STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2985

COMMENTARY

Comments on ‘Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond’ by M. J. Pencina *et al.*, *Statistics in Medicine* (this issue)

J. H. Ware*,† and T. Cai

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.

Although the C statistic is the most widely used measure of the performance of prediction models for cardiovascular disease, there is increasing concern that change in the C statistic may be an insensitive measure of the improvement in risk prediction when an additional risk factor is added to the model. Ridker *et al.* [1] have proposed that, rather than discrimination between those who will and will not have the event, the goal of risk prediction in the prospective setting should be to classify patients into risk strata requiring different management strategies. In measuring the incremental value of new risk factors for the prediction of coronary heart disease, they assess the extent to which individuals are reclassified by the extended model, and then evaluate whether the observed risk in groups of reclassified individuals is consistent with the predicted risk.

Although this approach has appeal, no global measure had been proposed to quantify and test the statistical significance of the improvement in risk classification achieved by adding a risk factor to a prediction model. Pencina *et al.* [2] now propose two measures that address this question, the ‘net reclassification improvement’ (NRI) and the ‘integrated discrimination improvement’ (IDI). As is also true of the C statistic, these measures do not depend on the prevalence of the outcome.

The NRI uses binary scoring to assess the degree to which an expanded risk prediction model ‘improves’ risk classification. For individuals subsequently observed to have the event, classification is improved if the individual is moved to a higher risk stratum and worsened if an individual is moved to a lower risk stratum. Similarly, for individuals who do not have the event, risk classification is improved if the individual is assigned to a lower risk stratum and worsened if the individual is moved to a higher risk stratum. With this scoring algorithm, the optimal rule would assign all those who subsequently have the event to the highest risk stratum and all those who do not have it to the lowest risk stratum. Thus, the NRI is based on improvement in discrimination

*Correspondence to: J. H. Ware, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.
†E-mail: ware@hsph.harvard.edu

Copyright © 2007 John Wiley & Sons, Ltd.

between those who will and will not have the event, not on accuracy of risk stratification. This is even more apparent in the definition of the IDI. Maximal improvement would be achieved by a rule that assigned a predicted risk of 1 to all those who had the event and 0 to those who did not. Thus, both measures assess binary discrimination, not risk classification. This is perhaps not surprising since membership in a risk stratum is not observable.

The notion of perfect discrimination suggests a variant of the NRI, namely that an individual would receive a numeric score corresponding to the number of categories of improvement in risk class achieved by the expanded model. For example, if individuals are assigned to four risk strata, then an individual who was reassigned from stratum 1 to stratum 4 (highest risk) by the extended rule and subsequently experienced the event would contribute +3 to the modified NRI. Scoring based on a difference calculated from the midpoints of the risk categories would be more closely related to the second measure, the IDI. To illustrate the relation, suppose that we define 100 risk categories; 0–1 per cent, 1–2 per cent, and so on. Then, assign a score to each individual corresponding to the number of categories the risk classification was increased or decreased by addition of a new covariate to the model. It is apparent that this variant of the NRI is a discrete version of the IDI.

Pencina *et al*. show that the IDI is based on a scalar measure of the performance of a risk prediction model, namely, the difference between the integrated sensitivity and the integral of 1 minus specificity. Equivalently, the metric is the difference in the mean predicted risk for individuals who do and do not develop the event. It may also be interpreted as a weighted area between the receiver operating characteristic (ROC) curve and the diagonal line. Pepe *et al.* [3] represent this quantity graphically as the region between the curves of sensitivity and 1 minus specificity. Their graphical representation demonstrates the intuitive appeal of the measure.

Although it has been suggested that the C statistic is insensitive to improvements in risk prediction, information about this issue has been limited. Pencina *et al*. found in their example that the C statistic is less sensitive to improvements in risk prediction than the IDI. One would expect the C statistic to be less sensitive than the likelihood ratio test, since the latter is optimal when the risk prediction model is correctly specified. Eguchi and Copas [4] show that the likelihood ratio statistic is directly related to the weighted area between the ROC curves of the two risk prediction models. Thus, all three statistics, the IDI, the C statistic, and the likelihood ratio statistic, may be interpreted as empirical versions of weighted areas between the ROC curves with different weight functions. It is not yet clear whether the IDI will dominate the C statistic in all settings. We hope that this issue can be clarified.

Finally, we note that these measures do not fully address the more difficult problem of assessing whether the improvement in risk prediction achieved by an additional risk factor justifies its clinical use. This would require consideration of the improvement in risk prediction achieved with a new risk factor, the effect of risk stratification and the resulting changes in patient management on outcome, the burden to the patient of risk factor assessment, and cost. It will be difficult to develop a unified method for considering these issues. In the meantime, improved methods for assessing the utility of risk factors are a step forward.

We thank the authors for their insights about this important problem.

REFERENCES

1. Ridker PM, Buring JE, Rifai N, Cook N. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women. *Journal of American Medical Association* 2007; **297**:611–619.

Copyright © 2007 John Wiley & Sons, Ltd.

2. 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*, in this issue. DOI: 10.1002/sim.2929.
3. Pepe MS, Feng Z, Huang Y, Longton GM, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. *UW Biostatistics Working Paper Series* #289, 2006. Available at: http://www.bepress.com/uwbiostat/paper289.
4. Eguchi S, Copas J. A class of logistic-type discriminant functions. *Biometrika* 2002; **89**:1–22.

Copyright © 2007 John Wiley & Sons, Ltd.