STRengthening Analytical Thinking for Observational Studies (STRATOS): TG6 – OVERVIEW AND GUIDANCE OF PERFORMANCE MEASURES FOR CLINICAL PREDICTION MODELS

Ben Van Calster (1, 2, 3), Gary S Collins (4), Laure Wynants (1, 2, 5), Kathleen F Kerr (6), Andrew J Vickers (7), Barreñada L (1, 2), Karel GM Moons (3), David J McLernon (8), Maarten van Smeden (3), Ewout W Steyerberg (3)

(1) Department of Development and Regeneration, KU Leuven, Leuven, Belgium

(2) Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium

(3) Julius Center for Health Sciences and Primary Care, University Medical Center (UMC) Utrecht, Utrecht, Netherlands

(4) Department of Applied Health Sciences, School of Health Sciences, College and Medicine and Health, University of Birmingham, Birmingham, UK

(5) Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, Netherlands

(6) Department of Biostatistics, University of Washington School of Public Health, Seattle, WA, USA

(7) Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

(8) Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK

With the increasing focus on machine learning algorithms, the number of risk prediction models for clinical diagnostic or prognostic outcomes is growing rapidly. Information on model performance in applied publications on clinical prediction models is chaotic with large variation in the number and selection of reported performance measures. Guidance around the relevance of available performance measures is important especially with clinical prediction models considered as medical devices. Topic group 6 from the STRATOS initiative provides a network to develop such guidance.

Together with experts from the machine learning field, we reviewed 32 key performance measures and accompanying plots [1]. This project updated an overview from 2010 that did not cover the machine learning perspective [2]. We focused on measures for prediction models with a binary diagnostic or prognostic outcome that are intended to be deployed in practice to support clinical decision-making. We illustrate claims using simulations and a case study in which the ADNEX risk model for ovarian cancer diagnosis is externally validated [3].

First, we outlined five domains of model performance under which individual measures can be classified. Discrimination addresses the extent to which the model can distinguish ('discriminate') between patients with the event and patients without the event. Calibration addresses the extent to which estimated probabilities are reliable: among individuals with an estimated probability of the event of 5%, do we observe that 5/100 have the event? Overall performance evaluates the closeness of event probabilities and event status (0 or 1). Overall performance reflects discrimination and calibration. Classification quantifies the extent to which actual

event outcomes correspond to classifications of individuals as low or high risk of the event using a probability threshold. Clinical utility assesses the extent to which these classifications may lead to better decisions. This decision-analytic performance goes beyond the other domains that focus on statistical performance.

Performance measures should exhibit two desirable characteristics: properness and focus. A proper measure yields the optimal value for the correct model, i.e. the model that yields correct probabilities conditional on the predictor variables. For focus, we consider that performance measures should either focus solely on a statistical aspect or decision-analytic performance by properly considering misclassification costs, i.e. the differential costs of a false positive (suggesting intervention in someone without the event) and a false negative (suggesting no intervention in someone with the event).

Fifteen measures violate one or both key characteristics. Thirteen of the 32 measures are improper: all 11 classification measures and 2 out of 9 overall performance measures (discrimination slope and mean absolute prediction error). Some classification measures are proper only if the probability threshold is 0.5 or equal to the event proportion, yet this is rarely a clinically relevant threshold.

Three measures have no clear focus because they poorly address misclassification costs (AUPRC or the area under the precision-recall curve, partial AUROC or the partial area under the receiver operating characteristic curve, and the F1 score). Notably, the F1 score violates both characteristics and is therefore the most misleading performance measure. This is remarkable since it is commonly reported in the machine learning literature.

In the case study, we externally validated the ADNEX model twice: the ADNEX model as is, and a recalibrated version of ADNEX. We observed that recalibration improved or did not change the value of proper measures. As expected, (semi-)proper measures whose value did not change were rank preserving measures such as AUROC. In contrast, recalibration worsened model performance for 8 of the 13 improper measures. This illustrates that improper measures can be misleading by favoring the poorer model.

For reporting, we suggest a set of measures and plots: (1) a risk distribution plot, (2) the AUROC (or the c statistic), (3) a flexible calibration plot, and (4) a clinical utility measure such as net benefit or expected cost in combination with a decision curve to evaluate model-based decision making for a range of relevant values for the differential costs of false positives and false negatives. Exception can be motivated, e.g. for the pair of sensitivity and specificity and the pair of positive predictive value and negative predictive value. These four measures are partial classification measures in that they deliberately condition on event status (sensitivity and specificity) or classification status (positive and negative predictive value). In isolation, these four measures are improper, but they may be valuable to report descriptively in pairs: sensitivity and specificity together, and/or positive and negative predictive value together.

We hope that focusing on a core set of informative performance measures with desirable characteristics will streamline reporting of clinical prediction models, thereby reducing arbitrariness and deception.

References

1. Van Calster B, Collins GS, Vickers AJ, Wynants L, Kerr KF, Barreñada L, Varoquaux G, Singh K, Moons KGM, Hernandez-Boussard T, Timmerman D, McLernon DJ, van Smeden M, Steyerberg EW. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. Lancet Digital Health 2025;7:100916. doi 10.1016/ j.landig.2025.100916.

2. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128-138. doi 10.1097/ EDE.0b013e3181c30fb2.

3. Landolfo C, Ceusters J, Valentin L, Froyman W, Van Gorp T, Heremans R, Baert T, Wouters R, Vankerckhoven A, Van Rompuy AS, Billen J, Moro F, Mascilini F, Neumann A, Van Holsbeke C, Chiappa V, Bourne T, Fischerova D, Testa A, Coosemans A, Timmerman D, Van Calster B. 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;130:934-940. doi 10.1038/s41416-024-02578-x.

Table. Recommendations and remarks for different measures and plots in the context of validating a prediction model to support clinical decision making. Reproduced with minor technical or grammatical adaptations from [1] (CC-BY license), no changes to the contents or recommendations were made.

Measure / Plot	Recommendation	Remark
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-analytic 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; for internal validation, quantifying calibration slope can be used to gauge overfitting (a need for 'shrinkage').
ECI, ICI, and ECE	Neither inadvisable nor essential	These measures summarize calibration plots, concealing the nature and direction of miscalibration, and struggle with statistical consistency.
Calibration plot (or reliability diagram)	Recommended	The most insightful approach to assess calibration, in particular when smoothing is used rather than grouping; for internal validation, reporting only the calibration slope is acceptable; for external validation, a calibration plot is strongly recommended, with indications of uncertainty (e.g. 95% confidence intervals).
OVERALL PERFORMANCE		
Loglikelihood, Brier, and R-squared measures (McFadden, Cox-Snell, Nagelkerke)	Neither inadvisable nor essential	We advise to evaluate discrimination and calibration separately. These measures are more relevant for model selection tasks, which are beyond the scope of this work.
Discrimination slope and MAPE	Inadvisable	These measures are improper; i.e. 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 behavior.
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 (or recall) and specificity	Not essential; can be descriptive if reported together	Although improper on their own, they can be reported descriptively if reported together. However, these measures are theoretical as they condition on the outcome.
PPV (or precision) and NPV	Not essential; can be descriptive if reported together	Although improper on their own, they can be reported descriptively if reported together. PPV and NPV are practical measures because they condition on the classification.
Classification plot	Neither inadvisable nor essential	Classification plots plot could be presented descriptively, showing either sensitivity and specificity or PPV and NPV by threshold.
CLINICAL UTILITY		
NB, standardized NB, or EC (with a decision curve)	Recommended	Important measures to quantify to what extent better decisions can be made with support of the model. Decision curves of NB allow one to show potential clinical utility at various clinically relevant decision thresholds relative to default decisions (and competing models).

AUROC, area under the receiver operating characteristic (ROC) curve; AUPRC, area under the precision-recall (PR) curve; pAUROC, partial AUROC; ; ECI, estimated calibration index; ICI, integrated calibration index; ECE, expected calibration error; R2, R-squared; MAPE, mean absolute prediction error; DOR, diagnostic odds ratio; MCC, Matthew's correlation coefficient; PPV, positive predictive value; NPV, negative predictive value; NB, net benefit; EC, expected cost.