Letters

COMMENT & RESPONSE

Use the Receiver Operating Characteristic to Assess Model Accuracy

To the Editor Wang et al[1] examined the ability of the receiver operating characteristic (ROC), which assesses discrimination (order of predictions) but not calibration (how close the predictions are to the observed value), to assess a model's predictive accuracy. They concluded that the ROC "does not directly quantify the prediction accuracy of the model"[1] because it does not assess calibration. They proposed quantifying a model's predictive accuracy using the mean absolute difference (MAD), which does use calibration.

A constitutive characteristic of MAD is that it is not scale invariant, which means that MAD's accuracy will vary with the time interval of the model (5-year model vs 10-year model) and that 2 models can only be directly compared if they have the same time interval. Another characteristic of MAD is that it cannot deal with unequal allocation of outcomes; when there is an unequal outcome frequency, the model can achieve a high accuracy by simply predicting the most frequent outcome. In addition, when using MAD to optimize calibration, it may change the order of the predictions, thus decreasing discrimination. Finally, MAD is sensitive to outliers. These MAD-related issues can result in models that select incorrect treatments and make inaccurate outcome predictions. In addition, if 2 randomized clinical trials have different time intervals, MAD may compromise our ability to compare the efficacy of their treatments.

There are a number of reasons why Somers,[2] who was well aware of MAD, developed the ROC in 1962. The ROC assesses a model's ability to discriminate between outcomes over the entire range of the model's predictions (all sensitivity/specificity pairs), and it allows us to compare different models. The ROC is scale invariant. Furthermore, it is not sensitive to unequal allocation of outcomes, the size of the population, skewed populations, outliers, and the prevalence of the disease in the population. Finally, the predictions of a poorly discriminating model trained with calibration cannot be improved, whereas the predictions of a highly discriminating model can be improved by a calibration postprocessor.[3] Thus, the proper approach to predictive accuracy is to optimize model discrimination and then, if necessary, apply a calibration postprocessor.

Harry B. Burke, MD, PhD

Author Affiliation: Uniformed Services University of the Health Sciences, Bethesda, Maryland.

Corresponding Author: Harry B. Burke, MD, PhD, Uniformed Services University of the Health Sciences, 2301 Jones Bridge Rd, Bethesda, MD 20914 (harry.burke@gmail.com).

Published Online: August 9, 2023. doi:10.1001/jamacardio.2023.2250

Conflict of Interest Disclosures: None reported.

Disclaimer: The views expressed in this Letter do not reflect those of the US Federal government or of the Uniformed Services University.

1. Wang X, Claggett BL, Tian L, Malachias MVB, Pfeffer MA, Wei LJ. Quantifying and interpreting the prediction accuracy of models for the time of a cardiovascular event—moving beyond C statistic: a review. *JAMA Cardiol*. 2023; 8(3):290-295. doi:10.1001/jamacardio.2022.5279

2. Somers RA. A new asymmetric measure of association for ordinal variables. *Am Sociol Rev*. 1962;27(6):799-811. doi:10.2307/2090408

3. Rosen DB, Burke HB, Goodman PH. Improving prediction accuracy using a calibration postprocessor. Paper presented at: World Congress on Neural Networks, International Neural Network Society 1996 Annual Meeting; July 16-19, 1996; San Diego CA. Lawrence Erlbaum Associates; 1996:1215-1220. Accessed June 30, 2023. https://harryburke.com/s/Rosen-1996-Calibration-postprocessor.pdf