Response

Margaret Sullivan Pepe

Affiliation of author: Fred Hutchinson Cancer Research Center, Public Health Sciences Division, Seattle, WA.
Correspondence to: Margaret Sullivan Pepe, PhD, Fred Hutchinson Cancer Research Center, Public Health Sciences Division, 1100 Fairview Ave N, M2-8500, Seattle, WA 98109-1024 (e-mail: mspepe@u.washington.edu).

It is well known that standard statistical methods are invalid when prediction models are evaluated using the same data set used to develop the models. This is true for all measures of prediction performance. The letter by Pencina, Neely, and Steyerberg considers primarily this scenario. However, the most surprising and alarming observation that we have made about the net reclassification index (NRI) statistic is that there is positive bias even when model evaluation is performed on a validation data set that is different from the development data set. That is, even with the best design that employs an independent validation data set to evaluate the performance of risk models, one is inclined to make incorrect positive conclusions about useless biomarkers with the NRI statistic, while this is not the case with use of other measures of prediction performance such as receiver operating characteristic (ROC) or net benefit statistics. Hilden and Gerds note that in statistical jargon the NRI statistic is therefore “improper” (1).

I disagree with the two final assertions made in the correspondence. First, the magnitude of bias in the NRI does not necessarily decrease with increasing sample size. We have recently shown in supplementary material to our article (2) that if the number of predictors increases proportionately with increasing sample size, the typical scenario in practice, then the problem with positive bias remains or gets worse. Second, the problem of positive bias may be less for the categorical NRI in certain scenarios such as the one considered by Pencina, Neely, and Steyerberg, but we have demonstrated (3) at least one setting where a two-category NRI leads to incorrect positive conclusions on independent validation data with 82% probability. That is, use of the categorical NRI can be highly misleading at least in certain settings. The categorical NRI therefore cannot be endorsed for general use, especially because there are no guidelines to identify settings where its use may be valid. I doubt that practical guidelines could be formulated for valid use of the NRI.

Finally, it is worth reiterating that the interpretation of the NRI statistic is not particularly helpful or meaningful in evaluating risk prediction markers (4, 5). The NRI is often misinterpreted as a “proportion of correct reclassifications,” while in fact it is not even a proportion and it can take values as large as 2. Another problem is that the NRI inappropriately counts all reclassifications of risk equally. More meaningful and more clinically relevant measures that have been proposed include the improvement in net benefit (6), also known as the change in the decision curve (7) or the relative utility (8). I encourage the use of such measures in part because of their relevant interpretation and in part because of their “proper” technical statistical properties (2) that guarantee correct statistical inference at least in independent validation data sets.

References