderate-to-high estimated probabilities, with less peaked distribution.

Classification measures

At the commonly recommended threshold of 10% in our case study,[23] ADNEX classified 578 patients as high risk, of whom 414 had a malignant tumour (true positive) and 164 had a benign tumour (false positive). The remaining 316 patients were classified as low risk, of whom 296 had a benign tumour (true negative) and 20 had a malignant tumour (false negative).

Classification measures are divided into summary measures and descriptive partial measures.[72] Common partial measures include sensitivity (or recall), specificity, PPV (or precision), and negative predictive value (NPV). Sensitivity and specificity assess classifications conditional on the observed outcome, which is unknown at prediction time. PPV and NPV are more clinically relevant, as they assess the outcome conditional on the risk classification.

As summary measures, we discuss classification accuracy, balanced accuracy, Youden index, kappa, diagnostic odds ratio, F1, and Matthew's correlation coefficient (MCC).[73–75] F1 and MCC were introduced to address the challenges caused by class imbalance, where classification accuracy can be inflated by classifying all patients as low risk when the event is rare.[76–80] F1 resembles AUPRC and shares several downsides: (1) F1 ignores true negatives, (2) F1 has no intuitive interpretation, and (3) the absolute value of F1 changes by simply switching the outcome labels (ie, when 1 becomes 0 and 0 becomes 1).[77,78] These issues hold for the more general F_{beta} class of measures, of which F1 is a special case.[78] Similar to F1, MCC has no intuitive interpretation.

At a given relevant decision threshold *t*, all classification measures are improper (Appendix pp 3–6). Some classification measures (balanced accuracy, Youden, and F1) are semi-proper when t=0·5 (classification accuracy) or when t equals the true prevalence; however, these values of t are rarely the most clinically relevant thresholds.[26] F1 is the only summary measure without a clear focus on statistical performance, as it conflates classification with clinical utility.

Plots relevant to classification performance include the ROC and PR curves, which show partial classification measures across all possible decision thresholds. A limitation of these plots is that the thresholds are not easily visible (figure 1).[38] An alternative plot is a classification plot, which has the probability threshold on the x-axis and one or more classification measures on the y-axis (appendix p 24).[38]

At a threshold of 10%, the ADNEX model showed a classification accuracy of 0·79 (95% CI 0·77–0·82), F1 score of 0·82 (0·79–0·84), and MCC of 0·63 (0·58–0·67) (table 1).

Clinical utility

In line with classic decision–analytical theory, clinical utility focuses on the quality of decisions based on model classifications that correspond to a clinically relevant threshold.[18,81] To assess utility, misclassification costs are explicitly considered. The most used measure for clinical utility in prediction studies in health care is net benefit.[21,82,83] The maximum value of net benefit equals the prevalence. Standardised net benefit equals net benefit divided by prevalence, and its maximum value is 1.[84] Net benefit sets the decision threshold based on the misclassification costs, following the classic connection between both.[18] Setting misclassification costs is not straightforward, and disagreements can occur about what the costs should be (appendix p 13).[85] Therefore, net benefit or standardised net benefit is plotted in a decision curve for a range of reasonable decision thresholds.[21,82] Net benefit and standardised net benefit are semi-proper; the probability estimates below the threshold can be anything as long as they are below the threshold, and the same goes for probability estimates above the threshold.[48]

A related measure is the expected cost.[86–89] In contrast to net benefit, expected cost searches for the decision threshold that minimises cost given the misclassification costs. Miscalibration of the model might thus be reflected in the decision threshold at which expected cost is minimised, whereas net benefit fixes the threshold such that miscalibration reduces the net benefit value.[89] Expected cost is semi-proper because it is insensitive to rank-preserving transformations of the probabilities. If we normalise the costs to sum to 1, we can plot expected cost for a range of reasonable normalised costs of a false positive or false negative.

Following decision theory, the key concern is to check whether the model has better utility than the reference strategies (to either treat everyone or treat no one) and, if relevant, competing models. For our ADNEX case study, if we accept to intervene in up to ten patients per true positive, we consider that the benefit of a true positive (or the harm of a false negative) is nine times higher than the harm of a

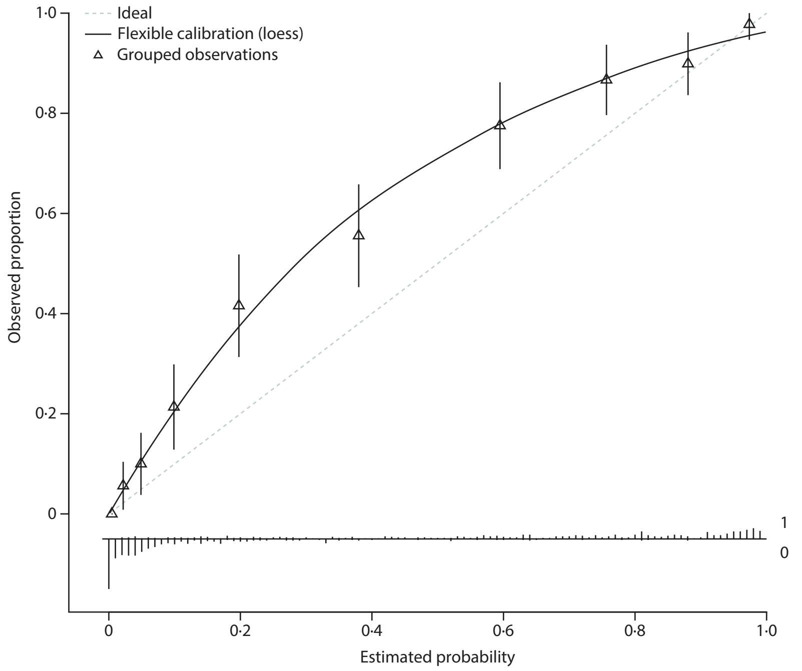

Figure 2: **Calibration plot for the ADNEX model using ten groups of equal sample size and using a loess smoother on the estimated probability**

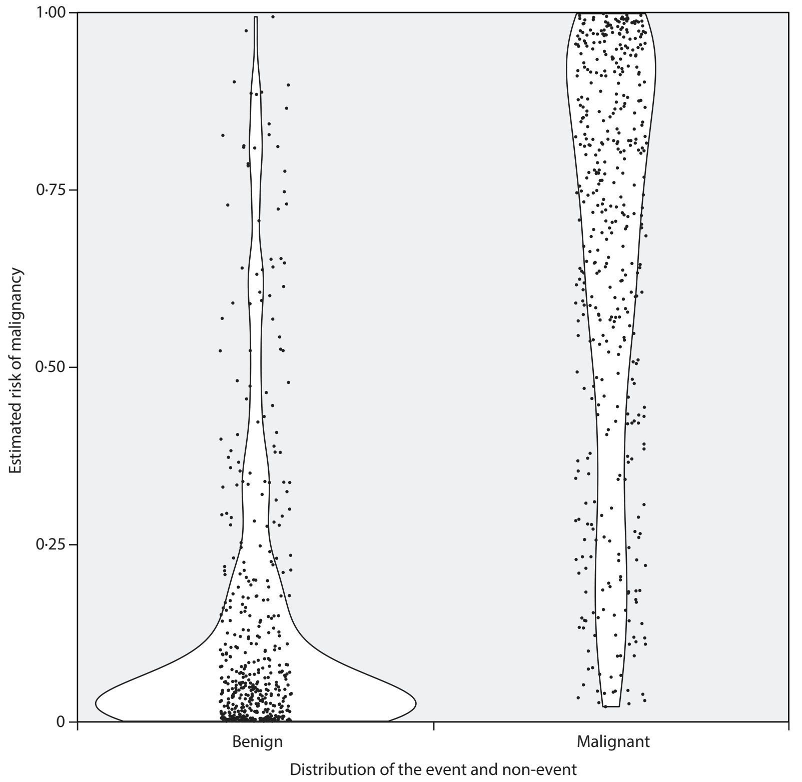

Figure 3: **Violin and dot plots of the estimated probabilities of malignancy based on the ADNEX model**

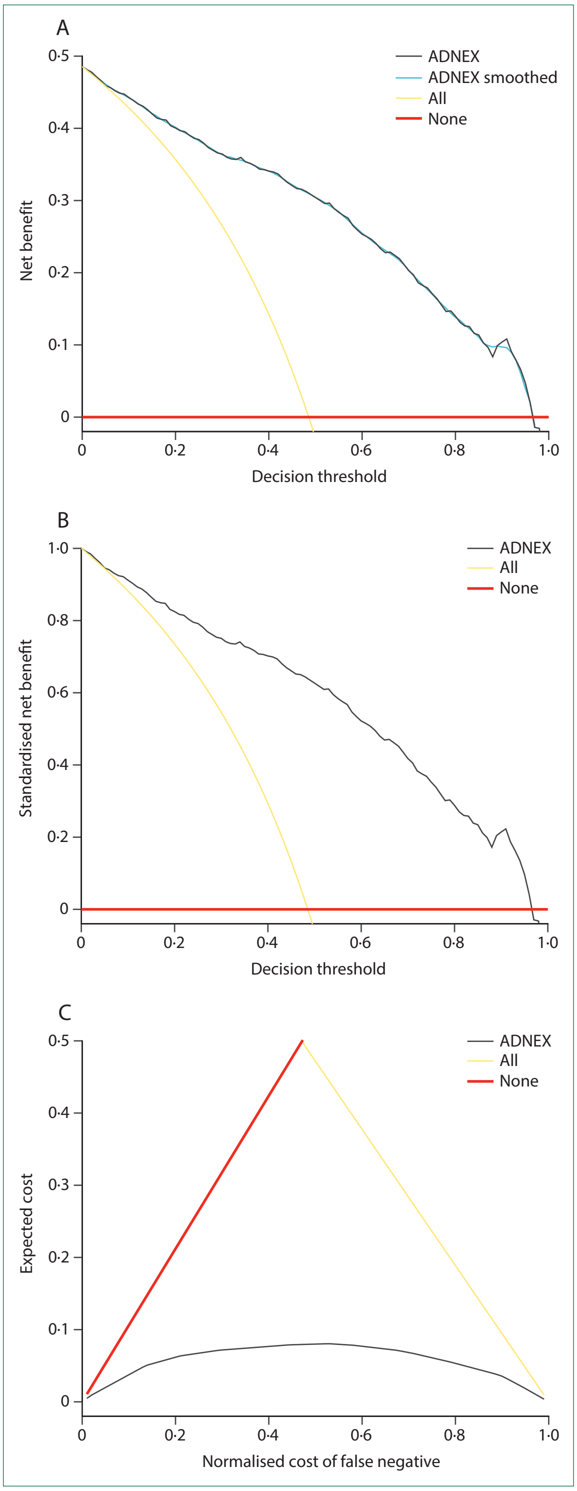

***Figure 4:* Decision curves with net benefit, standardised net benefit, and expected cost for the ADNEX model**

Decision curves with net benefit (A), standardised net benefit (B), and expected cost for our case study (C). We show the full x-axis range for educational purposes. As explained in the Clinical utility section, a reasonable range of decision thresholds is 0·05 to 0·40. This range corresponds to a range of normalised costs between 0·05 and 0·40 on the curve for expected cost. While plots (A), (B), and (C) shows results for decision thresholds or normalised costs between 0 and 1 for didactical reasons, it is recommended to restrict the x-axis to the reasonable range when presenting a decision curve in a validation study. (A) also shows a smoothed curve using central moving averages. All refers to the net benefit or expected cost of the default strategy to classify all individuals as high risk. None refers to the net benefit or expected cost of the default strategy to classify all individuals as low risk.

false positive. The associated decision threshold for these misclassification costs is 0·1.[18] For expected cost, the normalised cost of a false negative is 0·9 (*vs* 0·1 for a false positive). We further assume that the range of reasonable thresholds is 0·05 to 0·40 (normalised cost of false negatives between 0·60 and 0·95). The (standardised) net benefit was better for ADNEX than for the reference strategies across all reasonable decision thresholds (figure 4A–B). The expected cost curve gives the same impression (figure 4C). The net benefit of the model at *t*=0·1 was 0·44. For a normalised cost of a false negative of 0·9, expected cost was minimised to 0·35 at *t*=0·06.

Results after recalibration

Graphical displays of model performance for the recalibrated model are shown in the appendix (pp 25–29). The calibration plot is closer to the diagonal after recalibration in the validation dataset. Table 1 provides all performance measures for the ADNEX model before and after recalibration. All strictly proper measures improved after recalibration. Semi-proper measures either improved or remained unchanged. For example, owing to the rank-preserving updating method, recalibration cannot improve discrimination measures such as AUROC as they are based on ranks. The improper summary measures for classification (except diagnostic odds ratio) worsened remarkably. Some partial classification measures improved (sensitivity and NPV), whereas others worsened (specificity and PPV). The improper measures for overall performance improved. The worsening of most improper performance measures after recalibration illustrates the importance of the properness concept.

Discussion

We evaluated 32 classic and contemporary performance measures across five performance domains (discrimination, calibration, overall performance, classification, and clinical utility) for predictive AI models intended for medical practice. When validating the performance of a prediction model, we warn against the use of measures that are improper (13 measures) or that do not have a clear focus on either statistical or decision–analytical performance (three measures; table 2). Remarkably, F1 is the only measure violating both characteristics. Improper measures might mislead researchers instead of clarifying the performance of a model. Measures that conflate statistical and decision–analytical performance without properly accounting for misclassification costs are ambiguous and should be replaced with dedicated measures for clinical utility.

	Recommendation	Remarks
Discrimination		
AUROC	Recommended	This measure quantifies discrimination, which is a key component of statistical model performance.
AUPRC and pAUROC	Inadvisable	These measures attempt to move beyond a statistical assessment but violate decision–analytical principles.
ROC curve and PR curve	Neither inadvisable nor essential	These plots provide limited additional information over AUROC.
Calibration		
O:E ratio	Neither inadvisable nor essential	This measure is interpretable but provides only a partial assessment of calibration; O:E ratio is often 1 or close to 1 during internal validation.
Calibration intercept and calibration slope	Neither inadvisable nor essential	These measures are hard to interpret and provide a partial assessment of calibration; during internal validation, calibration slope can be used to gauge overfitting.[66]
ECI, ICI, and ECE	Not essential	These measures summarise calibration plots, concealing the nature and direction of miscalibration, and struggle with statistical consistency.
Calibration plot or reliability diagram	Recommended	This measure is the most insightful approach to assess calibration, particularly when smoothing is used rather than grouping; for internal validation, a plot is preferred but reporting only the calibration slope is acceptable; for external validation, a calibration plot is strongly recommended, with indications of uncertainty (eg, by 95% CIs).
Overall performance		
Loglikelihood, Brier, R^2 measures	Neither inadvisable nor essential	We advise to evaluate discrimination and calibration separately. These measures are highly relevant for model selection tasks, which are beyond the scope of this Viewpoint.
Discrimination slope and MAPE	Inadvisable	These measures are improper; ie, values can be better for incorrect models than for the correct model.
Risk distribution plots	Recommended	Displaying the distribution of the risk estimates for each outcome category provides valuable insights into a model's behaviour.
Classification		
Classification accuracy, balanced accuracy, Youden index, DOR, kappa, F1, and MCC	Inadvisable	These measures are improper at clinically relevant decision thresholds; in addition, some measures are hard to interpret.
Sensitivity (recall) and specificity	Not essential; can be descriptive if reported together	Although improper on their own, they can be presented descriptively if reported together. However, these measures are largely theoretical as they condition on the predicted outcome.
PPV (precision) and NPV	Not essential; can be descriptive if reported together	Although improper on their own, they can be presented descriptively if reported together. PPV and NPV are highly practical measures because they condition on the classification.
Classification plot	Neither inadvisable nor essential	Classification plots could be presented descriptively, showing either sensitivity and specificity or PPV and NPV by threshold.
Clinical utility		
NB or standardised NB (with a decision curve) and EC (with a cost curve)	Recommended	Important measures to quantify to what extent better decisions are made. Decision curves of NB allow one to show potential clinical utility at various clinically relevant decision thresholds relative to default decisions (and competing models).

AUPRC=area under the precision–recall curve. AUROC=area under the receiver operating characteristic curve. DOR=diagnostic odds ratio. EC=expected cost. ECE=expected calibration error. ECI=estimated calibration index. ICI=integrated calibration index. MAPE=mean absolute prediction error. MCC=Matthew's correlation coefficient. NB=net benefit. NPV=negative predictive value. O:E ratio=observed over expected ratio. pAUROC=partial AUROC. PPV=positive predictive value. PR=precision–recall. ROC=receiver operating characteristic.

Table 2: **Recommendations and remarks for different measures and plots in the context of validating a prediction model to support clinical decision making**

We argue that performance assessment of predictive AI models intended for medical practice should focus on discrimination, calibration, and clinical utility.[90] Discrimination and calibration aid the modeller and clinician to understand how a model can be improved. Poor discrimination indicates that other predictors facilitating improved distinction between individuals with and without the event could be selected. Miscalibration can compromise predictive AI application by leading to systematic overtreatment or undertreatment.[89] Miscalibration is often not just a problem of the model but a sign that we need to improve our understanding of the various contexts in which the model is validated and used.[91] Unfortunately, calibration measures are still under-reported.[92–94] Overall performance measures combine discrimination and calibration performance, making them less informative than separate assessments of discrimination and calibration performance. Clinical utility focuses on the decision maker and the patient by evaluating whether a model leads to improved clinical decisions on average.

We recommend the following core set of measures and plots that should be reported: AUROC, a smoothed calibration plot, a clinical utility measure such as net benefit with a decision curve, and a figure showing probability distributions for each outcome category (table 2). When internally validating a predictive AI model, calibration might be less important because model development and internal validation are based on individuals from the exact same population. Calibration is more important for external validation, when models are evaluated in different contexts and populations. Although a calibration plot is useful during internal validation, a limited assessment using calibration slope and perhaps O:E ratio can suffice; however, we would expect an O:E ratio close to 1 for well-developed models. In addition to the recommended core set, PPV in combination with NPV, or sensitivity in combination with specificity, can be reported descriptively. These measures are improper when used alone. Reported measures and plots should be accompanied by confidence intervals when possible, except for clinical utility measures, for which quantification of uncertainty is a topic of recent debate and research.[95–97]

Class imbalance has received a lot of attention for model development and performance assessment. We argue that class imbalance is not as problematic as often claimed. The extent of class imbalance is not mathematically proportional to the extent of imbalance in misclassification costs. Class imbalance is related to the target population as an epidemiological feature of the data, whereas misclassification costs are clinical concepts that relate to the context of decision making. Misclassification costs are informed by the nature and effect of the medical intervention at hand (eg, the decision to perform surgery or not).[21,82,85,98] Therefore, we advise against using F1, AUPRC, or pAUROC, in favour of a dedicated clinical utility measure.[39–46] Of note, we do not make claims regarding other situations in health care when true negatives are not well defined, such as lesion detection.[8]

Three topics related to performance assessment deserve emphasis: sample size, performance heterogeneity, and reporting transparency. First, adequate sample size is important to evaluate performance with sufficient precision. Previous recommendations were to include at least 100 to 200 individuals in the smallest outcome category.[52,99] More specific sample size calculations are now available for regression-based models.[100] Often, more data are needed when comparing calibration between models.[101] Second, heterogeneity in model performance should be expected based on differences in populations and measurement procedures between locations, settings, or time periods.[91,102–104] Meta-analysis and meta-regression methods can be used to quantify and understand heterogeneity in performance across external validation studies.[102,105,106] Naive comparison of models validated using different external datasets, reflecting different populations from different settings, can lead to wrong conclusions.[36] Third, comprehensive reporting of predictive AI modelling studies is imperative, which can be done by adhering to the TRIPOD+AI and related reporting guidelines.[107–109] To avoid performance hacking, increased attention should be paid to publishing protocols in advance, as well as to sharing of analysis code and data where reasonably possible.[110]

For more on the **code**, see https://github.com/benvancalster/PerfMeasuresOverview

A limitation of this Viewpoint is that we focused only on performance measures for binary outcomes. Nevertheless, the principles also hold for other types of outcomes, such as nominal, ordinal, time-to-event, or competing risk outcomes. A second limitation is that we could discuss several other topics in depth. We did not address counterfactual prediction (prediction under hypothetical interventions), which has deservedly gained traction recently.[111,112] Also, discussing all measures is impossible, and research on performance measures is ongoing. For example, calibration is an active area of research focusing on aspects such as strong calibration, quantifying the degree of miscalibration, and uncertainty.[62,113–115] Furthermore, we did not directly discuss model comparisons, although head-to-head comparisons of competing models on the same external validation dataset is of particular importance.[116] A specific topic related to model comparison is evaluating the incremental value of adding a new predictor to an existing model.[117] Although competing models can be evaluated using the same core set of measures and visualisations, proper overall measures become more interesting for tasks such as model selection and comparison. Dedicated measures, such as the widely used but improper net reclassification improvement, are available for evaluating competing models.[48,118,119]

In conclusion, we argue that performance measures should be proper and clearly focus on either purely statistical or decision–analytical evaluation. To evaluate predictive AI models for medical practice, the recommended core set of performance measures that is suitable for most circumstances include AUROC, calibration plot, a clinical utility measure such as net benefit with decision curve analysis, and a plot showing the distribution of risk estimates.

Contributors

BVC conceptualised the study. BVC and DT were responsible for resources and data curation. BVC, DT, and LB performed formal analysis. BVC and LB were responsible for software, validation, and visualisation. BVC and EWS wrote the original draft. All authors were responsible for investigation, methodology, writing and review, and editing. All authors agreed with submission of the manuscript for publication. BVC, LB, and DT accessed and verified the raw data. All authors had full access to all the data in the study and BVC and EWS had final responsibility for the decision to submit for publication.

Declaration of interests

BVC has stock options in Gynaia Inc, a company providing teaching, training, and decision support tools such as the ADNEX model for gynaecologists and sonographers. DT is a co-founder of, and has stock options in, Gynaia Inc. KFK received an honorarium for teaching a short course in the UW Summer Institutes for Statistics in Clinical and Epidemiological Research in July 2023 and July 2024. GV has stock options in Therapixel, a company providing imaging analysis software based on artificial intelligence (AI). TH-B received royalties from Coursera for courses on AI in health care, received consulting fees from Paul Hartmann AG and Grai-Matter, and has stock options in Verantos Inc. DJM received honoraria from Merck for a fertility prediction modelling workshop, and financial support from Merck and IVIRMA to present at scientific conferences. All other authors declare no competing interests.

Data sharing

The code developed and used in this study is freely available from the GitHub repository. The complete dataset is registered in the KU Leuven Research Data Repository but cannot be publicly shared owing to privacy considerations. However, we do provide risk estimates and outcomes (0 *vs* 1) on the GitHub page.

Acknowledgments

BVC received funding from Research Foundation–Flanders (FWO; project G097322N), Internal Funds KU Leuven (project C24M/20/064), and Kom Op Tegen Kanker (Stand up to Cancer; projects TransIOTA and BIOC). GSC received funding from Cancer Research UK (grant number C49297/A27294), Medical Research Council Better Methods Better Research (grant number MR/V038168/1), the Engineering and Physical Sciences Research Council grant for the project, and Artificial intelligence innovation to accelerate health research (grant number EP/Y018516/1). GSC is a senior investigator at National Institute for Health and Care Research (NIHR). The views expressed in this article are those of the author(s) and do not necessarily reflect those of the NIHR or the Department of Health and Social Care. AJV was supported in part by a grant from the National Cancer Institute (grant number P30CA008748). LW received funding from the Talent Programme financed by the Dutch Research Council (NWO) under the VIDI grant 9150172310023. KFK received funding from the National Institutes of Health (NIH)/National Institutes of Diabetes and Digestive and Kidney Diseases (grant number R01 DK093770). LB was supported by FWO (project G097322N). GV acknowledges funding from the PEPR SN SMATCH France 2030 (project ANR-22-PESN-0003). KS was supported by NIH (projects R42AI177108, R01DK133226, and R01EB030492). TH-B was supported by the National Center for Advancing Translational Sciences of the NIH (award Number UL1TR003142). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH. DT was supported by FWO (projects G049312N, G0B4716N, and 12F3114N), Internal Funds KU Leuven (project C24/15/037), and Kom Op Tegen Kanker (projects TransIOTA and BIOC). DT is a senior clinical investigator at FWO. The funders had no role in study design, data collection, data analysis, data interpretation, writing of the manuscript, or decision to publish.

For more on the **dataset**, see https://doi.org/10.48804/TXL95Z

References

1 Van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KGM. Clinical prediction models: diagnosis versus prognosis. *J Clin Epidemiol* 2021; **132:** 142–45.

2 Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer, 2019.

3 Zamaray B, Veld JV, Burghgraef TA, et al. Risk factors for a permanent stoma after resection of left-sided obstructive colon cancer - A prediction model. *Eur J Surg Oncol* 2023; **49:** 738–46.

4 Yuan N, Duffy G, Dhruva SS, et al. Deep learning of electrocardiograms in sinus rhythm from US veterans to predict atrial fibrillation. *JAMA Cardiol* 2023; **8:** 1131–39.

5 Kamran F, Tjandra D, Heiler A, et al. Evaluation of sepsis prediction models before onset of treatment. *NEJM AI* 2024; **1**.

6 Habbema JDF, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis. I. The problem, descriptive tools, and measures based on classification matrices. *Methods Inf Med* 1978; **17:** 217–26.

7 Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. *Pattern Recognit Lett* 2009; **30:** 27–38.

8 Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21:** 128–38.

9 Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: recommendations for image analysis validation. *Nat Methods* 2024; **21:** 195–212.

10 Hand DJ. Assessing the performance of classification methods. *Int Stat Rev* 2012; **80:** 400–14.

11 Hernández-Orallo J, Flach P, Ferri C. A unified view of performance metrics: translating threshold choice of into expected classification loss. *J Mach Learn Res* 2012; **13:** 2813–69.

12 Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 2024; **384:** e074819.

13 de Hond AAH, Shah VB, Kant IMJ, et al. Perspectives on validation of clinical predictive algorithms. *npj Digit Med* 2023; **6:** 86.

14 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.

15 Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. *Diagn Progn Res* 2022; **6:** 24.

16 Yates JF. External correspondence: decompositions of the mean probability score. *Organ Behav Hum Perform* 1982; **30:** 132–56.

17 Pencina MJ, Fine JP, D'Agostino RB Sr. Discrimination slope and integrated discrimination improvement - properties, relationships and impact of calibration. *Stat Med* 2017; **36:** 4482–90.

18 Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. *N Engl J Med* 1975; **293:** 229–34.

19 Elkan C. The foundations of cost-sensitive learning. In: Proceedings of the 17th international joint conference on artificial intelligence. Morgan Kaufmann Publishers, 2001: 973–78.

20 Wynants L, van Smeden M, McLernon DJ, et al. Three myths about risk thresholds for prediction models. *BMC Med* 2019; **17:** 192.

21 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; **26:** 565–74.

22 Adams NM, Hand DJ. An improved measure for comparing diagnostic tests. *Comput Biol Med* 2000; **30:** 89–96.

23 Timmerman D, Planchamp F, Bourne T, et al. ESGO/ISUOG/IOTA/ESGE consensus statement on preoperative diagnosis of ovarian tumors. *Ultrasound Obstet Gynecol* 2021; **58:** 148–68.

24 Hilden J. Commentary: on NRI, IDI, and "good-looking" statistics with nothing underneath. *Epidemiology* 2014; **25:** 265–67.

25 Choodari-Oskooei B, Royston P, Parmar MKB. A simulation study of predictive ability measures in a survival model I: explained variation measures. *Stat Med* 2012; **31:** 2627–43.

26 Buja A, Stuetzle W, Shen Y. Loss functions for binary class probability estimation and classification: structure and applications. November, 2005. http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf (accessed May 15, 2025).

27 Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 2007; **102:** 359–78.

28 Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med* 2014; **33:** 3405–14.

29 Kull M, Flach P. Novel decompositions of proper scoring rules for classification: score adjustment as precursor to calibration. In: Appice A, Rodrigues PP, Costa VS, et al, eds. Machine learning and knowledge discovery in databases. Springer, 2015: 68–85.

30 Linnet K. Assessing diagnostic tests by a strictly proper scoring rule. *Stat Med* 1989; **8:** 609–18.

31 Van Calster B, Van Hoorde K, Valentin L, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study. *BMJ* 2014; **349:** g5920.

32 Landolfo C, Ceusters J, Valentin L, et al. Comparison of the ADNEX and ROMA risk prediction models for the diagnosis of ovarian cancer: a multicentre external validation in patients who underwent surgery. *Br J Cancer* 2024; **130:** 934–40.

33 Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958; **45:** 562–65.

34 Ojeda FM, Jansen ML, Thiéry A, et al. Calibrating machine learning approaches for probability estimation: a comprehensive comparison. *Stat Med* 2023; **42:** 5451–78.

35 Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola AJ, Bartlett PJ, Schölkopf B, Schuurmans D, eds. Advances in large margin classifiers. MIT Press, 1999: 61–74.

36 Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Brief Bioinform* 2012; **13:** 83–97.

37 Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 1997; **30:** 1145–59.

38 Verbakel JY, Steyerberg EW, Uno H, et al. ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *J Clin Epidemiol* 2020; **126:** 207–16.

39 Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. Springer, 2018.

40 Adams NM, Hand DJ. Improving the practice of classifier performance assessment. *Neural Comput* 2000; **12:** 305–11.

41 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; **10:** e0118432.

42 Carrington AM, Fieguth PW, Qazi H, et al. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Med Inform Decis Mak* 2020; **20:** 4.

43 Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr* 2008; **17:** 145–51.

44 Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Cohen W, Moore A, eds. Proceedings of the 23rd international conference on machine learning. Association for Computing Machinery, 2006: 233–40.

45 Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol* 2015; **68:** 855–59.

46 Yuan Y, Su W, Zhu M. Threshold-free measures for assessing the performance of medical screening tests. *Front Public Health* 2015; **3:** 57.

47 Dodd LE, Pepe MS. Partial AUC estimation and regression. *Biometrics* 2003; **59:** 614–23.

48 Pepe MS, Fan J, Feng Z, Gerds T, Hilden J. The Net Reclassification Index (NRI): a misleading measure of prediction improvement even with independent test data sets. *Stat Biosci* 2015; **7:** 282–95.

49 McDermott MBA, Hansen LH, Zhang H, Angelotti G, Gallifant J. A closer look at AUROC and AUPRC under class imbalance. *Adv Neural Inf Process Syst* 2024; **37:** 44102–63.

50 Loh TP, Lord SJ, Bell K, et al. Setting minimum clinical performance specifications for tests based on disease prevalence and minimum acceptable positive and negative predictive values: practical considerations applied to COVID-19 testing. *Clin Biochem* 2021; **88:** 18–22.

51 de Hond AAH, Steyerberg EW, Van Calster B. Interpreting area under the receiver operating characteristic curve. *Lancet Digit Health* 2022; **4:** e853–55.

52 Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016; **74:** 167–76.

53 Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the “calibration slope” really measure? *J Clin Epidemiol* 2020; **118:** 93–99.

54 Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; **17:** 230.

55 Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Dzeroski S, De Raedt L, Wrobel S, eds. Proceedings of the 22nd international conference on machine learning. Association for Computing Machinery, 2005: 625–32.

56 Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med* 2014; **33:** 517–35.

57 Błasiok J, Nakkiran P. Smooth ECE: principled reliability diagrams via kernel smoothing. *arXiv* 2023; published online Sept 21. https://arxiv.org/abs/2309.12236 (preprint).

58 Naeini MP, Cooper GF, Hauskrecht M. Obtaining well calibrated probabilities using Bayesian binning. In: Proceedings of the AAAI conference on artificial intelligence. AAAI Press, 2015: 2901–07.

59 Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *J Biomed Inform* 2015; **54:** 283–93.

60 Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med* 2019; **38:** 4051–65.

61 Nattino G, Pennell ML, Lemeshow S. Assessing the goodness of fit of logistic regression models in large samples: a modification of the Hosmer-Lemeshow test. *Biometrics* 2020; **76:** 549–60.

62 Arrieta-Ibarra I, Gujral P, Tannen J, Tygert M, Xu C. Metrics of calibration for probabilistic predictions. *J Mach Learn Res* 2022; **23:** 1–54.

63 Bishop CM. Pattern recognition and machine learning. Springer, 2006.

64 Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 1950; **78:** 1–3.

65 Wilks DS. Statistical methods in the atmospheric sciences. Academic Press, 1995.

66 Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagn Progn Res* 2018; **2:** 7.

67 Menard S. Coefficients of determination for multiple logistic regression analysis. *Am Stat* 2000; **54:** 17–24.

68 Hu B, Palta M, Shao J. Properties of R^2 statistics for logistic regression. *Stat Med* 2006; **25:** 1383–95.

69 Nagelkerke NJD. A note on a general definition of the coefficient of determination. *Biometrika* 1991; **78:** 691–92.

70 Ferri C, Flach P, Hernandez-Orallo J, Senad A. Modifying ROC curves to incorporate predicted probabilities. In: Lachiche N, Ferri C, Macskassy S, Rakotomamonjy A, eds. Proceedings of the second workshop on ROC analysis in machine learning. International Conference on Machine Learning, 2005: 33–40.

71 Van Houwelingen HC, Putter H. Dynamic prediction in clinical survival analysis. CRC Press, 2012.

72 Varoquaux G, Colliot O. Evaluating machine learning models and their diagnostic value. In: Colliot O, ed. Machine learning for brain disorders. Springer, 2023: 601–30.

73 Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3:** 32–35.

74 Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: Proceedings of the 2010 20th international conference on pattern recognition. IEEE, 2010: 3121–24.

75 Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; **20:** 37–46.

76 Van Rijsbergen CJ. Information retrieval. Butterworths, 1979.

77 Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020; **21:** 6.

78 Christen P, Hand DJ, Kirielle N. A review of the F-measure: its history, properties, criticism, and alternatives. *ACM Comput Surv* 2024; **56:** 1–24.

79 Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000; **16:** 412–24.

80 Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min* 2021; **14:** 13.

81 Hunink M, Glasziou P, Siegel J. Decision-making in health and medicine: integrating evidence and values. Cambridge University Press, 2001.

82 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352:** i6.

83 Kerr KF, Brown MD, Marsh TL, Janes H. Assessing the clinical impact of risk models for opting out of treatment. *Med Decis Making* 2019; **39:** 86–90.

84 Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *J Clin Oncol* 2016; **34:** 2534–40.

85 Vickers AJ, Van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019; **3:** 18.

86 Margineantu DD, Dietterich TG. Bootstrap methods for the cost-sensitive evaluation of classifiers. In: Langley P, ed. Proceedings of the 17th international conference on machine learning. Morgan Kaufmann Publishers, 2000: 583–90.

87 Ferrer L. Analysis and comparison of classification metrics. *arXiv* 2022; published online Sep 12. https://arxiv.org/abs/2209.05355 (preprint).

88 Provost F, Fawcett T. Robust classification for imprecise environments. *Mach Learn* 2001; **42:** 203–31.

89 Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making* 2015; **35:** 162–69.

90 Kerr KF, Janes H. First things first: risk model performance metrics should reflect the clinical application. *Stat Med* 2017; **36:** 4503–08.

91 Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Med* 2023; **21:** 70.

92 Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ* 2020; **369:** m1328.

93 Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021; **375:** n2281.

94 Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol* 2022; **22:** 101.

95 Vickers AJ, Van Claster B, Wynants L, Steyerberg EW. Decision curve analysis: confidence intervals and hypothesis testing for net benefit. *Diagn Progn Res* 2023; **7:** 11.

96 Kerr KF, Marsh TL, Janes H. The importance of uncertainty and opt-in v. opt-out: best practices for decision curve analysis. *Med Decis Making* 2019; **39:** 491–92.

97 Sadatsafavi M, Lee TY, Wynants L, Vickers AJ, Gustafson P. Value-of-information analysis for external validation of risk prediction models. *Med Decis Making* 2023; **43:** 564–75.

98 Hand D, Christen P. A note on using the F-measure for evaluating record linkage algorithms. *Stat Comput* 2018; **28:** 539–47.

99 Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med* 2016; **35:** 214–26.

100 Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med* 2021; **40:** 4230–51.

101 Peek N, Arts DGT, Bosman RJ, van der Voort PHJ, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. *J Clin Epidemiol* 2007; **60:** 491–501.

102 van Leeuwen FD, Steyerberg EW, van Klaveren D, Wessler B, Kent DM, van Zwet EW. Instability of the AUROC of clinical prediction models. *Stat Med* 2025; **44:** e70011.

103 Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. *Nat Med* 2023; **29:** 2686–87.

104 Jenkins DA, Martin GP, Sperrin M, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res* 2021; **5:** 1.

105 Debray TPA, Damen JAAG, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res* 2019; **28:** 2768–86.

106 Barreñada L, Ledger A, Dhiman P, et al. ADNEX risk prediction model for diagnosis of ovarian cancer: systematic review and meta-analysis of external validation studies. *BMJ Med* 2024; **3:** e000817.

107 Moons KGM, 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.

108 Debray TPA, Collins GS, Riley RD, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster): explanation and elaboration. *BMJ* 2023; **380:** e071058.

109 Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024; **385:** e078378.

110 White N, Parsons R, Collins GS, Barnett A. Evidence of questionable research practices in clinical prediction models. *BMC Med* 2023; **21:** 339.

111 van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur J Epidemiol* 2020; **35:** 619–30.

112 Prosperi M, Guo Y, Sperrin M, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell* 2020; **2:** 369–75.

113 Perez-Lebel A, Morvan ML, Varoquaux G. Beyond calibration: estimating the grouping loss of modern neural networks. *arXiv* 2022; published online Oct 22. https://arxiv.org/abs/2210.16315 (preprint).

114 Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med* 2021; **4:** 4.

115 Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. *Biom J* 2023; **65:** e2200302.

116 Gupta S, Ko DT, Azizi P, et al. Evaluation of machine learning algorithms for predicting readmission after acute myocardial infarction using routinely collected clinical data. *Can J Cardiol* 2020; **36:** 878–85.

117 Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest* 2012; **42:** 216–28.

118 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* 2008; **27:** 157–72.

119 Leening MJG, Vedder MM, Witteman JCM, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med* 2014; **160:** 122–31.