Clinical Chemistry 54:1
17–23 (2008)

Mini-Review

Statistical Evaluation of Prognostic versus Diagnostic Models: Beyond the ROC Curve

Nancy R. Cook*

BACKGROUND: Diagnostic and prognostic or predictive models serve different purposes. Whereas diagnostic models are usually used for classification, prognostic models incorporate the dimension of time, adding a stochastic element.

CONTENT: The ROC curve is typically used to evaluate clinical utility for both diagnostic and prognostic models. This curve assesses how well a test or model discriminates, or separates individuals into two classes, such as diseased and nondiseased. A strong risk predictor, such as lipids for cardiovascular disease, may have limited impact on the area under the curve, called the AUC or c-statistic, even if it alters predicted values. Calibration, measuring whether predicted probabilities agree with observed proportions, is another component of model accuracy important to assess. Reclassification can directly compare the clinical impact of two models by determining how many individuals would be reclassified into clinically relevant risk strata. For example, adding high-sensitivity C-reactive protein and family history to prediction models for cardiovascular disease using traditional risk factors moves approximately 30% of those at intermediate risk levels, such as 5%–10% or 10%–20% 10-year risk, into higher or lower risk categories, despite little change in the c-statistic. A calibration statistic can asses how well the new predicted values agree with those observed in the cross-classified data.

SUMMARY: Although it is useful for classification, evaluation of prognostic models should not rely solely on the ROC curve, but should assess both discrimination and calibration. Risk reclassification can aid in comparing the clinical impact of two models on risk for the individual, as well as the population.

From the Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, and Department of Epidemiology, Harvard School of Public Health, Boston, MA.
* Address correspondence to the author at: ncook@rics.bwh.harvard.edu.
Received August 15, 2007; accepted October 29, 2007.
Previously published online at DOI: 10.1373/clinchem.2007.096529

Diagnostic and prognostic models are quite common in the medical field, and have several uses, including distinguishing disease states, classification of disease severity, risk assessment for future disease, and risk stratification to aid in treatment decisions. These two types of models, however, have different purposes. Diagnosis is concerned with determining the current state of the patient and accurately identifying an existing, but unknown, disease state. One may be interested, for example, in distinguishing cases of myocardial infarction from those with more minor symptoms, or those with early-stage cancer from those without.

Prognostic models add the element of time *(1)*. Although typically in medical terms prognosis refers to the most likely clinical course of a diseased patient, the term can also be applied to the prediction of future risk in a normal population. Except in rare instances, both of these settings include a stochastic element, one that is subject to chance *(2)*. Prognostication and prediction involve estimating risk, or the probability of a future event or state. The outcome not only is unknown, but does not yet exist, distinguishing this task from diagnosis. Clinically, prognostic models are most often used for risk stratification, or for assigning levels of risk *(3)*, such as high, intermediate, or low, which may then form the basis of treatment decisions. A well-known example of a prognostic model is the Framingham risk score, which predicts the 10-year risk of cardiovascular disease *(4)*.

Screening for early detection of disease is conducted for diagnostic purposes. In cancer screening, the aim of mammography or colonoscopy, for example, is to find evidence of small, but existing, tumors before clinical symptoms develop. In screening for cardiovascular disease, however, screening is often conducted to detect risk factors for disease. Blood pressure or cholesterol screening detects levels that lead to higher risk of later myocardial infarction or stroke. The results of the screening are then used in prognostic models for later cardiovascular events.

MODEL EVALUATION

Evaluation of models for medical use should take the purpose of the model into account. In diagnostic models, the goal is the accurate classification of individuals into their true disease states. In prognostic models, however, the goal is more complex. While correctly

predicting whether a future event will occur is of interest, it is more difficult owing to its stochastic nature. Accurately estimating the risk itself, and accurate classification into risk strata, is often the best that can be achieved in this setting.

Model accuracy has several aspects, and is often described by two components corresponding to the above goals, namely discrimination and calibration *(5)*. Discrimination is the ability to separate those with and without disease, or with various disease states. It is the ultimate goal of diagnostic models that aim to classify individuals into categories. Calibration, on the other hand, is the ability to correctly estimate the risk or probability of a future event. It measures how well the predicted probabilities, usually from a model or other algorithm, agree with the observed proportions later developing disease. The use of the term is analogous in clinical chemistry when laboratory measurements are compared to a known standard. In modeling, the standard is the observed proportion. Calibration is most suited to the prognostic setting where we would like to predict risk in the future. Although discrimination or accurate classification is of most importance in diagnosis, both discrimination and calibration are of prime interest in prognostication or risk prediction.

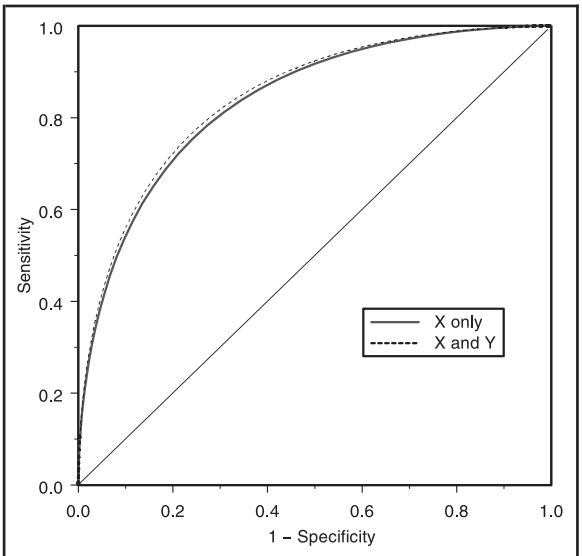

Fig. 1. ROC curves for model with a variable X with an odds ratio of 16 per 2 standard deviation units (solid line) and for a model with X and a second independent predictor Y with an odds ratio of 2 per 2 standard deviation units (dashed line).

DISCRIMINATION

When a single binary diagnostic test is used to predict disease or no disease, we can use a simple 2-by-2 table to assess how well the test classifies when the disease state is known by other means, generally by using a more invasive or expensive gold standard, such as a biopsy. The sensitivity (or the probability of a positive test among those with disease) and the specificity (or the probability of a negative test among those without disease) can easily be computed or assessed. In comparing tests, we prefer those that are higher in both sensitivity and specificity. Because one test may have higher sensitivity but lower specificity than another, the diagnostic likelihood ratio is sometimes used to combine these measures. The likelihood ratio for a positive test result is defined as LR(+)[1] = sensitivity/(1 − specificity), and the likelihood ratio for a negative test result is defined as LR(−) = (1 − sensitivity)/specificity. Although sensitivity and specificity are thought to be unaffected by disease prevalence, they may be related to such factors as case mix, severity of disease *(6)*, and selection of control subjects, as well as measurement technique and quality of the gold standard *(7)*.

More typically, however, the test is not a simple binary one, but may be a continuous measure, such as blood pressure or level of plasma protein. In addition, multimarker models can be used to develop a continuous score or function of a set of risk predictors. A positive test could be defined by classifying those with scores above a given cut point into one category, such as diseased, and those with lower scores into the other, such as nondiseased. Sensitivity and specificity can be defined for the given cut point.

An alternative is to consider the whole range of scores arising from the model. The most popular measure of discrimination using such a range is the receiver operating characteristic (ROC) curve, a plot of sensitivity vs 1 − specificity *(8)*. A typical ROC curve is shown in Fig. 1. Area under the curve (AUC) is also known as the c-statistic or c index, and can range from 0.5 (no predictive ability) to 1 (perfect discrimination). The c-statistic for models predicting 10-year risk of cardiovascular disease among a healthy population is often in the range 0.75 to 0.85.

The c-statistic is based on the ranks of the predicted probabilities and compares these ranks in individuals with and without disease. It is related to the Wilcoxon rank-sum statistic *(9)* and can be computed and compared using either parametric or nonparametric methods *(10)*. Other features of the ROC curve may be of interest in particular applications, such as

[1] Nonstandard abbreviations: LR, likelihood ratio; ROC, receiver operating characteristic; AUC, area under the curve; OR, odds ratio; NRI, net reclassification index.

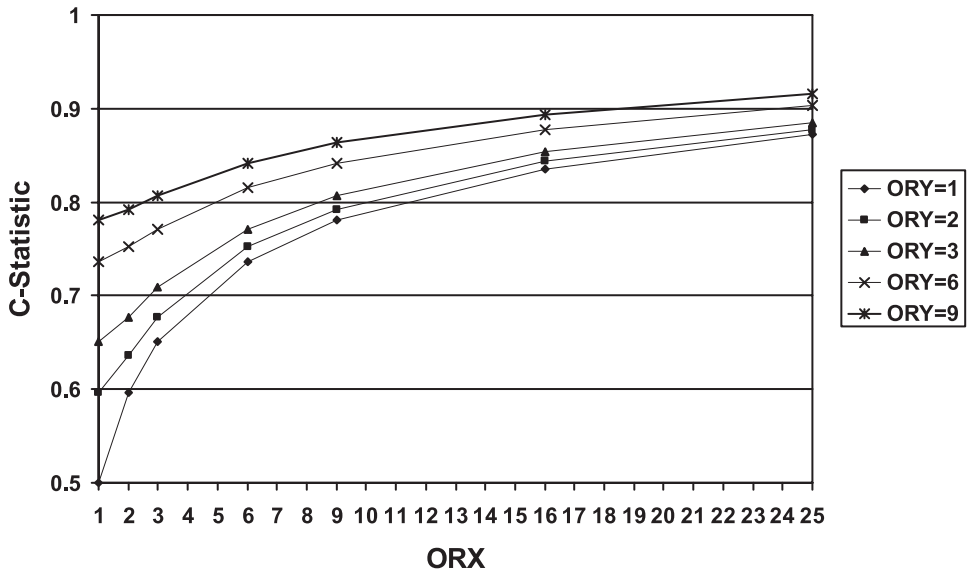

Fig. 2. The effect on the c-statistic of adding an independent variable Y to a model including variable or risk factor score X as a function of odds ratios per 2 standard deviation units for X (ORX) and Y (ORY).

the partial AUC *(11)*, which could be used, for example, when the specificity for a cancer screening test must be above a threshold to be clinically useful *(12)*. The curve may also be used to estimate an optimal threshold for clinical use, such as that which maximizes both sensitivity and specificity. The optimal threshold, however, should also be a function of the relative costs of misclassifying diseased and nondiseased individuals.

The ROC curve does not use the estimated probabilities themselves, however, so it may not be sensitive to differences in probabilities between models. For example, suppose that the predicted probabilities for a control with rank 10 and a case with rank 20 are 0.01 and 0.02, respectively, whereas those for a control with rank 80 and a case with rank 90 are 0.20 and 0.40, respectively. The influence of the two pairs on the c-statistic would be the same, despite the much larger difference in predicted probabilities in the latter pair.

ROC CURVES AND ODDS RATIOS

The odds ratio (OR), or alternatively, the rate ratio or hazards ratio, relating a predictor to a disease outcome, may have limited impact on the ROC curve and c-statistic *(13)*. To have an impact on the curve, the OR for an individual measure or score needs to be sizeable, such as 16 per 2 SD units, roughly corresponding to comparing upper and lower tertiles *(13)*. This size effect is achievable with a risk score, such as the Framingham risk score *(4)*, but is unlikely to be achievable for many individual biologic measures. In cardiovascular disease, the individual components of the Framingham score, such as total and low-density lipoprotein cholesterol, systolic blood pressure, or even smoking, all have far smaller hazard ratios, typically in the range 1.5 to 2.5 *(4)*, clinically important but unlikely individually to have an impact on an ROC curve. In Women's Health Study data, the hazard ratio per 2 SD of systolic blood pressure is only 2.2 given the other components of the score *(14)*. The hypothetical impact of such an effect can be seen in Fig. 1. While an OR of 2 is quite sizeable, there is little change in the curve.

The ROC curve and c-statistic are insensitive in assessing the impact of adding new predictors to a score or predictive model *(14)*. Suppose that there is a set of traditional markers that form a score denoted by X, and adding a new marker Y to the score is under consideration. The change in the ROC curve depends on both the predictive ability of the original set and the strength of the new marker, as well as the correlation between them. If X and Y are independent, the c-statistic is simply a function of the ORs (expressed here per 2 SDs) for each marker. In Fig. 1, the OR for X is 16, and that for Y is 2. The AUC is 0.84 for both the model with only X and the model with both X and Y. Fig. 2 shows the impact on the c-statistic for different combinations of ORs for X and Y. Whereas the c-statistic increases with the OR for Y, the change in the c-statistic decreases as the OR for X increases. Thus, the impact of a new predictor on the c-statistic is lower when other strong

predictors are in the model, even when it is uncorrelated with the other predictors. Lipid measures, which are accepted measures in cardiovascular risk prediction, have ORs closer to 1.7 *(4, 14)*, leading to very little change in the ROC curve. In a more extreme example, Wang et al. *(15)* examined a risk score for cardiovascular disease that was based on multiple plasma biomarkers. Estimates of 8-year risk of all-cause mortality in the high risk (top 20% of risk scores) and low risk (bottom 40% of risk scores) groups were 20% and 3%, respectively, indicating important differences in predicted risk. However, incorporation of these plasma biomarkers (with a multivariate hazard ratio of 4) into a risk function led to little improvement in the c-statistic compared with conventional risk factors alone.

CALIBRATION

Alternatively, calibration concerns itself directly with the estimated probabilities or predictive values. The positive predictive value is defined as the probability of disease given a positive test result, and the negative predictive value is the probability of no disease given a negative test result. When a risk score is used, the continuous analog is the probability of disease given the value or range of the score. An assessment of calibration directly compares the observed and predicted probabilities. Because "observed risk" or proportions can only be estimated within groups of individuals, measures of calibration usually form subgroups and compare predicted probabilities and observed proportions within these subgroups. The most popular measure of calibration, the Hosmer-Lemeshow goodness-of-fit test *(16)*, forms such subgroups, typically using deciles of estimated risk. Within each decile, the estimated observed proportion and average estimated predicted probability are estimated and compared. The statistic has a χ^2 distribution with $g - 2$ degrees of freedom, where g is the number of subgroups formed. Although deciles are most commonly used to form subgroups, other categories, such as those formed on the basis of the predicted probabilities themselves (such as 0 to <5%, 5 to <10%, etc.), may be more clinically useful.

Because groups must be formed to evaluate calibration, this test is somewhat sensitive to the way such groups are formed *(17)*. Ideally the predicted probability would estimate the underlying or true risk for each individual (perfect calibration). Since we cannot know the underlying risk, but can only observe whether the individual gets the disease, a stochastic event, the Hosmer-Lemeshow statistic is a somewhat crude measure of model calibration.

In diagnostic testing and modeling, calibration is typically not of as much interest as discrimination. Samples are often not population-based, and the predicted probabilities may be applicable only to the patients sampled. Predictive values depend on disease prevalence, so unless a population sample is used or a valid estimate of prevalence is available, the sensitivity and specificity are of greater interest. In estimating future risk, however, as in prognostic models, the actual risk itself is of greatest concern, and calibration, as well as discrimination, is important. The patient and clinician are interested in the future risk of disease rather than the probability of a positive test *(18)*.

CLINICAL RECLASSIFICATION

In clinical prognostic models, risk stratification is important for advising patients and making treatment decisions. The ATP III guidelines *(19)*, for example, suggest cholesterol-lowering medications for individuals with predicted risk scores above 20% based on Framingham risk models. In comparing models, we would prefer those that stratify individuals correctly into the various categories (i.e., those that are better calibrated). We would also prefer those that are able to classify more into the highest and lowest risk categories (i.e., that better discriminate), as long as these are accurate classifications.

The distribution of predicted values from each model separately, or the marginal distribution, can describe how many are classified into intermediate risk categories, but not whether this is done correctly. They also do not describe whether one model is better at classifying individuals, or if individual risk estimates differ between two models. One way of evaluating this is to examine the joint distribution through clinical risk reclassification *(14, 20)*. This method classifies predicted risk estimates into clinically relevant categories and cross-classifies these categories, such as in Table 1.

As an example, suppose that a model is formed using traditional risk factors with score X as above, and a new model includes the risk factors in X along with a new independent biomarker Y. In the example in Table 1, 10 000 simulated observations were generated using an initial risk score X with an odds ratio of 16 per 2 SDs, and a new uncorrelated biomarker Y with an OR of 2 per 2 SDs, with an overall risk of disease of 10%. The rows of Table 1 represent the model based on X only, and the columns represent the model including both X and Y. The categories represented are based on ones suggested for 10-year risk of cardiovascular disease *(19, 21)*.

The percent reclassified can be used as an indication of the clinical impact of a new marker, and will likely vary according to the original risk category. The total percentages reclassified into new risk categories in Table 1 were 6%, 38%, 35%, or 15%, depending on the

Table 1. Comparison of predicted risks in models including a variable or risk factor score X with an OR of 16 per 2 SD units with and without a new biomarker Y with an OR of 2, assuming an overall disease frequency of 10% in a simulated cohort of 10 000 individuals.

Model with X only		Model with X and Y				Total % reclassified
		0% to <5%	5% to <10%	10% to <20%	20%+	
0% to <5%	N	4728	314	3	0	
	% Reclassified	93.7	6.2	0.1	0.0	6.3
	Ave Pr(D|X), %	1.9	4.2	4.5	–	
	Ave Pr(D|X, Y), %	2.1	6.1	10.3	–	
	Observed risk, %	1.6	8.3	–	–	
5% to <10%	N	473	1285	296	1	
	% Reclassified	23.0	62.5	14.4	0.1	37.5
	Ave Pr(D|X), %	6.2	7.2	8.3	7.1	
	Ave Pr(D|X, Y), %	4.0	7.1	12.1	21.8	
	Observed risk, %	3.6	6.8	10.8	–	
10% to <20%	N	5	348	1035	213	
	% Reclassified	0.3	21.7	64.6	13.3	35.4
	Ave Pr(D|X), %	12.0	12.2	14.3	17.3	
	Ave Pr(D|X, Y), %	4.2	8.3	14.2	24.0	
	Observed risk, %	–	8.9	14.1	25.4	
20%+	N	0	2	190	1107	
	% Reclassified	0.0	0.2	14.6	85.3	14.7
	Ave Pr(D|X), %	–	20.3	23.7	36.1	
	Ave Pr(D|X, Y), %	–	9.5	17.1	36.8	
	Observed risk, %	–	–	17.9	36.9	

Percentage reclassified are those moved into a new risk category when Y is added to a model including X. Ave Pr(D|X) is the average predicted probability, expressed as a percent, in the model for X only for individuals in that cell. Ave Pr(D|X, Y) is the corresponding percent from the model including both X and Y.

initial risk category. In the two intermediate categories, some individuals moved up and some moved down with the new classification. Whereas in the example simulations here X and Y are uncorrelated, the degree of reclassification will lessen if the markers are highly correlated.

Also shown in the table are the average estimated risks from the two models for each cell. The average predicted risks based on the initial model are denoted by Ave Pr(D|X), and those from the new model are denoted by Ave Pr(D|X, Y). The change in estimated risk for individuals in the off-diagonal categories can be seen by comparing these two numbers. For example, for those initially in the 5 to <10% category, 14% are reclassified to the 10 to <20% category, and the average estimated risk changes from 8% to 12%, which could change recommended treatment under some guidelines.

Besides the percentage reclassified, it is important to verify that these individuals are being reclassified correctly, i.e., that the new risk estimate is closer to their actual risk. This can be examined by comparing the predicted risks from the models to the crude proportion developing events within each cell, or the observed risk. This is a form of calibration, and the numbers in the table show that the observed risk is closer to that estimated from the model with both X and Y than that with X alone. This can be formally tested using a Hosmer-Lemeshow test using the cross-classified rather than the marginal cells *(22)*. To avoid the effect of sparse data, only cells with at least 20 individuals are included. The observed proportions are compared to the predicted risks for each model separately. For the model using just X, the χ^2 statistic is 40.8 with 8 degrees of freedom and $P < 0.0001$, suggesting a lack of fit. For the model with both X and Y, the statistic is 5.8 with 8 degrees of freedom and $P = 0.67$, indicating acceptable fit. Thus Y seems to add important information despite little change in the ROC curve as seen in Fig. 1. This method has been demonstrated for binary outcomes, but it can be extended to survival data by using Kaplan-Meier estimates of survival at a given time point in each

cell along with the predicted survival probability at the same time *t*.

Pencina et al. *(23)* suggest a single measure to summarize the reclassification table. They describe the net reclassification index (NRI) as a measure of change in these clinical categories. They first form separate reclassification tables for cases and controls. They then examine the proportions moving up or down categories among cases and controls separately. The NRI is the difference in proportions moving up and down among cases vs controls, or NRI = [Pr(up | case) − Pr(down | case)] − [Pr(up | control) − Pr(down | control)]. In the example data, the NRI = 5.7% ($P = 0.0003$), indicating that 5.7% more cases appropriately move up a category of risk than down compared with controls. Because this statistic essentially compares rankings of cases and controls, it is a measure of discrimination rather than calibration, useful primarily for case-control data. It has the advantage over the ROC curve, however, that categories can be formed based on clinically important risk estimates.

For clinical use, it is often those in the intermediate-risk categories for whom treatment is questionable. For reasons of cost-effectiveness, it may be preferable to reserve the use of expensive markers or invasive procedures for this group. The middle 2 rows of Table 1 contain those in this gray area who are most likely to benefit from additional measures. In this subgroup, the NRI is 21% ($P < 0.0001$), suggesting that the reclassification may be more important in these individuals.

As an example, in data from the Women's Health Study, a model predicting cardiovascular disease risk that included high-sensitivity C-reactive protein and family history of myocardial infarction, in addition to traditional Framingham risk factors, led to an improvement in risk classification for individuals *(24)*. In those in the intermediate categories of 5%–10% or 10%–20% 10-year risk based on Framingham risk factors only, approximately 30% of individuals moved up or down a risk category with the new model. The overall NRI in test data was 4.7%, whereas that for those at intermediate risk was 12.0% *(22)*. The new risk model was also more accurate in terms of calibration ($P = 0.047$ for the Framingham variables vs $P = 0.56$ for the new model using the Hosmer-Lemeshow test with cross-classified categories), although there was little change in the c-statistic.

Conclusions

The purposes of diagnostic and prognostic models differ; the latter incorporate the added element of time and are stochastic in nature. Because prognostic models are created to predict risk in the future, the estimated probabilities are of primary interest. Instead of relying solely on the c-statistic, methods of model evaluation should accordingly focus on the predicted values and assess whether these are computed accurately. Besides examining these for a single model, when comparing models the joint distribution of risk estimates should be considered. An examination of clinical risk reclassification can describe how a new marker may add to predictive models for clinical use, and statistics such as the NRI and calibration test for the cross-classified categories can be used to more formally assess clinical utility.

Grant/funding Support: Supported by a research grant from the Donald W Reynolds Foundation (Las Vegas, NV).
Financial Disclosures: None declared.

References

1. Windeler J. Prognosis: what does the clinician associate with this notion? Stat Med 2000;19:425–30.
2. Coggon DIW, Martyn CN. Time and chance: the stochastic nature of disease causation. Lancet 2005;365:1434–7.
3. Bigger JT Jr, Heller CA, Wenger TL, Weld FM. Risk stratification after acute MI. Am J Cardiol 1978; 42:202–10.
4. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837–47.
5. Harrell FE Jr. Regression Modeling Strategies. New York: Springer; 2001.
6. Moons KGM, van Es G-A, Deckers JW, Habbema JDF, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiol 1996;8:12–7.
7. Phelps CE, Hutson A. Estimating diagnostic test accuracy using a "fuzzy gold standard." Med Decis Making 1995;15:44–57.
8. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 2007; 115:654–7.
9. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
10. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford; 2003.
11. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989;9:190–5.
12. Baker S. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. J Natl Cancer Inst 2003;95:511–5.
13. 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;159:882–90.
14. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circ 2007;115:928–35.
15. 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;355:2631–9.
16. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. Comm Stat 1980;A10:1043–69.
17. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 1997;16:965–80.
18. Moons KGM, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Acad Radiol 2003;10:670–2.
19. 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;285:2486–97.
20. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. Ann Intern Med 2006;145:21–9.
21. Greenland P, Smith SC Jr, Grundy SM. Improving coronary heart disease risk assessment in asymptomatic people: role of traditional risk factors and noninvasive cardiovascular tests. Circulation 2001;104:1863–7.
22. Cook NR. Comments on 'Evaluating the added predictive ability of a new biomarker: from area under the ROC curve to reclassification and beyond.' Stat Med 2007 Aug 1; Epub ahead of print.
23. 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 2007 Jun 13; Epub ahead of print.
24. 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;297:611–9.