Interpreting area under the receiver operating characteristic curve

Published Online
October 18, 2022
https://doi.org/10.1016/S2589-7500(22)00188-1

Clinicians often ask biostatisticians to provide a judgement on the quality of a prediction model. The prediction quality can be assessed with different performance criteria, with discriminative ability being a main issue: how well can we separate high-risk from low-risk patients? Discriminative ability is typically quantified by the area under the receiver operating characteristic curve (AUC) when we consider prediction of a binary event.

Models developed for COVID-19 patients could be an example of AUC in practice, providing risk estimates for the outcome of patients with COVID-19 (eg, 4C Mortality Score).[1] The 4C Mortality Score aims to predict 30-day mortality risk in patients with COVID-19. The AUC was 0·79 at validation in 6082 patients in 99 emergency departments in the USA.[1] The sample size was large enough to provide solid estimates of performance, with a 95% confidence interval of 0·77 to 0·80. How can we interpret this AUC?

We can refer to basic properties of the AUC. The AUC ranges between 0·5 and 1, with higher values indicating better discrimination between patients who are high risk and low risk. Some research papers label specific AUC values as poor, moderate, good, or excellent. The variety in AUC labelling systems is substantial (figure; appendix p 1). Most researchers consider AUC values lower than 0·6 as poor, but large variation exists for AUC values higher than 0·7. AUC values between 0·7 and 0·8 have been labelled as poor, moderate, fair, or good. We could not identify a justifiable reasoning on these value judgements in most papers; intuition and experience seem to have been dominant in initial texts, with simple referencing to published labelling systems in more recent literature.

One specific motivation of AUC labels is provided by comparison with the literature on effect size. Labels such as weak, moderate, and strong effect have been used often for effect sizes of 0·2 (associated AUC value 0·56), 0·5 (AUC=0·64), and 0·8 (AUC=0·71).[2] These values are associated with AUC values, under the assumption of two normal distributions of a risk score for patients with and those without the event (appendix p 2). Based on the literature on labelling systems and effect size, the 4C Mortality Score performance with AUC of 0·79 would be claimed to reflect strong discriminative ability. Again, no thorough justification is given for these effect size labels at these specific AUC values.

Literature from the past 5 years proposes labels ranging from unhelpful to very helpful for specific AUC values.[3] These labels associate AUC values to the actions that happen after deploying the prediction model. These actions typically include informing patients of their individual risk and support decision making. To reliably inform patients and physicians, we note that it is not discrimination, but calibration of risk predictions that is essential.[4] Hence, a helpful model with an AUC of more than 0·8 can be misleading if predictions are not well calibrated. For claims on making better decisions, dedicated measures beyond the AUC are needed. One such measure is net benefit.[5] Net benefit is calculated as:

$$NB=\frac{\textit{True positives}}{n}-\frac{\textit{False positives}}{n}\times\frac{p_t}{1-p_t}$$

where p_t is the decision threshold [0,1] above which a patient is positive. The true (false) positives are the number of patients with a risk prediction higher than the decision threshold that are truly positive (negative). Net benefit considers the clinical context, including the treatment decision for which the model is used and the relative importance of true-positive versus false-positive classifications. The clinical context determines what can be considered a reasonable decision threshold, not statistical criteria.[6] To evaluate the value of a model for decision making, a range of reasonable decision thresholds might be examined graphically in a decision curve. The curve goes beyond AUC as a summary measure of model quality because it allows comparison of the net benefit of a model with reference strategies (treat everyone or treat no one). When, at a given decision threshold, a model has higher net benefit than such default strategies, decision making based on the model is useful. If net benefit is lower than that of a default strategy, the model is harmful. This interpretation is lacking for performance measures related to discrimination or calibration. Net benefit could therefore serve as an initial assessment of clinical

See Online for appendix

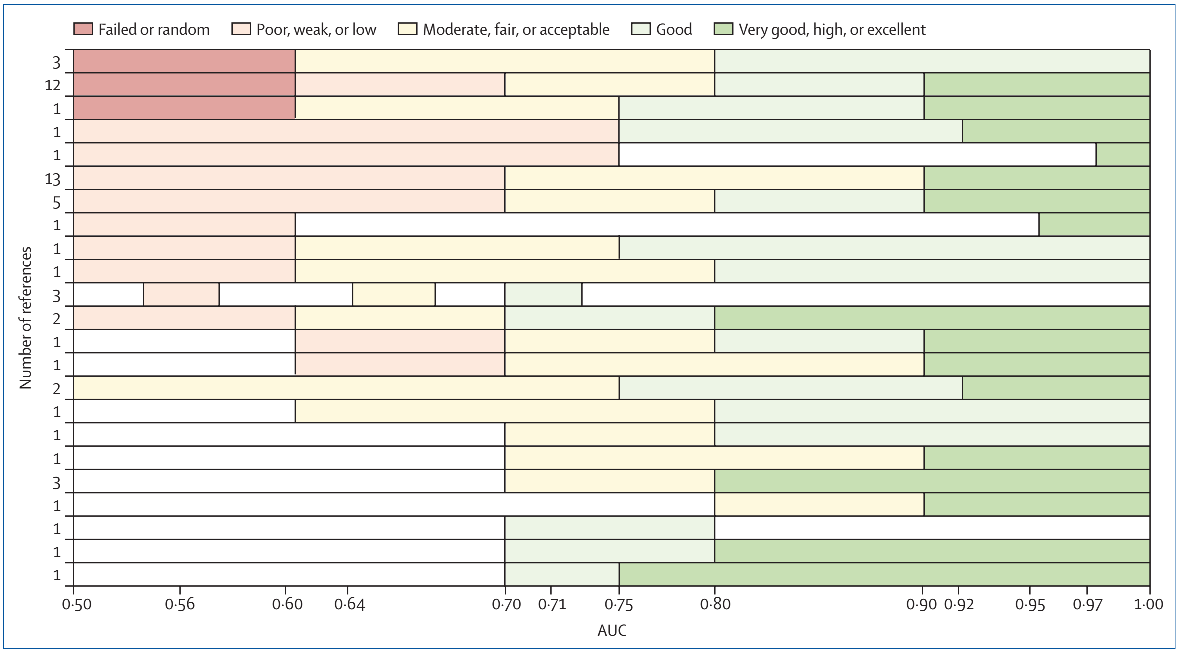

***Figure:* Different labelling systems of AUC from literature**
Number of references shown per AUC labelling system (appendix pp 4–8). AUC=area under the receiver operating characteristic curve.

usefulness, before proceeding with more time-intensive and costly evaluation steps, such as a health-economics evaluation or clinical trial for effects on patient or clinical outcomes.

Empirical research shows that models with higher AUC values are more often clinically useful.[7] When transporting a model from one setting to another, some miscalibration is typically observed, which reduces net benefit.[8] The lower the AUC, the more easily such miscalibration could make models harmful, depending on the type and extent of miscalibration, the decision threshold, and event rate (appendix pp 2-4). For the real-world 4C Mortality Score validation example, we assume that the event rate is 17%, and the AUC is 0·79. We show that quite some miscalibration can be tolerated for the model to remain helpful for decision making across a reasonable range of decision thresholds (appendix p 2), so a high AUC might compensate for poor calibration to some extent.

In conclusion, labelling systems for AUC are arbitrary. High discriminatory ability is not sufficient to claim positive potential effect of deploying a prediction model in clinical practice. The predominant purpose of AUC values might be to compare the discriminative ability of different models. If needed, the model with the best AUC can be updated to fix any observed miscalibration.[9] The implication is that we should report AUC values without using AUC labels and that we should look beyond AUC values when reporting model performance,[10] including measures for calibration and net benefit.[4,5]

We declare no competing interests.

**Anne A H de Hond, Ewout W Steyerberg, Ben van Calster*
a.a.h.de_hond@lumc.nl

Clinical AI Implementation and Research Lab (AAHdH, EWS) and Department of Biomedical Data Sciences (EWS, BvC), Leiden University Medical Centre, 2333 ZA Leiden, Netherlands; Department of Medicine (Biomedical Informatics), Stanford University, Stanford, CA, USA (AAHdH); Department of Development & Regeneration, KU Leuven, Leuven, Belgium (BvC)

1 Gordon AJ, Govindarajan P, Bennett CL, et al. External validation of the 4C Mortality Score for hospitalised patients with COVID-19 in the RECOVER network. *BMJ Open* 2022; **12:** e054700.

2 Pencina MJ, D'Agostino RB, Pencina KM, Janssens ACJW, Greenland P. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol* 2012; **176:** 473–81.

3 Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA* 2017; **318:** 1377–84.

4 Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; **17:** 230.

5 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.

6 Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980; **302:** 1109–17.

7 Gulati G, Upshaw J, Wessler BS, et al. Generalizability of cardiovascular disease clinical prediction models: 158 independent external validations of 104 unique models. *Circ Cardiovasc Qual Outcomes* 2022; **15:** e008487.

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

9 Su TL, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. *Stat Methods Med Res* 2018; **27:** 185–97.

10 Moons KG, Altman DG, Reitsma JB, Collins GS. New guideline for the reporting of studies developing, validating, or updating a multivariable clinical prediction model: the TRIPOD statement. *Adv Anat Pathol* 2015; **22:** 303–05.