The International Journal of Biostatistics

Volume 6, *Issue* 1 2010 *Article* 22

Comment: Measures to Summarize and Compare the Predictive Capacity of Markers

Nancy R. Cook*

*Harvard University, ncook@rics.bwh.harvard.edu

Comment: Measures to Summarize and Compare the Predictive Capacity of Markers*

Nancy R. Cook

Abstract

In their presentation on measures of predictive capacity Gu and Pepe say little about calibration. This comment distinguishes conditional and unconditional calibration and how these relate to the stated results.

*This work was supported by NHLBI BAA award contract number HHSN268200960011C.

The paper by Gu and Pepe (2009) provides a clear and comprehensive description of summary measures for evaluating predictive models. It is an excellent comparison of methods and how they relate. I have two comments, however, related to calibration of models.

First, what they call "Cook's reclassification percent" is not intended to serve as a measure of model accuracy. It merely describes differences in model classification without ascribing beneficial or detrimental effects. Models that have exactly the same overall predictive accuracy will exhibit some reclassification unless their predictions are identical. Indeed, even a model that performs worse would lead to reclassification but in the wrong direction. We have emphasized (Cook et al., 2006; Cook, 2007; Cook and Ridker, 2009) that it is not enough to describe reclassification but it is necessary to determine whether the new risk categories are more accurate. This led us to first heuristically (Cook et al., 2006) then more formally (Cook et al., 2009) compare the predicted risk in each cell of the reclassification table to the observed model-free risk obtained from either the proportion with disease or from a nonparametric survival curve such as the Kaplan-Meier curve for time to event data. A reclassification calibration statistic, similar to the Hosmer-Lemeshow statistic, is suggested for formal comparison.

Second, model calibration is not explicitly examined in the presentation. Gu and Pepe provide several results regarding metrics computed from the population distribution of risk. They do not, however, always distinguish the distribution of true risk from that predicted from a model. Gail and Pfeffer (2005) and others define a true underlying risk for prospective data that is inherent to the individual. While this may not be appropriate in diagnostic or retrospective data, prediction with prospective data is stochastic (Graf et al., 1999), and the distinction is relevant.

The concept of true risk has been entertained widely throughout the literature, including by Hilden et al. (1978), Spiegelhalter (1986), and Redelmeier et al. (1991). Let $\pi_i = Pr(Y_i = 1)$ represent the true probability of disease for the *i*th person. In the diagnostic setting where disease status is fixed but unknown (Graf et al., 1999; Cook, 2007), $\pi_i = 0$ or 1. In predicting risk for future events, this underlying probability is related to a stochastic event that one may or may not be able to determine even with all available information on covariates. Spiegelhalter (1986), and Redelmeier et al. (1991) used this concept to derive statistical tests for assessing the accuracy of models assuming that the observations Y_i are independent Bernoulli distributed random variables with probability π_i. It may not be possible to identify π_i at all with only baseline information. If predicting into the future, other intervening external events, for example, may influence an individual's risk.

While π_i is of ultimate interest, we typically only have access to predictions from a risk model based on covariates $X = x_i$, defined as $r(x)$. Gail and

Pfeiffer (2005) use the term "perfectly calibrated" to refer to the expectation conditional on the covariates x, such that a perfectly calibrated model has $r(x) = E(\pi \mid x)$. However, under the best of circumstances, we would like to be able to estimate π_i itself. Gail and Pfeiffer refer to this as a "perfect model" such that $r(X_i) = \pi_i$ for all i. The first definition could be thought of as a *conditional* calibration, given the covariates included in the model, while the latter definition could be called *unconditional* calibration, such that the model would estimate the true underlying probability rather than the conditional expectation. While the Brier score has several different decompositions, it can also be written as a function of these true probabilities. Gail and Pfeiffer write it as the following:

$$B = N^{-1}[\sum \{\pi_i - r(X_i)\}^2 + \sum (Y_i - \pi_i)^2].$$

Thus, the score can be decomposed into a term for a bias due to a lack of perfect unconditional calibration as well as variability in the outcome Y_i about the true π_i.

This distinction becomes most important when comparing two models, since only one, at most, can be perfectly calibrated in the unconditional sense. The reclassification calibration test compares models using either X alone or both X and Y within risk strata that could be considered important. It is not a test of whether r(X,Y) equals π_i, but rather a comparison of the two models, and whether r(X,Y) is closer to π_i than is r(X). The test for the smaller model r(X) is of primary interest, but we suggest the RC test be done for both the model with X only and the model with X and Y. If X and Y are both predictive, then the null hypothesis for the X,Y model should hold if the model is well-calibrated conditional on X and Y. That does not always occur in practice, however, especially if Y is not particularly predictive given X. The RC test should thus be considered for both models as a consequence. The RC test is not as sensitive as the test of association; it is intended to determine whether the effect of a predictor makes a substantial difference in risk prediction, such that individuals change within clinical risk strata or otherwise important categories. Note that the standard Hosmer-Lemeshow test is a test for conditional calibration using the marginal distribution of the predicted risk from a model. It has virtually no power against the alternative of an omitted variable (Hosmer and Hjort, 2002). The lack of (unconditional) calibration thus cannot be checked using the Hosmer-Lemeshow test, only that conditional on the variables included.

Even when they describe measures related to a single model, Gu and Pepe do not explicitly discuss calibration. Some of their results seem to rely on perfect calibration, at least in the conditional sense. For example, their proof of equation (6) assumes at least mean calibration, sometimes called calibration-in-the-large, if not perfect calibration in the tails of the risk distribution. In the extreme situation where a model assigns everyone a risk of 0 or 1, the predictiveness curve would

DOI: 10.2202/1557-4679.1257

be a step function and may look excellent. This would be misleading if the model is merely a random coin flip with no predictive power. Some formulations of the PEV and TG also seem predicated on perfect calibration since they are functions of the predicted risk only and not the observed outcomes. The second equation for PEV in Section 3.1 would show a large R^2, and the apparent total gain computed from equation (4) would be large in the coin flip example. Other formulations of both the PEV and total gain, such as the IDI formulation or the Kolmogorov-Smirnov distance between cases and controls, do not suffer from this problem, although some have argued that the IDI may not be a proper scoring rule (Hilden, personal communication). Alternatively, when the cumulative distributions for cases and controls are separated as in Figure 2, models can at least be compared via their discriminatory capacity. Gu and Pepe provide some very interesting results and are to be congratulated for their article, but their assumptions about calibration should be made more explicit.

References

Cook, N. R. (2007), "Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction," *Circulation*, 115, 928-935.

Cook, N. R., Buring, J. E., and Ridker, P. M. (2006), "The Effect of Including C-Reactive Protein in Cardiovascular Risk Prediction Models for Women," *Ann Intern Med*, 145, 21-29.

Cook, N. R., and Ridker, P. M. (2009), "Advances in Measuring the Effect of Individual Predictors of Cardiovascular Risk: The Role of Reclassification Measures," *Ann Intern Med*, 150, 795-802.

Gail, M. H., and Pfeiffer, R. M. (2005), "On Criteria for Evaluating Models of Absolute Risk," *Biostat*, 6, 227-239.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999), "Assessment and Comparison of Prognostic Classification Schemes for Survival Data," *Stat Med*, 18, 2529-2545.

Gu, W., and Pepe, M. (2009), "Measures to Summarize and Compare the Predictive Capacity of Biomarkers," *Int J Biostat*, 5.

Hilden, J., Habbema, D. F., and Bjerregaard, B. (1978), "The Measurement of Performance in Probabilistic Diagnosis III," *Meth Inform Med*, 17, 238-246.

Redelmeier, D. A., Bloch, D. A., and Hickam, D. H. (1991), "Assessing Predictive Accuracy: How to Compare Brier Scores," *J Clin Epidemiol*, 44, 1141-1146.

Spiegelhalter, D. J. (1986), "Probabilistic Prediction in Patient Management and Clincal Trials," *Stat Med*, 5, 421-433.

DOI: 10.2202/1557-4679.1257