American Journal of Epidemiology
Vol. 176, No. 6
DOI: 10.1093/aje/kws206
Advance Access publication:
August 8, 2012

Response to Invited Commentary

Pencina et al. Respond to "The Incremental Value of New Markers" and "Clinically Relevant Measures? A Note of Caution"

Michael J. Pencina*, Ralph B. D'Agostino, Olga V. Demler, A. Cecile J. W. Janssens, and Philip Greenland

* Correspondence to Dr. Michael J. Pencina, Department of Biostatistics, Boston Unviersity, Framingham Heart Study, Harvard Clinical Research Institute, Rm. 328, CrossTown, 3rd Floor, 801 Massachusetts Avenue, Boston, MA 02118 (e-mail mpencina@bu.edu).

Initially submitted March 30, 2012; accepted for publication April 10, 2012.

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

We thank Drs. Cook, Kerr, Bansal, and Pepe for their careful review of our work and insightful critiques. Several of the points they raise require further highlighting, discussion, or rebuttal.

In her commentary, Dr. Cook presents 2 interesting examples that shed additional light on some of the properties of the 3 measures of interest (1). On the basis of the developments presented by us (2), as well as those presented by Kerr et al. (3), the increments between the measures of interest must go in the same direction. Cook's example with diabetes shows that this need not be the case for binary variables. She explains the reason for this anomaly: "The problem is that counterintuitively, there are more cases among people without diabetes than among those with because nondiabetic participants comprise the majority of the cohort" (1, p. 488) and concludes that "[t]he new model may look worse because more are moving in the wrong direction, but the correct changes are larger and the incorrect changes are much smaller" (1, p. 488). This accentuates an important feature of the continuous net reclassification improvement (NRI(>0)); it focuses on the net numbers with altered risks without regard for the magnitude of the change. If we want to weight the movements by their magnitudes, we need to obtain the integrated discrimination improvement (IDI). Although we agree with Cook that helping many people very little may not be better than helping a few people a lot, obtaining such information can actually be valuable, and that information is easy to understand. One additional consideration that might improve the clinical interpretability of the NRI(>0) could be to require that only changes in risks greater than some minimal clinically significant amount be considered. This was the rationale behind our original notation, NRI(>x); x has usually been taken to be 0, but other numbers, for example 0.01 or 0.05, might sometimes be more appropriate.

This example shows the complementarity of the 3 measures in the general setting, an issue raised by Kerr et al (3). A number of other examples can be constructed for further illustration. We give only one here, based on Table 2 of our original article (2). A weak marker is added to a baseline model that has poor performance because of a limited range of some important predictors (for example, age). The observed increase in the area under the receiver operating characteristic curve (AUC) is 0.025, which would usually be considered somewhat promising. However, based on the NRI(>0) of 0.16, we would be able to conclude that it is a weak marker, the impact of which on the AUC will diminish if we extend the range of baseline predictors. Again, the complementarity of these 3 measures lies in their slightly different focuses: The AUC is primarily concerned with the risk model at hand and the NRI(>0) is primarily concerned with the novel marker at hand, whereas the IDI falls in between.

Cook's second example is perhaps even more important. It shows that it is possible (and in our experience quite likely) to have a predictor with no meaningful improvement in the AUC or the IDI but with a highly positive NRI(>0) that is statistically significant. The note of caution that Cook attaches to this observation needs to be highlighted and re-emphasized: Statistical significance of the NRI(>0) indicates practically nothing, and any inference based on this measure needs to be based on its magnitude. Hence, the benchmarks we provide are so important.

Here again, the above example can also serve as an answer to the criticisms of benchmarks by Kerr et al. (3). They 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" (3, p. 482). We disagree. First, benchmarks can be helpful whenever binary or ordinal assessment must be imposed on continuous phenomena. Furthermore, it is not clear what the authors mean by clinical interpretation. In our opinion, the distance between average risks for events and nonevents is a very natural and interpretable summary unless one specifically requires incorporation of clinical decision-making, which will necessitate the use of categories and thresholds. This approach has been explicitly endorsed by Dr. Cook in her commentary. In general, we agree that in situations in which established categories exist, their incorporation into the assessment of improvement in model performance affords an additional level of clinical interpretability. However, we do not think that this has to happen to the exclusion of the global measures.

Threshold-based inference may be premature unless improvement in the global performance measures has been established. First, reliance on global measures greatly improves the chances for successful replication of the results. Second, it might theoretically be possible to construct markers that do well based on threshold-based clinical rules but that do poorly when assessed with global measures. However, the chances of discovering such predictors in clinical practice are slim.

Kerr et al. devote a large portion of their commentary to the issue of correlation. They correctly point out that "the notions of incremental value and marginal strength are distinct concepts" (3, p. 482). However, they also use this distinction to criticize our choice of new markers with zero conditional correlation as the reference standard and ascribe to us the desire "to reinforce a common misconception that it is ideal for a new predictor to be uncorrelated with existing predictors" (3, p. 482). In our opinion, their reasoning confuses the issues of a theoretical relation between incremental value and marginal strength and the selection of reference standard and places an overtly large emphasis on the theoretical possibilities. We are aware that in some cases, large correlation can improve discrimination, and the increase in marginal effect size might actually be detrimental to the increase in incremental value. This was shown previously by Cochran (4) and Mardia et al. (5) and is illustrated in Figure 2 of the commentary by Kerr et al. (3). However, the fact that this "paradox" is possible does not mean that it represents the most likely scenario (unconditional correlations commonly observed in practice do not reach the levels required for them to improve discrimination), and it definitely does not represent one that is the easiest to conceptualize. The last 2 features are the key characteristics of a reference standard. Without the loss of generality, any correlated marker can be transformed into an uncorrelated one by a linear transformation that preserves the change in AUC, the IDI, and the NRI(>0). Once such a transformation is performed, predictors can be compared using their conditional effect sizes alone. Thus, even variables with more complex correlation structures can be brought to the plane of our reference standard. Furthermore, working with uncorrelated predictors makes the conceptualization of the increase in model performance easy: The Mahalanobis distance between the event and nonevent populations increases by the square of the effect size of each added variable. Of note, our choice of zero conditional correlation corresponds to a small (<0.1 for most situations) unconditional correlation that is likely to be encountered in practical applications.

Finally, we note that the ideas presented in our article (2) should be considered against the backdrop of the current state of the biomarker research, in which the focus is based solely on statistical significance. Weak predictors are hailed as big winners just because they can have a small enough *P* value. Our proposal is meant as a first step in the major reorientation of the field to have it rely on the magnitude of the observed effects rather than on their statistical significance. It is hard to imagine that such transformation could take place without the improved understanding of the magnitude of observed effects for which simulations and benchmarks serve as tools.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, Boston University, Boston, Massachusetts (Michael J. Pencina, Olga V. Demler); Department of Mathematics and Statistics, Boston University, Boston, Massachusetts (Michael J. Pencina, Ralph B. D'Agostino); Harvard Clinical Research Institute, Boston, Massachusetts (Michael J. Pencina, Ralph B. D'Agostino); Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands (A. Cecile J. W. Janssens); and Feinberg School of Medicine, Northwestern University, Chicago Illinois (Philip Greenland).

This work was supported by the National Institutes of Health/American Recovery and Reinvestment Act Risk Prediction of Atrial Fibrillation (grant 1 RC1HL101056; Michael J. Pencina); the National Heart, Lung, and Blood Institute's Framingham Heart Study (contract N01-HC-25195; Michael J. Pencina and Ralph B. D'Agostino); the Center for Medical Systems Biology in the framework of the Netherlands Genomics Initiative and the Netherlands Organisation for Scientific Research (A. Cecile J. W. Janssens); and the Northwestern University Clinical and Translational Sciences Institute (grant UL1RR025741; Philip Greenland).

Conflict of interest: none declared.

REFERENCES

1. Cook NR. Clinically relevant measures of fit? A note of caution. *Am J Epidemiol*. 2012;176(6):488–491.
2. 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.

3. Kerr KF, Bansal A, Pepe MS. Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context. *Am J Epidemiol*. 2012;176(6):482–487.
4. Cochran WG. On the performance of the linear discriminant function. *Technometrics*. 1964;6(2):179–190.
5. Mardia KV, Kent JT, Bibby JM. *Multivariate Analysis*. London, UK: Academic Press; 1979:78–79.