

COMMENTARY

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* (this issue)

Nancy R. Cook^{*,†}

*Brigham & Women’s Hospital, Harvard Medical School, 900 Commonwealth Ave. East,
Boston, MA 02215, U.S.A.*

Pencina *et al.* [1] are to be commended for their thoughtful and useful work. They extend the work on reclassification into clinical risk categories, first presented in Cook *et al.* [2], to develop two summary measures of overall fit. This kind of work is needed in the clinical literature, so that physicians and other researchers can more easily assess the broad array of new biomarkers now being evaluated, whether based on plasma, genetics, proteomics, or epidemiological data. New statistical tools that are both sensitive for detecting improvements and appropriate to clinical questions are needed. While the receiver operating characteristic (ROC) curve has been used most prominently for this purpose, particularly in clinical cardiology, this measure has limitations [3], and should not be the sole determinant of clinical utility.

The authors present two new summary measures of model fit based on the comparison of predicted values from two models. The first is the ‘net reclassification index’, or NRI, which compares the proportions moving up or down in clinical categories in cases *versus* controls. This measure seems to make sense and is an interpretable summary of the reclassification table. The net index offers some correction for chance variation as well as a correction for calibration, or reclassification among non-cases.

The test associated with the NRI is similar in spirit to the McNemar test, where discordance is defined as being off the diagonal. Pairs of predicted probabilities could be classified as being in the same, or in higher or lower categories with the two models. The test could then be based on the discordant pairs, and the proportions with disease could be compared among those who move higher *versus* those who move lower. Moskowitz and Pepe [4] described such a procedure for binary tests, comparing predictive values. Using the off-diagonal elements extends this classic idea beyond binary testing. Specifically, among the discordant pairs, i.e. those not lying in a diagonal

*Correspondence to: Nancy R. Cook, Brigham & Women’s Hospital, Harvard Medical School, 900 Commonwealth Ave. East, Boston, MA 02215, U.S.A.

†E-mail: ncook@rics.bwh.harvard.edu

cell, one would test whether $\Pr(D|\text{higher})$ was different from $\Pr(D|\text{lower})$. Logistic regression could also be used to determine the odds ratio for moving up in cases *versus* non-cases among discordant pairs. In the Framingham data presented in Table 2 of Pencina *et al.* [1], the odds ratio among these discordant pairs is 4.2, $p = 0.0004$, meaning that the odds of cases moving up *versus* those moving down are four times higher than among the non-cases.

For clinical purposes, physicians may be most interested in reclassification of those at intermediate risk, i.e. those in the 'grey zone' or for whom treatment decisions are unclear. In cardiovascular medicine, this would typically include those in the range of 6–20 per cent estimated risk, or those in the middle category in the Framingham data. Among those initially classified as at 6–20 per cent risk under the model without HDL, the percents moving up or down are very different among cases and non-cases. Among cases this reclassification index is 9.5 per cent, while among non-cases it is –13.3 per cent, for a NRI of 22.8 per cent. This indicates a more sizeable effect in the category of the most interesting clinically.

The second new summary measure, the integrated discrimination improvement (IDI), compares the integrals of sensitivity and specificity under the two models. While derived from integrals, this is essentially a comparison of the Yates slope under the two models, or the mean difference in predicted probabilities between cases and controls. As such, it is interpretable in terms of average differences in predicted probabilities. Because these differences tend to be rather small, this measure may be less important than the NRI clinically. The majority of individuals within a cohort may be at low risk, and changes in their predicted probabilities may be trivial, leading to a low overall IDI.

Inherently, as the authors suggest, both the NRI and IDI are measures of discrimination, and both condition on disease status. Model calibration, though, should also be considered in model assessment, and can be easily added to the arsenal of measures considered here. The Hosmer–Lemeshow test, while typically conducted within the deciles of risk, can also be used with categories based on risk estimates [5]. In particular, it could be used to examine model calibration within categories of the cross-classified, or reclassification, table. Deviations of the predicted probabilities from the observed risk within these categories could be then quantified. While the statistical properties of the test in this situation have not been fully characterized, using categories with a cell size of at least 20 generally leads to five or more expected outcomes with these clinically based categories.

Table I presents a rearrangement of data from Table II of Pencina *et al.* [1], showing the observed risks within each category of predicted risk. The model including HDL is better calibrated, since the observed risk is more accurately captured across rows than down columns. To formally test this observation, one can compare the observed risk with the average predicted risk within each cell of the table for each model separately. For illustration purposes, suppose that the average predicted risk within categories for each model is equal to the observed risk found in the diagonal cells of the table, i.e. 2, 11, and 23 per cent. The adjusted Hosmer–Lemeshow statistic [6], with 5 degrees of freedom, would then be 62.7 for the model without HDL ($p < 0.0001$) and 2.25 for the model including HDL ($p = 0.81$). Thus, the model without HDL demonstrates serious deviation from fit, while that including HDL appears to be well calibrated.

Previous publications have considered reclassification using data from the Women's Health Study (WHS) [7]. These include comparisons of models with and without high-sensitivity C-reactive protein [2], with and without HDL [3], and two models based on several measured biomarkers (the Reynolds Risk Score models) [8]. Various measures of model fit based on the published reclassification tables are shown for these in Table II, along with two-sided p -values. The NRI estimates range from 4.7 to 8.4 per cent, with smaller values, as expected, when a test

COMMENTARY

Table I. Observed risk (per cent) in reclassified categories for models with and without HDL in Framingham data.

Model with HDL	Model without HDL		
	<6 per cent	6–20 per cent	>20 per cent
<6 per cent	1.95	9.55	—
6–20 per cent	2.63	11.01	31.1
>20 per cent	—	10.71	22.8

set is used. The corresponding ORs based on discordant pairs (not shown) are all approximately 2 in the original training data and 1.6 in the test data. When the intermediate categories of most clinical interest are used, these ‘clinical’ or conditional NRI’s range from 15 to 31 per cent in training data and from 12 to 19 per cent in test data. All these are highly significant in the training data and at least of borderline significance in test data.

Tests for the IDI were based on *t*-tests comparing differences in predictions between the two models in cases *versus* controls, and used a conservative Satterthwaite approximation for degrees of freedom with unequal variances, due to larger variation among the cases. Values ranged from 0.26 to 0.70 per cent across models. While most of these differences were statistically significant, except when using a smaller test set, the size of the IDI may discourage its use among clinicians. As described above, because it compares predicted values across the entire range, it may be less useful for determining changes that are clinically relevant.

Also shown in Table II are the results of calibration tests comparing the observed and predicted risks within clinical categories, excluding cells with fewer than 20 individuals. As in the Framingham data, in all comparisons the reduced models demonstrate significant deviation from fit, while the full models generate predicted values much closer to those observed, with no significant deviation. This suggests that in all models considered, the addition of the new variables could substantially improve the ability to predict risk accurately.

The ROC curve has attained popularity in the medical literature, likely due to its interpretability and its easy-to-use 0.5–1 scale. Its main application, however, is in classification, rather than risk prediction. There is a difference between diagnosing disease that is already present and predicting disease in the future using current risk factors. In the latter situation, predictive values are ultimately most important. To replace the ROC curve, a measure that is also clinically interpretable would be preferred. The IDI, while based on a relatively simple measure, has values that may be too low to be interpretable. For generally low-risk populations, the average differences in slopes may be too modest to seem important. The NRI, in conjunction with a test of reclassification calibration, seems to be a step in the right direction. While these depend on the chosen categories, the categories can be ones that matter most clinically. If changes are not large enough to alter practice or medical advice, new predictors may not have much clinical impact. The method of model assessment should depend on the purpose of the model. If it is to be used in risk stratification, then how well it classifies into risk categories, determined by examining both discrimination and calibration, should be of primary importance.

Some questions still remain regarding the NRI and its use in model assessment. More statistical research is needed to determine its performance characteristics and the best use of the measure in particular applications. For example, how does the number of categories affect performance? Does

COMMENTARY

Table II. Comparison of published models based on NRI, IDI, and reclassification calibration.

Model comparison	H-L test of calibration											
	Clinical NRI					Reduced model					Full model	
	NRI (per cent)	<i>p</i>	NRI (per cent)*	<i>p</i>	IDI (per cent)	<i>p</i>	df	χ^2	<i>p</i>	χ^2	<i>p</i>	
WHS CRP model [2]	5.7	0.0001	15.0	<0.0001	0.26	<0.0001	8	25.2	0.0014	11.3	0.19	
Reynolds risk score [8] [†]												
Model A—training data	8.4	0.0025	31.1	<0.0001	0.70	0.0062	9	39.7	<0.0001	11.9	0.22	
Model B—training data	5.7	0.012	20.2	<0.0001	0.40	0.011	8	22.6	0.0039	8.3	0.41	
Model A—test data	6.0	0.12	19.0	0.013	0.61	0.11	10	22.0	0.015	11.7	0.31	
Model B—test data	4.7	0.17	12.0	0.078	0.34	0.21	10	18.5	0.047	12.3	0.27	
WHS HDL model [3]	7.3	0.0001	22.6	<0.0001	0.46	<0.0001	8	30.1	0.0002	6.8	0.56	
Framingham HDL model [1]	12.1	0.0003	22.8	<0.0001	0.9	0.008	5	62.7 [‡]	<0.0001	2.2 [‡]	0.81	

* Among those classified at intermediate risk in reduced model.

[†] Versus model with ATP III covariates refit to WHS data (from Table 4 of Ridker *et al.* [8]).

[‡] Approximate. Assumes average predicted risk in categories equal to observed risk in diagonal categories for each model.

COMMENTARY

this matter? What is a meaningful level of reclassification? Will we be able to develop guidelines to determine whether new variables are important? Can this be applied to case-control studies? Because the NRI is essentially a function of the rank and not the predicted probabilities, it may be useful in this setting even though the categories may not have the same clinical interpretation. Van der Steeg *et al.* [9] present a reclassification table for a case-control study and find that many cases and controls were incorrectly reclassified in models using the Apo B/Apo A-1 ratio *versus* the TC/HDL ratio. The NRI for their data is 1.7 per cent ($p = 0.20$), suggesting less improvement than in models considered here. All the above questions are pertinent to the adoption by the clinical community of the NRI or other methods based on reclassification. In the meantime, the test of calibration, a variation of the well-known Hosmer-Lemeshow test, is available as a useful tool to assess reclassification tables in prospective cohort studies.

The most basic and commonly used measure of biologic effect is the relative risk, expressed as a ratio of rates, proportions, odds, or hazards. However, many clinical and treatment decisions, including recommendations from the ATP-III [10], are made with absolute risk in mind. The relative risk and the ability to reclassify risk more accurately go hand in hand. When translating from relative to absolute risk, predictors by definition will have most impact at high- or intermediate-risk levels. Thus, despite its limited impact on the ROC curve [11], the relative risk estimate should not be dismissed out of hand. Because change in intermediate risk will also have the largest clinical impact, it makes sense to consider how these intermediate categories change and whether this is done correctly. Ultimately, this information can be used, together with cost-effectiveness analysis, to determine biomarkers and other predictors that have the largest potential for clinical impact.

REFERENCES

1. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new biomarker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, in this issue. DOI: 10.1002/sim.2929.
2. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Annals of Internal Medicine* 2006; **145**:21–29.
3. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; **115**:928–935.
4. Moskowitz CS, Pepe MS. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. *Clinical Trials* 2006; **3**:272–279.
5. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* 1997; **16**:965–980.
6. D'Agostino RB, Griffith JL, Schmidt CH, Terrin N. Measures for evaluating model performance. *Proceedings of the Biometrics Section*. Biometrics Section, American Statistical Association: Alexandria, VA, 1997; 253–258.
7. Ridker PM, Cook NR, Lee IM *et al.* A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. *New England Journal of Medicine* 2005; **352**:1293–1304.
8. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women. *Journal of the American Medical Association* 2007; **297**:611–619.
9. van der Steeg WA, Boekholdt SM, Stein EA *et al.* Role of the apolipoprotein B-apolipoprotein A-1 ratio in cardiovascular risk assessment: a case-control analysis in EPIC-Norfolk. *Annals of Internal Medicine* 2007; **146**:640–648.
10. 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). *Journal of the American Medical Association* 2001; **285**:2486–2497.
11. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* 2004; **159**:882–890.