

Advances in Measuring the Effect of Individual Predictors of Cardiovascular Risk: The Role of Reclassification Measures

Nancy R. Cook, ScD, and Paul M Ridker, MD

Models for risk prediction are widely used in clinical practice to stratify risk and assign treatment strategies. The contribution of new biomarkers has largely been based on the area under the receiver-operating characteristic curve, but this measure can be insensitive to important changes in absolute risk. Methods based on risk stratification have recently been proposed to compare predictive models. Such methods include the reclassification calibration statistic, the net reclassification improvement, and the integrated

discrimination improvement. This article demonstrates the use of reclassification measures and illustrates their performance for well-known cardiovascular risk predictors in a cohort of women. These measures are targeted at evaluating the potential of new models and markers to change risk strata and alter treatment decisions.

Ann Intern Med. 2009;150:795-802.

www.annals.org

For author affiliations, see end of text.

Risk prediction equations are used in various fields for risk stratification and to determine cost-effective and appropriate courses of treatment. The Framingham risk score, for example, has been used by the Adult Treatment Panel III (1) in guidelines for cholesterol-lowering therapy. Whether new risk predictors can add to a score in terms of clinical utility is an important question in many areas of research.

Traditionally, risk models have been evaluated by using the area under the receiver-operating characteristic curve (2), but this method has been criticized as being insensitive in comparing models (3) and for having little direct clinical relevance (4). New methods have recently been proposed to evaluate and compare predictive risk models. These are based primarily on stratification into clinical categories on the basis of risk and attempt to assess the ability of new models to more accurately reclassify individuals into higher or lower risk strata (5).

Since its first description in 2006 (6), much interest has been generated in reclassification, and although the approach is still in its infancy, there have been further methodological developments (7–9). Researchers in the fields of breast cancer (10), diabetes (11, 12), and genetics (12–14), as well as clinical cardiology (15–18), have published articles using these techniques. Our article is intended to be a guide for understanding this research, including the strengths and known limitations of and differences among the various new methods. We apply these to known predictors of cardiovascular disease in a cohort of women to describe how the new methods perform relative to more traditional ones.

CARDIOVASCULAR RISK EXAMPLE

Data are from the Women's Health Study (WHS), a large-scale, nationwide cohort of U.S. women 45 years or older who were free of cardiovascular disease and cancer at study entry beginning in 1992 (19). Women were followed annually for the development of cardiovascular disease, with an average follow-up of 10 years through March 2004. All reported cardiovascular disease outcomes, includ-

ing myocardial infarction, ischemic stroke, coronary revascularization procedures, and deaths from cardiovascular causes, were adjudicated by an end points committee after medical record review. During follow-up, 766 cardiovascular events occurred. All study participants provided written informed consent, and the institutional review board at Brigham and Women's Hospital, Boston, Massachusetts, approved the study protocol.

The baseline characteristics of the WHS sample are described elsewhere (20). We assayed baseline blood samples for total, high-density, and low-density lipoprotein cholesterol levels with direct-measurement assays (Roche Diagnostics, Basel, Switzerland). For C-reactive protein, we used a validated, high-sensitivity assay (Denka Seiken, Tokyo, Japan). Women eligible for the current analysis had adequate baseline plasma samples; had complete ascertainment of exposure data of interest, including age, blood pressure, current smoking status, diabetes, and parental history of myocardial infarction before age 60 years; and were used in the development and assessment of the Reynolds Risk Score for women ($n = 24\,558$) (20).

We fit models by using Cox proportional hazards models for cardiovascular risk. Predictors included components of the Framingham risk score (age [years], systolic blood pressure [mm Hg], current smoking status [yes or no], and total and high-density lipoprotein cholesterol levels [mg/dL]), as well as additional risk predictors included in the Reynolds Risk Score (hemoglobin A_{1c} [%] among patients with diabetes only, high-sensitivity C-reactive protein [mg/L], and parental history of myocardial infarction

See also:

Print

Glossary 796

Web-Only

Appendix
Appendix Table
Conversion of graphics into slides

Glossary

Traditional measures of model fit

R^2 : Nagelkerke generalized model R^2 , which is the fraction of the log likelihood explained by the predictors in the model, adjusted to a range of 0 to 1. Other definitions of R^2 are sometimes used.

Bayes information criterion: The value of the log likelihood with an added penalty for the number of parameters (variables) in the model. The penalty adjusts for the improvement due to adding variables. A lower number indicates better fit.

c-statistic: The area under the receiver-operating characteristic curve (or c-index for survival data). This is a measure of discrimination, or separation between case patients and control participants, based on ranks.

Hosmer–Lemeshow chi-square statistic: A test of goodness-of-fit of the model. It compares the observed number of events with that predicted from the model within categories (typically 10) on the basis of the predicted risk. It is often referred to as a test of calibration. A significant P value indicates lack of fit.

New measures based on reclassification

Reclassification calibration statistic: A test based on the Hosmer–Lemeshow statistic that compares the observed and expected number of events in each cell of a reclassification table. It should generally be restricted to cells with at least 20 observations. This is a test of calibration or whether the observed and predicted events are similar. A significant result indicates a lack of fit.

Net reclassification improvement: The net increase versus decrease in risk categories among case patients minus that among control participants.

Integrated discrimination improvement: The difference in Yates slopes between models, in which the Yates slope is the mean difference in predicted probabilities between case patients and control participants.

before age 60 years [yes or no]) assessed at baseline. We used the natural logarithm transformation for systolic blood pressure, total and high-density lipoprotein cholesterol levels, and C-reactive protein to linearize the relationship with outcome. We compared the full model with models without each of the risk predictors in turn but included all other factors. We estimated predicted probabilities at 8-year follow-up, and we based observed rates on the Kaplan–Meier survival estimates at 8 years. We extrapolated all rates to 10 years for presentation.

TRADITIONAL MEASURES OF MODEL FIT

Traditional measures of fit include measures of discrimination or the accurate separation into case patients and control participants, measures of calibration or how well the predicted probabilities compare with the observed (model-free) estimates, and global measures combining both. These criteria can be assessed for binary outcomes, such as from logistic models, or for survival outcomes, such as from the Cox model. These will be illustrated for survival data in the example data, although an important limitation to some of the measures is that they do not currently incorporate censored data.

First, only predictors that are statistically significant (for example, by using a likelihood ratio test) are typically used in predictive models. Overall model fit can be assessed by using the Nagelkerke R^2 (see Glossary), which is analogous to the percentage of variation explained for linear models, and compared by using the Bayes information criterion (see Glossary), a function of the log likelihood with an added penalty for the number of parameters that tends to select parsimonious models. Discrimination is usually assessed by using the c -statistic (see Glossary) or the area under the receiver-operating characteristic curve. The c -index is an analogous measure that incorporates censored data (3). Calibration within categories can be assessed by using the Hosmer–Lemeshow goodness-of-fit statistic (see Glossary) (21), with categories formed by deciles or by intervals of risk (for example, 0% to 2%, 2% to 4%, 4% to 6%, and so on).

Table 1 shows the overall measures of fit for the full model and the model leaving out each risk predictor one at a time in the WHS data. All variables were highly statistically significant, as indicated by the likelihood ratio chi-square test. As expected, the R^2 was highest and the Bayes information criterion was lowest in the full model. The c -statistic for the model without age was 0.76. All others ranged from 0.79 to

Table 1. Effect on Traditional Measures of Model Fit of Deleting Each Variable in Turn From the Reynolds Risk Score Model in the Women's Health Study

| Variable | Likelihood Ratio Chi-Square* | P Value Chi-Square | R^2 , % | Bayes Information Criterion | C-Index | Change in C-Index | P Value for C-Index | Chi-Square for HL1† | P Value HL1 | Chi-Square for HL2‡ | P Value HL2 |
|---|------------------------------|--------------------|-----------|-----------------------------|---------|-------------------|---------------------|---------------------|-------------|---------------------|-------------|
| Reynolds Risk Score | – | – | 8.58 | 14 416.8 | 0.797 | – | – | 10.3 | 0.25 | 9.6 | 0.30 |
| Age | 254.2 | <0.001 | 6.43 | 14 664.3 | 0.760 | 0.037 | <0.001 | 12.5 | 0.128 | 14.2 | 0.077 |
| HbA _{1c} in patients with diabetes | 108.8 | <0.001 | 7.67 | 14 518.8 | 0.786 | 0.010 | <0.001 | 7.4 | 0.49 | 9.1 | 0.34 |
| Current smoking status | 81.8 | <0.001 | 7.89 | 14 491.9 | 0.786 | 0.011 | 0.001 | 5.2 | 0.74 | 7.6 | 0.47 |
| ln(SBP) | 89.7 | <0.001 | 7.83 | 14 499.8 | 0.786 | 0.011 | <0.001 | 7.5 | 0.49 | 15.5 | 0.050 |
| ln(HDL) | 51.5 | <0.001 | 8.15 | 14 461.6 | 0.789 | 0.007 | 0.002 | 9.2 | 0.33 | 15.2 | 0.056 |
| ln(TC) | 33.7 | <0.001 | 8.30 | 14 443.8 | 0.793 | 0.004 | 0.028 | 10.6 | 0.23 | 10.5 | 0.23 |
| ln(hsCRP) | 26.1 | <0.001 | 8.36 | 14 436.2 | 0.794 | 0.002 | 0.110 | 6.9 | 0.54 | 9.6 | 0.30 |
| Parental history of MI | 15.6 | <0.001 | 8.45 | 14 425.7 | 0.795 | 0.002 | 0.098 | 11.5 | 0.175 | 12.0 | 0.152 |

HbA_{1c} = hemoglobin A_{1c}; HDL = high-density lipoprotein cholesterol; HL = Hosmer–Lemeshow statistic; hsCRP = high-sensitivity C-reactive protein; ln = natural logarithm; MI = myocardial infarction; SBP = systolic blood pressure; TC = total cholesterol.

* Likelihood ratio chi-square test for each variable separately.

† Hosmer–Lemeshow statistic using deciles of risk for survival probabilities.

‡ Hosmer–Lemeshow statistic using categories in 2% increments for survival probabilities.

Figure 1. Reclassification table comparing 10-year risk strata for models that include risk factors for cardiovascular disease in the Women's Health Study with and without SBP.

| Model Without SBP | Model With SBP | | | | Total, n (%) | Reclassified Into New Risk Category, % | | |
|--------------------------|----------------|-------------|------------|------------|----------------|--|--------|-------|
| | 0%–5% | 5%–10% | 10%–20% | ≥20% | | Lower | Higher | Total |
| 0% to <5% | | | | | | | | |
| Persons included, n (%) | 20 372 (96.6) | 696 (3.3) | 23 (0.1) | 0 (0.0) | 21 091 (85.9) | – | 3.4 | 3.4 |
| Case patients, %* | 218 (82.2) | 38 (14.8) | 0 (0.0) | 0 (0.0) | 256 (10.5) | – | 1.8 | 1.8 |
| Control participants, %* | 19 933 (96.8) | 642 (3.1) | 23 (0.1) | 0 (0.0) | 20 598 (87.2) | – | 3.2 | 3.2 |
| Observed risk, %† | 1.3 | 6.8 | 0.0 | – | 1.5 | – | – | – |
| 5% to <10% | | | | | | | | |
| Persons included, n (%) | 635 (26.6) | 1441 (60.3) | 307 (12.8) | 7 (0.3) | 2390 (9.7) | 26.6 | 13.1 | 39.7 |
| Case patients, %* | 22 (14.2) | 96 (61.9) | 36 (23.2) | 1 (0.6) | 155 (16.1) | 14.2 | 23.9 | 38.1 |
| Control participants, %* | 589 (27.4) | 1291 (60.1) | 263 (12.2) | 6 (0.3) | 2149 (9.1) | 27.4 | 12.5 | 39.9 |
| Observed risk, %† | 4.4 | 8.4 | 14.6 | 17.5 | 8.2 | – | – | – |
| 10% to <20% | | | | | | | | |
| Persons included, n (%) | 4 (0.5) | 204 (25.0) | 519 (63.5) | 90 (11.0) | 817 (3.3) | 25.5 | 11.0 | 36.5 |
| Case patients, %* | 0 (0.0) | 7 (7.8) | 59 (65.6) | 24 (26.7) | 90 (27.7) | 7.8 | 26.7 | 34.5 |
| Control participants, %* | 4 (0.6) | 188 (27.5) | 434 (63.4) | 58 (8.5) | 684 (2.9) | 28.1 | 8.5 | 36.6 |
| Observed risk, %† | 0.0 | 4.3 | 14.3 | 34.2 | 13.9 | – | – | – |
| ≥20% | | | | | | | | |
| Persons included, n (%) | 0 (0.0) | 2 (0.8) | 54 (20.8) | 204 (78.5) | 260 (1.1) | 21.5 | – | 21.5 |
| Case patients, %* | 0 (0.0) | 0 (0.0) | 11 (18.6) | 48 (81.4) | 59 (45.7) | 18.6 | – | 18.6 |
| Control participants, %* | 0 (0.0) | 2 (1.1) | 38 (21.1) | 140 (77.8) | 180 (0.8) | 22.2 | – | 22.2 |
| Observed risk, %† | – | 0.0 | 25.9 | 29.4 | 28.4 | – | – | – |
| Total | | | | | | | | |
| Persons included, n (%) | 21 011 (85.6) | 2343 (9.5) | 903 (3.7) | 301 (1.2) | 24 558 (100.0) | – | – | – |
| Case patients, %* | 240 (42.9) | 141 (25.2) | 106 (18.9) | 73 (13.0) | 560 (100.0) | – | – | – |
| Control participants, %* | 20 526 (86.9) | 2123 (8.9) | 758 (3.2) | 204 (0.9) | 23 611 (100.0) | – | – | – |
| Observed risk, %† | 1.4 | 7.6 | 14.7 | 30.5 | – | – | – | – |

Red shading indicates an increase in risk category; and blue shading indicates a decrease in risk category. Total reclassified = 2022 (8.23%); total reclassified in cells with at least 20 observations = 2009. SBP = systolic blood pressure.

* Case patients and control participants at 8 years of follow-up, ignoring censored observations.

† Observed risk at 10 years is estimated from Kaplan–Meier curve by using observations within each cell. The reclassification calibration statistic compares the observed risk with the average predicted risk within each cell. The chi-square statistic for the model without SBP is 68.3 ($P < 0.001$); for the model with SBP, chi-square is 22.9 ($P = 0.006$). Reclassification improvement is 10.5% among case patients (99 – 40 of 560), whereas classification worsened in control participants by 0.7% (821 – 992 of 23 611), leading to a net reclassification improvement of 9.8%.

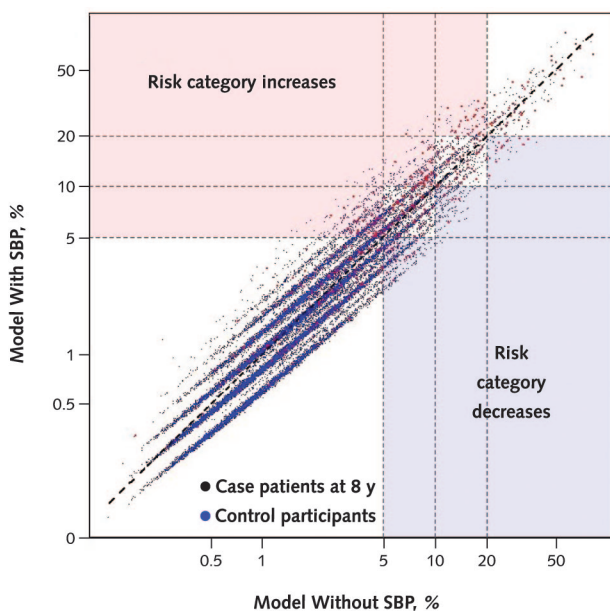
0.80 for the full model. The differences in the c-statistic were 0.01 or less for all variables except age, but were statistically significant for all except C-reactive protein and family history. When the optimism of this measure was assessed using 100 bootstrap samples (3), the value of the c-index was generally decreased by 0.002 or less. The Hosmer–Lemeshow test, within deciles or 2% risk intervals, demonstrated adequate fit, although some deviation from fit was suggested when 2% risk intervals were used.

RISK RECLASSIFICATION AND CALIBRATION

Risk reclassification for single factors can be examined by using models with and without each risk factor in turn, that is,

comparing a model without a given risk factor to the full model. For cardiovascular disease, relevant strata are 0% to less than 5%, 5% to less than 10%, 10% to less than 20%, and 20% or more for 10-year risk. **Figure 1** illustrates the risk reclassification for models with and without systolic blood pressure but includes all other risk factors. The model without systolic blood pressure categorized 86% of women into the lowest risk group (that is, 10-year risk was <5%). Ten percent were categorized into the 5% to less than 10% risk stratum, 3% were categorized into the 10% to less than 20% risk stratum, and 1% was categorized into the 20% or higher risk stratum. The same was approximately true for the model that included systolic blood pressure, such that the “marginal” proportions were very similar.

Figure 2. Plot of predicted 10-year risk from models including cardiovascular risk factors but with and without systolic blood pressure in the Women's Health Study.



The dashed diagonal line is the line of unity; horizontal and vertical lines represent risk strata cut-points. SBP = systolic blood pressure.

Figure 2 provides a continuous analogue of this table and shows the predicted values from both models, along with the category cut-points. Figure 2 shows the spread and difference in the values of the logarithm base 10 for the predicted risks from the 2 models, with the dashed diagonal line denoting the line of identity. Ideally, more case patients will be above than below the line of identity. The horizontal and vertical lines indicate the risk strata. The striated appearance is because of the use of categories for systolic blood pressure (9 categories of 10 mm Hg) from less than 110 mm Hg to 180 mm Hg or more. The lines show how much the predicted values can change when systolic blood pressure increases or decreases by 10 mm Hg.

The overall percentage reclassified gives some indication of how many persons would change risk categories, and possibly treatment decisions, under the new model. Of the 24 558 women, 2022 (8%) were classified into different risk strata. The overall percentage, however, is heavily influenced by the incidence of disease in the sample. In the WHS, most women were in the lowest category under both models. Those who were in the intermediate categories may be more clinically relevant and demonstrate more shift in risk category. For example, of those at 5% to less than 10% risk in the model without systolic blood pressure, 40% were reclassified into higher or lower categories. Of those at 10% to less than 20% risk, 36% were reclassified.

More important for model fit than the simple percentage reclassified, however, is a comparison of observed and expected rates of disease within each cross-classified category. This determines whether individuals are reclassified correctly or whether the changes are due to chance. Observations in a reclassified cell are considered “correctly” reclassified if the observed rate is closer to the new than to the old risk stratum. For example, 696 women were reclassified from less than 5% 10-year risk to 5% to less than 10% risk. The observed 10-year risk was 6.8% based on a Kaplan–Meier estimate for these 696 women, which falls into the 5% to less than 10% category. The average estimated risk for these women from the model without systolic blood pressure was 4.0%, whereas risk was 6.1% from the model with systolic blood pressure, which is closer to the 6.8% observed risk. Overall, 2022 women were reclassified; 2009 of these were in cells with at least 20 women, for whom the observed rate could be computed. Of these 2009 women, 1932 (96%) were reclassified correctly.

Observed and average predicted rates for cells with at least 20 observations can be compared on the basis of a chi-square goodness-of-fit test within reclassified categories for each model separately (9). This is simply the familiar Hosmer–Lemeshow goodness-of-fit statistic, but applied to reclassified categories, and we refer to it as the reclassification calibration statistic (see Glossary). It is calculated as follows:

$$X_{RC}^2 = \sum_{i=1}^K \frac{(O_k - n_k \bar{p}_k)^2}{n_k \bar{p}_k (1 - \bar{p}_k)} \sim X_{K-2}^2$$

where n_k is the number in cell k , O_k is the observed number of events in cell k , and \bar{p}_k is the average predicted risk in cell k for the model under consideration. Survival data can be incorporated by using the observed events and predicted risk for a given time, such as 10 years. The Kaplan–Meier estimates of the observed risk can be used to accommodate censored data. The statistic follows an approximate chi-square distribution with $K-2$ degrees of freedom, where K is the number of cells with at least 20 observations in the table. In Figure 1, $K = 11$. As with the usual Hosmer–Lemeshow test, a statistically significant result indicates poor fit. The test for systolic blood pressure found that the model without systolic blood pressure had a strong lack of fit (chi-square test, 68.3; $P < 0.001$). For the model with systolic blood pressure, the chi-square test yields 22.9 ($P = 0.006$), which still indicated some lack of fit but to a much lesser extent. We did all computations for this and other reclassification measures by using SAS, version 9.1 (SAS Institute, Cary, North Carolina). SAS macros to compute the reclassification measures are available in the Appendix (available at www.annals.org).

Table 2 examines risk reclassification from the initial reduced model eliminating each predictor individually compared with the full model. The overall percentage re-

Table 2. Reclassification Measures for Deleting Variables From the RRS Model in the Women's Health Study*

| RRS Variables | Reclassification Percentages | | | Fit to Survival Probabilities | | | | |
|--|------------------------------|--------|---------|-------------------------------|----------------------------|------------------------|---------------------|-----------------|
| | Overall | 5%–10% | 10%–20% | Correct, % | Chi-Square for Exclusionst | P Value for Exclusions | Chi-Square for RRS‡ | P Value for RRS |
| Age | 13.4 | 61.9 | 61.2 | 100.0 | 161.0 | <0.001 | 7.0 | 0.80 |
| HbA _{1c} levels in patients with diabetes | 5.9 | 27.2 | 38.8 | 100.0 | 154.2 | <0.001 | 23.2 | 0.010 |
| Current smoking status | 6.8 | 32.3 | 31.9 | 96.3 | 62.6 | <0.001 | 16.2 | 0.040 |
| ln(SBP) | 8.2 | 39.7 | 36.5 | 96.2 | 68.3 | <0.001 | 22.9 | 0.006 |
| ln(HDL) | 6.2 | 29.8 | 28.3 | 96.1 | 27.8 | <0.001 | 12.3 | 0.137 |
| ln(TC) | 4.8 | 22.9 | 21.0 | 95.8 | 16.1 | 0.041 | 7.3 | 0.51 |
| ln(hsCRP) | 4.0 | 19.2 | 16.1 | 97.0 | 31.8 | <0.001 | 18.5 | 0.017 |
| Parental history of MI | 2.8 | 13.5 | 12.7 | 96.4 | 22.2 | 0.005 | 13.3 | 0.101 |

HbA_{1c} = hemoglobin A_{1c}; HDL = high-density lipoprotein cholesterol; hsCRP = high-sensitivity C-reactive protein; IDI = integrated discrimination improvement; ln = natural logarithm; MI = myocardial infarction; NRI = net reclassification improvement; RI = reclassification improvement; RRS = Reynolds Risk Score; SBP = systolic blood pressure; TC = total cholesterol.

* Risk strata use 4 categories: 0% to <5%, 5% to <10%, 10% to <20%, and ≥20%. Calibration statistics are based on observed 8-year rates from Kaplan–Meier curves. Net reclassification improvement statistics use 8-year case–control status, excluding those who are censored.

† Reclassification calibration (Hosmer–Lemeshow) statistic and *P* value for RRS model omitting each variable in turn. Significance indicates lack of fit.

‡ Reclassification calibration statistic and *P* value for full RRS model.

classified ranged from 3% for models with and without parental history of myocardial infarction to 13% for models with and without age. The percentages that were reclassified within the intermediate risk categories of 5% to less than 10% and 10% to less than 20% were much higher, ranging from at least 13% to 62% for age, suggesting more substantial changes within these risk strata. For each model comparison, more than 95% of those reclassified were reclassified correctly when the variable was included. In comparing the observed with the expected rates, the reclassification calibration statistic showed significant lack of fit in models that excluded each variable. Although the full model sometimes demonstrated a lesser degree of lack of fit within these cross-classified categories, the full model provided better fit to the observed rates in each comparison. Thus, each of these variables improved the fit of the model to the observed rates of cardiovascular disease.

MEASURES OF IMPROVEMENT IN DISCRIMINATION

Some other proposed measures of improvement include the net reclassification improvement (7) (see Glossary) and the integrated discrimination improvement (7) (see Glossary). The net reclassification improvement (*NRI*) assesses risk reclassification and is the difference in proportions moving up and down risk strata among case patients versus control participants, that is, those who did or did not develop the disease during follow-up, or, where *Pr* stands for probability:

$$NRI = [Pr(up|cases) - Pr(down|cases)] - [Pr(up|controls) - Pr(down|controls)].$$

The net reclassification improvement is similar to the simple percentage reclassified, but it distinguishes move-

ments in the correct direction (up for case patients and down for control participants). Ideally, the predicted probabilities would move higher (up a category) for case patients and lower (down a category) for control participants. The net reclassification improvement can be rearranged to reflect improvement in both cases and controls as follows:

$$\begin{aligned} NRI &= [Pr(up|cases) - Pr(down|cases)] \\ &+ [Pr(down|controls) - Pr(up|controls)] \\ &= \text{relative improvement for cases} \\ &+ \text{relative improvement controls.} \end{aligned}$$

The net reclassification improvement is then the sum of improvements for case patients and control participants.

Figure 1 also shows the data representation for systolic blood pressure for case patients and control participants separately in the WHS. In this cohort study, control participants were defined as those who did not develop disease at 8-year follow-up. Of the case patients, 38 + 36 + 1 + 24 = 99 (17.6%) correctly moved up a risk category and 22 + 7 + 11 = 40 (7.1%) incorrectly moved down when adding systolic blood pressure to the model, resulting in a relative improvement for case patients (percentages moving up – those moving down) of 10.5%. For the control participants, 821 (3.5%) correctly moved down, whereas 992 (4.2%) incorrectly moved up, yielding an overall change of –0.7%. This shows a slight upward movement among control participants or a worsening of classification for control participants. The net reclassification improvement is the sum of the 2, or 9.8%. This means that compared with control participants, case patients were almost 10% more likely to move up a category than down. Table 2 shows the results for other variables in the WHS. The net reclassification improvement was highest for age, at 19.5%,

Table 2—Continued

| Fit Conditional on Case–Control Status | | | | | |
|--|--------------------------------|--------|-----------------|--------|-----------------|
| RI for Case Patients, % | RI for Control Participants, % | NRI, % | P Value for NRI | IDI, % | P Value for IDI |
| 21.8 | −2.3 | 19.5 | <0.001 | 1.79 | <0.001 |
| 9.3 | 1.7 | 11.0 | <0.001 | 1.33 | <0.001 |
| 8.9 | −0.4 | 8.5 | <0.001 | 0.56 | <0.001 |
| 10.5 | −0.7 | 9.8 | <0.001 | 0.49 | 0.001 |
| 4.5 | −0.4 | 4.0 | 0.021 | 0.39 | <0.001 |
| 3.6 | −0.4 | 3.2 | 0.032 | 0.13 | 0.159 |
| 5.4 | −0.2 | 5.2 | <0.001 | 0.16 | 0.026 |
| 3.4 | −0.1 | 3.2 | 0.005 | 0.11 | 0.040 |

and ranged down to 3.2% for total cholesterol levels. All values were statistically significant in these data.

The integrated discrimination improvement is the difference in Yates, or discrimination, slopes between 2 models, in which the Yates slope is the mean difference in predicted probabilities between case patients and control participants. The integrated discrimination improvement (IDI) is defined as:

$$IDI = (ave \hat{p}_{cases} - ave \hat{p}_{controls})_{new \ model} - (ave \hat{p}_{cases} - ave \hat{p}_{controls})_{old \ model}$$

where \hat{p} is the predicted probability. The terms in parentheses are the Yates slopes for the 2 models; ideally, we would like the case patients to have a higher average probability than the control participants. The difference in slopes is a measure of improvement in the model. The integrated discrimination improvement can also be thought of as a percentage of variance explained (8). In the model without systolic blood pressure, the average predicted probability was 7.4% for the case patients and 2.2% for the control participants, yielding a slope of 5.2%. In the model that included systolic blood pressure, the averages were 7.9% for case patients and 2.2% for control participants, yielding a slope of 5.7%. The integrated discrimination improvement for systolic blood pressure is the difference in these, or 0.5% (Table 2). This means that the difference in average predicted probabilities between case patients and control participants increased by 0.005 when systolic blood pressure was added to the model. For the models that omitted age, hemoglobin A_{1c} levels, and all other predictors, the integrated discrimination improvement was 1.8%, 1.3%, and less than 0.6%, respectively.

Both the net reclassification improvement and integrated discrimination improvement condition on case–control status, and neither of these measures assess model calibration (7). Because they depend on outcome status, they are not available for censored data. Status at 8-year follow-up was used ad hoc in these analyses because most women had been followed for at least this duration, and observations censored before 8 years of follow-up were ex-

cluded. Of the total 766 cardiovascular events, 560 (73%) occurred by 8 years of follow-up and could be used for calculation of the net reclassification improvement and integrated discrimination improvement. An additional 387 women were censored before 8 years and were excluded from calculation of these measures. This is an important current limitation of these measures.

When reclassification into risk strata is considered, the particular categories used can affect the estimates. For example, the Appendix Table (available at www.annals.org) shows the results when 3 categories were used with cut-points of 5% and 20% only (7). The percentages that were reclassified were lower, as would be expected. The value of the net reclassification improvement, an adjusted percentage reclassified, was also lower and ranged from 14% for age to 1.5% for parental history of myocardial infarction. This was reduced further when only 2 categories (with a 10% cut-point) were used, with a net reclassification improvement for age of only 9.2%; the reclassification chi-square statistic was also reduced, but so was its expectation, which is based on the number of cells. In a 4 × 4 table, the number of cells with more than 20 observations is often 10; for the 3 × 3 table, it is usually 7. The values of the chi-square statistic divided by its expectation, as well as the levels of statistical significance, were relatively similar whether 3 or 4 categories were used.

Whenever fit of a new model is evaluated, the estimated fit in the derivation data may be too optimistic. Use of a test data set, bootstrapping, or cross-validation can adjust measures for optimism (3). When 10-fold cross-validation was applied to these data, there was little change in the estimated effects or in the test statistics (data not shown). The net reclassification improvement for age increased from 19.5% to 20.5%, and for family history decreased from 3.2% to 2.2%. The reclassification chi-square statistics were also relatively similar.

DISCUSSION

These data illustrate the use of several newly proposed reclassification measures and demonstrate their magnitude and variation for well-known cardiovascular risk factors in a cohort of women. The net reclassification improvement ranged from 3.2% for parental history of myocardial infarction before age 60 years to almost 20% for age in these data using clinically meaningful risk strata. Other studies have presented similar results for some of these variables (7, 9, 22, 23). Of note, a significant statistical association does not necessarily lead to an improvement in risk stratification. For example, although a polymorphism at the 9p21 gene was associated with increased risk for cardiovascular disease, it did not improve calibration, and the estimated net reclassification improvement was negative (13). Also, as noted by Pepe and colleagues (8), testing the integrated discrimination improvement is equivalent to testing whether the regression coefficient in a model is 0. It can be

represented as a change in R^2 or the proportion of variance explained. Whether this can translate to clinical utility thus remains questionable.

The net reclassification improvement and the integrated discrimination improvement both condition on case-control or later disease status. As such, they do not provide information on calibration of the estimated risks. As for the receiver-operating characteristic curve (5), they do not measure how close the predicted observations fall to the actual probabilities. Alternatively, the reclassification calibration statistic directly compares the observed and predicted probabilities within each cell of the reclassification table, and assesses this calibration directly. In particular, the reclassification calibration statistic for the model excluding the variable of interest examines whether the model without this variable provides adequate fit; it is therefore a useful adjunct to measures of discrimination.

Because the net reclassification improvement and integrated discrimination improvement condition on outcome status, how to assess these measures with survival data is not yet clear. In our analysis, some women were censored before 8 years and were simply excluded from the calculations. The reclassification calibration measures, however, can use survival analysis to determine the observed rate within each cell while allowing for such censoring. They are thus more readily generalizable to survival or prospective data, particularly when length of follow-up differs among individuals.

A limitation of the net reclassification improvement and other reclassification measures is that they depend on the particular categories used. The calibration test seems to depend somewhat less on the number of categories because it is adjusted for the number of categories. The best choice of risk categories for reclassification tables remains an important question, however. For best interpretation, the choice should have clinical meaning. As Greenland (24) suggests, "predictive values, costs, and cut-points must be considered together to make informed decisions." For cost-effectiveness and public policy considerations, categories based on the absolute predicted risk, rather than ranks, would be of most interest. If a particular category is important, such as a 10% treatment threshold, one may wish to include categories both above and below this number. In preventive cardiology, the 5%, 10%, and 20% cut-points have been proposed as clinical choices for treatment decisions (1, 25), which ideally should be based on considerations of cost-effectiveness. Net benefit curves (26) or relative utility estimates (27) can help determine whether it is cost-effective to measure a new marker in a sample or a subset at intermediate risk.

Even if clinical or treatment categories are not widely available, reclassification measures, particularly the reclassification calibration statistic and net reclassification improvement, may be useful in demonstrating the ability of new models and markers to change risk strata and alter treatment decisions. Although their sta-

tistical properties are still being explored, they are gaining popularity as other means of comparing the accuracy of absolute risk estimates.

From the Donald W. Reynolds Center for Cardiovascular Research and the Center for Cardiovascular Disease Prevention, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts.

Grant Support: By the Donald W. Reynolds Foundation (Dr. Ridker), and the Leducq Foundation (Dr. Ridker). The overall Women's Health Study is supported by grants from the National Heart, Lung, and Blood Institute and the National Cancer Institute (HL-43851 and CA-47988).

Potential Financial Conflicts of Interest: *Consultancies:* P.M. Ridker (AstraZeneca, Schering-Plough, Sanofi Aventis, ISIS, Siemens, Merck, Novartis, Vascular Biogenics). *Grants received:* P.M. Ridker (National Heart, Lung, and Blood Institute, National Cancer Institute, Donald W. Reynolds Foundation, Leducq Foundation, AstraZeneca, Merck, Novartis, Abbott, Roche, Sanofi Aventis). *Patents received:* P.M. Ridker (Brigham and Women's Hospital). *Royalties:* P.M. Ridker (Brigham and Women's Hospital).

Requests for Single Reprints: Nancy R. Cook, ScD, Division of Preventive Medicine, Brigham and Women's Hospital, 900 Commonwealth Avenue East, Boston, MA 02215; e-mail, ncook@rics.bwh.harvard.edu.

Current author addresses are available at www.annals.org.

References

1. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *JAMA*. 2001;285:2486-97. [PMID: 11368702]
2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29-36. [PMID: 7063747]
3. Harrell FE, Jr. Regression Modeling Strategies. New York: Springer; 2001.
4. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med*. 2008;149:751-60. [PMID: 19017593]
5. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115:928-35. [PMID: 17309939]
6. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann Intern Med*. 2006;145:21-9. [PMID: 16818925]
7. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27:157-72; discussion 207-12. [PMID: 17569110]
8. Pepe MS, Feng Z, Gu JW. Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., *Statistics in Medicine* (DOI: 10.1002/sim.2929). *Stat Med*. 2008;27:173-81. [PMID: 17671958]
9. Cook NR. Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., *Statistics in Medicine* (DOI: 10.1002/sim.2929). *Stat Med*. 2008;27:191-5. [PMID: 17671959]
10. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann Intern Med*. 2008;148:337-47. [PMID: 18316752]
11. Chien K, Cai T, Hsu H, Su T, Chang W, Chen M, et al. A prediction model for type 2 diabetes risk among Chinese people. *Diabetologia*. 2009;52:443-50. [PMID: 19057891]
12. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, et al.

Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med.* 2008;359:2208-19. [PMID: 19020323]

13. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. *Ann Intern Med.* 2009;150:65-72. [PMID: 19153409]

14. Kathiresan S, Melander O, Anevski D, Guiducci C, Burt NP, Roos C, et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med.* 2008;358:1240-9. [PMID: 18354102]

15. Ingelsson E, Schaefer EJ, Contois JH, McNamara JR, Sullivan L, Keyes MJ, et al. Clinical utility of different lipid measures for prediction of coronary heart disease in men and women. *JAMA.* 2007;298:776-85. [PMID: 17699011]

16. Holme I, Aastveit AH, Jungner I, Walldius G. Relationships between lipoprotein components and risk of myocardial infarction: age, gender and short versus longer follow-up periods in the Apolipoprotein MOrtality RiSk study (AMORIS). *J Intern Med.* 2008;264:30-8. [PMID: 18298486]

17. Rietbrock S, Heeley E, Plumb J, van Staa T. Chronic atrial fibrillation: Incidence, prevalence, and prediction of stroke using the Congestive heart failure, Hypertension, Age >75, Diabetes mellitus, and prior Stroke or transient ischemic attack (CHADS2) risk stratification scheme. *Am Heart J.* 2008;156:57-64. [PMID: 18585497]

18. Ankle Brachial Index Collaboration. Ankle brachial index combined with Framingham Risk Score to predict cardiovascular events and mortality: a meta-analysis. *JAMA.* 2008;300:197-208. [PMID: 18612117]

19. Ridker PM, Cook NR, Lee IM, Gordon D, Gaziano JM, Manson JE, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascu-

lar disease in women. *N Engl J Med.* 2005;352:1293-304. [PMID: 15753114]

20. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *JAMA.* 2007;297:611-9. [PMID: 17299196]

21. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. *Commun Stat* 1980;A10:1043-69.

22. Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. *Circulation.* 2008;118:2243-51, 4p following 2251. [PMID: 18997194]

23. Wilson PW, Pencina M, Jacques P, Selhub J, D'Agostino R, O'Donnell CJ. C-reactive protein and reclassification of cardiovascular risk in the Framingham Heart Study. *Circ Cardiovasc Qual Outcomes.* 2008;1:92-97.

24. Greenland S. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., *Statistics in Medicine* (DOI: 10.1002/sim.2929). *Stat Med.* 2008;27:199-206. [PMID: 17729377]

25. Greenland P, Smith SC Jr, Grundy SM. Improving coronary heart disease risk assessment in asymptomatic people: role of traditional risk factors and non-invasive cardiovascular tests. *Circulation.* 2001;104:1863-7. [PMID: 11591627]

26. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26:565-74. [PMID: 17099194]

27. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. *J R Stat Soc [Ser A].* 2009. [Forthcoming].

INFORMATION FOR AUTHORS

The *Annals* Information for Authors section is available at www.annals.org/shared/author_info.html. All manuscripts must be submitted electronically using the manuscript submission option under the Information for Authors/Reviewers item at www.annals.org.

Current Author Addresses: Drs. Cook and Ridker: Division of Preventive Medicine, Brigham and Women's Hospital, 900 Commonwealth Avenue East, Boston, MA 02215.

Author Contributions: Conception and design: N.R. Cook, P.M. Ridker. Analysis and interpretation of the data: N.R. Cook, P.M. Ridker. Drafting of the article: N.R. Cook.

Critical revision of the article for important intellectual content: N.R. Cook, P.M. Ridker.

Final approval of the article: N.R. Cook, P.M. Ridker.

Provision of study materials or patients: P.M. Ridker.

Statistical expertise: N.R. Cook.

Administrative, technical, or logistic support: N.R. Cook, P.M. Ridker.

Collection and assembly of data: P.M. Ridker.

Appendix Table. Reclassification Measures for Deleting Variables From the RRS Model in the Women's Health Study*

| RRS Variables | Reclassification Percentages | | Fit to Survival Probabilities | | | | Fit Conditional on Case–Control Status | | | | |
|--|------------------------------|--------|-------------------------------|---------------------------|------------------------|---------------------|--|-------------------------|--------------------------------|--------|-----------------|
| | Overall | 5%–20% | Correct, % | Chi-Square for Exclusion† | P Value for Exclusions | Chi-Square for RRS‡ | P Value for RRS | RI for Case Patients, % | RI for Control Participants, % | NRI, % | P Value for NRI |
| Age | 10.7 | 39.3 | 100.0 | 130.5 | <0.001 | 6.2 | 0.29 | 16.1 | –1.6 | 14.4 | <0.001 |
| HbA _{1c} levels in patients with diabetes | 4.2 | 18.9 | 100.0 | 105.0 | <0.001 | 17.8 | 0.003 | 8.2 | 1.1 | 9.3 | <0.001 |
| Current smoking status | 5.1 | 18.7 | 95.1 | 49.9 | <0.001 | 14.9 | 0.011 | 6.1 | –0.2 | 5.9 | <0.001 |
| ln(SBP) | 6.2 | 22.9 | 96.3 | 43.1 | <0.001 | 13.0 | 0.024 | 5.4 | –0.4 | 5.0 | 0.005 |
| ln(HDL) | 4.6 | 16.5 | 94.5 | 25.5 | <0.001 | 11.3 | 0.046 | 3.2 | –0.2 | 3.0 | 0.034 |
| ln(TC) | 3.5 | 12.3 | 100.0 | 12.1 | 0.033 | 6.4 | 0.27 | 1.8 | –0.2 | 1.6 | 0.181 |
| ln(hsCRP) | 2.9 | 10.4 | 96.0 | 13.4 | 0.020 | 7.3 | 0.20 | 2.1 | –0.1 | 2.0 | 0.053 |
| Parental history of MI | 2.0 | 7.2 | 95.0 | 14.2 | 0.014 | 9.5 | 0.092 | 1.6 | –0.1 | 1.5 | 0.093 |

HbA_{1c} = hemoglobin A_{1c}; HDL = high-density lipoprotein cholesterol; hsCRP = high-sensitivity C-reactive protein; ln = natural logarithm; MI = myocardial infarction; NRI = net reclassification improvement; RI = reclassification improvement; RRS = Reynolds Risk Score; SBP = systolic blood pressure; TC = total cholesterol.

* Risk strata use 3 categories: 0%–5%, 5%–20%, and ≥20%. Calibration statistics are based on observed 8-year rates from Kaplan–Meier curves. Net reclassification improvement statistics use 8-year case–control status, excluding those who are censored.

† Reclassification calibration statistic and *P* value for model omitting variable. Significance indicates lack of fit.

‡ Reclassification calibration statistic and *P* value for RRS model.