American Journal of Epidemiology
© The Author 2012. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 176, No. 6
DOI: 10.1093/aje/kws210
Advance Access publication:
August 8, 2012

Invited Commentary

Further Insight Into the Incremental Value of New Markers: The Interpretation of Performance Measures and the Importance of Clinical Context

Kathleen F. Kerr*, Aasthaa Bansal, and Margaret S. Pepe

* Correspondence to Dr. Kathleen F. Kerr, Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195 (e-mail: katiek@u.washington.edu).

Initially submitted November 4, 2011; accepted for publication February 10, 2012.

In this issue of the *Journal*, Pencina and et al. (*Am J Epidemiol*. 2012;176(6):492–494) examine the operating characteristics of measures of incremental value. Their goal is to provide benchmarks for the measures that can help identify the most promising markers among multiple candidates. They consider a setting in which new predictors are conditionally independent of established predictors. In the present article, the authors consider more general settings. Their results indicate that some of the conclusions made by Pencina et al. are limited to the specific scenarios the authors considered. For example, Pencina et al. observed that continuous net reclassification improvement was invariant to the strength of the baseline model, but the authors of the present study show this invariance does not hold generally. Further, they disagree with the suggestion that such invariance would be desirable for a measure of incremental value. They also do not see evidence to support the claim that the measures provide complementary information. In addition, they show that correlation with baseline predictors can lead to much bigger gains in performance than the conditional independence scenario studied by Pencina et al. Finally, the authors note that the motivation of providing benchmarks actually reinforces previous observations that the problem with these measures is they do not have useful clinical interpretations. If they did, researchers could use the measures directly and benchmarks would not be needed.

area under curve; biomarkers; bivariate binomial distribution; receiver operating characteristic; risk assessment; risk factors

Abbreviations: AUC, area under the receiver operating characteristic curve; ΔAUC, change in area under the receiver operating characteristic curve; IDI, integrated discrimination improvement; NRI(>0), continuous net reclassification improvement.

Pencina et al. present an interesting study of the behavior of measures of incremental value (1). They pay particular attention to the change in the area under the receiver operating characteristic curve (ΔAUC), integrated discrimination improvement (IDI) (2), and continuous net reclassification improvement (NRI(>0)) (3). ΔAUC is a classic measure that might reasonably be described as one that "everyone uses and no one likes." IDI and NRI(>0) are newer measures thought to be more sensitive than ΔAUC. The NRI(>0) and its category-based variants have recently become very popular (4).

As previously noted, none of these measures has a useful clinical interpretation (5–8). This is arguably less of a problem for ΔAUC because people have some experience with it and some ability to judge whether a particular value of ΔAUC is large in specific clinical applications based on this experience. This is not true for IDI and NRI(>0). The primary goal of the article by Pencina et al. was to "derive heuristic benchmarks for small, medium, and large incremental values" (1, p. 473) for these newer measures. The context is early biomarker development, a context in which investigators may wish to perform an initial screening to identify the most promising new markers for further evaluation. In our opinion, however, the fundamental problem with measures of incremental value, including ΔAUC, IDI, and NRI(>0), is that they are not clinically meaningful. If they were, we would not need benchmarks because we would have a meaningful scale on which to judge whether an observed value of a prediction measure was large or small. Moreover, one must consider the population and the intended clinical application when assessing incremental value. Different applications of a risk model will typically have different costs and benefits. A new marker may have adequate value for one application but not another. Gail and Pfeiffer (9) gave a nice example of this phenomenon. Therefore, benchmarks for small, medium, and large incremental value should depend on the intended clinical application for the risk model.

The main results of the article by Pencina et al. were calculations and simulation studies for situations in which "new" predictors with marginal small, medium, or large effect sizes were added to predictive models. Tables 2 and 3 in their article nicely summarized the results. If we look across a row for ΔAUC, we see that this measure indicates less incremental value for a conditionally independent new marker with fixed marginal effect size as the strength of the baseline model increases. At the other extreme, NRI(>0) is constant within a row. In the scenarios the authors considered, NRI(>0) depended only on the marginal strength of the new predictor and did not vary with the strength of the baseline model. However, this invariance was particular to the situation the authors considered, in which *Y* was conditionally independent of *X*. This can be seen using the formula for NRI(>0) in the authors' Web Appendix 3, in which a little algebra shows that NRI(>0) reduces to a function of μ_Y when (and only when) ρ, the correlation between *X* and *Y* conditional on event status, is 0. When $\rho \neq 0$, NRI(>0) varies with the strength of the baseline model; we show this explicitly later. On the other hand, it is not clear why such invariance would be desirable, as the authors seemed to suggest. An index that merely reflects marginal strength and, in particular, does not depend on the baseline model is only a measure of marginal strength, not a measure of incremental value. We posit that baseline strength is an important consideration. For example, if the baseline model is almost perfect, incremental value cannot be large and any statistical measure of incremental value should be small.

One conclusion of Pencina et al. is that ΔAUC, IDI, and NRI(>0) offer complementary information, and the authors recommended reporting all 3 in studies of new markers. However, this conclusion was belied by the authors' own example, in which there was complete agreement among the 3 measures as to which of 2 candidate predictors is more promising for improving prediction in a cardiovascular disease setting. We do not think the authors have justified their conclusion that the measures are complementary or that it is helpful to report all 3.

We present some additional results that generalize the scenarios studied by Pencina et al. to illustrate our points. Similar to Pencina et al., we considered normally distributed predictors. We used the bivariate normal equal correlation

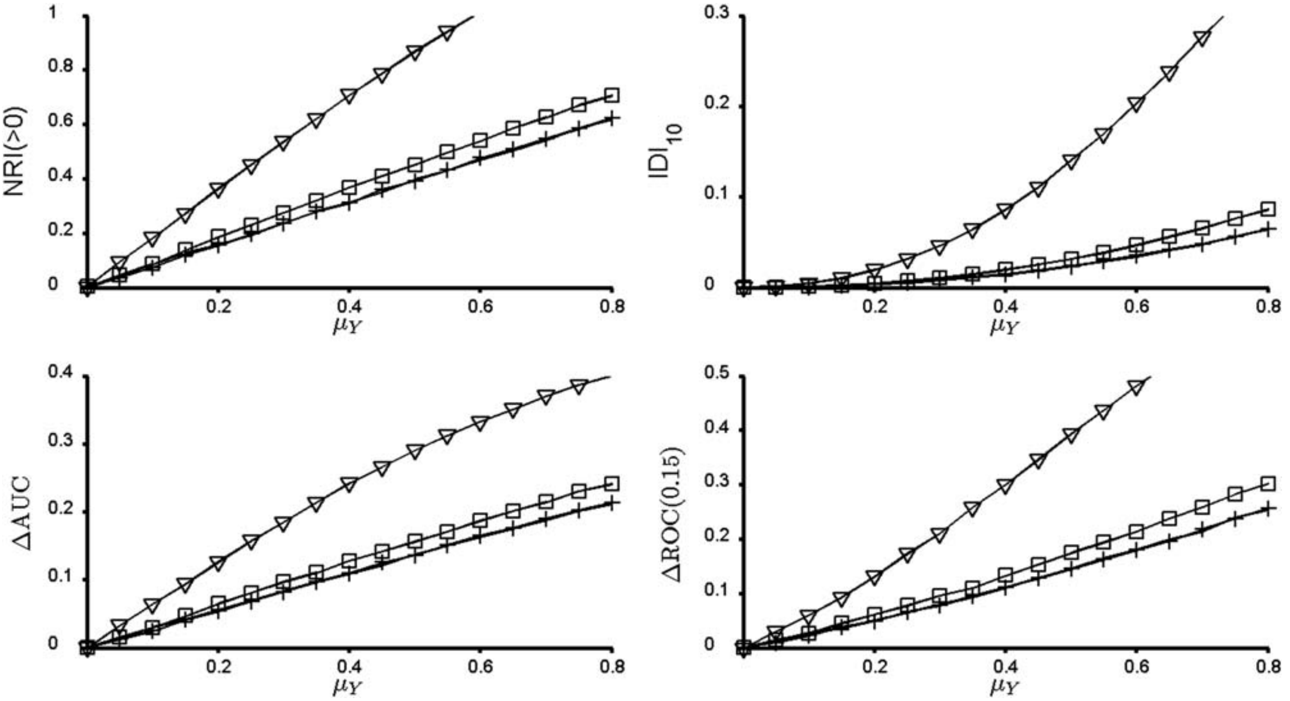

Figure 1. Measures of incremental value for the bivariate normal model with $\mu_X = 0$ and $AUC_X = 0.5$. All 4 measures of incremental value agree that for a given effect size μ_Y, the most promising new marker has high correlation with the existing marker *X*. Solid line, conditional correlation $\rho = 0$; +, $\rho = 0.1$; □, $\rho = 0.5$; ∇, $\rho = 0.9$. NRI(>0), continuous net reclassification index; IDI_{10}, integrated discrimination improvement index for 10% event rate; ΔAUC, change in the area under the receiver operating characteristic curve; ΔROC(0.15), change in sensitivity at 15% false positive rate.

model from reference 10. Although the model is limited in its generalizability, it extends the model considered by Pencina et al. by allowing for correlation between the existing marker X and the new marker Y. The distribution of the existing marker X and the new marker Y is

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \text{ in controls (nonevents),}$$

and

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{MVN}\left(\begin{pmatrix} \mu x \\ \mu y \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \text{ in cases (events).}$$

In this model, the marginal strength of the new predictor Y is completely captured by μ_Y, and the strength of the baseline model (which uses only X) is completely captured by μ_X. Note that $\mu_Y = \log(\text{odds}(\text{event}|Y = y + 1)/\text{odds}(\text{event}|Y = y))$, with a similar equation for μ_X. The incremental value of Y depends on both μ_Y and ρ. Pencina et al. only considered the case $\rho = 0$, wherein they refer to X and Y as "independent." However, when $\rho = 0$, there is only conditional independence between X and Y (conditional on event or nonevent status). That is, the new marker Y is independent of the existing marker X among controls (nonevents) and similarly among cases (events). However, in the population as a whole, X and Y are correlated because of their associations with the event outcome.

Figures 1–3 illustrate the addition of new predictors with different values of μ_Y and ρ. In each figure, the strength of the baseline marker μ_X is fixed. By displaying results this way, we emulated the context considered by the authors, in which some baseline model exists and investigators wish to identify the most promising new markers from a set of candidates. The only difference among Figures 1–3 is the strength of the baseline model. Consider Figure 1, in which X is actually a useless predictor on its own. Suppose we had 4 candidate markers with the same marginal strength (μ_Y), differing only in their correlations with X. For any given value of μ_Y, all 4 metrics presented agree on which of these markers has the highest incremental value. As Pencina et al. remarked, this is to be expected in a model like the one used here. However, the point remains: It is not clear that these metrics are complementary in any sense. The same observation holds for Figure 2, in which the baseline model has an AUC of 0.7, and Figure 3, in which the baseline model has an AUC of 0.9. The figures suggest, and it turns out to be true, that the value of μ_Y where any 2 curves intersect is the same for all 4 metrics in the figures. This implies that for any given value of μ_Y, all 4 metrics agree on which marker has the highest incremental value. In fact, using the formulas in Web Appendix 3 of the article by Pencina et al., we can make a stronger statement. Let Y_1 and Y_2 be 2 candidate predictors in the bivariate

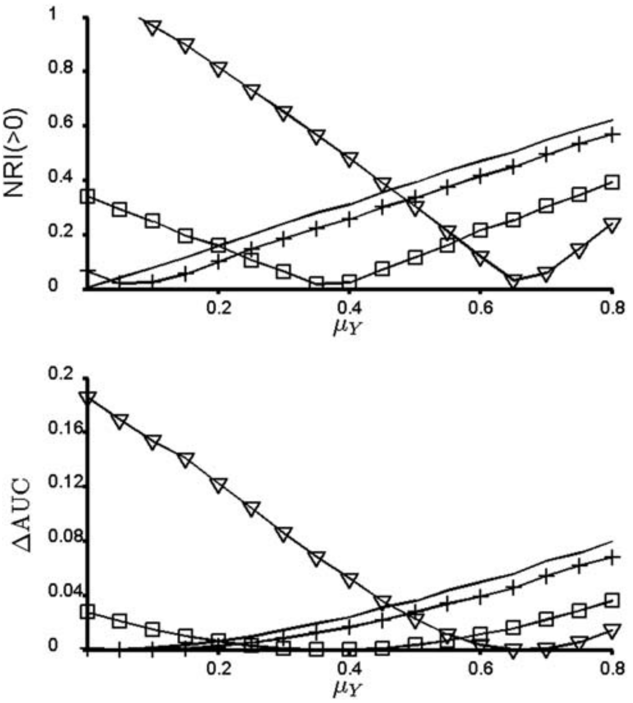

Figure 2. Measures of incremental value for the bivariate normal model with $\mu_X = 0.742$ and $\text{AUC}_X = 0.7$. For $\rho > 0$ there is a nonmonotone relation between the incremental value of a new marker Y and its marginal effect size μ_X. Solid line, conditional correlation $\rho = 0$; +, $\rho = 0.1$; □, $\rho = 0.5$; ∇, $\rho = 0.9$. NRI(>0), continuous net reclassification index; IDI$_{10}$, integrated discrimination improvement index for 10% event rate; ΔAUC, change in the area under the curve; ΔROC(0.15), change in sensitivity at 15% false positive rate.

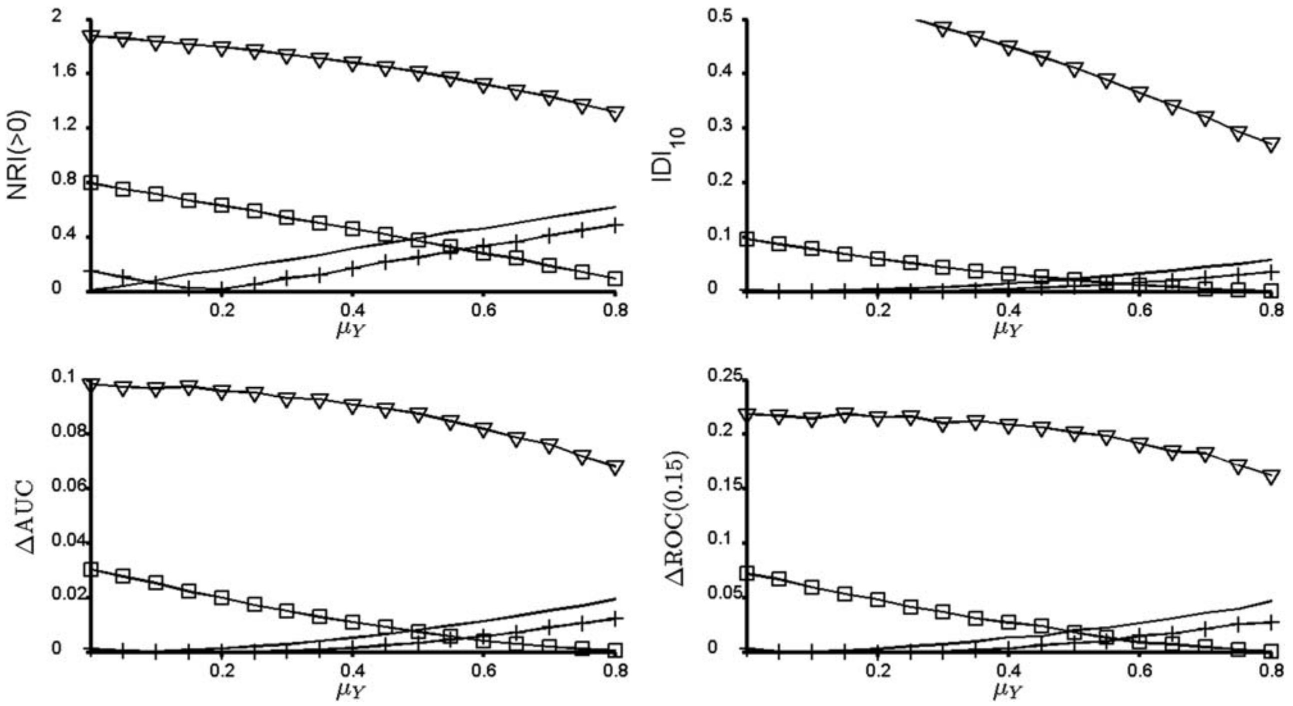

Figure 3. Measures of incremental value for the bivariate normal model with $\mu_X = 1.812$ and $\text{AUC}_X = 0.9$. Solid line, conditional correlation $\rho = 0$; +, $\rho = 0.1$; □, $\rho = 0.5$; ∇, $\rho = 0.9$. NRI(>0), continuous net reclassification index; IDI_{10}, integrated discrimination improvement index for 10% event rate; ΔAUC, change in the area under the curve; ΔROC(0.15), change in sensitivity at 15% false positive rate.

normal model with parameters μ_1 and ρ_1 for Y_1 and μ_2 and ρ_2 for Y_2. Define M_1^2 as the squared Mahalanobis distance for the model that includes Y_1 and M_2^2 as the squared Mahalanobis distance for the model that includes Y_2. Then $M_1^2 > M_2^2$. $M_1^2 > M_2^2$. In other words, ΔAUC_1 and NRI(>0) will always agree on which candidate new marker is preferred. Therefore, there is a sense in which the differences among the metrics can be considered a matter of scaling in these scenarios as far as judging the relative promise of a new biomarker. Although we considered more general data structures than the authors, we have not shown that the measures will always give the same rank order to new markers. Still, we are not convinced that it is useful to consider all 3 measures, and we don't understand the sense in which they offer complementary information.

Figures 1–3 bring up a few additional important points. In Figure 1, X is marginally useless for prediction. Perhaps counterintuitively, X can become useful when combined with a correlated Y, as pointed out recently in references 10 and 11, although this is not a new discovery (12). Interestingly, across all values of μ_Y, the best prediction is achieved when Y is highly correlated with X conditional on event status. However, Pencina et al. comment that the NRI(>0) "captured the impact of correlation and penalized those markers that might be strongly associated with the outcome but also correlated with variables already in the model" (1, p. 473). This statement appears to reinforce a common misconception that it is ideal for a new predictor to be uncorrelated with existing predictors. Unfortunately, this misconception leads some researchers to pursue a potential new marker Y only if the correlation between Y and an existing predictor X is small. The top row of Figure 4 shows that markers that are correlated with markers already in the model can be some of the most promising new predictors and should not be "penalized" for their correlation. The bottom row of Figure 4 displays the same data in a different way and illustrates our earlier point that NRI(>0) is not invariant to the strength of the baseline model when $\rho \neq 0$.

Figure 3 and especially Figure 2 illustrate another important, and perhaps counterintuitive, fact that is not evident from the article by Pencina et al.: There can be a nonmonotone relation between a new marker's incremental value and its marginal strength. This reminds us that the notions of incremental value and marginal strength are distinct concepts.

The framework of the study by Pencina et al. is that we can learn to interpret new measures, such as IDI and NRI(>0), by relating their magnitudes to the magnitudes of conditionally independent predictors of various marginal effect sizes. The strategy that is implied is that investigators should relate an observed value of NRI(>0) (or IDI, etc.) to the solid curves in Figures 1–3. For example, if the baseline model has an AUC ≈ 0.7 and the NRI(>0) is 0.40, one can consider that Y is like a conditionally independent predictor with effect size $\mu_Y \approx 0.5$. But why is the solid curve the appropriate reference? Clearly, it is much easier to understand marginal predictive strength than incremental value, and it is convenient when there is a monotone relation between

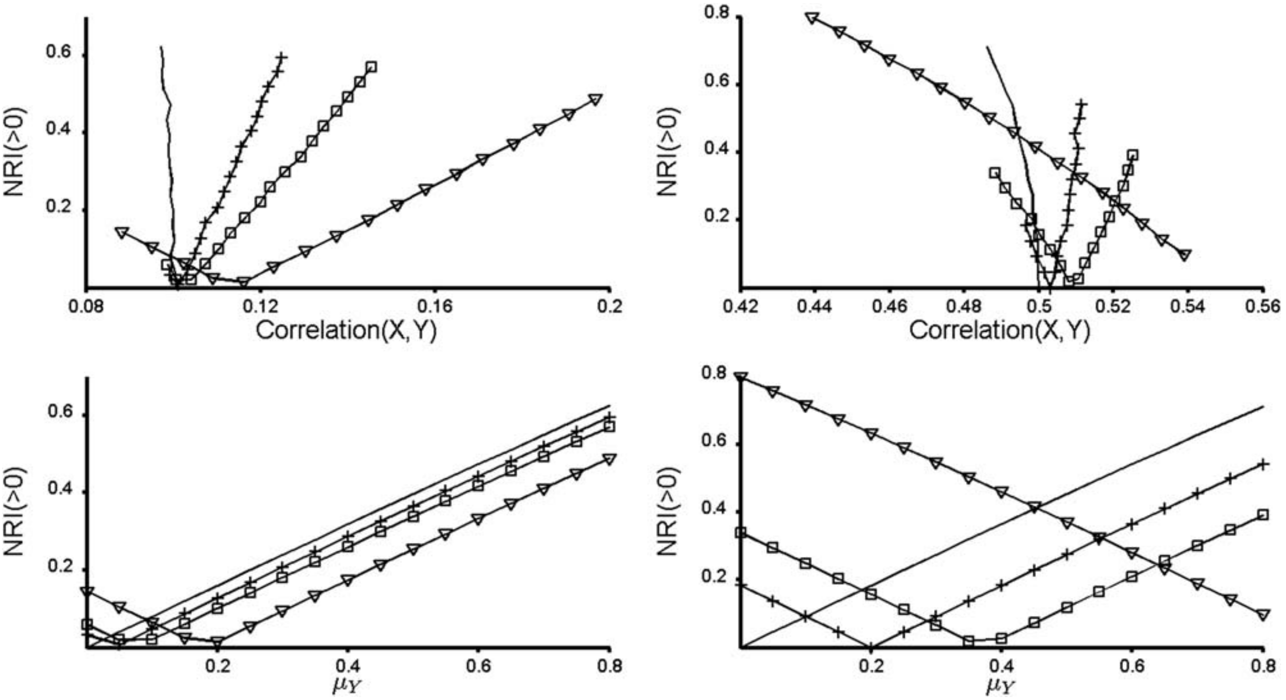

Figure 4. The behavior of continuous net reclassification index (NRI(>0)). The left panels show results for $\rho = 0.1$ and the right panels show results for $\rho = 0.5$, where ρ is the conditional correlation between the markers *X* and *Y*. The event rate is 0.10 in both cases. The top row shows that NRI(>0) can increase or decrease as the unconditional correlation between *X* and *Y* increases. The bottom row shows the same data in a different way to illustrate that NRI(>0) depends on the strength of the baseline model when the new marker *Y* is not conditionally independent of the existing marker *X*. Solid line, $\mu_X = 0$; +, $\mu_X = 0.4$; □, $\mu_X = 0.742$; ∇, $\mu_X = 1.812$.

incremental value and marginal strength. However, readers should be cautioned against equating the two. Incremental value also depends on the correlation between the marker and the baseline predictors. High correlation can be a very good thing as demonstrated, for example, in Figure 2.

We conclude by highlighting points of agreement with the article. First, we concur with the authors' choice to disregard issues of statistical significance. Experience shows that the additional predictive ability that is required for a biomarker to be useful dominates what is required for statistical significance. If clinical utility is a possibility, then statistical significance is typically not in question. Second, we endorse the authors' comment that ultimately the value of a new marker must be evaluated in terms of the costs and benefits relating to how the risk model will be used. Finally, the authors' results remind us that there is no such thing as "intrinsic" incremental value—we must always keep the population and application in mind. Two researchers assessing a new marker might get very different results because the strengths of the baseline models are very different in the 2 populations the researchers are studying. For example, suppose the baseline model uses age, which is much more predictive in older populations than in younger populations, and one researcher has data on a relatively young population whereas the other is studying an older population. We can envision that the researcher with the younger population would estimate a higher incremental value than the researcher with the older population. This should not be considered a contradiction; the marker may in fact have utility in the younger population and not in the older population.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, University of Washington, Seattle, Washington (Kathleen F. Kerr, Aasthaa Bansal, Margaret S. Pepe); and Program in Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, Washington (Margaret S. Pepe).

This work was funded by grants R01-GM054438 and U24-CA086368 to Margaret S. Pepe.

The authors thank Dr. Ying Huang for valuable comments.

Conflict of interest: none declared.

REFERENCES

1. Pencina MJ, D'Agostino RB, Pencina KM, et al. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol*. 2012;176(6):473–481.

2. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):157–172.
3. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30(1):11–21.
4. Tzoulaki I, Liberopoulos G, Ioannidis JP. Use of reclassification for assessment of improved prediction: an empirical evaluation. *Int J Epidemiol*. 2011;40(4):1094–1105.
5. Greenland S. 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 M. J. Pencina et al., Statistics in Medicine. *Stat Med*. 2008;27(2):199–206.
6. Mihaescu R, van Zitteren M, van Hoek M, et al. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. *Am J Epidemiol*. 2010;172(3): 353–361.
7. Chi YY, Zhou XH. 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. *Stat Med*. 2008;27(2):182–184.
8. Pepe MS, Janes H. Commentary: reporting standards are needed for evaluations of risk reclassification. *Int J Epidemiol*. 2011;40(4):1106–1108.
9. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics*. 2005;63(2):227–239.
10. Bansal A, Pepe MS. *When Does Combining Markers Improve Classification Performance and What Are Implications for Practice*? *University of Washington Biostatistics Working Paper 387*. Berkley, CA: The Berkeley Electronic Press; 2011.
11. Kerr KF, Pepe MS. Joint modeling covariate adjustment, interaction: contrasting notions in risk prediction models risk prediction performance. *Epidemiology*. 2011;22(6): 805–812.
12. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3:1157–1182.