Published in final edited form as:
Int J Clin Pract. 2010 December ; 64(13): 1723–1727. doi:10.1111/j.1742-1241.2010.02469.x.

Methods for Evaluating Novel Biomarkers – a New Paradigm

Nancy R. Cook, ScD
Division of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

In late 2006 the New England Journal of Medicine (NEJM) published a paper that evaluated a series of novel plasma biomarkers for the prediction of cardiovascular disease (CVD) using data from the Framingham Heart Study.[1] The authors concluded that the use of 10 novel biomarkers resulted in only small increases in the ability to assess risk. More recently in July of 2009, some of the same authors analyzed a set of six biomarkers using data from Malmo, Sweden.[2] The more recent paper again concluded that gains over conventional risk factors in predicting CVD were minimal. What may not be obvious is that between the first and second papers, a sea change in the evaluation of biomarkers occurred. While the papers both concluded that novel biomarkers add little to prediction, the statistical methods they used were very different; new methods were developed to compare risk prediction models in the interim. In this perspective the change in methodology will be described and whether the conclusions are justified will be discussed.

ROC Curves and Model Discrimination

For many years the standard for comparing risk prediction models has been the Receiver Operating Characteristics (ROC) curve. This is a plot of sensitivity versus 1-specificity across a range of cut points for a biomarkers or model. It shows how well a model can discriminate or separate the cases and controls. The area under the curve (AUC) has a value between 0.50 and 1.0, with 0.5 indicating no discrimination and with 1.0 indicating perfect discrimination. A good fitting model, at least in the field of cardiovascular disease risk prediction, has an AUC close to 0.8. This would indicate that the probability of a case being assigned a higher risk score than a control is 80%.[3]

Pepe et al showed that it is difficult for an individual biomarker to improve the AUC even if it has a strong association with disease.[4] A factor would have to have a relative risk near 16 per 2 standard deviations to offer adequate discrimination, and this threshold may be higher when adding biomarkers to a set of established markers. Extremely strong associations are needed to show improvement based on the AUC.

The limitations of the ROC as a tool for model evaluation have recently been addressed, however.[5] The AUC is, in fact, an insensitive measure of improvement in model accuracy. Even traditional well-accepted risk factors for cardiovascular disease, such as smoking, cholesterol measures and blood pressure, would individually have little impact on the AUC. If these traditional measures were being evaluated using the ROC curve, we wouldn't conclude that they added to predictive ability and wouldn't be treating patients based on these measures.

Address for correspondence: Division of Preventive Medicine, 900 Commonwealth Ave. East, Boston, MA 02215. Phone: 617-278-0796; Fax: 617-264-9194; ncook@rics.bwh.harvard.edu..

Supported by grants from the Donald W. Reynolds Foundation (Las Vegas, NV) and NHLBI BAA award contract number HHSN268200960011C. The overall Women's Health Study is supported by grants (HL-43851 and CA-47988) from the National Heart Lung and Blood Institute and the National Cancer Institute, both in Bethesda, MD.

Risk Reclassification

Since publication of the 2006 NEJM paper, a new paradigm for comparing models has been developed based on the concept of risk reclassification.[5-7] Risk strata are frequently used in clinical research and practice to determine the course of therapy or to determine the cost effectiveness of various interventions. The ATP III guidelines for cholesterol-lowering agents use predicted risk strata based on the Framingham risk score.[8] The new methods compare risk strata formed from models with and without the biomarkers or other factors of interest and determine which model leads to the most accurate classification of risk. The categories used to form strata should be meaningful such that changes in category reflect important changes or potential differences in treatment.

An example of a risk reclassification table is shown in Table 1.[6] This compares models predicting CVD with and without systolic blood pressure (SBP) in data from the Women's Health Study (WHS).[9] Even though blood pressure is one of the strongest and most important risk factors for CVD, it too has little effect on the ROC curve. The Reynolds Risk Score,[10] which included blood pressure along with several other risk factors for cardiovascular disease, had an overall AUC of 0.80. A model including all terms except for SBP had an AUC of 0.79, a difference of only 0.01. If SBP were a novel biomarker, one could argue that there is no need to measure it due to the small change in the AUC.

The risk reclassification table, however, leads us to a different conclusion. When the women are classified into categories of <5%, 5 to <10%, 10 to <20% and 20% or greater 10-year risk, many women fall into different categories depending on which model is used. Among those at the lowest risk, only 3% move up into a higher category. More changes occur among those at intermediate risk based on the model without SBP, with over 35% in the two middle categories changing risk stratum. Among those in the highest stratum of risk, 21% moved down to a lower category. Thus, a substantial proportion of women at intermediate or high risk would change risk strata and would potentially receive different treatments.

Evaluating Reclassification

The proportion reclassified by itself, however, is not a sufficient measure of improvement. To determine whether such changes in risk are more accurate, several measures have been developed. The first assesses calibration, or how closely the predicted risk agrees with the observed risk within categories. We can first roughly evaluate how well these agree by examining the observed proportions that have the event within each cell of the table. For example, in those who were predicted to have <5% risk by both models the actual proportion with events was 1.3%, which falls within the 0-5% range. In those who were predicted to be at 5-<10% risk by the model without SBP but who were reclassified into the 5-10% category by the model including SBP, 6.8% of women experienced the event, so the observed risk fit into the reclassified 5-10% category. With some exceptions women who were reclassified had an observed risk closer to that predicted by the new risk stratum.

The agreement can more formally be assessed by the reclassification calibration (RC) statistic, which is a chi-square test of how well the average predicted risk within each cell agrees with the observed proportion of individuals who experience the event in cells with at least 20 people.[6] For the model without SBP the value of the RC statistic was 68.3, leading to a p-value of <0.001. This indicates that this model did not provide a good fit since the observed and predicted risks were significantly different. On the other hand, the model including SBP had a chi-square value of 22.9 (p=0.006). While this still exhibits a deviation in fit, the observed and average predicted values were much closer and the value of the chi-square statistic much smaller.

Another way to compare models based on the reclassification table breaks up the data into cases and controls as in Table 2. If a model is better, we would expect cases to move up in risk category and controls or non-cases to move down. In this table, 99 cases moved to a higher risk category but 40 moved down, leading to a net improvement of 50/560 = 10.5% among the cases. Among the controls, 821 moved down, which is an improvement, but 992 moved up, leading to a net worsening of 71/23,611 = 0.7%. Altogether the net improvement is 10.5% - 0.7% = 9.8% ($p<0.001$), which is called the net reclassification improvement, or NRI.[7]

Unlike the RC statistic which assesses calibration, the NRI is a measure of discrimination, or separation of cases and controls. Only the ranks of the categories are needed in its calculation. Some authors have used quantiles, such as quartiles to define the risk strata; the NRI could be computed using these in a case-control study. Since estimates of absolute risk are generally not available in a case-control study, however, the RC statistic cannot be used in that setting.

On the other hand, the RC statistic is more readily adapted to survival data. Cox models and Kaplan-Meier survival curves can be used to assess calibration. It is more difficult to adapt the NRI since it inherently separates the data into cases and controls, and censored observations are not taken into account. Some corrections for this have recently been proposed.[11]

An alternative measure not based on categories is called the integrated discrimination improvement (IDI).[7] This measure is based on the Yates slope, or the difference in predicted probabilities for cases and controls. Ideally the predicted risk would be much higher in cases. The IDI is the difference in such slopes between models. While this measure has some attractive statistical properties,[12] it is difficult to interpret because the size of effect is typically very small. In the WHS data, the IDI comparing models with and without SBP was 0.5% ($p=0.001$),[6] meaning that the net difference in predicted probabilities of cardiovascular disease was only 0.005 between the two models.

Do Novel Biomarkers Make a Difference?

In evaluating the ten novel biomarkers, the analysis of Wang et al used differences in the AUC to determine whether the new markers improved risk prediction. As described above, however, this measure does not directly translate into clinical use.[13] In the same paper the authors showed survival curves by level of the multi-marker score. There was a clear separation of curves, and those with high scores had risks that were at least four times that of those with the lowest scores; risk of mortality over eight years varied from approximately 3% to 20% across score groups, a substantial difference in risk.

The 2009 paper from Malmo instead used the newer methods of comparing models. Besides computing the AUC, it also evaluated discrimination using the NRI and IDI. They again found that there was no overall improvement with the addition of six biomarkers, two of which remained statistically significant in multivariable models.

Can we conclude that novel biomarkers are not effective in risk prediction based on these data? The answer is yes and no. While the methods used now directly relate to clinical utility in terms of potential treatment of patients, there are some caveats to consider when interpreting these analyses. First, the power of the study to detect differences is unclear. While 418 total CVD events occurred in the Malmo study, only 364 of these had all biomarkers, and only 238 of these (65%) were included in the analysis of the NRI and IDI. This may explain some of the discrepancy in the significance of results from the Cox model and the IDI, which should usually be quite similar.[12]

Second, it would be preferable to also report the NRI and IDI associated with the traditional risk factors, such as blood pressure and cholesterol measures. The purpose would not be to judge the importance of these well-established risk factors or to provide substitutes, but rather to provide a basis of comparison for the new markers.[14] If an analysis cannot detect the usual strong importance of these well-known and established factors, then there is little hope for novel markers, which are likely to be less strong.

Third, the categories used can make a difference in the analyses. The original papers discussing risk reclassification[5, 10] used category cut points at 5, 10, and 20%, while others use three categories with cutoffs at 5 and 20%.[2, 7] The category definitions can change the estimated effects, especially the percent reclassified and the NRI.[6] For example, the percent reclassified in models with and without SBP was 8.2% using four categories and 6.2% using three. The NRI was 9.8% ($p<0.001$) using four categories as described above, but half that, or only 5.0% ($p=0.005$) using three.[6] The reclassification calibration test is less affected by these changes since it is a test rather than an effect estimate, and the degrees of freedom are adjusted along with the number of cells. Whether using four or three categories, the test provided evidence that the model with SBP was better calibrated for individuals within these categories. Ideally the categories would correspond to those used for clinical decisions and could aid in cost-effectiveness analyses. However, even if categories change, the RC statistic can assess calibration within categories of potential clinical importance.

Finally, the results found by Melander et al for the set of biomarkers in Malmo is in contrast to results from other studies, including the WHS,[10] the Physicians' Health Study,[15] and the Framingham Heart Study.[16] Similar reclassification analyses were conducted for CRP alone in each of these studies and found a significant improvement in reclassification, whether assessed with the RC statistic or the NRI. It is thus surprising that a similar analysis involving CRP along with other biomarkers showed no improvement overall. This is perhaps due to the smaller size of the Malmo Study along with an associated lower power.

Conclusion

Risk reclassification measures are now being used in a variety of fields to assess the addition of markers or other predictors to risk prediction equations. While the ROC curve and its summary AUC can provide useful information and should not be abandoned, the new measures can more directly assess the clinical implications and consequences of a new model. They may be particularly useful for cost-effectiveness analyses since they can easily show how many patients would fall into differing treatment groups with a new model along with the actual absolute risk among each group of patients. The ultimate goal of risk prediction is to assign the right therapy to the right group of patients. Risk reclassification measures offer an advance in model evaluation by more directly addressing this goal.

References

1. Wang TJ, Gona P, Larson MG, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. N Engl J Med 2006;355:2631–2639. [PubMed: 17182988]
2. Melander O, Newton-Cheh C, Almgren P, et al. Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. JAMA 2009;302:49–57. [PubMed: 19567439]
3. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29–36. [PubMed: 7063747]
4. 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. Am J Epidemiol 2004;159:882–890. [PubMed: 15105181]

5. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928–935. [PubMed: 17309939]
6. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Ann Intern Med 2009;150:795–802. [PubMed: 19487714]
7. Pencina MJ, D'Agostino RBS, D'Agostino RBJ, Vasan RS. Evaluating the added predictive ability of a new biomarker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27:157–172. [PubMed: 17569110]
8. 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 May 16;2001 285(19):2486–2497. [PubMed: 11368702]
9. Ridker PM, Cook NR, Lee IM, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. N Engl J Med 2005;352:1293–1304. [PubMed: 15753114]
10. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women. JAMA 2007;297:611–619. [PubMed: 17299196]
11. Steyerberg EW, Pencina M. Reclassification calculations with incomplete follow-up (letter). Ann Intern Med. Available at http://www.annals.org.ezp-prod1.hul.harvard.edu/cgi/eletters/150/11/795.
12. Pepe MS, Feng Z, Gu JW. Comments on 'Evaluating the added predictive ability of a new biomarker: from area under the ROC curve to reclassification and beyond'. Stat Med 2008;27:173–181. [PubMed: 17671958]
13. Moons KGM, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Acad Radiol 2003;10:670–672. [PubMed: 12809422]
14. Cook NR. Biomarkers for prediciton of cardiovascular events (letter). JAMA 2009;302(19):2089. [PubMed: 19920231]
15. Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. Circulation 2008;118:2243–2251. [PubMed: 18997194]
16. Wilson PWF, Pencina M, Jacques P, Selhub J, D'Agostino R, O'Donnell CJ. C-reactive protein and reclassification of cardiovascular risk in the Framingham Heart Study. Circ Cardiovasc Qual Outcomes 2008;1:92–97. [PubMed: 20031795]

Table 1

Reclassification table comparing 10-year risk strata for models including risk factors for cardiovascular disease in the Women's Health Study but with and without systolic blood pressure.*

Model without SBP		Model Including SBP						Total % reclassified into new risk
		0-<5%	5-<10%	10-<20%	20%+	Total	%	
0-<5%	N	20372	696	23	0	21091	85.9	
	%	96.6	3.3	0.1	-			3.4
	Observed risk (%)	1.3	6.8	0.0	-			
5-<10%	N	635	1441	307	7	2390	9.7	
	%	26.6	60.3	12.8	0.3			39.7
	Observed risk (%)	4.4	8.4	14.6	17.5			
10-<20%	N	4	204	519	90	817	3.3	
	%	0.5	25.0	63.5	11.0			36.5
	Observed risk (%)	0.0	4.3	14.3	34.2			
20%+	N	0	2	54	204	260	1.1	
	%	-	0.8	20.8	78.5			21.5
	Observed risk (%)	-	0.0	25.9	29.4			
Total	N	21011	2343	903	301	24558	100.0	
	%	85.6	9.5	3.7	1.2	100.0		

*Adapted from Cook and Ridker.[6] SBP = systolic blood pressure.

Int J Clin Pract. Author manuscript; available in PMC 2011 December 1.

Table 2

Reclassification table comparing 10-year risk strata for models including risk factors for cardiovascular disease in the Women's Health Study but with and without systolic blood pressure, by case-control status.*

Model without SBP	Model Including SBP					
	0-<5%	5-<10%	10-<20%	20%+	Total	%
Cases as of 8 yrs						
0-<5%	218	38	0	0	256	45.7
5-<10%	22	96	36	1	155	27.7
10-<20%	0	7	59	24	90	16.1
20%+	0	0	11	48	59	10.5
Total	240	141	106	73	560	
%	42.9	25.2	18.9	13.0		
Non-cases as of 8 yrs						
0-<5%	19933	642	23	0	20598	87.2
5-<10%	589	1291	263	6	2149	9.1
10-<20%	4	188	434	58	684	2.9
20%+	0	2	38	140	180	0.8
Total	20526	2123	758	204	23611	
%	86.9	9.0	3.2	0.9		

*Adapted from Cook and Ridker.[6] SBP = systolic blood pressure. Uses case-control status as of 8 years, excluding those who are censored.