# Special Report

# Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction

Nancy R. Cook, ScD

***Abstract*—**The *c* statistic, or area under the receiver operating characteristic (ROC) curve, achieved popularity in diagnostic testing, in which the test characteristics of sensitivity and specificity are relevant to discriminating diseased versus nondiseased patients. The *c* statistic, however, may not be optimal in assessing models that predict future risk or stratify individuals into risk categories. In this setting, calibration is as important to the accurate assessment of risk. For example, a biomarker with an odds ratio of 3 may have little effect on the *c* statistic, yet an increased level could shift estimated 10-year cardiovascular risk for an individual patient from 8% to 24%, which would lead to different treatment recommendations under current Adult Treatment Panel III guidelines. Accepted risk factors such as lipids, hypertension, and smoking have only marginal impact on the *c* statistic individually yet lead to more accurate reclassification of large proportions of patients into higher-risk or lower-risk categories. Perfectly calibrated models for complex disease can, in fact, only achieve values for the *c* statistic well below the theoretical maximum of 1. Use of the *c* statistic for model selection could thus naively eliminate established risk factors from cardiovascular risk prediction scores. As novel risk factors are discovered, sole reliance on the *c* statistic to evaluate their utility as risk predictors thus seems ill-advised. **(*Circulation*. 2007;115:928-935.)**

**Key Words:** cardiovascular diseases ■ epidemiology ■ follow-up studies ■ prevention ■ risk factors ■ statistics ■ risk

Downloaded from http://circ.ahajournals.org/ by guest on June 24, 2017

Models for prognostic risk prediction have been widely used in the cardiovascular field to predict risk of future events or to stratify apparently healthy individuals into risk categories.[1] Appropriate model assessment is critical to the determination of clinical impact and to guideline development. In particular, as novel risk factors, which include blood-based biomarkers and those derived from genomics or proteomics, are discovered, whether these factors can contribute to overall risk prediction becomes an important question.

The accuracy of models can be assessed in several ways. Two major components are calibration and discrimination.[2] Calibration is a measure of how well predicted probabilities agree with actual observed risk. When the average predicted risk within subgroups of a prospective cohort, for example, matches the proportion that actually develops disease, we say a model is well calibrated. The Hosmer-Lemeshow statistic[3] compares these proportions directly and is a popular, though imperfect,[4] means to assess model calibration.

Discrimination is a measure of how well the model can separate those who do and do not have the disease of interest. If the predicted values for cases are all higher than for non-cases, we say the model can discriminate perfectly, even if the predicted risk does not match the proportion with disease. Discrimination is of most interest when classification into groups with or without prevalent disease is the goal, such as in diagnostic testing.[5] Discrimination is most often measured by the area under the receiver operating characteristic (ROC) curve, or *c* statistic, as described below.

In the diagnostic setting the outcome is already determined but unknown to the investigator, and the estimated classification can often be compared with a more expensive or invasive gold standard. In prognostic modeling or risk stratification, however, the outcome has not yet developed at the time that predictors are assessed. Future disease status remains to be determined by stochastic process, and can only be estimated as a probability or risk.[6] Measures of discrimination are nonetheless commonly emphasized in such settings, which ignores the random nature of the outcome. Calibration, as well as discrimination, is important in accurate risk prediction. More global measures of fit that combine calibration and discrimination exist, such as likelihood statistics, $R^2$, and the Brier score.[2,7] The performance of risk prediction models in the cardiovascular literature, however, is often judged solely on the basis of the *c* statistic,[8–14] despite the existence of large prospective cohort studies from which risk can be estimated directly.

## The ROC Curve and the *c* Statistic

The most popular measure of model fit in the cardiovascular literature has been the *c* statistic, a measure of discrimination

From the Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Mass, and the Department of Epidemiology, Harvard School of Public Health, Boston, Mass.

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

© 2007 American Heart Association, Inc.

*Circulation* **is available at http://www.circulationaha.org** **DOI: 10.1161/CIRCULATIONAHA.106.672402**

also known as the area under the ROC curve,[15] or the c index, its generalization for survival data.[2,16] The ROC curve and its associated *c* statistic are functions of the sensitivity and specificity for each value of the measure or model. The sensitivity of a test is the probability of a positive test result, or of a value above a threshold, among those with disease (cases). The specificity is the probability of a negative test result, or a value below a threshold, among those without disease (noncases). It is commonly believed that sensitivity and specificity are properties of a test and are not subject to alteration by prevalence of disease, as are the positive and negative predicted values. This has been shown to be false, however, both theoretically and clinically.[17,18] Both sensitivity and specificity can be influenced by case mix, disease severity, or risk factors for disease. For example, a test is likely to be more sensitive among more severe than among milder cases of disease. Similarly, specificity can depend on characteristics of noncases, such as gender, age, or prevalence of concomitant risk factors.[18–20]

The ROC curve is a plot of sensitivity versus 1−specificity (often called the false-positive rate) that offers a summary of sensitivity and specificity across a range of cut points for a continuous predictor. The area under the curve, or *c* statistic, ranges from 0.5 (no discrimination) to a theoretical maximum of 1. Perfect discrimination corresponds to a *c* statistic of 1 and is achieved if the scores for all the cases are higher than those for all the non-cases, with no overlap. The *c* statistic is equivalent to the probability that the measure or predicted risk is higher for a case than for a noncase.[15] Note that the *c* statistic is not the probability that individuals are classified correctly or that a person with a high test score will eventually become a case. The latter is closer in meaning to the predictive value, or the probability of disease given the test result.

The *c* statistic also describes how well models can rank order cases and noncases, but is not a function of the actual predicted probabilities. For example, a model that assigns all cases a value of 0.52 and all noncases a value of 0.51 would have perfect discrimination, although the probabilities it assigns may not be helpful. The actual predicted probabilities do matter, however, in clinical risk prediction models such as those commonly used for the assessment of global cardiovascular risk.

In a prospective cohort that is considered generally low-risk, such as many population-based cohorts, there may be a small proportion of individuals who are at high risk, with a preponderance of those at low or very low risk. Rank-based measures such as the *c* statistic do not take this distribution into account. Differences between 2 individuals who are at very low risk (eg, 1.0% versus 1.1%) have the same impact on the *c* statistic as 2 individuals who are at moderate versus high risk (eg, 5% versus 20%) if their differences in rank are the same. Clinically, however, it may be more important to separate the latter 2 individuals, particularly if treatment decisions are based on predicted probabilities, such as those used by the Adult Treatment Panel III.[1]

## *c* Statistics and Model Selection

Because the *c* statistic is based solely on ranks, it is less sensitive than measures based on the likelihood or other global measures of model fit.[2] This characteristic may make it a poor choice for the selection of variables to be used in a predictive model. As an example, consider data from the Women's Health Study,[20a] a prospective cohort including 26 901 initially healthy nondiabetic women followed for the development of future vascular events over an average period of 10 years. Table 1 (top) shows the results from traditional Framingham risk factors for future cardiovascular disease (CVD) that were modeled with a Cox proportional hazards model. The likelihood ratio statistic tests the significance of the addition of each variable separately to a predictive model that included age only (Table 1, top). Each variable is highly significant statistically, and the $\chi^2$ statistics indicate that, after age, systolic blood pressure (SBP) is the strongest predictor of risk, followed by smoking and lipids. To directly compare the magnitude of effects in this population, the rate (hazard) ratios per 2 SD units are shown, roughly comparable to a comparison of risks for extreme tertiles. The rate ratios lie in the same order as the likelihood ratio statistics for the continuous variables.

The *c* statistic in the model that included only age is 0.70 (Table 1), based on the generalized c index for censored data. This means that the probability is 70% that a case is older than a noncase. When SBP is added to the models, the *c* statistic improves to 0.74, which means that the probability that the predicted risk is higher in cases than noncases is 74%. Although the likelihood ratio statistics and the rate ratios suggest that SBP is the strongest predictor after age, the *c* statistic is 73% to 74% for models with SBP, smoking, and high-density lipoprotein cholesterol (HDL-C), and is unable to distinguish between these 3 factors. The *c* statistic is only 0.71 for the model that includes age and low-density lipoprotein cholesterol (LDL-C). This failure to improve the *c* statistic would suggest that LDL-C is not predictive of CVD, even though it is highly statistically significant in these data, the effect size is moderate, and we know from many studies and trials that LDL-C is an important modifiable risk factor for CVD. An example of improved predictive accuracy despite little change in the *c* statistic is given below.

In a similar manner, with age, SBP, and smoking together in the model for future CVD risk, the *c* statistic was 0.76, and improved only slightly to 0.77 when any of the lipids were added individually (Table 1, middle). Despite this, the likelihood statistics and the rate ratios indicate that HDL-C is a strong cardiovascular risk predictor, followed by total and LDL cholesterol. Finally, when each variable was in turn removed from the full model (Table 1, bottom), the *c* statistic dropped from 0.78 to only 0.77 or 0.76 for all variables except age. Thus, in this example, the likelihood-based measures of model fit were able to distinguish the importance of several established risk factors, whereas the *c* statistic could not. Indeed, if improvement in the *c* statistic was used as the criterion for model inclusion, then neither LDL-C, HDL-C, nor total cholesterol would have been included in risk models after accounting for age, blood pressure, and smoking. In an example from the Framingham Heart Study, family history of premature atherosclerosis was found to be an independent predictor of cardiovascular events, with a relative risk of 2.0 for men and 1.7 for women.[21] However,

Downloaded from http://circ.ahajournals.org/ by guest on June 24, 2017

**TABLE 1. Contributions to Cardiovascular Disease Prediction in the Women's Health Study***

| Variable | Variable $\chi^2$ | *P* | RR per 2 SD† | *c* Statistic |
|---|---|---|---|---|
| Effect of adding variables to model with age only | | | | |
| Ln(age) only | 395.9 | <0.0001 | 4.0 | 0.70 |
| +Ln(SBP) | 148.8 | <0.0001 | 2.5 | 0.74 |
| +Current smoking† | 121.1 | <0.0001 | 2.9 | 0.73 |
| +Ln(HDL) | 85.7 | <0.0001 | 1/2.0 | 0.73 |
| +Ln(TC) | 33.3 | <0.0001 | 1.6 | 0.72 |
| +LDL | 28.8 | <0.0001 | 1.5 | 0.71 |
| Effect of adding variables to model with age, SBP, and smoking | | | | |
| Ln(age), Ln(SBP), smoking | ... | ... | ... | 0.76 |
| +Ln(HDL) | 45.3 | <0.0001 | 1/1.7 | 0.77 |
| +Ln(TC) | 21.5 | <0.0001 | 1.4 | 0.77 |
| +LDL | 18.6 | <0.0001 | 1.4 | 0.77 |
| Effect of deleting variables from full model | | | | |
| Ln(age), Ln(SBP), smoking, Ln(TC), Ln(HDL)− | ... | ... | ... | 0.78 |
| −Ln(TC) | 39.8 | <0.0001 | 1.6 | 0.77 |
| −Ln(HDL) | 63.6 | <0.0001 | 1/1.7 | 0.77 |
| −Current smoking | 99.7 | <0.0001 | 2.6 | 0.76 |
| −Ln(SBP) | 114.1 | <0.0001 | 2.2 | 0.76 |
| −Ln(Age) | 257.4 | <0.0001 | 3.2 | 0.73 |

+ indicates the addition of each variable separately to the model with age only, or age, SBP, and smoking only; −, the deletion of each variable separately from the full model; Ln, natural logarithms; RR, relative risk; and TC, total cholesterol. $\chi^2$ is the likelihood ratio statistic for each single variable when added to the model (top and middle) or subtracted from the full model (bottom). Ln were used to normalize distributions and improve the fit for individual predictors.

*Estimated from Cox proportional hazards models.

†RR is relative risk when each variable is included in models shown in the top and middle sections, and is relative risk in full model in the bottom section; RR compares risk across 2 SD units, except for smoking, which is yes versus no.

Downloaded from http://circ.ahajournals.org/ by guest on June 24, 2017

the *c* statistic increased only from 0.80 to 0.81 in men and 0.81 to 0.82 in women, which may inappropriately limit enthusiasm for this variable.

In these examples, sole reliance on the *c* statistic would seem ill-advised because discrimination is only one aspect of model performance. Likelihood-based measures, such as the likelihood ratio statistic or the Bayes information criterion, which adjusts for the number of variables in the model, are alternatives that are more sensitive and more global measures of model fit.[2] Use of these criteria would have selected age, SBP, smoking, total cholesterol, and HDL-C from the variables considered in Table 1, and reached a final model similar to that developed from the Framingham data.[1,22]

## Odds Ratios and Predictive Values

In epidemiological studies, the most common choices of effect measures are ratios, expressed either as a relative risk (risk ratio), an odds ratio (OR), a rate ratio, or a hazards ratio. Pepe et al[23] describe the relation between the OR and the *c* statistic, and show that an OR as large as 3.0 may have little impact on the ROC curve or the *c* statistic. This may be pertinent when classification into 2 groups is the objective, such as in diagnostic testing. A relative risk of this size, however, could lead to a clinically important improvement in risk prediction for future disease.

Consider, for example, a novel biomarker that has a relative risk of 3.0, but leads to little or no improvement in the *c* statistic given traditional risk factors. For some patients, a high level of the biomarker would shift estimated 10-year risk from 1% to only 3%, a clinically unimportant difference. For others, the same high biomarker level could alter the estimated 10-year risk of a cardiovascular event from 8% to 24%, and lead to different treatment recommendations under current Adult Treatment Panel III guidelines. Thus, for risk prediction, the actual or absolute predicted risk, which is not captured by the *c* statistic, is of primary clinical interest.

Figure 1 shows the distribution of a hypothetical normally distributed risk factor X among cases and controls. Among controls the mean is 0, with a SD of 0.5. Suppose that the OR per 2 SD units equals 3.0, which corresponds to a *c* statistic of 0.65. Despite this moderately large OR, there is a great deal of overlap between the distributions for cases and noncases. This extent of overlap often occurs in practice, as evidenced by the distributions of total cholesterol among cases and noncases of coronary heart disease in the Framingham study.[24] Thus, total cholesterol by itself would be considered a poor classifier for cardiovascular risk, even though it is known to be a pathophysiological determinant of disease.

The OR does, however, relate to the predictive value, or the probability of disease given a positive test, or a value above a threshold. This is a direct function of the OR. Figure 1 also plots the probability of disease given the value of risk factor X in a population with an overall disease probability of 10%.

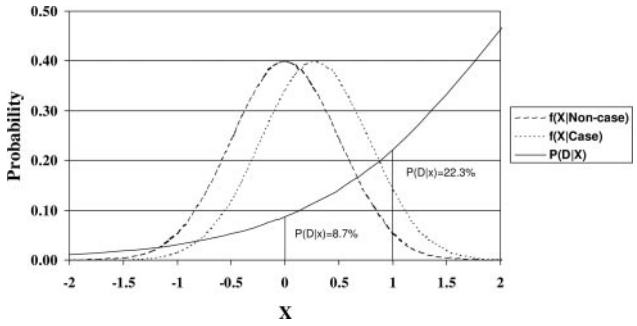


**Figure 1.** Plot of probability density functions for noncases [f(X|Non-case) = dashed line] and cases [f(X|case) = dotted line] with an assumption of normal distributions for hypothetical predictor X, with an OR of 3.0 per 2 SD units. Also shown on the same scale is the probability of disease given the value of X [P(D|X)], with an assumption of an overall prevalence of disease in the population of 10%.

Despite the overlap in the distributions, the predicted probabilities range from <5% to >25%, a difference that may be clinically important. Although the actual numbers may depend on overall disease incidence, the revised risk could cross risk strata in treatment guidelines and lead to different treatment decisions.

SBP provides a more concrete example. Among men in National Health and Nutrition Evaluation Study II data, the estimated mean SBP is 129 mm Hg (SD 17.7).[25] An OR of 3.0 would correspond to a mean of 139 mm Hg among cases, or a difference of 10 mm Hg between cases and noncases (Table 2). Corresponding differences in other measures would be 6 mm Hg for diastolic blood pressure, 24 mg/dL for total cholesterol, 21 mg/dL for LDL-C, and 8 mg/dL for HDL-C,[25,26] all of which would appear to be clinically important differences. None of these by itself, however, would lead to substantial improvement in the area under the ROC curve.

Pepe[23] suggests that an OR of about 16, which corresponds to a *c* statistic of about 0.84, may be needed to achieve reasonable discrimination, or classification into cases and noncases. As shown in Table 2, this would require much larger differences in means between cases and noncases, as large as 24 mm Hg in SBP and 62 mg/dL for total cholesterol. It is unlikely that a single marker could achieve such levels of separation and discrimination. It is, however, possible for a score composed of several of these predictors together to achieve this target. Thus, the Framingham score, which includes several traditional risk factors, has been shown to discriminate reasonably well.[22]

**TABLE 2. Difference in Means Between Cases and Noncases Associated With Specified Odds Ratio in Men**

| | | | Odds Ratio per 2 SD | | | |
|---|---|---|---|---|---|---|
| Risk Factor | Mean | SD | 1.5 | 3 | 9 | 16 |
| SBP (mm Hg) | 129 | 17.7 | 3.6 | 9.7 | 19.4 | 24.5 |
| DBP (mm Hg) | 81 | 11.3 | 2.3 | 6.2 | 12.4 | 15.7 |
| TC (mg/dL) | 211 | 44.5 | 9.0 | 24.4 | 48.9 | 61.7 |
| LDL (mg/dL) | 140 | 37.8 | 7.7 | 20.8 | 41.5 | 52.4 |
| HDL (mg/dL) | 45.2 | 14.0 | 2.8 | 7.7 | 15.4 | 19.4 |

Mean and SD estimated from data from the National Health and Nutrition Evaluation Survey II (NHANES II). DBP indicates diastolic blood pressure; TC, total cholesterol.

Thus, if a relative risk >3.0 were required as a strict criterion for inclusion of each additional biomarker in risk prediction, then, except for age, few of the components of the Framingham risk score would be eligible for inclusion. In the Framingham model,[22] none of the risk factors besides age, which include blood pressure, smoking, or lipids, individually achieves a rate ratio higher than 1.9 for men or 2.2 for women. Although the Framingham score as a whole, including age, increased the *c* statistic from 0.5 (ie, from chance alone with no predictors used) to 0.74 in men and 0.77 in women, the individual clinical risk factors could not do so based on their conditional relative risks. Use of an improvement in the ROC curve for each individual biomarker as a criterion, then, would eliminate most risk factors currently in use for cardiovascular risk prediction, which would include lipids, blood pressure, and smoking.

## Predictive Values and Calibration

Calibration has largely been overlooked in discussions of model fit in the cardiovascular field.[8,9,11,13,14] Although a model can be recalibrated to the overall risk in a new population,[2] we prefer that predicted risk match observed risk within all subgroups. There is, in fact, a trade-off between discrimination and calibration, and a model typically cannot be perfect in both. Diamond[27] showed that a perfectly calibrated model, in which the predicted risk equals the observed risk for all subgroups, cannot achieve a *c* statistic equal to 1 in usual settings. With the assumption of a uniform distribution of risk in the population, the maximum *c* statistic is 0.83. Gail and Pfeiffer demonstrated that this upper limit varies with the distribution of risk in the population.[28] Figure 2 shows the maximum *c* statistic that can be achieved with perfectly calibrated models with various distributions of population risk. The 3 distributions in Figure 2 (left) have an average 10-year risk of 10%. If there is relatively little spread, then risk is centered around 10%, and the maximum *c* statistic with a perfectly calibrated model is only 0.63. If the average risk is the same, but there is more spread, the limit for the *c* statistic increases, and could reach 0.76 or even closer to 1.0. The same is true for the distributions in Figure 2 (right), which all have an average risk of 50%.

In a population similar to the Women's Health Study cohort, with a low average 10-year risk of 2.5% and a 99th percentile of 23%, the maximum *c* statistic is 0.89. In a population with a higher overall risk of 10% and a 99th percentile of 40%, closer to that for coronary heart disease (inclusive of angina) in the Framingham cohort,[22] the maximum *c* statistic is 0.76. Both perfect calibration and perfect discrimination could be achieved only if the true risk as well as the estimated risk were either 0 or 1, similar to the U-shaped distribution in Figure 2 (right). This may be true in the diagnostic setting where the outcome is determined, and the individual truly does or does not have disease. In the prediction of 10-year CVD risk in population-based cohorts,

Downloaded from http://circ.ahajournals.org/ by guest on June 24, 2017

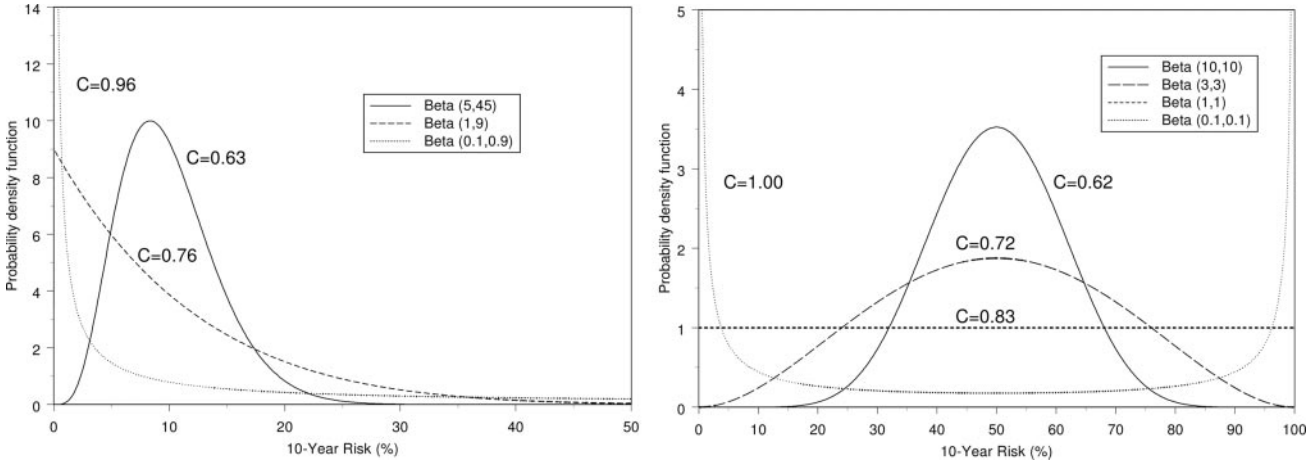


**Figure 2.** Hypothetical risk distributions based on the beta distribution with an average 10% (left) or 50% (right) 10-year risk, and the maximum attainable *c* statistic for a perfectly calibrated model.

Downloaded from http://circ.ahajournals.org/ by guest on June 24, 2017

however, the maximum *c* statistic for perfectly calibrated models appears to be ≈0.75 to 0.90.

A related way to compare models is to examine curves of predicted values or estimated risk (as opposed to true risk, which is unknown).[29] In theory, a stronger model should lead to a wider spread of predicted values and, consequently, stronger discrimination. A plot of the predicted risk versus the risk percentile has been used to compare models.[29,30] Such distributions may, however, also be insensitive in distinguishing between models. An example is shown in Figure 3, which plots the predicted risk from models in the Women's Health Study that include age, smoking, total and HDL cholesterol, but with and without SBP in the model. Transposition of the x and y axes yields a plot of the cumulative probability distribution functions. As shown, there is little difference in these curves even though SBP is the strongest risk predictor after age. The populations in the plot also do not tell us whether 1 model estimates risk more accurately or how the predicted risks differ for individuals with the 2 models. A look at conditional or joint distributions of predicted risk, as described below, may give more insight.

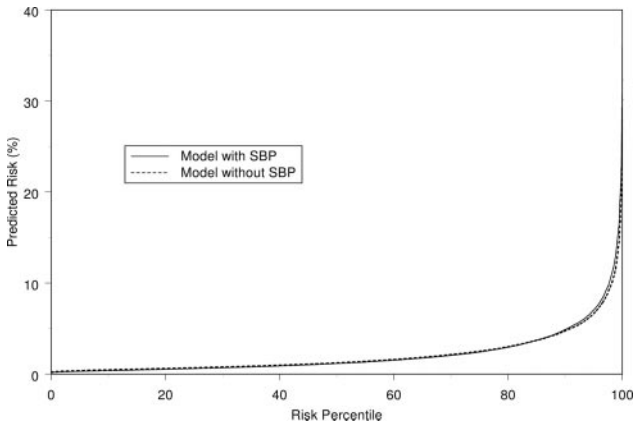


**Figure 3.** Plot of predicted risk (%) versus the risk percentile in the population for models that include age, smoking, total and HDL-C with and without SBP. Data derived from the Women's Health Study.[20a]

## Clinical Risk Reclassification

Most important for clinical risk prediction is whether a new model can more accurately stratify individuals into higher or lower risk categories of clinical importance. The Adult Treatment Panel III, for example, uses estimated 10-year risk categories in its treatment guidelines.[1] If such risk stratification can be made more accurate, the model will be improved.

Table 3 presents the results of risk stratification with models that include age, SBP, smoking, and total cholesterol, but with and without HDL-C. The *c* statistic only changed from 0.77 to 0.78 when HDL-C was included with the other variables (Table 1, bottom), despite a relative risk of 0.54. However, of women classified at 5% to 20% 10-year risk in the model without HDL-C, >34% changed risk category when HDL-C was included in the model. More important, the new risk estimate that used HDL-C was a more accurate representation of actual risk for all but 6 of 1920 women reclassified. A similar result has been shown for the addition of C-reactive protein to models that include traditional risk factors.[31] If HDL-C was not useful in risk prediction, this reclassification would occur randomly. When a completely random variable was added to the full model above, <1% were reclassified in each risk category, and roughly half of these were more accurate. When all the traditional risk factors were added to a model with age only, 7% of those at <5% risk and >60% of those at 5% risk to 20% risk were reclassified more accurately.

To assess the potential for reclassification, risk could be estimated over a range of values of a new biomarker to determine whether it may be important to measure in an individual. Suppose that a woman's age, SBP, smoking status, and total cholesterol are known, but that her HDL-C is not. Suppose that, with the assumption of a reference value for HDL-C of 50 mg/dL, her estimated 10-year risk is 8%. This could vary from about 5% for an HDL-C of 80 mg/dL to about 14% for an HDL-C of 30 mg/dL. Figure 4 shows how a woman's absolute risk estimate may vary with changes in HDL-C compared with risk at the reference HDL-C of 50 mg/dL given her other risk factors. For those at low risk, the additional information is minimal,

**TABLE 3. Comparison of Observed and Predicted Risks Among Women in the Women's Health Study***

| | Model With HDL 10-Year Risk (%) | | | | |
|---|---|---|---|---|---|
| Model Without HDL 10-Year Risk (%) | 0 to <5% | 5 to <10% | 10 to <20% | 20%+ | % Reclassified |
| 0% to <5% | | | | | |
| Total, n | 22655 | 696 | 6 | 0 | ... |
| %† | 97.0 | 3.0 | 0.0 | 0.0 | 3.0 |
| Observed 10-year risk (%)‡ | 1.5 | 5.9 | 0.0 | ... | ... |
| 5% to <10% | | | | | |
| Total, n | 593 | 1712 | 291 | 0 | ... |
| % | 22.8 | 66.0 | 11.2 | 0.0 | 34.0 |
| Observed 10-year risk (%) | 3.7 | 7.6 | 14.7 | ... | ... |
| 10% to <20% | | | | | |
| Total, n | 3 | 214 | 512 | 76 | ... |
| % | 0.4 | 26.6 | 63.6 | 9.4 | 36.4 |
| Observed 10-year risk (%) | 0.0 | 7.5 | 10.7 | 23.3 | ... |
| 20%+ | | | | | |
| Total, n | 0 | 0 | 41 | 102 | |
| % | 0.0 | 0.0 | 28.7 | 71.3 | 28.7 |
| Observed 10-year risk (%) | ... | ... | 15.8 | 32.5 | ... |

*This comparison uses models that include Framingham risk factors with and without HDL. All estimated and observed risks represent 10-year risk of cardiovascular disease.

†Percent classified in each risk stratum by the model with HDL.

‡Observed proportion of participants developing cardiovascular disease in each category.

whereas for those at higher risk the impact on risk of disease is more substantial. If the difference in risk over the range of HDL-C is clinically important, then a test could be ordered to obtain the woman's actual posttest probability. Although other factors such as age[32] must be considered, such focus on intermediate categories of risk is an option for novel predictors or biomarkers, which are difficult or expensive to obtain.

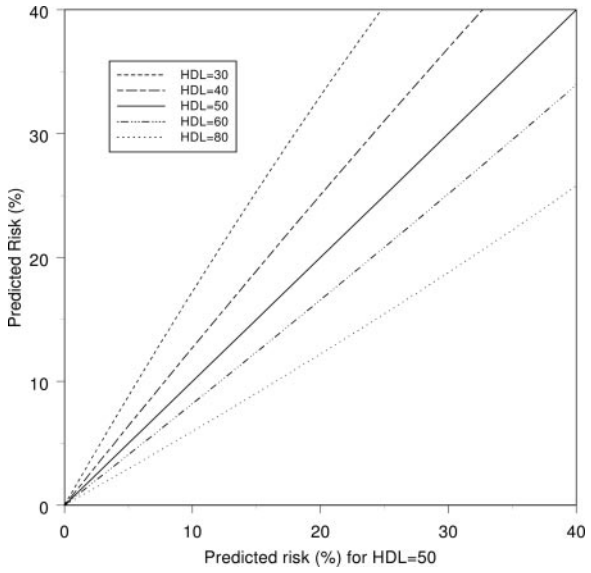


**Figure 4.** Plot of predicted risk over a range of levels of HDL-C versus predicted risk at the reference HDL-C of 50 mg/dL. The lines compare risk at different levels of HDL-C given a woman's other risk factors. Data derived from the Women's Health Study.[20a]

The estimated risk or predicted values, and how well these predict actual risk, may be a more important aspect of a prognostic model than sensitivity and specificity, on which the ROC curve is based. Even in diagnostic testing, patients (and examining physicians) are interested in whether they have the disease given a test result rather than their probability of having a positive test given the presence or absence of disease, as expressed by the sensitivity and specificity.[33,34] If a patient has hypertension, he or she may not be interested in whether everyone with a myocardial infarction has hypertension, but rather his/her chances of having a myocardial infarction. The predictive value, or posttest probability, can thus be more relevant for patient care. It may be especially important for prognostic models in which the clinical question is the chance of disease development in the future given current risk factors.

## Conclusion

When the goal of a predictive model is to categorize individuals into risk strata, the assessment of such models should be based on how well they achieve this aim. Inclusion of novel risk factors in risk prediction equations could lead to more accurate risk classification, despite little change in the *c* statistic. The use of a single, somewhat insensitive, measure of model fit such as the *c* statistic can erroneously eliminate important clinical risk predictors for consideration in scoring algorithms.

A full discussion of model fitting and validation is beyond the scope of this paper (see Harrell[2]), but some simple suggestions for comparison of predictive models

Downloaded from http://circ.ahajournals.org/ by guest on June 24, 2017

**TABLE 4. Suggestions for Comparison of Models for Risk Prediction**

1. To compare global model fit, use a measure based on the log likelihood function, such as the Bayes Information Criterion, in which lower values indicate better fit and a penalty is paid if the number of variables is increased.
2. Compare general indices of calibration (such as the Hosmer-Lemeshow statistic, which compares the observed and predicted risk within categories) and discrimination (such as the c-statistic).
3. If global fit is better for 1 model but general calibration and discrimination are similar, fit may be better among some individuals (for instance, those at higher risk). Determine how many individuals would be reclassified in clinical risk categories and whether the new risk category is more accurate for those reclassified.
4. For clinical use of a new invasive or expensive biomarker, determine if a higher or lower estimated risk would change treatment decisions for the individual patient.

are shown in Table 4. First, a sensitive measure, such as the likelihood ratio test, or the Bayes information criterion, should be used to determine global model fit. The Bayes information criterion applies a penalty for the number of variables and can compare models that are not nested. It is related to the posterior probability that the model is correct, and is a conservative criterion for model selection. Second, measures of calibration and discrimination, such as the Hosmer-Lemeshow statistic and the *c* statistic, can be informative and should also be assessed. When these statistics give different answers, it may be that fit is better for a subset of individuals, such as those at higher risk, and predicted risks for individuals should be compared. One can determine the extent of reclassification in clinically important risk categories, and which model classifies more accurately. Finally, an important criterion for a new marker's usefulness in practice is whether its measurement could lead to different treatment decisions.

Ultimately the decision whether to include a new risk factor in prediction models depends on relative costs, both in terms of dollars and in the potential for illness prevented and lives saved.[28] In the setting of prospective risk prediction, the proportion of patients reclassified correctly, rather than the *c* statistic, would seem to have more relevance for such calculations. Currently, several potential biomarkers for cardiovascular risk have been proposed by various investigators. Although individual predictors may add incremental value to risk prediction, the possibilities for model improvement are greater for combinations of markers. The most promising of these novel risk factors should thus be examined rigorously and simultaneously to evaluate their potential role in improved models for clinical risk prediction.

## Acknowledgment

The author wishes to thank Fran Cook for helpful comments and suggestions.

## Sources of Funding

This work was supported by a grant from the Donald W. Reynolds Foundation (Las Vegas, Nevada). The Women's Health Study cohort 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.

## Disclosures

None.

## References

1. 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.
2. Harrell FE Jr. *Regression Modeling Strategies.* New York: Springer; 2001.
3. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. *Comm Stat.* 1980;A10:1043–1069.
4. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med*. 1997; 16:965–980.
5. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiol*. 2003;229:3–8.
6. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med*. 1999;18:2529–2545.
7. D'Agostino RB, Griffith JL, Schmidt CH, Terrin N. Measures for evaluating model performance. In: *Proceedings of the Biometrics Section.* Alexandria, VA: American Statistical Association, Biometrics Section; 1997:253–258.
8. Greenland P, O'Malley PG. When is a new prediction marker useful? a consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk. *Arch Intern Med*. 2005;165:2454–2456.
9. Lloyd-Jones DM, Liu K, Tian L, Greenland P. Narrative review: assessment of C-reactive protein in risk prediction for cardiovascular disease. *Ann Intern Med*. 2006;145:35–42.
10. Blankenberg S, McQueen MJ, Smieja M, Pogue J, Balion C, Lonn E, Rupprecht HJ, Bickel C, Tiret L, Cambien F, Gerstein H, Münzel T, Yusef S, for the HOPE Study Investigators. Comparative impact of multiple biomarkers and N-terminal pro-brain natriuretic peptide in the context of conventional risk factors for the prediction of recurrent cardiovascular events in the Heart Outcomes Prevention Evaluation (HOPE) study. *Circ.* 2006;114:201–208.
11. Lloyd-Jones DM, Tian L. Predicting cardiovascular risk: so what do we do now? *Arch Intern Med*. 2006;166:1342–1344.
12. de Lemos JA. The latest and greatest new biomarkers: which ones should we measure for risk prediction in our practice? *Arch Intern Med*. 2006; 166:2428–2430.
13. Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, Benjamin EJ, D'Agostino RB, Vasan RS. Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med*. 2006;355:2631–2639.
14. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med*. 2006;355:2615–2617.
15. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29–36.
16. Harrell FEJ, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. 1982;247:2543–2546.
17. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med*. 1997;16: 981–991.
18. 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;8:12–17.
19. Hlatky MA, Pryor DB, Harrell FEJ, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. *Am J Med*. 1984;77:64–71.
20. Levy D, Labib SB, Anderson KM, Christiansen JC, Kannel WB, Castelli WP. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. *Circulation*. 1990;81:1144–1146.

20a. Ridker PM, Cook NR, Lee IM, Gordon D, Gaziano JM, Manson JE, Hennekens CH, Buring JE. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. *N Engl J Med.* 2005;352:1293–1304.

21. Lloyd-Jones DM, Nam B-H, D'Agostino RB, Levy D, Murabito JM, Wang TJ, Wilson PWF, O'Donnell CJ. Parental cardiovascular disease as

Downloaded from http://circ.ahajournals.org/ by guest on June 24, 2017

a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring. *JAMA*. 2004;291:2204–2211.
22. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97:1837–1847.
23. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159:882–890.
24. Rose G. Sick individuals and sick populations. *Int J Epidemiol*. 1985;14:32–38.
25. Drizd T, Dannenberg AL, Engel A. Blood pressure levels in persons 18–74 years of age in 1976–80, and trends in blood pressure from 1960 to 1980 in the United States. *Vital Health Stat 11*. 1986;234:1-68.
26. Carroll M, Sempos C, Briefel R, Gray S, Johnson C. Serum lipids of adults 20–74 years: United States, 1976–80. *Vital Health Stat 11*. 1993;242:1–107.
27. Diamond GA. What price perfection? calibration and discrimination of clinical prediction models. *J Clin Epidemiol*. 1992;45:85–89.
28. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostat*. 2005;6:227–239.
29. Pepe MS, Feng Z, Huang Y, Longton GM, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. *UW Biostatistics Working Paper Series, #289*. 2006. Available at: http://www.bepress.com/uwbiostat/paper289. Accessed October 20, 2006.
30. Folsom AR, Chambliss LE, Ballantyne CM, Coresh J, Heiss G, Wu KK, Boerwinkle E, Mosley THJ, Sorlie P, Diao G, Sharrett R. An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk markers: the Atherosclerosis Risk in Communities study. *Arch Intern Med*. 2006;166:1368–1373.
31. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann Intern Med*. 2006;145:21–29.
32. Ridker PM, Cook NR. Should age and time be eliminated from cardiovascular risk prediction models? rationale for the creation of a new national risk detection program. *Circ*. 2005;111:657–658.
33. Moons KGM, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Acad Radiol*. 2003;10:670–672.
34. Guggenmoos-Holzmann I, van Houwelingen HC. The (in)validity of sensitivity and specificity. *Stat Med*. 2000;19:1783–1792.

Downloaded from http://circ.ahajournals.org/ by guest on June 24, 2017

Circulation

American Heart Association®

# Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction

Nancy R. Cook

*Circulation.* 2007;115:928-935
doi: 10.1161/CIRCULATIONAHA.106.672402
*Circulation* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2007 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/115/7/928

Downloaded from http://circ.ahajournals.org/ by guest on June 24, 2017

**Permissions:** Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Circulation* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

**Reprints:** Information about reprints can be found online at:
http://www.lww.com/reprints

**Subscriptions:** Information about subscribing to *Circulation* is online at:
http://circ.ahajournals.org//subscriptions/