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/kws208
Advance Access publication:
August 8, 2012

Invited Commentary

Clinically Relevant Measures of Fit? A Note of Caution

Nancy R. Cook*

* Correspondence to Dr. Nancy R. Cook, Division of Preventive Medicine, Brigham and Women's Hospital, 900 Commonwealth Ave. East, Boston, MA 02215 (e-mail: ncook@rics.bwh.harvard.edu).

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

Risk reclassification methods have become popular in the medical literature as a means of comparing risk prediction models. In this issue of the *Journal*, Pencina et al. (*Am J Epidemiol.* 2012;176(6):492–494) present further results for continuous measures of model discrimination and describe their characteristics in nested models with normally distributed variables. Measures include the change in the area under the receiver operating characteristic curve, the integrated discrimination improvement, and the continuous net reclassification improvement. Although theoretically interesting, these continuous measures may not be the most appropriate to assess clinical utility. The continuous net reclassification improvement, in particular, is a measure of effect rather than model improvement and can sometimes exhibit erratic behavior, as illustrated in 2 examples. Caution is needed before using this as a measure of improvement. Further, the test of the continuous net reclassification improvement and that for the integrated discrimination improvement are similar to the likelihood ratio test in nested models and may be overinterpreted. Reclassification in risk strata, while requiring thresholds, may be more relevant clinically with its ability to examine potential changes in treatment decisions.

calibration; discrimination; model fit; risk prediction

Abbreviations: AUC, area under the receiver operating characteristic curve; BMI, body mass index; CI, confidence interval; IDI, integrated discrimination improvement; NRI, net reclassification improvement; NRI(>0), continuous net reclassification improvement.

After the introduction of clinical reclassification in risk strata (1), Pencina et al. presented the net reclassification improvement (NRI) as a measure of reclassification conditional on case-control status (2). They also presented the integrated discrimination improvement (IDI) and subsequently the continuous NRI (NRI(>0)) and discussed how these methods can be applied to survival data (3). They have continued their interesting work in the current article (4), focusing on continuous measures that do not require prespecified categories of predicted risk and that are thus less arbitrary. What they add to the existing panoply of measures of fit, however, remains unclear.

Pencina et al. have shown that the change in the area under the receiver operating characteristic curve (AUC) and the NRI(>0) are related (3, 4). Both of these measures examine the rank order of probabilities among cases and controls in one model versus another. Absolute differences in estimated risk, however, cannot be determined from simple rankings. In a typical population with a majority at low risk, changes in rank may not lead to meaningful differences. Changes are more meaningful if they are large or they occur among those of at least moderate risk. They are clinically important if they are enough to change treatment decisions.

In the current article by Pencina et al. (4), the authors derived results for a situation in which the predictors have normal distributions, providing a useful reference. In the simple case of a single normally distributed variable, the continuous measures reduce to functions of the Mahalanobis distance between cases and controls. Pencina et al. discovered the surprising fact that NRI(>0) was not a function of the strength of other variables in the model if the variables were independent. It is essentially a measure of effect rather than a measure of model improvement. At least

when adding a single normal variable, it contains the same information as the odds ratio. What it adds to standard measures, at least in this simple setting, is not clear.

In contrast to the NRI(>0), the change in the AUC is dependent on the strength of the reference model and is thus a measure of actual improvement. The IDI, while also a function of the Mahalanobis distances, is dependent on the prevalence of disease. Under relatively strong assumptions, the relative IDI can be compared to $1/p$, where p is the number of predictors already in the model, representing the average effect of the other variables. This relative measure thus compares a new marker with the other variables in the model. Whether the new marker adds information in an absolute sense that is clinically meaningful still cannot be determined.

Although these results are theoretically interesting, how the methods work in practice is illustrative. An interesting anomaly occurs in data on diabetes as a predictor of cardiovascular disease in the Women's Health Study, similar to that previously described for hemoglobin A1c (5). To simplify, consider first a single predictor Z that is a composite of traditional cardiovascular risk factors, namely age, smoking, systolic blood pressure, and total and high density lipoprotein cholesterol levels standardized to a mean of 0 and a standard deviation of 1. Its hazard ratio for incident cardiovascular disease is 2.90 per each standard-deviation unit (95% confidence interval (CI): 2.69, 3.13) in Women's Health Study data.

When a binary predictor for history of diabetes is added to the model, both variables are highly predictive. The hazard ratio for Z is now 2.66 (95% CI: 2.47, 2.88) and that for diabetes is 3.62 (95% CI: 2.96, 4.42). The difference in the AUC is 0.015 ($P < 0.0001$), and the IDI is 0.014 (95% CI: 0.013, 0.016; $P < 0.0001$), with a large relative IDI of 0.401 (Table 1). When risk is stratified into 3 categories with cut points of 5% and 20% 10-year risk, the reclassification measures support a strong effect (Table 2). The categorical NRI is 0.086 (95% CI: 0.049, 0.123; $P < 0.0001$), and the reclassification calibration (RC) statistics suggest that the model with diabetes fits the observed probabilities more closely. After adjustment for optimism with bootstrapping, the NRI is 0.082 (95% CI: 0.039, 0.125; $P = 0.0002$), and the RC statistics are 133.9 and 13.5. All of these measures are highly statistically significant and support a strong positive effect of diabetes on cardiovascular disease.

In contrast, the NRI(>0) is negative, equal to −0.236 (95% CI: −0.308, −0.165; $P < 0.0001$), implying a worsening rather than an improvement in fit. More individuals move in the wrong direction, particularly cases, because the relative improvement for cases is −0.528 and that for noncases is 0.292, both of which are highly significant. When residuals for diabetes independent of Z are computed using binary regression, the hazard ratios become 2.74 (95% CI: 2.54, 2.96) for Z and 1.23 (95% CI: 1.19, 1.27) for residual diabetes. However, the predicted values from the model with residuals are identical to those using diabetes itself, so all measures of fit stay the same, indicating that the anomaly is not due to the correlation of Z and diabetes.

Table 1. Continuous Measures of Improvement in Predicting Risk of Cardiovascular Disease for Variables Added to the Model With Traditional Risk Factors, Women's Health Study ($n = 24{,}551$)

Variable	HR	*P* Value	*c* Statistic		*P* Value for Difference	IDI	*P* Value	NRI(>0)	*P* Value	RI(>0)			
			Without Variable	With Variable						Cases	*P* Value	Noncases	*P* Value
Diabetes	3.62	<0.0001	0.782	0.797	<0.0001	0.014	<0.0001	−0.236	<0.0001	−0.528	<0.0001	0.292	<0.0001
Log BMI	1.05	0.17	0.782	0.782	0.84	−0.0002	0.13	0.157	0.0002	−0.230	<0.0001	0.386	<0.0001

Abbreviations: BMI, body mass index; HR, hazard ratio; IDI, integrated discrimination improvement; NRI(>0), continuous net reclassification improvement; RI(>0), continuous reclassification improvement.

Table 2. Categorical[a] Measures of Improvement in Predicting Risk of Cardiovascular Disease for Variables Added to the Model With Traditional Risk Factors, Women's Health Study (n = 24,551)

Variable	Reclassification Calibration Chi-Square				NRI	*P* Value	Reclassification Improvement			
	Without Variable	*P* Value	With Variable	*P* Value			Cases	*P* Value	Noncases	*P* Value
Diabetes	152.0	<0.0001	16.8	0.005	0.086	<0.0001	0.073	<0.0001	0.013	<0.0001
Log BMI	5.2	0.16	5.2	0.16	−0.0014	0.83	−0.0037	0.56	0.0023	0.0002

Abbreviations: BMI, body mass index; NRI, net reclassification improvement.
[a] Categories of 10-year risk of cardiovascular disease with cut points of 5% and 20%.

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. Estimated risk increases among virtually all diabetic participants and generally decreases slightly among nondiabetic participants. As the persons without diabetes move slightly down in risk, their cases move down too, leading to more nondiabetic cases moving down than up. The change in risk estimates for diabetic participants, however, is much larger. The average change in estimated risk among the diabetic subjects is 0.092, whereas that among the nondiabetic subjects is −0.003. This illustrates the problem with relying on ranks without regard to the size of the changes in the probabilities themselves. The 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.

Another example using body mass index (BMI, measured as weight in kilograms divided by height in meters squared) shows that a null result can have a positive NRI (>0). When BMI is normalized using the log transform and added to a model with the same variable *Z*, its hazard ratio is 1.05 per each standard-deviation unit (95% CI: 0.98, 1.13; $P = 0.17$). As with other Women's Health Study data (6), BMI is no longer predictive given the traditional risk factors. The same is true using residuals of log BMI, regressing out *Z*; the hazard ratio remains 1.05 (95% CI: 0.98, 1.12; $P = 0.17$), whereas that for *Z* increases from 2.88 to 2.92. This example thus follows the assumptions of Pencina et al. regarding normality and lack of correlation.

The change in AUC, the IDI, and the categorical reclassification measures all show no improvement in prediction when adding BMI (Tables 1 and 2). The NRI(>0), however, is strongly positive at 0.157 (95% CI: 0.074, 0.240; $P = 0.0002$). Although this would be considered a "weak" improvement by Pencina et al., it is of size comparable to that of total and high density lipoprotein cholesterol and C-reactive protein and stronger than that for family history of myocardial infarction (5). Only 39% of cases move up compared with 31% of noncases, leading to the NRI(>0) of 15.7%. Although this suggests a large improvement, the actual change in risk estimates is very small, only −0.0005 in cases and −0.0003 in noncases, leading to the small and nonsignificant IDI. Even though one could require the test of association to be significant before examining model improvement, this example suggests that the NRI(>0) may give unusual results even in such conditional analyses. Results for diabetes and BMI seen here may seem like anomalies; however, these situations seem to occur relatively frequently in practice.

Although Pencina et al. (4) focus on effect size, it is useful to also consider power. The NRI(>0) is much more likely to be statistically significant than is the categorical NRI. Cook and Paynter examined the power for tests of various measures when adding a single predictor variable to a logistic model using similar assumptions regarding normality (5, 7). The power for the likelihood ratio test, the IDI, and the NRI(>0) are virtually the same. They are basically different versions of the same test, as they test the same hypothesis, using difference in R^2 versus likelihood ratio test, difference in predicted risk using means versus using ranks. The test for the continuous reclassification improvement among cases or controls separately is simply a sign test, and the NRI(>0) is a combination of these (5). If one test (8) is all we want, then we could stop at the likelihood ratio test. Power for the categorical NRI and reclassification calibration test are lower (7). The reclassification tests essentially raise the bar higher. Not only do the new markers need to be predictive, but enough people also need to cross important thresholds to matter.

Continuous measures such as those proposed may not be necessary in the simple situation with a normal independent biomarker. If we can estimate an odds ratio or relative risk, do we need the NRI(>0), especially if it performs like another measure of effect size? Likewise, if we can estimate a difference in model R^2 measures, do we need the IDI? These new measures have promise, however, in situations that are more complex. If the continuous NRI can summarize the predictive power in a group of variables of all types and distributions and if it remains interpretable as an analog of the odds ratio, then perhaps it could be useful as an overall summary of effect size, although not model improvement. More needs to be discovered about its properties in other situations with nonnormal data, including binary variables, groups of variables, and correlated variables. Because it is based on predicted values, the IDI can be generalized to any model, whether linear function, tree, or neural net and whether or not a formal likelihood can be defined. More work on how the measures perform in these general situations is needed.

On the other hand, none of these continuous measures may be appropriate to assess clinical utility. Although they may have limitations, the reason that reclassification tables and the NRI have "taken off like wildfire" (9, p. 1107) is that clinicians find them useful. Clinicians treat patients, and risk stratification helps them by suggesting courses of

action (10). When proposed in 2006 for cardiovascular disease prevention (11), clinical reclassification was the focus. By using risk strata, the reclassification table focuses on ranges of risk with clinical meaning. Guidelines for therapies in many other fields, including oncology and infectious disease (12–15), are similarly based on the absolute risk of disease with associated cut points for risk. It thus makes sense to consider whether individuals would cross established thresholds and be assigned different therapeutic regimens. The reclassification table directly shows how treatment decisions might change with a different model or a new test and bridges the gap from model to clinical usefulness.

A drawback of the categorical measures is that they can be arbitrary if not based on prespecified cut points. In the absence of clinical guidelines, such thresholds can be derived using relative costs (16). Cook and Paynter suggested using the population incidence or prevalence as a default threshold when no information was available, with additional cut points at half and twice this proportion (7). Mealiffe et al., however, found that the NRI is only weakly dependent on the placement of cut points (17). Whether to use 3 or 4 categories is also open for debate. Although some advocate 3 categories to correspond to 2 thresholds for lifestyle or medical interventions (3), 4 narrower categories are more amenable to calibration. Using 4 categories would also allow areas of uncertainty on either side of a central threshold and would highlight changes in treatment strategies among those in a gray area. Often, however, the definition of clinical utility must depend on the specific application under consideration. Because standards of practice have not been established, researchers must at minimum clearly describe the criteria they use and justify their choice of any cut points used when no established thresholds exist.

The categorical NRI has become popular in the medical literature, and the NRI(>0) is now starting to be embraced by the clinical community, perhaps because it is more likely to show a large improvement. In general, the NRI(>0) is much larger and more often statistically significant than the categorical NRI, and it may be overinterpreted. As Pencina et al. (4) show, however, the NRI(>0) is not a measure of improvement but of association. Its behavior can also sometimes be erratic, as seen in the examples above. Clearly, caution is needed along with more clarity regarding the performance and interpretation of these measures.

ACKNOWLEDGMENTS

Author affiliation: Division of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts (Nancy R. Cook).

Dr. Cook and the Women's Health Study were supported by National Institutes of Health grants CA047988 and HL080467 from the National Heart, Lung, and Blood Institute and the National Cancer Institute, both in Bethesda, Maryland.

Conflict of interest: none declared.

REFERENCES

1. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007; 115(7):928–935.
2. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new biomarker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):157–172.
3. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30(1):11–21.
4. 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.
5. Cook NR, Paynter NP. Comments on 'Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers' by M. J. Pencina, R. B. D'Agostino Sr and E. W. Steyerberg, Statistics in Medicine 2010;30(1):11–21. *Stat Med*. 2012;31(1):93–95.
6. Ridker PM, Buring JE, Rifai N, et al. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women. *JAMA*. 2007;297(6): 611–619.
7. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. *Biom J*. 2011;53(2):237–258.
8. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC Med Res Methodol*. 2011;11(1):13. (doi:10.1186/1471-2288-11-13).
9. Pepe MS, Janes H. Commentary: reporting standards are needed for evaluations of risk reclassification. *Int J Epidemiol*. 2011;40(4):1106–1108.
10. Bigger JT Jr., Heller CA, Wenger TL, et al. Risk stratification after acute myocardial infarction. *Am J Cardiol*. 1978;42(2): 202–210.
11. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann Intern Med*. 2006;145(1):21–29.
12. Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). *JAMA*. 2001;285(19): 2486–2497.
13. Munshi NC, Anderson KC, Bergsagel L, et al. Consensus recommendations for risk stratification in multiple myeloma: report of the International MyelomaWorkshop Consensus Panel 2. *Blood*. 2011;117(18):4696–4700.
14. Viswanathan K, Chlebowski RT, Hurley P, et al. American society of clinical oncology clinical practice guideline update on the use of pharmacologic interventions including tamoxifen, raloxifene, and aromatase inhibition for breast cancer risk reduction. *J Clin Oncol*. 2009;27(19): 3235–3258.
15. Worth LJ, Lingaratnam S, Taylor A, et al. Use of risk stratification to guide ambulatory management of neutropenic fever. *Intern Med J*. 2011;41(1b):82–89.
16. Pauker SG, Kassirer JP. Therapeutic decision making: a cost–benefit analysis. *N Engl J Med*. 1975;293(5):229–234.
17. Mealiffe ME, Stokowski RP, Rhees BK, et al. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. *J Natl Cancer Inst*. 2010; 102(21):1618–1627.