

COMMENTARY

The need for reorientation toward cost-effective prediction: Comments on ‘Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond’ by Pencina *et al.*, *Statistics in Medicine* (DOI: 10.1002/sim.2929)

Yueh-Yun Chi¹ and Xiao-Hua Zhou^{*,†,1,2}

¹*Department of Biostatistics, University of Washington, Seattle, WA 98101, U.S.A.*

²*HSR&D VA Puget Sound Health Care System, Seattle, WA 98101, U.S.A.*

Discrimination and calibration have been two major components in the evaluation of model performance. Discrimination measures the ability of a model to separate patients with different outcomes. In the case of a binary outcome, good discrimination indicates adequate distinction in the distributions of predicted values, based on the model, between the two classes, defined by the binary outcome. Calibration quantifies how closely the predicted values agree with the observed outcomes. Often, good calibration tends to correspond to good discrimination and *vice versa*; however, there are exceptions in which a model is strong in one measure and weak in the other [1]. For instance, a model that predicts all positive outcomes to occur with probability 0.51 and all negative outcomes to occur with probability 0.49 has perfect discrimination but bad calibration. If prediction is the goal, it is generally recommended that we should choose the model with good discrimination over the one with good calibration. If a predictive model has poor discrimination, no adjustment or calibration can correct the model. On the other hand, if discrimination is good, but calibration is poor, the model can be re-calibrated without sacrificing the discrimination.

The area under the ROC curve (AUC) is the commonly used measure to evaluate model discrimination by calculating the probability that among all possible pairs of individuals with two different outcomes, the predicted value for the one with positive outcome is higher than for the one with negative outcome. However, this index is a global measure. With numerical simulations, Pepe *et al.* [2] demonstrated the relation between association, measured in odds ratios, and classification, depicted by AUC. They concluded that statistical significance in association by itself does not characterize the discriminatory capacity of a marker, and that

*Correspondence to: Xiao-Hua Zhou, Department of Biostatistics, University of Washington, Seattle, WA 98101, U.S.A.

†E-mail: azhou@u.washington.edu

a meaningful AUC requires an association with a magnitude rarely seen in epidemiological studies.

The work by Pencina *et al.* [3] focuses on the evaluation of the added discrimination of a new marker. They proposed two interesting indices, the net reclassification improvement (NRI) and integrated discrimination improvement (IDI), to supplement the improvement in AUC, which may be too stringent to achieve. As demonstrated in the application for evaluating the incremental value of HDL cholesterol in heart disease risk models, both NRI and IDI were highly significant in suggesting a significant improvement in performance, while the change in AUC disagreed. It could be worthwhile to numerically quantify and compare the degree of separation in predicted values needed to reach a significant NRI/IDI and an improvement in AUC. One way to achieve this goal is to examine the relation between odds ratios and the NRI/IDI index in the way of Pepe *et al.* [2], and compare the magnitude of odds ratios required to obtain a same values of the NRI/IDI and AUC change.

The IDI is equivalent to an integrated difference in Youden's indices, which are defined for every possible cut-off value [4, 5]. For $c \in (0, 1)$, we have

$$\begin{aligned} Y(c) &= P(\hat{p} > c \mid D = 1) + P(\hat{p} > c \mid D = 0) - 1 \\ &= \text{sensitivity}(c) + \text{specificity}(c) - 1 \end{aligned}$$

where \hat{p} denotes the predicted probability. The IDI may suffer from the drawback of Youden's indices, which impose equal weight on sensitivity and specificity. The relative importance of sensitivity and specificity may vary with the scientific objectives. The integrated Youden's indices of the IDI implicitly assume equal weight across all cut-offs. A subject-matter utility function may help address the relative importance of different cut-off values. In comparison with the weights for sensitivity in the calculation of the AUC, the decline rate of the derivative of specificity shall depend on the functional form of specificity. The graph in Figure 3 of Pencina *et al.* [3] was just one of the many cases. The partially integrated difference in Youden's indices may be of a good use when some cut-offs are irrelevant.

Evaluating model performance solely from the data that are used to develop the model can be misleading [1, 6]. The most stringent test of a model is an external validation—using new data for the evaluation. In the absence of a second data set, other alternatives include cross-validation and bootstrapping. Cross-validation repeatedly splits data into the training (for model development) and test samples, while bootstrapping involves taking samples with replacement from the original data. It has been shown that bootstrapping provides nearly unbiased estimates of predictive accuracy that are of relatively low variance. A bootstrap validation would be useful to correctly determine the added predictive ability of a new marker, based on either the NRI or the IDI.

REFERENCES

1. Schmid CH, Griffith JL. Multivariate classification rules: calibration and discrimination. *Encyclopedia of Biostatistics*. Wiley: Chichester, U.K., 1998; 2844–2850.
2. Pepe MS, Janes H, Longton G, Leisenring W, Polly N. Limitations of the odds ratios in gauging the performance of a diagnostic, prognostic or screening marker. *American Journal of Epidemiology* 2004; **159**:882–890.
3. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* 2007; DOI: 10.1002/sim.2929.

COMMENTARY

4. Youden WJ. An index for rating diagnostic tests. *Cancer* 1950; **3**:32–35.
5. Hilden J, Glasziou P. Regret graphs, diagnostic uncertainty and Youden's index. *Statistics in Medicine* 1996; **15**:969–986.
6. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**:361–387.

DISCLAIMER

“The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs.”