Published in final edited form as:

Biom J. 2011 March ; 53(2): 237–258. doi:10.1002/bimj.201000078.

Performance of Reclassification Statistics in Comparing Risk Prediction Models

Nancy R. Cook and Nina P. Paynter

Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA

Abstract

Concerns have been raised about the use of traditional measures of model fit in evaluating risk prediction models for clinical use, and reclassification tables have been suggested as an alternative means of assessing the clinical utility of a model. Several measures based on the table have been proposed, including the reclassification calibration (RC) statistic, the net reclassification improvement (NRI), and the integrated discrimination improvement (IDI), but the performance of these in practical settings has not been fully examined. We used simulations to estimate the type I error and power for these statistics in a number of scenarios, as well as the impact of the number and type of categories, when adding a new marker to an established or reference model. The type I error was found to be reasonable in most settings, and power was highest for the IDI, which was similar to the test of association. The relative power of the RC statistic, a test of calibration, and the NRI, a test of discrimination, varied depending on the model assumptions. These tools provide unique but complementary information.

Keywords

Calibration; Discrimination; Model accuracy; Prediction; Reclassification

1 Introduction

For many years model comparison in the medical literature has used the area under the receiver operating characteristic (ROC) curve (Hanley and McNeil, 1982), whose analog in the survival context is the c index or c-statistic (Harrell, 2001). This measure, while valuable, can be insensitive to changes in absolute risk estimates (Cook, 2007). It is a measure of discrimination that is directly applicable to the setting of classification, but is less useful in evaluating risk prediction of future events since it is a function only of ranks, not the predicted probabilities themselves (Moons and Harrell, 2003; Janes, *et al.*, 2008).

Within the past few years, new methods for evaluating and comparing the fit of predictive models have been proposed, and have generated considerable interest in both the statistical and medical literature. The new measures are based on reclassification tables, which stratify individuals into risk categories and examine changes in categories under a new model. Measures based on both calibration (Cook, 2008) and discrimination (Pencina, *et al.*, 2008) have been proposed. While methods of model assessment should depend on the model's

Correspondence to: Nancy R. Cook.

Conflict of Interest

The authors have declared no conflict of interest.

Supporting Information for this article is available from the author or on the WWW under <http://dx.doi.org/10.1022/bimj.XXXXXXX>.

intended use, when the model is intended to guide treatment decisions, evaluation of risk strata may be more clinically useful than comparing overall ranks.

The performance of these measures has been demonstrated in applied examples, and they have been described further elsewhere (Gu and Pepe, 2009; Steyerberg, *et al.*, 2009; Cook, 2010; Whittemore, 2010), but a rigorous examination of their distributions and performance characteristics has not yet been done. While not sufficient to determine clinical utility, questions concerning type I error and power relative to commonly used measures are of interest. In addition, there are unique questions in the setting of reclassification in categories. These include the effects of the number of categories and of the category composition on estimates and power. Some of these effects have been explored in simulations (Mihaescu, *et al.*, 2010) and applied examples (Cook and Ridker, 2009; Mealiffe, *et al.*, 2010) and differences due to category definition have been found in some measures.

The current paper explores these characteristics for binary outcomes in a series of simulations using logistic models, and examines the performance of the measures in both the null model and under varying effect sizes. The definitions of the new measures and their variations are presented in Section 2, with a worked example. Simulation results are presented in Section 3, and an example based on data from the Women's Health Study is presented in Section 4.

2 Definitions

The analyses here assume a binary outcome, such as the development of cardiovascular disease (CVD) within a ten-year period of follow-up. Interest centers on the predicted risk, and whether a new marker or model can improve the fit of the model. Often one model is contained in the other, and one or more additional variables are included to the usual predictors. More generally, however, any two models could be compared even if not nested.

2.1 Standard measures

Standard analyses include the evaluation of two components of model accuracy: discrimination and calibration. Discrimination is typically assessed using the c-statistic, which calculates the area under the ROC curve for a continuous predictor and binary outcome. Other properties of the ROC curve may be of interest in particular applications, such as the partial area under the curve given a threshold of specificity (McClish, 1989). When comparing models, changes in the c-statistic have been used to evaluate "clinical utility," although this use is questionable in the setting of risk prediction (Janes *et al.*, 2008). In these analyses the c-statistic was estimated using the Wilcoxon rank sum statistic (Hanley and McNeil, 1982), and the difference in c-statistics was tested using two variance estimators, one based on DeLong, DeLong and Clarke-Pearson (1988), and the other based on Rosner and Glynn (2009).

Calibration is often assessed using the Hosmer-Lemeshow goodness-of-fit test (Hosmer and Lemeshow, 1980). Decile categories are typically used, and the observed risk (proportion) is compared to the average predicted risk within each category using a chi-square statistic. Alternative categories may be used, particularly those based on the estimated risk itself, similar to the H statistic examined by Hosmer and Lemeshow (1980). In the Women's Health Study, more than 90% of the women had risk less than 5%, and interest centered on the fit among those at higher risk. Thus categories based on values of the predicted risk in 2% increments up to 20% were used (Cook, *et al.*, 2006). Direct comparisons of model calibration are not typically done, but the statistic is computed for each model separately. Alternative measures based on smoothed residuals have been proposed (Le Cessie and van Houwelingen, 1991), but are not evaluated here.

2.2 Risk reclassification

Reclassification derives from placing individuals within risk strata, as is common in medical applications. These may be pre-determined based on cost-effectiveness considerations, or on intuitive definition of important risk levels. A reclassification table is a cross-tabulation of risk categories for two models, and shows how individuals move between categories and fall within the same or different risk strata (Cook, 2007). A hypothetical example is shown in Table 1 which shows four estimated risk strata formed from two models fit in a sample of 10,000 individuals. These data were generated from a logistic model with overall probability of disease equal to 10%. Outcomes were generated from the true model, denoted Model XY, which included terms for normally-distributed independent variables X and Y with an odds ratio for X (ORX) of 16.0 per 2 standard deviation units and an OR for Y (ORY) of 3.0 per 2 standard deviation units. Model X included only the term for X and would thus be expected to provide poorer fit. The most obvious summary of the table is the percent of individuals who change categories, here 27% overall. This is a function of the agreement between the two models, and is an estimate of how applying a new model may alter decisions. Clinicians may also be interested in the number changing risk category by initial risk strata, which are 9%, 54%, 52%, and 23% in the four strata in Table 1.

2.3 Reclassification calibration

Whether the changes in category are appropriate, however, is more important than the percent reclassified. Cook (2008) suggested a reclassification calibration statistic using a variation of the Hosmer-Lemeshow chi-square goodness-of-fit test. This compares the observed risk to the average estimated risk within each cross-classified category, and can be considered an assessment of calibration (D'Agostino, *et al.*, 1997). Let p_{1ij} represent the predicted risk in the cell in the i th row and the j th column from the first, or standard, model, and let p_{2ij} represent predicted risk from the new model. The reclassification calibration (RC) statistic for Model X has the form

$$X_{RC1}^2 = \sum_i \sum_{\substack{j \\ n_{ij} \geq 20}} \frac{(O_{1ij} - n_{1ij} \bar{p}_{1ij})^2}{n_{1ij} \bar{p}_{1ij} (1 - \bar{p}_{1ij})} \sim X_{K-2}^2$$

where K is the number of cells. To improve the large-sample characteristics, only cells containing at least 20 individuals are included in the calculation. In addition, we examined the statistic using only those cells containing an average expected number of cases of at least 5, where the mean probability for the two models was averaged in each cell. This test could be performed for each model separately, with an analogous definition for X_{RC2}^2 , the test for Model XY. The test for Model X is of primary interest, but that for Model XY should be conducted for comparison. In Table 1, using the estimated probabilities from each model, the test statistic for Model X is $X_{RC1}^2 = 102.3$ ($p < 0.0001$) and that for Model XY is $X_{RC2}^2 = 14.9$ ($p = 0.24$) both with 12 degrees of freedom. When cells were restricted to those with an average expected value at least five, the degrees of freedom were reduced to 10 and the statistics became 70.7 ($p < 0.0001$) for Model X, and 9.3 ($p = 0.50$) for Model XY.

D'Agostino *et al* (1997) proposed an adjustment to the classic Hosmer-Lemeshow statistic that corrects for small predicted values by adding a small term to each cell. The analog for reclassification would be

$$X^2_{RCAdj1} = \sum_i \sum_{j, n_{ij} \geq 20} \frac{(O_{1ij} - n_{1ij} \bar{p}_{1ij})^2}{n_{1ij} (\bar{p}_{1ij} + 1/n_{ij})(1 - \bar{p}_{1ij} + 1/n_{ij})} \sim X^2_{K-2}$$

which may be more likely to follow a chi-square distribution when there are small expected numbers within cells. Using data from Table 1, the statistic is 86.8 for Model X ($p < 0.0001$) and 13.0 for Model XY ($p = 0.37$) for cells of at least size 20, and 68.5 ($p < 0.0001$) and 8.8 ($p = 0.55$) for average expected numbers of at least 5 for models X and XY, respectively. Pigeon and Heyse (1999) proposed an alternative variation of the classic Hosmer-Lemeshow test, called J^2 with $K-1$ degrees of freedom, which accounts for the under-dispersion within each cell by adding a variance correction φ_{ij} to each cell. Its analog for reclassification is

$$J^2_1 = \sum_i \sum_{j, n_{ij} \geq 20} \frac{(O_{1ij} - n_{1ij} \bar{p}_{1ij})^2}{\varphi_{ij} n_{1ij} \bar{p}_{1ij} (1 - \bar{p}_{1ij})} \sim X^2_{K-1}$$

where

$$\varphi_{ij} = \sum_{k=1}^{n_{ij}} \frac{\widehat{p}_{ijk} (1 - \widehat{p}_{ijk})}{n_{ij} \bar{p}_{ij} (1 - \bar{p}_{ij})}$$

and the summation is over all individuals within the ij th cell. Hosmer and Lemeshow (2000) originally examined this form of the statistic, and concluded that the correction was not necessary for practical purposes. The setting of reclassification, however, would typically uses only 3 or 4 risk strata, and thus the variation of predicted probabilities within strata may be greater. Using data from Table 1, $J^2 = 102.8$ ($p < 0.0001$) for Model X and 15.0 ($p = 0.31$) for Model XY for cell sizes of at least 20. For cells with average expectation of at least five, the statistics were 71.1 ($p < 0.0001$) and 9.3 ($p = 0.59$) for Models X and XY, respectively. The performance of all three forms of this statistic, with both types of size restrictions, was examined in simulations.

2.4 Discrimination improvement

Pencina *et al* (2008) suggested an alternative approach that divides the reclassification table into cases and controls. Ideally, predicted risk for cases would move up and that for controls would move down. They described the net reclassification improvement (NRI) which sums the improvement among cases and controls. It is computed as

$$NRI = [\Pr(up|D=1) - \Pr(down|D=1)] - [\Pr(up|D=0) - \Pr(down|D=0)]$$

where $D = 1$ stands for cases and $D = 0$ stands for controls. This measures discrimination or separation of predicted risk for cases and controls. The margins display the sensitivity and specificity using the category cutoffs as risk thresholds (Janes *et al.*, 2008). Table 1 shows the reclassification among cases and controls necessary to compute the NRI. In these data the NRI using these four risk categories is 8.7% (standard error = 1.8%, $p < 0.0001$).

The value of the NRI depends strongly on the categories used to form risk strata. If the two middle categories in Table 1 are combined, the NRI is 7.1% (standard error = 1.5%, $p < 0.0001$). With only two risk strata defined by risk above and below 10%, the NRI is 2.5% (standard error = 1.1%, $p = 0.030$). In such a setting with only two risk strata, the NRI represents the improvement in sensitivity plus the improvement in specificity. It would then be equivalent to the difference in the Youden index, which is sensitivity plus specificity minus 1. When there are more than 2 risk strata, the NRI adds terms for the incremental sensitivity and specificity at the additional cut points, and would increase with more categories. With infinite strata, this is equivalent to testing whether the predicted value is higher in cases and lower in controls for Model XY vs. model X. This leads to a simple two-by-two table of case status vs. an indicator of whether the probability for Model 2 is higher or lower than that for Model 1. In the example data, the NRI computed in this way is 35.1% (standard error = 3.4%, $p < 0.0001$).

When risk strata are based on clinical cut points, there may be interest in sequential testing or measurement of biomarkers. For example, physicians may be most interested in further testing patients at intermediate risk of disease, whose clinical course is not yet determined. In this situation, a ‘clinical’ NRI (cNRI) has been suggested by restricting the measure to the intermediate risk strata based on Model X (Cook, 2008). However, due to the symmetric nature of the table, this measure may be biased in its simple form. Under the null situation, the table would be expected to be symmetric about the axis of identity, which could lead to an expected value greater than zero. That is, under the null hypothesis of symmetry or equivalence of the two models, the i,j th and j,i th cells would be expected to be equal, both overall and among cases and controls, and the expectation for each cell would be $(c_{ij} + c_{ji})/2$. To correct the bias in the cNRI, one can subtract its expectation under this null hypothesis.

In Table 1 for cases, for example, the expected number for those reclassified from 5–<10% to 0–<5% would be $(25+22)/2 = 23.5$ under the null hypothesis of symmetry. The expected numbers for all cells are shown in Supplementary Table 1. The expected percent of cases in the 5–<10% and 10–<20% strata for Model X who would move up under symmetry is 27.4 and who would move down is 19.2, for an expected reclassification improvement of 8.1%. The expected reclassification improvement among controls is 11.4%, leading to an expected cNRI of 19.6% under the null hypothesis of symmetry. Thus, the original cNRI is 32.2% and the adjusted cNRI is 12.6%. Using the non-null variance estimate, the standard error is 3.8% ($p = 0.0008$).

Pencina *et al.* (2008) also describe the integral of sensitivity and specificity over all possible cutoff values. This leads to the Integrated Discrimination Improvement (IDI) which can be viewed as the sum of differences in these terms between models. The measure is equal to the difference in Yates, or discrimination, slopes between models, where the Yates slope is the difference in average estimated risk between cases and controls in a model. The IDI can then be written as

$$IDI = (\text{ave } \widehat{p}_{D=1} - \text{ave } \widehat{p}_{D=0})_{\text{Model 2}} - (\text{ave } \widehat{p}_{D=1} - \text{ave } \widehat{p}_{D=0})_{\text{Model 1}}$$

where the average is over all predicted probabilities for cases or controls. Pepe, *et al.* (2008a) showed that this is equivalent to change in an R^2 measure representing percent of variance explained in terms of probabilities. While this is true asymptotically (Hu, *et al.*, 2006), Tjur (2009) shows that the Yates slope (which he calls the coefficient of discrimination) is exactly equal to a combination of R^2 measures. In the example data, the IDI, or change in R^2 , is 2.6% (standard error = 0.17%, $p < 0.0001$).

2.4 Predictiveness curves

Pepe *et al.* (2008b) suggested using predictiveness curves to help quantify the improvement in prediction. The curve is a plot of the predicted risks from the model as a function of its ranks, as illustrated in Figure 1 (left) for the example data. It is a cumulative distribution plot of the predicted risk with the axes transposed. The diagram is useful in displaying the full risk distribution as well as the cumulative fractions falling above or below any given risk threshold along the continuum of risk. The curve itself does not use the observed but only the predicted outcomes, however. The model is assumed to be perfectly calibrated, which may not hold exactly, especially in the tails of the distribution.

It is also possible to compare models using predictiveness curves, as illustrated in Figure 1 (left) for the example data. Although the Y variable shows a strong association with the outcome, is more closely calibrated to the outcome proportions in the reclassification table, and has a highly significant NRI and IDI, the curves are remarkably close. Like the ROC curve, the predictiveness curve which is related (Huang and Pepe, 2009), seems insensitive in comparing models. Because it focuses on the marginal distributions, it cannot determine changes in risk for individuals, or potential changes in clinical risk strata. This will be further illustrated with data from the Women's Health Study in Section 4.

3 Simulations

We ran simulations based on a logistic model with two variables, X and Y. We assumed that X was either a single previously known strong predictor or a linear combination of traditional predictor variables. Y was considered a new variable to potentially be added to a predictive model. Both X and Y were assumed to have standard normal distributions, and the odds ratios for X (ORX) were 4, 8, or 16 per two standard deviations (SD), and for Y (ORY) ranged from 1 up to 10 per 2 SD units. The overall proportion PD developing disease was set to 2.5, 5, 10, 20, and 50%, with a sample size of 5000. Other sample sizes corresponding to an average of 500 expected events were considered (i.e., 5,000 for PD = 0.10; 10,000 for PD = 0.05; and 20,000 for PD = 0.025). X and Y were assumed to be independent and uncorrelated predictors, but additional runs allowed correlations of 0.1, 0.2 and 0.4. The latter models defined ORX and ORY as the conditional effects with both terms in the model. One thousand simulations were run under each scenario.

Standard measures of model fit were assessed, including the estimated beta coefficient, the c-statistic, and the Hosmer-Lemeshow test statistic. Reclassification measures were assessed using several different sets of categories, including those previously suggested for 10-year risk of cardiovascular disease with either 4 categories (0–<5%, 5–<10%, 10–<20% and 20%+) or 3 categories (0–<5%, 5–<20%, 20%+). We also considered alternative risk cut-offs, including tertiles, quartiles, and quintiles of risk, as well as an ad hoc categorization with cut points at the average incidence, half the incidence, and twice the incidence, or ½ PD, PD, and 2PD. For PD = 0.10, this was the same as the 4 categories of cardiovascular risk.

3.1 Null model

To examine the estimates and type I error when Y had no effect, we set the OR for Y to 1. Table 2 shows the average estimated statistics in models with X only and with X and Y together, and the 95% intervals over all 1000 simulations for PD = 0.10. The estimated odds ratio for Y was close to 1.0 in all instances, and the average c-statistic was 0.68, 0.76, and 0.81 for ORX = 4, 8, and 16, respectively, with an average difference near 0 across all simulations. The Wald, likelihood ratio and Hosmer-Lemeshow tests displayed a type I error close to 0.05, but the test of the difference in c-statistics was conservative using both the DeLong *et al.* (1988) and Rosner and Glynn (2009) variance estimates (Table 3). The computational time required for the DeLong variance estimate was greater, requiring 38

seconds vs. 0.08 seconds for the Rosner estimate for the sample size of 10,000 in the example data in Table 1. Estimated values of the IDI were close to 0.0 on average, appropriately indicating no improvement in prediction, with a 95% interval of 0.0 to 0.1 or 0.2. As noted by Pepe *et al.* (2008a), the IDI was very close to the difference in R^2 when computed as the difference in correlations of the predicted risk and observed outcome (Hu *et al.*, 2006). The type I error for the IDI was slightly under 5% in the situations considered (Table 3).

When the four CVD categories defined as <5%, 5%–<10%, 10%–<20%, and 20% or more were used, the percent reclassified was 2.7 to 3.5 overall in the situations considered, with about half reclassified correctly in the null models. In the null setting, there was no reclassification in some tables, and reclassification measures were assessed only in tables with reclassification. The NRI could not be computed due to little or no reclassification in 6–9% of simulations. Among those with reclassification, the mean NRI was typically 0.1 to 0.2%, with 95% intervals of width 3 or 4%. The NRI had type I error from 5 to 8%. When three categories were used, the NRI could not be computed in approximately 10% of tables, but the null estimates and type I error were similar. The adjusted cNRI displayed similar estimates and size, but could not be computed in 7–22% of simulations due to lack of reclassification in the middle categories under the null.

The reclassification calibration statistics using cell sizes of 20 or more ranged from about 3.6 to 5.1 under the various scenarios. The type I error was about 5% for the RC statistic but lower for the adjusted version. That for J^2 was too conservative, reaching less than 3% in all situations considered. When three CVD categories were used, the results were similar, although the values of the calibration statistics were lower due to the lower degrees of freedom in this setting. The distribution of p-values and the type I error were similar to those using four categories. In addition, when a correlation was allowed between X and Y, the type I error was very similar for the null model (data not shown).

Additional simulations assessed the fit of the RC statistics to the chi-square distribution in the null situation. When $ORY = 1$, the extent of reclassification was often small; the overall percent reclassified was 3–5% and that reclassified in the two middle categories was also small, ranging from 6–8%. The degrees of freedom associated with the RC test were correspondingly small, sometimes as low as 1 or 2, reflecting the limited amount of reclassification. Such a test may not be justified in this setting. In the simulations approximately 6–9% of null tables displayed no reclassification, depending on ORX.

Adherence to a chi-square distribution under the null setting is displayed in Figure 2 for cells with at least 20 observations and for an average expected value of at least 5. Three scenarios were considered, including tests of Models X and XY when $ORX=8$ and $ORY=1$ and the test of the correct Model XY when $ORX=8$ and $ORY=3$. The percent of p-values falling below 10% or 5% corresponded roughly to 10 or 5%, respectively, for the unadjusted calibration statistic (Figure 2) using either the restriction on cell size or the expected number. The adjusted chi-squares were lower when cell sizes of at least 20 were used, but performance improved for expected values of at least five. Values for the J^2 statistic were conservative throughout.

3.2 Power

Additional simulations assessed the statistics for $ORX = 8$ and under various assumptions about ORY (Table 4). The difference in the c-statistic was small when ORY was less than 2.0. The average difference was 0.005 for $ORY = 1.5$, and 0.03 for $ORY = 3$. The power to detect a difference was 43% for $ORY = 1.5$, 82% for $ORY = 1.75$, and 97% for $ORY = 2$ (Figure 3). The classic Hosmer-Lemeshow statistic did not detect any lack of calibration

when the model was mis-specified using X only, and the power remained close to 5% throughout as in the null model considered above. The IDI ranged from 0.2% to 3.2% for $OR_Y = 1.25$ to 3, and reached 12% for $OR_Y = 10$ (not shown), again mimicking the change in R^2 between the two models. The power for detecting a difference was 54% for $OR_Y = 1.25$ and reached 96% for $OR_Y = 1.5$. The power was high but slightly lower than that for the likelihood ratio test for the regression coefficient throughout. [See Supplementary Table 2 for more detail.]

Using the four CVD categories, the overall percents reclassified ranged from 8% to 35% for $OR_X = 8$ (Table 4), and the percent reclassified correctly ranged from 58% to 92%. The NRI was 1% for $OR_Y = 1.25$ and ranged up to 16% for $OR_Y = 3$. Power for the NRI increased from 15% for $OR_Y = 1.25$ to 73% for $OR_Y = 1.75$ and to 92% for $OR_Y = 2.0$ (Figure 3). Values for the adjusted cNRI were generally similar, but slightly higher than those for the NRI, but power was lower due to the restricted sample size. On average from one-third to half of those in the two intermediate groups were reclassified. When three categories of risk, defined as <5%, 5–<10%, and 20% or higher, were used, the values of the NRI were 22 to 30% lower, ranging from 0.7% for $OR_Y = 1.25$ to 12.6% for $OR_Y = 3$. Power was also somewhat lower for higher values of OR_Y , ranging from the same 15% for $OR_Y = 1.25$ to 68% for $OR_Y = 1.75$ and 87% for $OR_Y = 2$. In this scenario, estimates and power for the adjusted cNRI were much lower, reflecting the more limited movement with only three categories. Finally, when categories were formed based on tertiles, quartiles, and quintiles of the predicted risk estimates, the NRI for $OR_Y = 2$ ranged from 5% to 9% and for $OR_Y = 3$ from 11% to 18%. Estimates using the same number of categories but different cut points were more similar, especially for higher values of OR_X . For example, for $OR_X = 4$, the mean NRI was 11.1% for $OR_Y = 2$ using the 4 CVD categories, but 9.5% using quartiles of predicted risk. For $OR_X = 8$, the NRI changed from 7.3% to only 7.0% in these situations.

The calibration statistics stayed fairly constant when testing the correct XY model, and power equaled the type I error of 5% expected. The statistics increased with OR_Y when testing the model with X only, with power ranging from 9% for $OR_Y = 1.25$ to 62% for $OR_Y = 1.75$ and to 87% for $OR_Y = 2.0$ (Figure 3). Power for J^2 was low, only reaching 56% for $OR_Y = 1.75$ and 84% for $OR_Y = 2.0$, although this may be partially due to the conservative type I error. The same pattern held when using three categories. Power for the calibration tests was also affected by the category definition; it was the same 9% for $OR_Y = 1.25$, 53% for $OR_Y = 1.75$, and 78% for $OR_Y = 2.0$ when three categories were used.

As shown in Figure 3, power for the NRI and the c-statistic crossed for $OR_X = 4$; power for the NRI was generally higher for low values of OR_Y . As the OR for Y increased, the power became closer and that for the c-statistic became higher, especially when X was stronger, such as for $OR_X = 8$ or 16. The lower power for c at low values of OR_Y is likely due to the conservative alpha error. Power for the RC X^2 statistic was lower when $OR_X = 4$, but higher than that for the NRI when $OR_X = 16$. Power for these two statistics was nearly identical for $OR_X = 8$. Power for the J^2 calibration statistic and the adjusted clinical NRI were consistently lower. Note that effect estimates as well as power would be expected to be lower in an external validation set.

When the probability of disease was allowed to vary, the values of the NRI stayed consistent except for the more extreme assumption of $PD = 0.40$ when it was reduced, for example, for $OR_X = 8$ and $OR_Y = 3$ from 16% for $PD = 0.10$ to 12% for $PD = 0.40$. Values for the IDI increased, from 3% for $PD = 0.10$ to 5% for $PD = 0.40$. As expected, power increased for each test with the proportion of cases, as shown in Figure 4. When the sample size was increased to yield an equivalent number of cases, the power was nearly identical (data not shown).

When a correlation between X and Y was included in the simulations using the same size effects, the estimated ORX and ORY were centered around the assumed true conditional values, but in the model with X only the estimated ORX was higher due to its correlation with Y. For ORX=8 and a correlation of 0.4, this ranged from 8.8 for ORY=1.25 to 11.5 for ORY=3. The effect of X was thus stronger in the model with X only. Values of the c-statistics increased for both models, but the difference between them decreased and power to detect a difference was reduced. For example, for a true ORX = 8, ORY=1.5 where X and Y had a correlation of 0.4 and P(D)=0.10, the estimated c-statistic for the XY model increased from 0.762 to 0.775, the difference between models decreased, and the power to detect the difference decreased 43% to 33% using the DeLong method. The overall percent reclassified decreased from 14.5% with no correlation to 12.5% with a correlation of 0.4, the median NRI decreased from 2.7% to 2.1%, the IDI from 0.48 to 0.40, and the RC X² from 12.8 to 11.6. Power for all measures correspondingly decreased by 5–10 percentage points in this situation.

Finally, when a binary predictor Y was considered, the relative power for the various statistics changed with the prevalence of the exposure. Power for the likelihood ratio test and IDI were again highest throughout (Figure 5). For a rarer exposure (prevalence = 0.05 or 0.10), power for the reclassification calibration statistic was higher than that for the NRI or c-statistic. Power for the c- and RC statistics were similar for a prevalence of 0.20, and that for the c-statistic was higher for a prevalence of 0.50. Power for the NRI was lower than either of these tests, except for the most common exposure (prevalence = 0.50) where it equaled that of the RC statistic, while that for the adjusted cNRI was lowest throughout.

4 Women's Health Study Data

Data from the Women's Health Study (WHS) (Ridker, *et al.*, 2005), a nationwide cohort of women aged 45 and older, were used to examine the clinical utility of adding high-sensitivity C-reactive protein (hsCRP) to a predictive model for cardiovascular disease that included traditional risk factors. Women were followed for an average of 10 years for development of cardiovascular disease (CVD). All women were followed for at least 8 years, and the binary outcome is defined as CVD as of 8 years for this example. These data have been described in more detail elsewhere (Cook *et al.*, 2006; Cook and Ridker, 2009). All study participants provided written informed consent, and the study protocol was approved by the institutional review board of Brigham and Women's Hospital (Boston, MA).

Baseline blood samples were analyzed for traditional CVD risk biomarkers, including total and high-density lipoprotein (HDL) cholesterol, as well as hsCRP. A total of 24 171 women had complete data and were not censored as of 8 years. Cox proportional hazards regression models were fit including the traditional CVD risk factors of age, systolic blood pressure, diabetes, current smoking, total and HDL cholesterol, and family history of myocardial infarction, both with and without hsCRP. Risk of CVD as of eight years was computed from the Cox models, and compared to the observed outcome as of eight years of follow-up.

Risk strata with cutoffs of 4%, 8%, and 16% at 8 years were formed to coincide with traditional cutoffs of 5%, 10%, and 20% for 10-year risk. The reclassification table is displayed in Table 5. A total of 4% of women were reclassified, 19% and 16% in the two intermediate risk groups. The observed risks corresponded more closely to the predicted risk categories for the model with hsCRP in most cells, except for the 106 women who moved from 8–<16% to 4–<8% with an observed risk of 1%, and the 31 who move from 16+ to 8–<16%, with an observed risk on the border of 16%. Using the criterion of cells size of at least 20 (df=8), the RC X² statistic for the model without hsCRP was 32.6 (p<0.0001), and

that for the model with hsCRP was 18.9 ($p=0.015$), reflecting the imperfect calibration within these two cells. When only cells with an average expected value of at least 5 were used ($df=6$), these statistics were 32.4 ($p<0.0001$) and 18.8 ($p=0.004$) for models without and with hsCRP, respectively. The latter tests left out the 24 and 31 women moving between the two highest categories, which may contribute important information regarding fit.

The NRI for these data was 5.2% ($p<0.0001$), with a 95% confidence interval of 2.6% to 7.8%. The IDI was 0.1% ($p=0.026$). When the two intermediate risk categories were considered, the unadjusted clinical NRI was 15.8%, but 9.2% ($p=0.0002$) after correction with a 95% confidence interval of 4.4% to 14.1%.

The predictiveness curves for models with and without hsCRP are displayed in Figure 1 (middle). While these are virtually the same, the tests of association as well as the reclassification table above suggested important differences. Further insight may be gained by examining predictiveness curves for models fit in the same data but with and without systolic blood pressure in Figure 1 (right). These are also virtually the same. Since systolic blood pressure is one of the strongest known CVD risk factors after age, has been found to be related to cardiovascular disease in many different cohorts around the world including the WHS, and its treatment has been shown to substantially lower risk, the lack of a difference in curves suggests limited power to detect important differences. Similar to the related ROC curve (Huang and Pepe, 2009), predictiveness curves appear to be conservative when comparing models.

4 Discussion

As noted previously, statistical tests of association, including likelihood ratio tests or odds ratio estimates, cannot determine whether markers or models will be useful in risk prediction or have clinical utility (Harrell, 2001; Pepe, *et al.*, 2004). Model fit statistics, which examine discrimination, calibration, or overall model accuracy, can aid in determining the performance of risk prediction models. The simulations described above examined both classic measures of model fit as well as more recent ones based on reclassification to directly compare models. While the c-statistic, or area under the ROC curve, has often been used to quantify model fit, the difference between c-statistics is typically small, even for established predictors (Pepe *et al.*, 2004; Cook, 2007). This is also true in the simulations where it was usually no more than 0.01 for odds ratios of 1.5 or 2.0. Power for testing the difference in c-statistics, however, was conversely relatively high for a continuous Y. While power was less than 50% for an ORY of 1.5, it reached 90% when ORY was 2 for a sample size of 5000 with a 10% probability of disease. Thus, even small changes in c were statistically significant. Power for the c-statistic with a binary predictor was lower when the prevalence was low.

We considered two methods of computing the variance of the difference between c-statistics for correlated data: that described by DeLong *et al.* (1988) and another proposed by Rosner and Glynn (2009). Both methods are based on ranks and the Wilcoxon statistic, but Rosner and Glynn use the bivariate normal distribution function to estimate joint probabilities given a shift alternative, while DeLong *et al.* estimate these empirically, leading to a more computer intensive method. The variance estimates were similar for the two methods, along with the corresponding power, although the Rosner-Glynn estimate was usually a couple of percentage points lower. The advantage of the Rosner-Glynn estimate was that it was much quicker to run. With a very large data set, the difference in time could be substantial. Both methods, though, suffered from a conservative type I error.

The IDI statistic, as previously noted (Pepe *et al.*, 2008a; Tjur, 2009), corresponded with the difference in R^2 based on the correlation between the observed outcome and predicted

probabilities. Results were similar for the Nagelkerke version of R^2 . While R^2 has been promoted as a means to assess fit in logistic models (Ash and Schwartz, 1999), it remains difficult to interpret, especially since the size is typically low for binary outcomes. Pepe *et al.* (2008a) have shown that a test of the IDI is asymptotically equivalent to a test of statistical significance. As expected given these theoretical considerations, the power for the IDI was relatively high and comparable to the likelihood ratio test for the regression coefficient throughout, indicating that the IDI is sensitive at detecting model differences. A significant regression coefficient, or even IDI, again may not be enough to determine clinical utility.

The classic Hosmer-Lemeshow test was found to have low power in all scenarios as illustrated in Figure 3, which has been previously demonstrated in extensive simulations (Hosmer and Hjort, 2002). In fact, none of the twelve goodness-of-fit tests considered by Hosmer and Hjort had adequate power to detect an omitted covariate. Le Cessie and van Houwelingen (1991) suggest that partitioning on the covariate or 'x' space as suggested by Tsiatis (1980), rather than the outcome or 'y' space, may help circumvent the problem. Difficulties in partitioning remain, especially when considering multiple variables. The reclassification table alternatively considers and partitions on a model including the omitted variables. Partitioning can then be done in important categories of the 'y' space for models both with and without the variable(s) in question.

A perfect model should display agreement of the observed and expected risk throughout the distribution or however categories are defined (Diamond, 1992; Gail and Pfeiffer, 2005). The usual definition of calibration often refers to the expected risk conditional on variables in the model. Ideally, however, we would like a model to reflect the true underlying probability, regardless of whether all predictor variables are in the model (Cook, 2010). While the Hosmer-Lemeshow test has little power to detect an omitted variable, the RC test can at least indirectly test this 'unconditional' calibration by comparing observed and expected proportions against an omitted variable.

Relevant determination of clinical significance can be found through reclassification measures since they more directly address changes in risk strata and potential treatment. The simple percent reclassified is not a useful measure on its own. An assessment of the accuracy of the reclassifications is necessary, whether through the RC test or the NRI. If the new categories are found to be more accurate, then the percents reclassified may be helpful for clinical purposes. The percent reclassified into the various categories could be used to evaluate the impact of screening strategies or could be used for cost-benefit analyses.

We found that the type I error for the NRI and the RC statistic were reasonable, but the relative power varied according to the model assumptions. The primary difference between these two tests lies in their interpretation. The NRI is a test of discrimination, or separation of cases and controls. It does not assess whether the predicted probabilities are accurate or calibrated, but whether cases are higher than controls in one model vs. another. Alternatively, the RC statistic directly assesses calibration rather than discrimination within categories defined by both models. These tools thus appear to provide complementary but unique information

As an estimate, the NRI is dependent on the particular categories used. For example, the use of three vs. four risk categories can lead to a change in the NRI, which is more pronounced than in the test of calibration (Cook and Ridker, 2009). In the example in section 2.4., the NRI ranged from 2.5% with two categories to 35.1% with cut points at each observation. The RC statistic is less sensitive to the number of categories due to the change in degrees of freedom. If a model accurately reflects the observed outcomes, it should do so however the

categories are defined. As with the Hosmer-Lemeshow statistic upon which it is based, however, there will be variability of the RC statistic with changing category definition. We believe that this represents random variability rather than the systematic bias that is displayed by the NRI and other versions of the proportion reclassified. A sensible strategy to use in the absence of established clinical guidelines may be to use quartiles or multiples of the base probability, which yielded similar results here. In particular, it may be reasonable to use category cut points that are one-half, equal to, and double the average population risk.

For clinical use, individuals who fall into the intermediate categories, or “gray zone,” may be of most interest. While the original cNRI is biased, the adjusted version can be used to determine the net improvement in this group. In the simulations, estimates were slightly higher than those for the NRI when four risk categories were used. For three risk categories, there was limited reclassification in the intermediate group, so this statistic is not as useful.

We examined three versions of the reclassification calibration statistic, and found that the adjusted version led to a conservative type I error and lower power when cells were restricted to those with at least 20 observations. It was more similar to the unadjusted RC statistic when the average expected values were at least five. This latter restriction, though, may ignore important information in the smaller categories, as in the example data. The J^2 statistic was conservative throughout, likely due to the extra degree of freedom charged. While theoretically such adjustments could improve the fit to the chi-square distribution, this was found not to be the case. The unadjusted RC statistic exhibited good test characteristics and general adherence to the chi-square distribution in the null situation. Thus, we recommend using the unadjusted statistic with cell sizes of at least 20.

A limitation to these analyses is that all of the models presented here were fit and compared in the same data and thus the overall estimates of performance will be optimistic. Since over-fitting will likely be somewhat greater in a larger model, the estimates of power for testing differences in models will also be somewhat over-estimated. However, the underlying or true models were theoretical and no selection rules were applied, so we would not expect the degree of optimism to be very large, as previously demonstrated (Cook and Ridker, 2009).

Note that the majority of these simulations assumed no correlation between the predictors. If a new variable is highly correlated with those already in the model, it will add little to prediction, as demonstrated by lower amounts of reclassification, lower estimates of all measures of model difference, and lower power in our simulations. If predictors are suspected of being correlated, performance measures could alternatively be estimated using residuals given the other predictors and estimates of the conditional effect on the outcome.

While use of categories in assessing calibration has been discouraged (Hosmer, *et al.*, 1997), such risk strata serve two purposes. First, use of risk stratification tables can illustrate the potential for changes in treatment along important points on the continuum of risk. They help assess changes in risk that may prove useful to clinicians in guiding treatment decisions, so are readily interpretable in terms of clinical impact or utility. Reclassification was originally proposed in a clinical setting where treatment guidelines were established according to levels of projected risk (Cook *et al.*, 2006). While such clinical categories are not always available, the categories should be ‘important’ such that changes in category should have potential clinical impact. Second, using categories is a rough means of weighting the importance of changes in prediction. In healthy cohorts, such as the Framingham study (Pencina *et al.*, 2008) or the Women’s Health Study (Ridker *et al.*, 2005) for example, the bulk of the cohort may fall at very low risk. Doubling or even tripling risk within these low-risk individuals may lead to small changes in absolute risk with little

clinical impact. Weighting by the observed distribution may thus not be optimal for assessing clinical utility. Strata that give more weight to those at higher risk may be more relevant to decision-making, whether or not the particular categories used follow established clinical criteria.

This paper has focused on the statistical characteristics of the various performance measures, particularly type I error and power. The most powerful test, however, is not necessarily the best to determine clinical utility. If it were, we would use the test of association based on the estimated log odds ratio (Pepe *et al.*, 2004). Similarly the c-statistic was often found to have good power even when changes were very small. Whether small differences in the c-statistic translate to clinical utility, however, remains questionable (Cook, 2007). The focus should be on how the model will be used in practice and a corresponding performance measure. Since treatment guidelines are sometimes based on predicted risk strata as in the ATP III report (2001), a method that compares the accuracy of the predicted strata for two models would seem to be appropriate in many situations.

Ultimately, what matters most in assessing predictive models should be related to actual clinical use. Will the new model change practice or change treatment decisions for physicians? Is the cost of additional screening, whether financial or otherwise, justified by an ultimate benefit to the patient? The NRI implicitly assigns a cost of $1/p$ for the cases and $1/(1-p)$ for the controls, but could be generalized to other costs (Pencina, *et al.*, 2009). Since costs may change over time or by personal preference, flexible decision rules such as net benefit (Vickers and Elkin, 2006) or relative utility curves (Baker, *et al.*, 2009), may be helpful in determining the best model for a given cost structure. Such considerations could also assist in determining the optimum cut points for clinical use as well as comparing one-stage vs. two-stage testing (Baker *et al.*, 2009). While such cost considerations, however, may be the ultimate criteria for adopting a model into clinical practice, determination of model fit and whether treatments would change remain critical first steps in evaluating clinical usefulness (Hlatky, *et al.*, 2009).

Acknowledgments

This work was 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.

References

- Adult Treatment Panel III. 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; 285:2486–2497. [PubMed: 11368702]
- Ash A, Shwartz M. R^2 : a useful measure of model performance when predicting a dichotomous outcome. Statistics in Medicine. 1999; 18:375–384. [PubMed: 10070680]
- Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. Journal of the Royal Statistics Society, Series A. 2009; 172:729–748.
- Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007; 115:928–935. [PubMed: 17309939]
- Cook NR. Comments on 'Evaluating the added predictive ability of a new biomarker: from area under the ROC curve to reclassification and beyond'. Statistics in Medicine. 2008; 27:191–195. [PubMed: 17671959]
- Cook NR. Comment: Measures to summarize and compare the predictive capacity of markers. International Journal of Biostatistics. 2010; 6:Article 22. [PubMed: 20737021]

- 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. [PubMed: 16818925]
- Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Annals of Internal Medicine*. 2009; 150:795–802. [PubMed: 19487714]
- D'Agostino, RB.; Griffith, JL.; Schmidt, CH.; Terrin, N. Proceedings of the Biometrics Section. Alexandria, VA: American Statistical Association, Biometrics Section; 1997. Measures for evaluating model performance; p. 253-258.
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988; 44:837–845. [PubMed: 3203132]
- Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. *Journal of Clinical Epidemiology*. 1992; 45:85–89. [PubMed: 1738016]
- Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics*. 2005; 6:227–239. [PubMed: 15772102]
- Gu W, Pepe M. Measures to summarize and compare the predictive capacity of markers. *The International Journal of Biostatistics*. 2009; 5:Article 27. [PubMed: 20224632]
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143:29–36. [PubMed: 7063747]
- Harrell, FE, Jr. Regression Modeling Strategies. Springer; New York: 2001.
- Hlatky MA, Greenland P, Arnett DK, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation*. 2009; 119:2408–2416. [PubMed: 19364974]
- Hosmer DW, Hjort NL. Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine*. 2002; 21:2723–2738. [PubMed: 12228887]
- 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. [PubMed: 9160492]
- Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics: Theory and Methods*. 1980; A10:1043–1069.
- Hosmer, DW.; Lemeshow, S. Applied Logistic Regression. Wiley; New York: 2000.
- Hu B, Palta M, Shao J. Properties of R^2 statistics for logistic regression. *Statistics in Medicine*. 2006; 25:1383–1395. [PubMed: 16059870]
- Huang Y, Pepe MS. A parametric ROC model-based approach for evaluating the predictiveness of continuous markers in case-control studies. *Biometrics*. 2009; 65:1133–1144. [PubMed: 19459841]
- Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Annals of Internal Medicine*. 2008; 149:751–760. [PubMed: 19017593]
- Le Cessie S, van Houwelingen JC. A goodness-of-fit test for binary regression based on smoothing methods. *Biometrics*. 1991; 47:1267–1282.
- McClish DK. Analyzing a portion of the ROC curve. *Medical Decision Making*. 1989; 9:190–195. [PubMed: 2668680]
- Mealiffe ME, et al. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. *Journal of the National Cancer Institute*. 2010; 102:1618–1627. [PubMed: 20956782]
- Mihaescu R, et al. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. *American Journal of Epidemiology*. 2010; 172:353–361. [PubMed: 20562194]
- Moons KGM, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Academic Radiology*. 2003; 10:670–672. [PubMed: 12809422]
- 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. *Statistics in Medicine*. 2008; 27:157–172. [PubMed: 17569110]

- Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Response to 'Net reclassification improvement and decision theory' by Vickers et al. *Statistics in Medicine*. 2009; 28:526–528.
- 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'. *Statistics in Medicine*. 2008a; 27:173–181. [PubMed: 17671958]
- Pepe MS, et al. Integrating the predictiveness of a marker with its performance as a classifier. *Anm J Epidemiol*. 2008b; 167:362–368.
- 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. *Anm J Epidemiol*. 2004; 159:882–890.
- Pigeon JG, Heyse JF. An improved goodness of fit statistics for probability prediction models. *Biometrical Journal*. 1999; 41:71–82.
- Ridker PM, 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. [PubMed: 15753114]
- Rosner B, Glynn RJ. Power and sample size estimation for the Wilcoxon rank sum test with application to comparisons of C statistics from alternative prediction models. *Biometrics*. 2009; 65:188–197. [PubMed: 18510654]
- Steyerberg EW, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2009; 21:128–138. [PubMed: 20010215]
- Tjur T. Coefficients of determination in logistic regression models - A new proposal: the coefficient of discrimination. *American Statistician*. 2009; 63:366–372.
- Tsiatis AA. A note on the goodness-of-fit test for the logistic regression model. *Biometrika*. 1980; 67:250–251.
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*. 2006; 26:565–574. [PubMed: 17099194]
- Whittemore AS. Evaluating health risk models. *Statistics in Medicine*. 2010; 29:2438–2452. [PubMed: 20623821]

Figure 1. Predictiveness curves for models with and without variable Y (left), high sensitivity C-reactive protein (CRP) (middle), and systolic blood pressure (SBP) (right). Horizontal lines indicate the prevalence of the outcome.

Figure 2.

Null distribution of reclassification calibration statistics: observed proportions with p-values < 0.10 (top) or < 0.05 (bottom) for test of correct model X or XY when ORY=1 and for model XY when ORY=3 with cell size > 20 and for average expectation ≥ 5 , for $P(D)=0.10$.

Figure 3.
Power for measures of model fit, with $P(D) = 0.10$.

Figure 4. Power for IDI (top), NRI (middle) and reclassification calibration test (bottom) by probability of disease $P(D)$ and ORY, using $ORX = 8$, $N=5000$, and category cut points of $0.5 \cdot P(D)$, $P(D)$, and $2 \cdot P(D)$.

Figure 5. Power for measures of model fit, with ORX = 8, $P(D) = 0.10$, and varying prevalence (PY) of a binary predictor Y.

Table 1

Hypothetical example of reclassification data table with four risk strata.

All Model 1	Model 2					Total
	0<5%	5<10%	10<20%	20%+		
0<5%	N	4564	416	36	2	5018
5<10%	Observed risk	.020	0.053	0.222	-	
	N	738	977	375	32	2122
10<20%	Observed risk	.034	0.075	0.123	0.125	
	N	76	451	763	288	1578
20% +	Observed risk	0.053	0.086	0.166	0.243	
	N	0	25	266	991	1282
	Observed risk	-	0.160	0.158	0.387	
Total		5378	1869	1440	1313	10000
Cases Model 1	Model 2					Total
	0<5%	5<10%	10<20%	20%+		
0<5%	N	93	22	8	0	123
5<10%	N	25	73	46	4	148
10<20%	N	4	39	127	70	240
20% +	N	0	4	42	384	430
Total		122	138	223	458	941
Controls Model 1	Model 2					Total
	0<5%	5<10%	10<20%	20%+		
0<5%	N	4471	394	28	2	4895
5<10%	N	713	904	329	28	1974
10<20%	N	72	412	636	218	1338
20% +	N	0	21	224	607	852
Total		5256	1731	1217	855	9059

Table 2

Average estimated statistics in models with X only and X plus Y for ORY = 1, based on 1000 simulations with outcome proportion p=0.10 and sample size 5000.

Estimates	Mean (95% interval)		
	ORX = 4	ORX = 8	ORX = 16
Effect Estimate			
OR Y	1.00 (0.83, 1.20)	1.00 (0.82, 1.21)	1.00 (0.81, 1.23)
C-statistic			
X Model	0.684 (0.659, 0.710)	0.758 (0.735, 0.779)	0.816 (0.797, 0.834)
XY Model	0.684 (0.659, 0.710)	0.758 (0.735, 0.779)	0.816 (0.797, 0.835)
Difference	0.000 (−0.002, 0.000)	0.000 (−0.002, 0.002)	0.000 (−0.001, 0.000)
Model R² (%)			
X Model	4.4 (3.1, 6.1)	9.8 (7.7, 12.1)	16.0 (13.1, 18.8)
XY Model	4.5 (3.1, 6.1)	9.8 (7.7, 12.2)	16.0 (13.1, 18.9)
Difference	0.02 (−0.01, 0.14)	0.03 (−0.02, 0.18)	0.03 (−0.03, 0.22)
Nagelkerke R² (%)			
X Model	4.5 (3.3, 5.9)	9.5 (7.8, 11.3)	15.4 (13.2, 17.7)
XY Model	4.5 (3.3, 5.9)	9.5 (7.8, 11.3)	15.5 (13.1, 18.9)
Difference	0.02 (0.00, 0.12)	0.02 (0.00, 0.13)	0.03 (0.00, 0.14)
Reclassification Statistics			
IDI (%)	0.02 (0.00, 0.13)	0.03 (0.00, 0.16)	0.03 (−0.01, 0.18)
Categories = 0–<5%, 5–<10%, 10–<20%, 20%+			
% Reclassified	3.5 (0.1, 9.9)	2.9 (0.1, 8.5)	2.7 (0.06, 6.6)
% Reclassified correctly	51.3 (5.9, 97.2)	51.6 (7.7, 91.5)	52.0 (0.0, 100.0)
NRI (%)	0.2 (−1.4, 2.4)	0.2 (−1.5, 2.0)	0.1 (−1.4, 1.7)
Adj cNRI (%)	0.2 (−1.5, 2.5)	0.2 (−2.2, 2.9)	0.2 (−2.4, 3.1)
Categories = 0–<5%, 5–<20%, 20%+			
% Reclassified	1.8 (0.04, 5.3)	1.8 (0.06, 5.1)	1.4 (0.04, 4.2)
% Reclassified correctly	50.5 (0.0, 100.0)	51.8 (0.0, 100.0)	51.3 (0.0, 100.0)
NRI (%)	0.2 (−1.0, 1.7)	0.1 (−1.1, 1.6)	0.1 (−1.2, 1.4)
Adj cNRI (%)	0.1 (−0.7, 0.12)	0.1 (−1.0, 1.6)	0.1 (−1.4, 1.8)

Table 3

Type I error ($p < 0.05$) for statistics in models with X only and X plus Y for ORY = 1, based on 1000 simulations with outcome proportion $p = 0.10$ and sample size 5000.

Estimates	Type I Error (%)		
	ORX = 4	ORX = 8	ORX = 16
Tests of Association for Y			
Wald Test	4.3	4.8	5.1
C-statistic			
Difference (DeLong)	0.2	0.4	0.1
Difference (Rosner)	0.2	0.4	0.0
Hosmer-Lemehow statistics			
Deciles X Model			
HL X^2	5.6	4.8	5.5
HL X^2 adjusted	4.8	3.5	3.5
Deciles XY Model			
HL X^2	3.6	4.7	7.2
HL X^2 adjusted	2.9	3.6	4.8
Reclassification Statistics –			
IDI	3.8	3.4	3.5
Categories = 0–<5%, 5–<10%, 10–<20%, 20%+			
NRI	4.6	4.3	3.2
Adj cNRI	3.7	2.8	3.0
Reclassification calibration (n = 20)			
X Model			
RC X^2	5.7	4.7	5.9
RC X^2 adjusted	3.8	3.1	4.0
J^2	2.9	2.9	2.6
XY Model			
RC X^2	4.7	5.0	5.5
RC X^2 adjusted	3.3	2.9	3.8
J^2	2.8	2.2	2.5
Categories = 0–<5%, 5–<20%, 20%+			
NRI	4.4	4.5	4.3
Adj cNRI	3.1	3.9	5.0
Reclassification calibration (n = 20)			
X Model			
RC X^2	5.8	5.2	5.0
RC X^2 adjusted	4.0	3.0	3.1
J^2	2.3	2.4	2.1
XY Model			
RC X^2	6.0	4.9	5.1
RC X^2 adjusted	3.8	2.9	2.9

Estimates	Type I Error (%)		
	ORX = 4	ORX = 8	ORX = 16
J^2	1.9	2.3	1.7

Table 4

Average estimated statistics in models with ORX = 8, based on 1000 simulations with outcome proportion 0.10 and sample size 5000.

Estimates	Mean (95% Interval)			
	ORY = 1.25	ORY = 1.50	ORY = 2.0	ORY = 3.0
Effect Estimate				
OR Y	1.25 (1.04, 1.51)	1.50 (1.23, 1.80)	2.0 (1.63, 2.43)	3.0 (2.48, 3.73)
C-statistic				
X Model	0.758 (0.734, 0.779)	0.757 (0.735, 0.778)	0.756 (0.733, 0.777)	0.751 (0.727, 0.773)
XY Model	0.760 (0.737, 0.782)	0.762 (0.740, 0.783)	0.769 (0.745, 0.790)	0.782 (0.759, 0.804)
Difference	0.002 (0.000, 0.006)	0.005 (0.001, 0.010)	0.013 (0.006, 0.022)	0.031 (0.019, 0.044)
Model R²				
X Model	9.8 (7.6, 12.0)	9.6 (7.5, 12.0)	9.5 (7.4, 11.8)	8.9 (6.9, 11.1)
XY Model	9.9 (7.7, 12.2)	10.1 (8.0, 12.5)	10.8 (8.6, 13.3)	12.2 (9.7, 14.8)
Difference	0.17 (-0.01, 0.55)	0.48 (0.09, 1.04)	1.32 (0.56, 2.19)	3.21 (2.06, 4.63)
Nagelkerke R²				
X Model	9.5 (7.7, 11.4)	9.4 (7.7, 11.3)	9.3 (7.5, 11.2)	8.9 (7.2, 10.7)
XY Model	9.7 (7.9, 11.6)	9.8 (8.0, 11.9)	10.5 (8.5, 12.4)	11.7 (9.7, 13.8)
Difference	0.14 (0.00, 0.42)	0.42 (0.10, 0.84)	1.15 (0.58, 1.84)	2.82 (1.90, 3.91)
Reclassification Statistics				
IDI (%)	0.16 (0.00, 0.50)	0.48 (0.12, 0.95)	1.30 (0.62, 2.13)	3.18 (2.12, 4.51)
Categories = 0-<5%, 5-<10%, 10-<20%, 20%+				
% Reclassified	8.1 (1.4, 15.2)	14.5 (7.1, 21.3)	24.0 (17.6, 30.3)	35.0 (29.4, 40.3)
% Reclassified correctly	58.0 (18.5, 100.0)	66.5 (24.6, 100.0)	81.5 (45.0, 100.0)	92.0 (66.8, 100.0)
NRI (%)	1.0 (-1.5, 4.1)	2.7 (-1.0, 6.7)	7.3 (2.0, 12.9)	16.2 (10.0, 22.2)
Adjusted cNRI (%)	1.2 (-2.2, 5.5)	3.2 (-1.8, 8.8)	8.4 (2.1, 14.9)	16.9 (10.3, 23.3)
Categories = 0-<5%, 5-<20%, 20%+				
% Reclassified	4.9 (0.8, 9.3)	8.9 (4.3, 13.1)	14.9 (10.8, 19.2)	23.2 (18.9, 27.4)
% Reclassified correctly	57.1 (0.0, 100.0)	64.9 (14.9, 100.0)	80.7 (29.9, 100.0)	95.3 (54.2, 100.0)
NRI	0.7 (-1.4, 3.3)	2.0 (-1.0, 5.3)	5.4 (1.3, 10.0)	12.6 (7.0, 18.3)
Adjusted cNRI	0.7 (-1.4, 3.1)	1.8 (-1.0, 4.8)	4.8 (1.1, 8.9)	10.3 (5.8, 14.6)

Table 5

Data from the Women's Health Study of 24,558 women followed for cardiovascular disease. Risk is estimated as of eight years of follow-up with corresponding changes in category definitions.

All Traditional Risk Factors	Traditional Risk Factors Plus hsCRP			
	0<4%	4<8%	8<16%	16%+
0<4%	N 20471	342	0	0
Observed risk	1.1%	5.6%	-	-
4<8%	N 295	1816	139	0
Observed risk	2.0%	6.7%	13.7%	-
8<16%	N 0	106	694	24
Observed risk	-	0.9%	11.8%	16.7%
16% +	N 0	0	31	253
Observed risk	-	-	16.1%	27.3%
Total	20766	2264	864	277
Cases Model 1	Total			
	0<4%	4<8%	8<16%	16%+
0<4%	N 234	19	0	0
4<8%	N 6	121	19	0
8<16%	N 0	1	82	4
16% +	N 0	0	5	69
Total	240	141	106	73
Controls Model 1	Total			
	0<4%	4<8%	8<16%	16%+
0<4%	N 20237	323	0	0
4<8%	N 289	1695	120	0
8<16%	N 0	105	612	20
16% +	N 0	0	26	184
Total	20526	2123	758	204
				23611