

REJOINDER

Comments on 'Integrated discrimination and net reclassification improvements—Practical advice'

We begin with thanking the authors for their insightful critiques. We believe that our original paper [1] and the seven commentaries [2–8] form a comprehensive survey of statistical (and beyond) issues underlying the problem of evaluating the usefulness of a new marker in risk prediction models. Motivated by the questions and opinions raised in the commentaries we undertook further evaluation of the proposed methods and would like to share the results. We will also take the opportunity to comment on some of the issues highlighted in the critiques.

We agree with Pepe *et al.* [2] that the relationship between the integrated discrimination improvement (IDI) and logistic regression R-square as described by Hu *et al.* [9] is interesting, compelling and re-assuring. It is important to note, however, that in practice this relationship requires two important assumptions. The first one, pointed out by the authors, is that the model on which the predicted probabilities of event are based possesses what we call *perfect basic calibration*, which means that the average of all predicted probabilities of event is equal to the event incidence in the sample on which we evaluate the performance. The second, even more restrictive assumption is that the predicted probabilities of event based on both samples (development and validation) are the same. Both of these are satisfied automatically if the model is evaluated on the same sample on which it is developed (that was the case with our high-density lipoprotein (HDL) example). Large samples and large numbers of events are another way to ascertain that the IDIs estimated as differences in discrimination slopes and logistic R-squares will be close. If the necessary assumptions are met, Pepe *et al.* [2] show that the following hypotheses are equivalent: (1) IDI different than zero; (2) change in logistic R-square different than zero; (3) regression coefficient in the logistic model different than zero. Hence, in this case, the IDI can be seen as the most natural measure of model performance, mirroring the well-known relationship from linear regression.

As correctly pointed out by Chi and Zhou [3] the actual improvement in performance should be established based on a validation and not on a development set. If a validation set is not available, bootstrapping the original sample might provide a sense of potential over-optimism. Either approach might create a calibration problem. We point out in the original paper that IDI (and by extension net reclassification improvement (NRI) as well) depends on model calibration. To illustrate the issue, imagine an extreme situation in which the incidence in the development sample is twice the incidence in the validation sample but all other relationships remain the same. In this case the predicted probabilities of event will be twice what they should be and hence the IDI will be doubled. The problem is even more pronounced for NRI where differences in incidence much less extreme than doubling might lead to higher or lower values. That is why it is important to ascertain that incidence rates in the development and validation sets are similar. If this is not the case, some form of recalibration should be considered.

Following the recommendation of Chi and Zhou [3] and Pepe *et al.* [2] we constructed 95 per cent confidence intervals for IDI and NRI based on 999 bootstrap samples and compared them

with intervals obtained based on the development sample using asymptotic methods outlined in the original paper. The results are as follows:

	IDI	NRI
Original	0.0085 (0.0016, 0.0154)	12.05 (5.52, 18.59)
Bootstrap	0.0086 (0.0009, 0.0148)	12.01 (5.81, 18.33)

The bootstrap results are very close to those obtained using asymptotic methods. Interestingly, the point estimates do not seem to indicate the presence of over-optimism. This might be partially explained by the fact that IDI (and also NRI) can be seen as a difference of differences in which any over-optimism cancels out. Moreover, the bootstrap approach might be an imperfect substitute for a formal validation on a separate sample.

We note that the above-mentioned calibration issue does not apply to the area under the curve (AUC) or improvement in the AUC (IAUC) as these are scale invariant and independent from model calibration. The difference in metrics between IAUC and IDI can be seen in the following representation, which was motivated by Ware and Cai's [4] question of dominance between these two measures. Denote by X_1 the random variable describing the predicted probabilities of event based on the model without the new marker for those who experience events and by X_0 the corresponding quantity for non-events. Further, denote by Y_1 and Y_0 the same random variables based on the model with the new marker. The IAUC and IDI are given by

$$\text{IAUC} = P(Y_1 > Y_0) - P(X_1 > X_0) = P(Y_1 - Y_0 > 0) - P(X_1 - X_0 > 0) \quad (1)$$

$$\text{IDI} = (EY_1 - EY_0) - (EX_1 - EX_0) = E(Y_1 - Y_0) - E(X_1 - X_0) \quad (2)$$

The differences $Y_1 - Y_0$ and $X_1 - X_0$ can be seen as random variables indicating the spread in probabilities between events and non-events and hence we can interpret the IAUC as a difference in probabilities of these spreads being positive and IDI as a difference in expected value between these spreads. This representation allows us to see the difference in metrics between the IAUC and IDI, shows how the former is and the latter is not invariant to uniform transformations of scale and enables us to answer the question of dominance raised by Ware and Cai [4]. It turns out that neither is the IDI always above the IAUC nor *vice versa*. If X_1 and X_0 come from a completely random model and the incidence of the condition is, say, 0.2, then $P(X_1 - X_0 > 0) = 0.5$ and $E(X_1 - X_0) = 0$. If the new marker spreads all the predicted probabilities of event to 0.24 for events and 0.19 for non-events we have $E(Y_1 - Y_0) = 0.05$ but at the same time $P(Y_1 - Y_0 > 0) = 1$. Thus, $\text{IAUC} = 0.5$ while $\text{IDI} = 0.05$ giving a case in which $\text{IAUC} \gg \text{IDI}$. But the opposite is not only possible but also more likely in practical applications. As an example, consider a model with only one but very strong predictor, yielding $\text{AUC} = 0.827$. Consider the candidate marker to be weaker but also quite strong, producing an AUC of 0.665 on its own. Adding this new marker to our model raises the AUC to 0.857, yielding IAUC of only 0.030. The IDI is much more pronounced as it increases the discrimination slope from 0.240 to 0.311, giving $\text{IDI} = 0.071$. Considered on a relatively increased scale the differences are even more visible.

The problem of scale of the IDI has been raised by both P. Greenland [5] and Cook [6]. It is true that the magnitude of discrimination slopes and their differences will tend to be small (0.009 in our example). Formula (2) provides a good insight on why this must be true, especially when

the incidence of the condition of interest is relatively low. Moreover, the invariance to uniform transformations of scale possessed by AUC and IAUC is a desirable property. Hence, it appears more meaningful to consider relative IDI instead of absolute IDI. Using the same notation as above we define it as

$$\text{Relative IDI} = \frac{EY_1 - EY_0}{EX_1 - EX_0} - 1 \quad (3)$$

In our example the relative IDI is estimated as 13.52 per cent with bootstrap-based 95 per cent confidence interval (1.66, 25.54 per cent). The advantage of the relative increase is that it is scale free; hence, it will not be influenced by basic calibration. Expressed as a percentage, it still remains meaningful.

Calculating the IDI in cases similar to ours, one might want to apply special treatment to age. Many correctly argue that one could or should treat age as a timescale [10, 11] and thus including age as a candidate risk factor might unnecessarily inflate the denominator in the calculation of relative IDI. In our example, the discrimination slope for age is 0.0230 and the IDI offered by the addition of all other variables except HDL to the model with only age is 0.0400. Using the last number as the denominator in the calculation of relative IDI offered by the addition of HDL to the model with all other variables we obtain the relative IDI of 21 per cent. Given five risk factors in the model (sex, diabetes, smoking, systolic blood pressure and total cholesterol), we see how HDL cholesterol alone offers an increase comparable with the average of the five other variables.

Both Kraemer [7] and S. Greenland [8] stress the importance of false positives and false negatives in the selection and evaluation of new markers. Kraemer suggests a method based on the kappa statistic [12] and S. Greenland stresses the importance of cost considerations. We agree that both approaches provide valuable tools that can offer additional insights; however, we would not advocate a full replacement of the more 'traditional' methods. Cut-off-based algorithms (like the approach proposed by Kraemer) are yet to be proven to be transportable and sole reliance on cost would be too subjective given the technological and scientific progress, leading to a steady decrease in costs of many medical therapies.

It is of interest to note that under certain assumptions (properly calibrated model and costs of true positives and true negatives being zero) the integrand of the IDI, i.e. the difference in Youden's indices can be expressed in a form similar to the expected cost equation given by S. Greenland [8]. Denoting the probability of event by p , numbers of false negatives for a given cut-off u for the model with (new) and without (old) the new marker by $b_{\text{new}}(u)$ and $b_{\text{old}}(u)$, respectively, and the corresponding numbers of false positives by $c_{\text{new}}(u)$ and $c_{\text{old}}(u)$, we can write the integrand of the IDI as

$$Y_{\text{new}}(u) - Y_{\text{old}}(u) = \frac{1}{p}(b_{\text{old}}(u) - b_{\text{new}}(u)) + \frac{1}{1-p}(c_{\text{old}}(u) - c_{\text{new}}(u)) \quad (4)$$

The reciprocals of event and non-event rates serve as costs for false negatives and false positives. We agree with S. Greenland [8] that it would be only by chance that these would correspond exactly to the actual costs but they do form 'reasonable guesses'. These, unlike actual costs, change only with changing incidence of the condition of interest—the rarer the disease becomes the more focus on false negatives *versus* false positives. It is not difficult to show that formula (4) can also be viewed as a difference in kappa statistics as given by Kraemer [12] and calculated at the $1-p$ point, again indicating much more weight given to false negatives for small p .

In their commentaries, Kraemer [7] and Cook [6] raise the possibility that given the predicted probabilities based on a model without the new marker, we might focus only on a subset of participants, possibly those in the middle of the risk range. Such an approach might be very helpful in reducing unnecessary burden and cost. However, it needs to be stressed that it can be implemented only if dictated by an *a priori* design of a method. Kraemer [7] suggests such a method and we apply it to our example. On the other hand, *post hoc* approaches based on models that utilize new marker information on all individuals but then focus only on their subset for model evaluation can be misleading. As correctly pointed out by Pepe *et al.* [2], people tend to leave a given risk category at a similar rate at which other people move into this very same category. If a given person with a given value of the new marker contributed to model construction, they should also be taken into account in model evaluation.

Following Kraemer's [7] suggestion we calculated predicted probabilities of event based on the model without the new marker and used them together with HDL cholesterol to determine the risk-based subset on which HDL should be measured. We used maximization of kappa at 0.90 as our criterion (incidence rate in our sample was about 6 per cent; kappa at 0.90 was the largest value available in our software package) [cf. 12]. The algorithm divided the predicted probabilities into three categories: 0–5, 5–8 and > 8 per cent and HDL entered the model (was found to be informative beyond the risk based on a model without HDL) for all of them. The HDL cut-offs were 30 mg/dL in the 0–5 per cent group, 44 mg/dL in the 5–8 per cent group and 41 mg/dL in the > 8 per cent group. This result indicates that HDL should be measured regardless of the risk level calculated on other predictors. Even though we and Kraemer [7] set out to answer different questions, the conclusion is the same: HDL cholesterol is an important predictor of coronary heart disease (CHD) risk. If the judgment was to be made based on our sample, Kraemer's method would provide additional reassurance that it is useful to measure HDL on everyone.

Our discussion focused mainly on IDI but several of our points apply to NRI. As mentioned earlier NRI depends on model calibration. It can also be heavily influenced by the number and extent of the risk categories selected. It is easy to construct extreme cases in which negligible NRI will be observed despite employment of meaningful variables in the model (take for example < 20, 20–40, > 40 per cent as risk categories when incidence is only 5 per cent). It is crucial that the risk categories used to build NRI correspond to the incidence rate of the condition. For example, < 6, 6–20 and > 20 per cent might be meaningful for CHD but not necessarily for cardiovascular disease as it has a higher incidence. Cook's [6] application of NRI to the Women's Health Study data reveals that NRI is not 'automatically' significant on large data sets. In her example NRI does not achieve statistical significance on the validation set, suggesting that more is needed than just a large sample size. These and other properties of NRI should be explored further; for example, it would be helpful to quantify the effect of the number and choice of risk categories and pursue some of the modifications suggested by Pepe *et al.* [2] and Ware and Cai [4].

We agree with Cook [6] that calibration should be an important criterion when evaluating model performance. However, when calculated on our data, the modified Hosmer–Lemeshow [13] chi-square statistic decreased only from 12.73 for the model without HDL to 11.01 for the model with HDL. Since calibration chi-square does not need to be a monotone function of the number of variables and because models with and without the new marker should have the same basic calibration, no substantial improvements in calibration chi-square can be expected even with the addition of powerful markers. Similar to the *c* statistic, calibration chi-square is a necessary component of model performance evaluation but it may not be the most informative in assessing the utility of a new marker. To further illustrate this point we go back to the example of calculating

REJOINDER

the added utility of adding additional risk factors to a model with only age. The c statistic increases from 0.677 to 0.774, a 14 per cent relative increase or 55 per cent if calculated from 0.500. The change in calibration chi-square is even more disappointing—it decreases from 13.83 to 11.01. On the other hand, the discrimination slope increases from 0.023 to 0.071, yielding a relative IDI of 207 per cent. The NRI reaches 31.6 per cent. While this example does not constitute a proof and simulation studies are needed to better understand the behavior of both IDI and NRI in various scenarios, it does provide an illustration of the limitations of the current performance measures when applied to judging the utility of new markers. Valid, useful measures capable of evaluating the contribution of a new variable are needed. We believe that our work contributes useful insights into this important field.

MICHAEL J. PENCINA and RALPH B. D'AGOSTINO Sr
Department of Mathematics and Statistics
Framingham Heart Study, Boston University
111 Cummington St., Boston, MA 02215, U.S.A.

RALPH B. D'AGOSTINO Jr
Department of Biostatistical Sciences
Wake Forest University School of Medicine
Medical Center Boulevard, Winston-Salem, NC 27157, U.S.A.

RAMACHANDRA S. VASAN
Framingham Heart Study
Boston University School of Medicine
73 Mount Wayte Avenue, Suite 2
Framingham, MA 01702-5803, U.S.A.

REFERENCES

1. 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.
2. Pepe MS, Feng Z, Gu JW. Commentary on '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.2991.
3. Chi YY, Zhou XH. Commentary on 'Evaluating the added predictive ability of a new marker'. *Statistics in Medicine* 2007; DOI: 10.1002/sim.2986.
4. Ware JH, Cai T. Commentary on '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.2985.
5. Greenland P. Commentary on '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.2976.
6. Cook NR. Commentary on 'Evaluating the added predictive ability of a new marker'. *Statistics in Medicine* 2007; DOI: 10.1002/sim.2987.
7. Kraemer HC. Commentary on 'Evaluating the added predictive ability of a new marker'. *Statistics in Medicine* 2007; DOI: 10.1002/sim.2948.
8. Greenland S. Evaluating the added predictive ability of a new marker: the need for reorientation toward cost-effective prediction. *Statistics in Medicine* 2007; DOI: 10.1002/sim.2995.
9. Hu B, Palta M, Shao S. Properties of R^2 statistics for logistic regression. *Statistics in Medicine* 2006; **25**: 1383–1395.
10. Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of time scale. *American Journal of Epidemiology* 1997; **145**(1):72–80.

REJOINDER

11. Pencina MJ, Larson MG, D'Agostino RB. Choice of time scale and its effect on significance of predictors in longitudinal studies. *Statistics in Medicine* 2007; **26**:1343–1359.
12. Kraemer H. *Evaluating Medical Tests: Objective and Quantitative Guidelines*. Sage Publications: Newbury Park, CA, 1992.
13. D'Agostino RB, Nam BH. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of Statistics* 2004; **23**:1–25.