Published in final edited form as:

Stat Med. 2010 October 15; 29(23): 2438–2452. doi:10.1002/sim.3991.

Evaluating Health Risk Models

Alice S. Whittemore

Department of Health Research and Policy Stanford University School of Medicine, 259 Campus Drive, Stanford, CA 94305-5405 alicesw@stanford.edu Tel: 650-723-5460 Fax: 650-725-6951

SUMMARY

Interest in targeted disease prevention has stimulated development of models that assign risks to individuals, using their personal covariates. We need to evaluate these models and quantify the gains achieved by expanding a model to include additional covariates. This paper reviews several performance measures and shows how they are related. Examples are used to show that appropriate performance criteria for a risk model depend upon how the model is used. Application of the performance measures to risk models for hypothetical populations and for US women at risk of breast cancer illustrate two additional points. First, model performance is constrained by the distribution of risk-determining covariates in the population. This complicates the comparison of two models when applied to populations with different covariate distributions. Second, all summary performance measures obscure model features of relevance to its utility for the application at hand, such as performance in specific subgroups of the population. In particular, the precision gained by adding covariates to a model can be small overall, but large in certain subgroups. We propose new ways to identify these subgroups and to quantify how much they gain by measuring the additional covariates. Those with largest gains could be targeted for cost-efficient covariate assessment.

Keywords

absolute risk; Brier score; calibration; concordance; discrimination; personalized disease prevention; precision; resolution; risk model

INTRODUCTION

A risk model is a statistical procedure for assigning to an individual a probability of developing a future adverse outcome in a given time period. The assignment is made by combining his or her values for a set of risk-determining covariates with incidence and mortality data and published estimates of the covariates' effects on the outcome. Such risk models are playing increasingly important roles in the practise of medicine, as clinical care becomes more tailored to individual characteristics and needs. The recent literature contains useful discussion of risk models and their evaluation [1-5].

The utility of a risk model is assessed with respect to two attributes. Its *accuracy* (also called its calibration) reflects how well the model predictions agree with outcome prevalences within subgroups of the population. Its *precision* (also called its discrimination or resolution) reflects its ability to discriminate those with different true risks. An important issue is to quantify how much a model's attributes can be improved by expanding it with additional covariates. We describe several scores used to assess and compare models with respect to

the two attributes. The relevance of these scores depends strongly on how the models will be used [6], as shown by the following three applications.

The first application is to determine eligibility for a randomized clinical trial involving a treatment with serious side effects. For example, eligibility for the National Surgical Adjuvant Breast Prevention (NSABP) trial of tamoxifen to prevent breast cancer was determined by the Breast Cancer Risk Assessment Tool (BCRAT) [7]. Because tamoxifen increases the risk of endometrial cancer and deep-vein thrombosis, and because its value in preventing breast cancer was unknown, eligibility was restricted to women whose breast cancer risks were deemed high enough to warrant risking these side effects. Thus eligible women were those whose BCRAT-assigned five-year breast cancer risk exceeded 1.67% [8]. Here a good risk model is one that yields a low proportion of women whose assigned risk exceeds 1.67% but whose actual risk is lower. A model without this attribute could expose women needlessly to the side effects of tamoxifen.

The second application is to improve the cost-efficiency of a preventive intervention. For example, screening with breast magnetic resonance imaging (MRI) detects breast cancers earlier but costs more and produces more false positive scans, compared to mammography. Costs can be reduced by restricting MRI to women at high breast cancer risk [9]. The risk model most useful for this application is one with high sensitivity and specificity when “high risk” is defined as having an assigned risk exceeding some fixed cut-off value.

The third application is to facilitate personal health care decisions. Consider, for example, a postmenopausal woman with osteopenia who must choose between two drugs, raloxifene and alendronate, to prevent hip fracture. Because she has a family history of breast cancer, raloxifene would seem a good choice, since it also reduces breast cancer risk. However she also has a family history of stroke, and raloxifene is associated with increased stroke risk. To make a rational decision, she needs a risk model that provides accurate and precise information on her own risks of developing three adverse outcomes (breast cancer, stroke, hip fracture), and the effects of the two drugs on these risks.

The first two applications involve dichotomizing risks into “high” and “low” categories; thus they require risk models with low false positive and/or false negative probabilities with respect to the outcome. In contrast, the third application involves balancing risks for multiple adverse outcomes, and thus it requires risk models whose assigned risks are accurate enough *at the individual level* to facilitate rational health care decisions.

This paper reviews and synthesizes several performance measures with respect to a model's accuracy and precision. We begin by formalizing the notions of a population risk distribution and a risk model. We then describe several measures of model performance developed for forecasting problems in meteorology, economics and psychology [10-13], and we use hypothetical data to illustrate their use in medical applications. We note important limitations of the measures, such as their loss of useful information on model performance in specific population subgroups. This is followed by a discussion of how the measures are used to compare performance for different risk models, particularly when one model expands another with additional covariates. We propose a new method for evaluating performance improvement within subgroups of the population, which helps identify the individuals who profit most from the additional covariates. We illustrate the method by application to data on risk of breast cancer, and include a glossary in the Appendix.

POPULATION RISKS AND RISKS MODELS

Population Risks

Consider a population of women at risk of developing breast cancer within a ten year period. We assume that each woman has an unknown probability p of this outcome, which depends on her values for a set of covariates z through an unknown relation $p = \xi(z)$. The covariates z may include continuous and/or discrete components, and thus the risks p may assume continuous or discrete values. In practise, however, continuous covariates and risks are grouped into finitely many discrete categories, and here we shall represent all risks as discrete.

The distribution $\varphi(z)$ of covariates in the population determines the population risk distribution $f(p)$ via the relation

$$f(p) = \sum_{z: \xi(z)=p} \varphi(z). \quad (1)$$

The mean π of the risk distribution f specifies the prevalence of the outcome in the population. In symbols, $\pi = E[Y] = \Pr(Y = 1)$, where the random variable Y is an indicator assuming the value one if a woman develops breast cancer in ten years, and zero if she dies from other causes or survives the period without breast cancer. The variance σ^2 of f specifies the degree of risk heterogeneity in the population.

Plate 1 shows two risk distributions, both with mean 10%, that provide bounds on population risk heterogeneity for all populations with outcome prevalence $\pi = 10\%$. At one extreme, the “constant” distribution c (left panel) assigns all mass to the mean risk π , with $c(p) = 1$ if $p = \pi$, and $c(p) = 0$ otherwise. Under this distribution, all in the population have the same risk π , and the risk variance $\sigma^2 = 0$. The variance of this distribution gives a lower bound on that of any distribution with mean π .

At the other extreme, the deterministic distribution d (right panel) assigns mass $1 - \pi$ to $p = 0$ and mass π to $p = 1$, with $d(p) = 1 - \pi$ if $p = 0$, and $d(p) = \pi$ if $p = 1$. Under this distribution, the outcome occurs deterministically, with probability one in a fraction π of the population, and with probability zero in the remainder. The risk variance is $\sigma^2 = \pi(1 - \pi)$, which gives an upper bound on the variance of any risk distribution with mean π . When the mean risk is $\pi = 10\%$, for example, the maximum standard deviation of risks is 30%. In summary, the variance σ^2 of any risk distribution f of equation (1) with mean π is bounded below by the zero variance of the constant distribution, and bounded above by the variance $\pi(1 - \pi)$ of the deterministic distribution:

$$0 \leq \sigma^2 \leq \pi(1 - \pi). \quad (2)$$

Moreover the standard decomposition of a variance into the variance of a conditional expectation plus the expectation of a conditional variance gives the variance of outcomes as the sum

$$\begin{aligned} \text{var}[Y] &= \text{var}_p \left[E_{Y|p}(Y|p) \right] + E_p \left[\text{var}_{Y|p}(Y|p) \right] \\ &= \sigma^2 + \left[\pi(1 - \pi) - \sigma^2 \right]. \end{aligned} \quad (3)$$

The term $\pi(1 - \pi) - \sigma^2$ in equation (3) is the nonlinear analogue of the residual sum of squares in linear regression [14]. We shall see that this term produces an unknown and inflexible bound on the precision of any risk model.

Risk Models

We don't know the complete covariates z that determine a woman's breast cancer risk, much less their population distribution $\varphi(z)$ or the relation $p = \zeta(z)$ between covariates and risks. Instead, we use measured covariates $x = k(z)$, and epidemiological data to develop a statistical model that assigns to all women with covariates x a risk $\gamma(x) = r$, $0 < r < 1$. Thus a woman's unknown true risk p is approximated by her assigned risk r .

The unknown distribution $\varphi(z)$ of complete covariates determines the distribution of a model's assigned risks by the relation $g(r) = \sum \varphi(z)$, where the summation is taken over the set of covariates z such that $\gamma[k(z)] = r$. Similarly $\varphi(z)$ determines the joint distribution of true risk P and assigned risk R as $f_{PR}(p, r) = \sum \varphi(z)$, with summation taken over the set of covariates z such that $\zeta(z) = p$ and $\gamma[k(z)] = r$. The distribution of true risks among individuals assigned risk r is

$$f_{PR}(p|r) = \frac{f_{PR}(p, r)}{g(r)}. \quad (4)$$

The mean $\pi(r)$ of this distribution is the outcome prevalence in the subgroup of individuals with assigned risk r [15].

PERFORMANCE MEASURES FOR RISK MODELS

The performance of a risk model is best evaluated by applying it to a cohort of individuals independent of those used to develop the model. The outcome prevalences $\pi(r)$ are estimated by the observed cumulative incidence of the outcome among subjects in the various assigned risk groups r . To simplify the presentation and focus on the probabilistic properties of the performance measures, I assume throughout that the cohorts are so large that the sampling error of the estimates $\widehat{\pi}(r)$ can be ignored. Thus $\widehat{\pi}(r)$ is identified with $\pi(r)$.

What attributes do we want in a risk model? The answer depends on how the model will be used. If used to determine eligibility for clinical trials, a model should have high positive predictive power, to ensure that participants satisfy the eligibility criteria. If used to improve the cost-efficiency of a preventive intervention, high values for both positive and negative predictive power are needed. If used to facilitate personal health care decisions, the requirements are more stringent: we want the model to assign *each individual* a risk that agrees well with his or her true risk.

How do we measure accuracy when we don't know people's true risks? Since we can't assess accuracy at the individual level, we instead assess it at the group level, by measuring agreement between a woman's assigned risk r and the outcome prevalence $\pi(r)$ (i.e., the mean true risk) among all those who are assigned the same risk r as she. Measures of model *calibration* describe this agreement.

However even a perfectly calibrated model has limited utility for any application if it assigns a single risk to a group of individuals whose true risks vary substantially. Such a group might consist of two subgroups, one containing individuals at high risk and the other

individuals at low risk. Overall, their outcome prevalence might agree with their assigned risk, but important risk-determining covariates would not be reflected in the model. Thus the second desirable attribute of a model is its *precision* (also called *resolution* or *discrimination*), which reflects its ability to sort the population into subgroups with different true risks.

In considering a model's calibration and precision, it is useful to compare it to a theoretical, *perfect* model (Model P) that assigns each individual his or her true risk: $r = p$. The risks of Model P are those people would receive if we could measure their complete covariates z , and could correctly specify the relation $p = \xi(z)$ between covariates and risk. The distribution of assigned risks for Model P is just $g(r) = f(r)$.

Model Calibration

A risk model's calibration describes how well it agrees with outcome prevalences in subgroups of the population. The *calibration bias* in a group of individuals assigned a given risk r is the difference $r - \pi(r)$ between assigned risk and outcome prevalence in the subgroup. A risk model is said to be *well-calibrated* if its calibration bias is zero for all assigned risks r . Calibration biases are displayed in a *calibration plot* or *attribute diagram* [16], which is a plot of the points $(r, \pi(r))$ for given assigned risks r . For well-calibrated models, such as the perfect Model P, these points lie on the 45-degree line.

To illustrate calibration with a simple hypothetical example, consider the risk distributions in panels (a) and (b) of Plate 2 for two populations with different distributions of four risk-determining covariates $z = z_0, z_1, z_2, z_3$ (see Appendix I for details). Both risk distributions have mean $\pi = 10\%$, and both sets of risks range from 2% to 74%. However the variance of risks is larger for Population B ($\sigma^2 = 2.83\%$) than for Population A ($\sigma^2 = 0.30\%$). Now consider a model that assigns risks using only the first two of the covariates, z_0 and z_1 . Panels (c) and (d) show the distributions $g(r)$ of outcome prevalences $\pi(r)$ within the five risk groups of this model. The model can assign each of these five subgroups any risk between 0 and 1. However it is well-calibrated to a population only if it assigns each risk group a risk r that equals the outcome prevalence in that group. Thus a model that is well calibrated to Population A assigns the five risks shown on the abscissa of panel (c) of Plate 2. If these same risks were assigned to individuals in Population B, the model would show the calibration biases displayed in the attribute diagram in the upper panel of Plate 3. These biases occur because Populations A and B have different distributions of the two unmeasured risk-determining covariates z_2, z_3 , and thus they have different outcome prevalences within assigned risk groups.

A model's accuracy is often summarized by its overall calibration bias, whose square is the average of the squared biases $r - \pi(r)$ (vertical distances of points from diagonal line in Plate 3), weighted by the proportions $g(r)$ of individuals in the assigned risk groups:

$$Bias_g^2 = E_R \{ [r - \pi(r)]^2 \} = \sum_{0 \leq r \leq 1} g(r) [r - \pi(r)]^2. \quad (5)$$

The overall calibration bias for the data in the upper panel of Plate 3 is $Bias_g = 8.9\%$. This value reflects the discrepancy between assigned risk and outcome prevalence in the entire population. In practise, risks are assigned to a sample of individuals who are followed for outcome occurrence. In this circumstance the Hosmer-Lemeshow (HL) statistic [17], which is based on an estimate of $Bias_g^2$, can be used to test the null hypothesis that the model is well-calibrated. Approximate significance levels can be obtained by referring the HL statistic to a chi-square distribution on n degrees of freedom, where n is the number of risk

groups with at least 20 events. However a P-value for statistical significance (suggesting poor calibration) is less informative than an attribute diagram, such as those in Plate 3. Such a P-value depends strongly on the size of the sample, and it tells little about the overall magnitude of the bias and nothing about population subgroups for whom the model is particularly biased.

Model Precision

The overall precision (also called discrimination or resolution) of a model reflects its ability to discriminate among those with different true risks. Several closely related summary measures of precision have been proposed. Most are simple functions of the population outcome prevalence π and the model variance $var_R [\pi(r)]$. The latter is the variance of outcome prevalences across subgroups assigned different risks r , i.e., the variance of the conditional risk distribution $f_{P/R}(p/r)$ of equation (4). (Note that a model's precision does not depend on the variance of the actual assigned risks, which, if biased, can vary quite differently from the outcome prevalences). As shown in Appendix II, the model variance has a useful heuristic interpretation as the covariance between the outcome prevalences within assigned risk groups and the individual outcome indicators Y :

$$var_R [\pi(r)] = cov_{R,Y} [\pi(r), y]. \quad (6)$$

A model's *precision loss* is the difference

$$PL_g = \pi(1 - \pi) - var_R [\pi(r)] \quad (7)$$

between the maximum risk variance $\pi(1 - \pi)$ for a deterministic outcome and the model variance. The larger the model variance $var_R [\pi(r)]$, the smaller the precision loss.

Similar to a model's precision loss is its *prevalence-outcome (PO) correlation coefficient* ρ_g , which measures the extent of correlation between individual outcomes and the outcome prevalences within risk groups:

$$\rho_g = \frac{cov_{R,Y} [\pi(r), y]}{\sqrt{var_Y(y) var_R [\pi(r)]}} = \sqrt{\frac{var_R [\pi(r)]}{\pi(1 - \pi)}}. \quad (8)$$

The second equality in (8) follows from (6). The squared correlation ρ_g^2 is also called the Yates slope [18], and the fraction of outcome variation explained by the model, a generalization of R^2 from linear to binary outcomes [14,19]. Equations (7) and (8) show that the precision loss is proportional to the fraction of outcome variation *unexplained* by the model:

$$PL_g = (1 - \rho_g^2) \pi(1 - \pi). \quad (9)$$

The most widely used summary precision measure is the *concordance* (also called the C-statistic or area under the received operating characteristic (ROC) curve (AUC)). To describe it, we introduce the conditional distributions of assigned risk among those who do and do not develop the outcome:

$$h_y(r) = \Pr(R=r|Y=y) = \frac{g(r) [\pi(r)]^y [1 - \pi(r)]^{1-y}}{\pi^y (1 - \pi)^{1-y}}, \quad y=0, 1. \quad (10)$$

The concordance ζ is the probability that an assigned risk chosen from the distribution h_1 exceeds one chosen from $\zeta = \sum_{0 \leq r \leq 1} H_1(r) h_0(r)$ [18,20]. Here $H_1(r)$ is the cumulative distribution corresponding to $h_1(r)$. As indicated by its AUC acronym, a model's concordance is the area under its ROC curve. This curve is a plot of points $(x(r), y(r))$ as r varies over the model's assigned risks, increasing from 0 to 1. The abscissa $x(r)$ and ordinate $y(r)$ are, respectively, the proportions $H_0(r)$ and $H_1(r)$ of individuals without and with the outcome who are assigned a risk exceeding r . The concordance is invariant under any rank-preserving transformation of the assigned risks. Thus any set of assigned risks whose ranks agree with those of the outcome prevalences $\pi(r)$ has the same concordance as that of a well-calibrated risk model.

Choice of precision measure—One yardstick for evaluating a summary measure concerns how easily it can be interpreted by patients and clinicians. As a function of the conditional distributions $h_1(r)$ and $h_0(r)$ of (10), the concordance is most easily interpreted as a retrospective assessment of how well a model discriminates the risks of those with and without the outcome. The other measures also can be interpreted this way. To see this, note from equations (10) and (8) that the squared PO correlation coefficient can be written as the mean outcome prevalence among the assigned risk subgroups of those who develop the outcome, minus the corresponding mean in those who do not:

$\rho_g^2 = E_{RY} [\pi(r) | y=1] - E_{RY} [\pi(r) | y=0]$. This representation has motivated the name *Integrated Discrimination* for ρ_g^2 [21]. Equations (8) and (9) imply that as functions of ρ_g^2 , a model's variance and precision loss share this retrospective interpretation. Nevertheless the interpretation has limited utility for the applications considered earlier, particularly that of facilitating personal health care decisions [22-25]. To their advantage, the model variance, the precision loss and the PO correlation also have prospective interpretations as measures of the extent to which individuals with different true risks are grouped together and assigned a common risk.

Overall Performance Score

Accuracy and precision measures can be combined into a single score for model performance. For example, a model's *Brier score* [10], also called its mean probability score [12], is the mean squared error between individual outcomes and assigned risk:

$BS_g = E_{R,Y} [(y - r)^2]$. A decomposition of the Brier score, due to Murphy [11], shows that

$BS_g = Bias_g^2 + PL_g$, where $Bias_g^2$ is the model's squared calibration bias (5), and PL_g is its precision loss (7). However accuracy and precision reflect two different aspects of model performance, and some have argued that they should not be combined [26].

Limitations of performance measures

Calibration measures have limited utility for evaluating a risk model designed to facilitate personal health care decisions. The woman who needs to know her breast cancer risk requires a model that accurately assesses her own risk, rather than the mean risk in the subgroup of women with the same assigned risk as she. In the example shown in Plate 2, for instance, the subgroup of population A with assigned risk of 33% contains some individuals whose true risk is 26%, others whose true risk is 42% and still others whose true risk is 74%. These differences reflect unmeasured covariates omitted from the model. Moreover no

model can be well-calibrated to two populations with difference distributions of such unmeasured covariates.

Precision measures also have limitations. They are bounded by the unknown distribution of measured and unmeasured risk factors in the population under study. Indeed, as shown in Appendix II, a model's variance is bounded by the variance σ^2 of the perfect model, which by (2) is itself bounded by the variance of a deterministic outcome:

$$\text{var}_R [\pi(r)] \leq \sigma^2 \leq \pi(1 - \pi). \quad (11)$$

The other precision measures of the previous section are similarly bounded. They approach their optimal levels as the risk model captures increasingly many of the risk-determining covariates. Nevertheless they all are constrained by the distribution of these covariates in the population to which the model is applied (which the investigator cannot control). Consider, for example, the risk model based on the first two of the four risk-determining covariates z_0, z_1, z_2, z_3 (Model 1) as applied to populations A and B in Plate 2. Table I shows that the model variance is $\text{var}_R [\pi(r)] = 0.24\%$ in Population A, which by inequality (11) is bounded by the variance $\sigma^2 = 0.30\%$ of true risks. In contrast, the model variance in Population B is $\text{var}_R [\pi(r)] = 2.57\%$, with the more relaxed upper bound of 2.83%. Table I also shows the values of other precision measures for the model. For example, the PO correlation coefficients for Model 1 (16.3% for Population A and 53.4% for Population B) are not far from their optimal values 18.1% and 56.1% for these two populations, given by the perfect model. Similarly, the concordance of Model 1 is 59.4% for Population A and 80.6% for Population B. For comparison, concordances of the perfect model for these two populations are 60.3% and 83.3%, respectively.

This example shows that a model's poor precision may be caused more by population covariate homogeneity than by loss of covariate information, a fact that has been noted in the literature [22], but that is not widely appreciated when precision is reported in medical applications. Indeed, a risk model that uses the same covariates in the same way can have very different precision when applied to populations with different covariate distributions. Because of this strong dependence of precision measures on the population covariate distribution, comparing summary measures of model performance in different populations requires caution. For example, Table I shows that if we applied Model 1 to Population B and Model 2 to Population A, we would conclude that Model 1 is superior to Model 2 by any precision measure, despite the consistently better precision of Model 2 when both models are applied to the same population.

Finally, all summary performance measures are aggregates that obscure important information on subgroups of the population, which may limit their utility for some applications. Suppose, for example, that a model is needed to improve the cost-efficiency of a preventive intervention. Then the goal is to split the population into individuals at high and low risk, with the risk cutpoint separating the two groups determined by considerations of efficacy and cost. A risk model's usefulness for this goal depends on the coordinates $(x(r), y(r))$ of its ROC curve at the predetermined risk cutpoint r of interest, rather than the area under the entire curve, as given by its concordance. In the following section we provide additional examples showing how summary performance measures can obscure model features that are important for particular subgroups of the population, and we suggest solutions to the problem.

Comparing Risk Models

It is often desirable to evaluate two risk models applied to the same population. Of special interest is the case when one model expands the other with additional covariates [27-28]. An important issue is how to compare the two models, which may differ in both calibration accuracy and precision, with respect to their utility for a given clinical application.

Comparing accuracy—The overall calibration of two models can be compared by evaluating the difference or ratio of their squared biases, given by equation (5). Consider, for example, the two populations in Plate 2. Suppose we expand Model 1 by including one additional covariate z_2 (see Appendix A.1.). Panels (e) and (f) of Plate 2 show the distributions of outcome prevalences within the seven assigned risk groups of the expanded Model 2. The overall calibration bias incurred by Model 2, if calibrated to Population A but applied to Population B, is 5.5%, an overall improvement in accuracy compared to the value 8.9% for Model 1. In general, however, expanding a model with additional covariates can actually decrease its accuracy. In population B, for instance, Model 2 might assign biased risks to each of the two subgroups with outcome prevalences 38.8% and 67.6% (panel (f) in Plate 2), whereas Model 1 might assign an accurate risk of 61.8% to the combined group in panel (d).

Comparison of the models' attribute diagrams provides useful additional information about the extent of bias associated with specific assigned risks. The upper and lower panels of Plate 3 give attribute diagrams for the two models as calibrated to Population A and applied to Population B. It is evident that individuals with assigned risks exceeding the population mean of 10% enjoy the most bias reduction from Model 2 compared to Model 1.

A more informative comparison would classify individuals jointly according to the risks they receive by the two models, and determine the outcome prevalence in each of the resulting subgroups. For example, Table II shows outcome prevalences among 100,000 women from Population B who are classified according to the risk groups of two models that are poorly calibrated to this population because they were fit to the covariate distribution of Population A, as described in the Appendix A.1.. Plate 4 compares these outcome prevalences to the risks assigned by the two models which are shown in the row and column headings of the table. Both models exhibit substantial downward bias in the group with the highest outcome prevalence, with at most small bias in the five groups at lowest risk. Moreover compared to Model 1, the expanded Model 2 is less biased in groups 7 and 9, but more biased in groups 6 and 8. In practise, when risks are assigned to a sample of individuals who are followed for outcome occurrence, calibration can be tested using the Hosmer-Lemeshow (HL) statistic [17] as applied to the subgroups determined by Model CC. In this context the HL statistic has been called the *Reclassification Calibration Statistic*, and Table II has been called a *reclassification table* [29].

Table II also shows that some individuals assigned different risks by Model 1 can receive the same risk by Model 2. Thus adding covariates to a model need not yield a more refined partition of the population; in this sense, nested sets of covariates need not yield nested sets of risks.

Comparing precision—While expanding a model with additional covariates may actually decrease its accuracy, generally such expansion will increase its precision. The issue is whether the gain is large enough to warrant the cost of measuring the additional covariates.

Precision gains are usually measured as either the difference or ratio of model-specific precision scores, and most of these relative measures are closely related. For example,

equations (7) and (9) show that the difference in precision loss for two models is just the difference between the models' variances, and that both are proportional to the difference between the models' squared PO correlation coefficients:

$$PL_1 - PL_2 = \text{var}_{R_2} [\pi(r_2)] - \text{var}_{R_1} [\pi(r_1)] = \pi(1 - \pi) (\rho_{g_2}^2 - \rho_{g_1}^2). \quad (12)$$

Here the subscripts 1 and 2 refer to the two risk models. The difference $\rho_{g_2}^2 - \rho_{g_1}^2$ in squared PO correlations has been called the *Integrated Discrimination Improvement* (IDI) [21].

Equation (8) also shows that the ratio $\rho_{g_2}^2 / \rho_{g_1}^2$ of squared PO correlations for the two models is just their variance ratio $\text{var}_{R_2} [\pi(r_2)] / \text{var}_{R_1} [\pi(r_1)]$.

Recently Pencina et al [21] proposed a new measure of precision gain called the Net Reclassification Index (NRI). The NRI is the probability that Model 2 provides a correct reclassification of risks (i.e., increasing the risk of one who develops the outcome or decreasing it for one who does not), minus the probability of an incorrect reclassification (increasing the risk of one who avoids the outcome or decreasing the risk of one who develops it). specifically

$$NRI = \Pr(R_2 > R_1 | Y=1) + \Pr(R_2 < R_1 | Y=0) - [\Pr(R_2 < R_1 | Y=1) + \Pr(R_2 > R_1 | Y=0)]. \quad (13)$$

Table I shows NRIs comparing a well-calibrated Model 1 to a well-calibrated expanded Model 2 and to the perfect Model P, for the example of Plate 2. For Model 2, the NRI is -0.102 in Population A and 0.806 in Population B. The corresponding values for Model P are 0.034 in Population A and 0.602 in Population B. These NRIs illustrate two anomalies of this index. First, as seen for Model 2 in Population A, the NRI can be negative even when the expanded model captures additional population variation in outcome prevalences. Second, the NRI does not increase monotonically with the variance of outcome prevalences across model risk groups: in Population B, for example, the NRI for Model 2 exceeds that of Model P, despite their variances of 2.70×10^{-2} and 2.83×10^{-2} , respectively.

All of these comparative precision scores provide summary measures of the models' relative ability to sort individuals into groups having different true risks. However comparison of model precision (no matter how measured) requires caution, for two reasons. First, the precision gained by one model relative to another depends on the covariate distribution in the population, and thus can vary across populations. This variation can be seen in Table I, where relative to Model 1, Model 2 provides an AUC increase of 1.5% for Population B but only 0.5% for Population A. Similarly, the NRI for Model 2 is 0.806 for Population B but -0.102 for Population A.

Second, the overall precision gain from adding new covariates can be negligible, even though the expanded model may be considerably more precise in certain subgroups of the population [24]. This anomaly occurs because individuals with small risks, who typically comprise most of the population, tend to benefit little from increased precision, while the few with large risks have much to gain. Since the precision measures are averages over the entire population, the small gains for the majority tend to dominate the larger gains for a few. For example, Table I shows that the overall precision gain from expanding Model 1 with an additional covariate is small by any measure, and it tells us little about subgroups who may gain appreciable risk information by adding the new covariate.

This problem can be addressed by creating subgroup-specific measures of relative precision, where each subgroup consists of individuals assigned a common risk by a given model (say,

Model 1). Consider for instance the outcome prevalences in Table II, which cross-classifies individuals according to risks assigned by Models 1 and 2. We wish to determine the Model 1 risk groups who gain most from the additional covariate used by Model 2. The last columns of Table II show several measures of precision gain within each of the five risk groups of Model 1 (rows), based on variation in outcome prevalences across risk groups of Model 2 (columns). These include ranges, standard deviations (SDs), NRIs and slopes of outcome prevalences. The SD for the i^{th} Model 1 risk group is the square root of the variance of expected outcome prevalences π_{ij} across the columns, which differs from the sampling error variances of sample estimates $\hat{\pi}_{ij}$. The NRI for the i^{th} Model 1 risk group is obtained by conditioning the probabilities in (13) on membership in this risk group:

$$NRI_i = \Pr(R_2 > r_{1i} | Y=1, R_1=r_{1i}) + \Pr(R_2 < r_{1i} | Y=0, R_1=r_{1i}) - [\Pr(R_2 < r_{1i} | Y=1, R_1=r_{1i}) + \Pr(R_2 > r_{1i} | Y=0, R_1=r_{1i})]. \quad (14)$$

Finally, the slope for the i^{th} Model 1 risk group is defined as the rate of increase of cross-classified outcome prevalence per unit increase in Model 2 risk. (This measure differs from the slope suggested by Rosner et al [30], which is the rate of increase of log prevalence per decile of cross-classified risk.)

Table II shows that the standard deviation of outcome prevalences across the rows corresponding to the risks of Model 1 ranges from zero (in those assigned a risk of 10% by Model 1) to 11.5% (in those assigned the highest risk of 61.8%). Similarly, the NRI ranges from zero to 39%. According to these measures, the individuals at highest risk gain substantially more precision from the additional covariate than do others, and this information can be important for applications involving personal health decisions. If the additional covariates are difficult to measure, it might be cost-efficient to do so only for individuals in the risk groups of Model 1 with most to gain, as suggested in the management of cardiovascular disease [31,32]. In contrast to the standard deviations and NRIs, the subgroup-specific slopes in Table II are similar, despite considerable variation in their risk variances. This similarity reflects a problem with the slope as a measure of subgroup-specific precision gain: it fails to reflect the range of Model 2 risks across a row. Thus two rows can have the same slope but represent substantially different variation in actual risk.

The row-specific SDs and NRIs in a reclassification table are closely related to the corresponding summary measures of precision gain for Model 2 compared to Model 1. For example, the overall difference $var_{R_2}[\pi(r_2)] - var_{R_1}[\pi(r_1)]$ in model variances equals the difference between two sets of row-specific variances, each averaged over the relevant risk groups:

$$var_{R_2}[\pi(r_2)] - var_{R_1}[\pi(r_1)] = E_{R_1} \left\{ var_{R_1, R_2 | R_1}[\pi(r_1, r_2) | r_1] \right\} - E_{R_2} \left\{ var_{R_1, R_2 | R_2}[\pi(r_1, r_2) | r_2] \right\}. \quad (15)$$

Here for example, $var_{R_1, R_2 | R_1}[\pi(r_1, r_2) | r_1]$ is the variance of outcome prevalences across the assigned risk groups of Model 2, within the group of individuals assigned risk r_1 by Model 1. (see Appendix for a proof of equation (15)). Similarly, each of the four probabilities in the summary NRI of equation (13) is an average of the corresponding conditional probabilities in equation (14). For example,

$$\Pr(R_2 > R_1 | Y=1) = \sum_{r_1} \Pr(R_1=r_1 | Y=1) \Pr(R_2 > r_1 | Y=1, R_1=r_1).$$

APPLICATION TO DATA

We illustrate the precision measures of the previous sections by application to two models for risk of estrogen-receptor-positive (ER+) breast cancer among postmenopausal women. These models were developed by Rosner et al [30] using data from the prospective Nurses Health Study (NHS). Model 1 is based on a woman's current age and breast cancer risk factors, and Model 2 includes one additional covariate representing estimated serum levels of endogenous estradiol. The investigators identified 1559 ER+ breast cancer cases in 746,590 person-years of follow-up among postmenopausal women, giving a crude annual incidence rate of $1559/746,590 = 0.0021$ cases per woman per year. To illustrate the methods described here, we consider the ten-year breast cancer risks in a hypothetical cohort of postmenopausal women aged 50 years with the same covariate distribution and overall incidence as the NHS women. The annual death rate for US white women aged 50-59 years is 0.0053 deaths per woman per year [33]. This rate, combined with the annual breast cancer incidence rate of 0.0021, yields a mean risk $\pi = 2.02\%$ of developing ER+ breast cancer within ten years. Rosner et al [30] cross-classified case counts, person-years of follow-up, and incidence rates by joint deciles of the risks assigned by Models 1 and 2.

Here we used these data to compute the ten-year breast cancer risks shown in Table III. We assume that the calibration bias is zero, i.e., that the probabilities in the joint risk groups of Table III represent the actual ten-year breast cancer prevalences. To evaluate the relative precision of the two models, we use the summary column and summary row of Table III to

calculate standard deviations of $\left\{ \sum_{r_1=1}^{10} 0.1[\pi(r_1) - 0.0202]^2 \right\}^{1/2} = 1.06\%$ for Model 1 risks and 1.26% for Model 2 risks. These values correspond to PO correlation coefficients of 7.52% and 8.98%, respectively. The summary NRI comparing Model 2 to Model 1 is 44.07%, indicating that Model 2 more precisely classifies women, relative to their subsequent outcome occurrence.

Which risk groups of Model 1 benefit most from adding endogenous estradiol levels to Model 1? The last columns of Table III show the standard deviation, NRI and slope of the cross-classified outcome prevalences within each of the Model 1 risk groups. (To calculate these quantities, we assumed that within each decile of Model 1 risk, women were distributed in the same proportions as the person-years of follow-up given by Rosner et al [30].) The row-specific SDs tend to increase with outcome prevalence; however this trend is not seen in the row-specific NRIs or slopes. In summary, there are clear precision benefits associated with adding a serum estradiol measurement to the breast cancer risk assignments of Model 1 for postmenopausal women. These benefits are evident overall and in specific risk groups of Model 1.

DISCUSSION

We have reviewed several performance measures for risk models that use personal covariates to assign individual risks of a future adverse outcome. Application of the measures to risk models for hypothetical populations and for women at risk of breast cancer illustrates the following points. First, as noted by others [6,29], the characteristics needed in a risk model depend on how it will be used. Some applications use a model to split the population into subgroups at high and low risk of the outcome; these applications require risk models with low false positive and/or false negative probabilities. In contrast, other applications involve patients' need to balance risks for multiple adverse outcomes, and thus they require models whose assigned risks are accurate enough at the individual level to facilitate rational health care decisions. This type of application places high demands on a risk model.

Second, performance measures have limitations that warrant consideration when using them to evaluate and/or compare models. For example, measures of both accuracy and precision depend on the distribution of risk factors in the population of interest [1]. No model can be more precise than an ideal perfect model that assigns the true risks to all in the population. Because a model's precision bounds are unknown in practise, the precision gained by adding new covariates can be informative: a large precision gain indicates that the expanded model accounts for substantial variation in risk. However a small precision gain is less informative.

Summary performance measures can obscure model features of importance to subsets of the population. This limitation can be addressed by focusing on subgroup-specific performance measures, which can help identify those with the most to gain by enlarging a model to include covariates that distinguish their risks. These individuals could then be targeted for the additional covariate assessment [31,32]. We have reviewed methods for evaluating subgroup-specific gains in both calibration and precision, and we have proposed a new method for evaluating the precision gain within each of the risk groups of the reduced model, by quantifying the spread of outcome prevalences across the joint risk groups of both models.

Evaluating performance in population subgroups also helps assess a model's value for facilitating personal health decisions. An individual who needs to know his own risk does not care how a model performs for others in the population; yet summary performance measures involve the distribution of covariates in the entire population to which he belongs. Restricting the measures to the subgroup with risks similar to his own provides performance measures more relevant to his needs.

Finally, sampling error issues have largely been suppressed here, in order to simplify the presentation and focus on the probabilistic properties of the performance measures. However the bias & efficiency of estimated performance measures and optimal designs for moderately sized cohorts are areas in need of further research.

Acknowledgments

This research was supported by NIH grant CA094069. I am grateful to Joseph B. Keller for useful discussions, to Nicole Ng and Jerry Halpern for help with the calculations, and to the reviewers and an Associate Editor for comments that greatly improved earlier versions of the manuscript.

APPENDIX

A.1. Risk distributions of Plate 2

The risks shown in panels (a) and (b) of Plate 2 are determined by four covariates z_0, z_1, z_2, z_3 , with z_0 coded as $-1, 0, 1$, and z_j coded as $0, 1, j = 1, 2, 3$. These covariates determine risk according to the rule

$$\begin{aligned} p = \xi(z) &= 0.1 + 0.01 z_0 2^{z_1+z_2+z_3}, z_0 = -1, 0, \\ &= 0.1 + 0.08 z_0 2^{z_1+z_2+z_3}, z_0 = 1. \end{aligned}$$

The distributions $f(p)$ in the two panels are determined by the relation (1), where the covariate distribution $\varphi(z) = \tau(z_0) \prod_{i=1}^3 \alpha^{z_i} (1 - \alpha)^{1-z_i}$. Here $\tau(z_0) = 0.8, 0.1, 0.1$, for $z_0 = -1, 0, 1$, respectively. The value $\alpha = 0.2$ gives the distribution for Population A and $\alpha = 0.8$ gives the one for Population B. The mean risk is 10% and the variance is $7.2 \times 10^{-4} (1 + 3\alpha)^3$.

Panels (c) and (d) of Plate 2 show outcome prevalences within population subgroups assigned common risks based on two of the four covariates: z_0, z_1 . These prevalences are

$$\pi [r_1 (z_0, z_1)] = \begin{cases} 0.1 - .01(1+\alpha)^2 \times 2^{z_1}, & z_0 = -1 \\ 0.1, & z_0 = 0 \\ 0.1 + .08(1+\alpha)^2 \times 2^{z_1}, & z_0 = 1 \end{cases} \quad \text{if } \begin{matrix} z_0 = -1 \\ z_0 = 0 \\ z_0 = 1 \end{matrix}.$$

Panels (e) and (f) of Plate 2 give outcome prevalences in the seven subgroups determined by expanding a risk model based on z_0, z_1 (Model 1) to include z_2 . These prevalences are

$$\pi [r_1 (z_0, z_1, z_2)] = \begin{cases} 0.1 - .01 (1+\alpha) \times 2^{z_1+z_2}, & z_0 = -1 \\ 0.1, & z_0 = 0 \\ 0.1 + .08 (1+\alpha) \times 2^{z_1+z_2}, & z_0 = 1 \end{cases} \quad \text{if } \begin{matrix} z_0 = -1 \\ z_0 = 0 \\ z_0 = 1 \end{matrix}.$$

A.2. Model variance

To see that the variance of any risk model is bounded above by the variance σ^2 of the perfect model, we use the standard decomposition of a variance into the expectation of a conditional variance plus the variance of a conditional expectation:

$$\sigma^2 = E_R [\sigma^2(r)] + \text{var}_R [\pi(r)] \geq \text{var}_R [\pi(r)],$$

which implies equation (11).

We next show that the variance of a risk model equals the covariance of its grouped outcome prevalences with the individual outcomes. Since $E_R [\pi(r)] = \pi = E_Y [y]$, we have

$$\begin{aligned} \text{cov}_{R,Y} [y, \pi(r)] &= E_{R,Y} [y\pi(r)] - \pi^2 \\ &= E_R \{E_{Y|R} [y\pi(r)|r]\} - \pi^2 \\ &= E_R \{[\pi(r)]^2\} - \pi^2 = \text{var}_R [\pi(r)]. \end{aligned} \quad (\text{A1})$$

A.3. Proof of equation (15)

We write the left side of (15) as

$$\text{var}_{R_2} [\pi(r_2)] - \text{var}_{R_1} [\pi(r_1)] = \text{var}_{R_1, R_2} [\pi(r_1, r_2)] - \text{var}_{R_1} [\pi(r_1)] - \{\text{var}_{R_1, R_2} [\pi(r_1, r_2)] - \text{var}_{R_2} [\pi(r_2)]\}, \quad (\text{A2})$$

where $\text{var}_{R_1, R_2} [\pi(r_1, r_2)]$ is the variance of outcome prevalences in the cross-classified model. We write this variance as

$$\begin{aligned} \text{var}_{R_1, R_2} \{\pi(r_1, r_2)\} &= E_{R_1} [\text{var}_{R_1, R_2|R_1} \{\pi(r_1, r_2)|r_1\}] + \text{var}_{R_1} \{E [\pi(r_1, r_2)|r_1]\} \\ &= E_{R_1} [\text{var}_{R_1, R_2|R_1} \{\pi(r_1, r_2)|r_1\}] + \text{var}_{R_1} \{\pi(r_1)\}, \end{aligned} \quad (\text{A3})$$

with a similar decomposition with respect to the risk groups of Model 2. Substituting (A3) into the right side of equation (A2) for each of the two models gives equation (15).

IV. Glossary

accuracy (also called calibration)- extent of agreement between assigned risks and outcome prevalences among

those assigned the same risk.

area under ROC curve (AUC) - see concordance

attribute diagram - plot of outcome prevalences versus assigned risks. If a model is perfectly calibrated, these

points lie on the diagonal line $y=x$.

bias - square root of the mean-squared difference between assigned and true risks.

Brier score - mean squared difference between outcomes and assigned risks.

calibration - see accuracy

concordance (also called AUC or C-statistic) - probability that the assigned risk of a randomly selected

individual who develops the outcome exceeds that of a randomly selected individual who does not.

cross-classified (CC) model - obtained by cross-classifying individuals into cells according to the risks assigned

by each of two models, and assigning each member of a cell the outcome prevalence among those in that cell.

deterministic outcome –one that occurs with probability one in some members of a population and

probability zero in the remaining members.

Hosmer-Lemeshow (HL) statistic - standardized sum of squared differences between assigned risks and

outcome prevalences. Under the null hypothesis that a model is perfectly calibrated, the statistic has a

chi-squared distribution.

integrated discrimination - the square of the PO correlation coefficient, which equals the fraction R^2 of

outcome variance explained by the model [14] and the Yates slope [12].

integrated discrimination improvement (IDI) - difference in squared PO correlation coefficients between two

models when one is obtained from the other by including additional covariates.

model variance - variance of outcome prevalences across risk groups of the model.

net reclassification index (NRI) – probability that an expanded model correctly reclassifies a person's risk

(relative to outcome occurrence) minus the corresponding probability of an incorrect classification.

outcome prevalence – expected proportion of individuals in a group who develop the outcome, equal to the

mean risk of individuals in the group.

perfect model – assigns each individual the risk that would result if we knew all risk-determining covariates

and their joint effects on outcome probability.

precision (also called discrimination, resolution) – extent to which a model assigns different risks to individuals

with substantially different true risks.

precision loss – difference between Bernoulli variance of outcomes and the model variance.

prevalence-outcome (PO) correlation coefficient - correlation between actual outcomes and outcome

prevalences in assigned risk groups of a model.

reclassification calibration statistic – Hosmer-Lemeshow test statistic applied to the risk groups determined by

cross-classifying individuals according to risks assigned by two models.

receiver operating characteristic (ROC) curve - plot of points $(x(r), y(r))$ as r varies from 0 to 1, where $x(r)$ is

the probability that the model assigns risks to individuals who do not develop the outcome, and $y(r)$ is the

corresponding probability for those who develop the outcome.

REFERENCES

1. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clinical Chemistry*. 2008; 54:17–23. [PubMed: 18024533]
2. Pepe MS, Feng Z, Anes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *Journal of National Cancer Institute*. 2008; 100:1432–1438.
3. 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]
4. Pepe MS, Gu JW, Morris DE. The potential of genes and other markers to inform about risk. *Cancer epidemiology, biomarkers & prevention*. 2010; 19:655–665.
5. 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]

6. Cook NR. Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond'. *Statistics in Medicine*. 2008; 27:191–195. [PubMed: 17671959]
7. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mutvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of National Cancer Institute*. 1989; 81:1879–1886.
8. Fisher B Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, Vogel V, Robidoux A, Dimitrov N, Atkins J, Daly M, Wieand S, Tan-Chiu E, Ford L, Wolmark N. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *Journal of National Cancer Institute*. 1998; 90:1371–1388.
9. Plevritis SK, Kurian AW, Sigal BM, Daniel BL, Ikeda DM, Stockdale FE, Garber AM. Cost-effectiveness of screening BRCA1/2 mutation carriers with breast magnetic resonance imaging. *Journal of the American Medical Association*. 2006; 295:2374–2384. [PubMed: 16720823]
10. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. 1950; 78:1–3.
11. Murphy AH. A new vector partition of the probability score. *Journal of Applied Meteorology*. 1973; 12:595–600.
12. Yates JF. External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*. 1982; 30:132–156.
13. Wilks, DS. *Statistical methods in the Atmospheric Sciences*. Academic Press; London: 1995.
14. Hu B, Palta M, Shao J. Properties of R(2) statistics for logistic regression. *Statistics in Medicine*. 2006; 25:1383–1395. [PubMed: 16059870]
15. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics*. 2005; 6:227–239. [PubMed: 15772102]
16. Hsu WR, Murphy AH. The attributes diagram A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*. 1986; 2:285–293.
17. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics*. 1980; 9:1043–1069.
18. Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; 2003.
19. 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'. *Statistics in Medicine*. 2008; 27:173–181. [PubMed: 17671958]
20. 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]
21. 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. *Statistics in Medicine*. 2008; 27:157–172. [PubMed: 17569110]
22. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007; 115:928–935. [PubMed: 17309939]
23. Ridker PM, Buring JE, Rifar N, Cook N. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women. *Journal of the American Medical Association*. 2007; 297:611–619. [PubMed: 17299196]
24. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *Journal of National Cancer Institute*. 2008; 100:978–979.
25. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*. 1999; 18:2529–2545. [PubMed: 10474158]
26. Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. *Journal of clinical epidemiology*. 1992; 45:85–89. [PubMed: 1738016]
27. Chen J, Pee D, Ayyagari R, Graubard B, Schairer C, Byrne C, Benichou J, Gail MH. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *Journal of National Cancer Institute*. 2006; 98:1215–1226.

28. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *Journal of National Cancer Institute*. 2008; 100:1037–1041.
29. 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]
30. Rosner B, Colditz GA, Iglehart JD, Hankinson SE. Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the Nurses' Health Study. *Breast Cancer Res*. 2008; 10:R55. [PubMed: 18598349]
31. Greenland P, Lloyd-Jones D. Defining a rational approach to screening for cardiovascular risk in asymptomatic patients. *Journal of the American College of Cardiology*. 2008; 52:330–332. [PubMed: 18652939]
32. Greenland P, Bonow RO, Brundage BH, Budoff MJ, Eisenberg MJ, Grundy SM, Lauer MS, Post WS, Raggi P, Redberg RF, Rodgers GP, Shaw LJ, Taylor AJ, Weintraub WS, American College of Cardiology Foundation Clinical Expert Consensus Task Force (ACCF/AHA Writing Committee to Update the 2000 Expert Consensus Document on Electron Beam Computed Tomography); Society of Atherosclerosis Imaging and Prevention; Society of Cardiovascular Computed Tomography. ACCF/AHA 2007 clinical expert consensus document on coronary artery calcium scoring by computed tomography in global cardiovascular risk assessment and in evaluation of patients with chest pain: a report of the American College of Cardiology Foundation Clinical Expert Consensus Task Force (ACCF/AHA Writing Committee to Update the 2000 Expert Consensus Document on Electron Beam Computed Tomography) developed in collaboration with the Society of Atherosclerosis Imaging and Prevention and the Society of Cardiovascular Computed Tomography. *Journal of the American College of Cardiology* 2007. 49:378–402.
33. Arias E. United States Life Tables 2004. *National Vital Statistics Reports*. 2007; 56:1–39. [PubMed: 18274319]

Plate 1.

Two extreme distributions for a population with outcome prevalence 10%. a) all individuals have common risk of 10 per cent, so that the risk variance is $\sigma^2 = 0$ b) 90 per cent of individuals have risk zero and the remaining have risk one. The variance of risks is $\sigma^2 = 0.10 \times (1 - .10) = 0.09$

Plate 2.

Risk distributions for two hypothetical populations, A and B. Risks depend on four covariates. The first row shows the distributions of true risk, with mean = 10 per cent. The standard deviation of risks is 5.4 per cent for population A and 16.7 per cent for Population B. The second row shows the Distributions of outcome prevalences in subgroups determined by a risk model based on two of the four covariates (Model 1). The third row shows outcome prevalences in subgroups obtained by expanding Model 1 with an additional covariate (Model 2). (See Appendix A.1. for details).

Plate 3.

Panel (a): attribute diagram for a model that uses two of four covariates to partition Population B into five risk groups, and then assigns each group the outcome prevalence observed in Population A. Panel (b): attribute diagram for an expanded model that includes an additional covariate, and again is calibrated to Population A and applied to Population B.

Plate 4.

Outcome prevalences and assigned risks in cross-classified risk groups of Models 1 and 2 when both models are calibrated to Population A and applied to Population B.

Table 1

Precision Measures^a for Risk Models and Populations A and B of Plate 2

	Model Variance ^b		Precision Loss		PO Correlation		Concordance (AUC) ^e		NRI ^f	
	Pop A / Pop B	Pop B	Pop A / Pop B	Pop B	Pop A / Pop B	Pop B	Pop A / Pop B	Pop B	Pop A / Pop B	Pop B
Model 1 ^c	0.24 / 2.57		8.76 / 6.43		16.3 / 53.4		59.4 / 80.6		--- / ---	
Model 2 ^c	0.27 / 2.70		8.73 / 6.30		17.2 / 54.7		59.9 / 82.1		-10.2 / 80.6	
Model p ^d	0.30 / 2.83		8.70 / 6.17		18.1 / 56.1		60.3 / 83.3		3.4 / 60.2	

^aall values are multiplied by 100

^bvariance of outcome prevalences across assigned risk groups

^cbased on two (Model 1) or three (Model 2) covariates

^dperfect model, based on all covariates

^earea under the receiver operating characteristic curve

^fNet Reclassification Index comparing each model to Model 1

Outcome Prevalence^a among 100,000 Women from Population B, Cross-classified according to Risks Groups of Models 1 & 2^b

Table II

Model 1 Group / Risk (%) ^c		Model 2 Group / Risk (%) ^c					All	Range	SD ^d	NRI ^e	Slope ^f
1/7.1	Women, <i>n</i>	1/5.2	2/7.6	3/8.8	4/10.0	5/19.6	6/29.2	7/48.4			
	Events, <i>n</i>	51,200	12,800	0	0	0	0	0	64,000		
	Outcome Prevalence	1,434	819	0	0	0	0	0	2,253		
2/8.6	Women, <i>n</i>	2.8	6.4	-	-	-	-	-	3.5	2.8-6.4	1.4
	Events, <i>n</i>	0	12,800	3,200	0	0	0	0	16,000		
	Outcome Prevalence	0	819	260	0	0	0	0	1,079		
3/10.0	Women, <i>n</i>	-	6.4	8.2	-	-	-	-	6.8	6.4-8.2	0.7
	Events, <i>n</i>	0	0	0	10,000	0	0	0	10,000		
	Outcome Prevalence	0	0	0	1,000	0	0	0	1,000		
4/22.8	Women, <i>n</i>	-	-	-	10.0	-	-	-	10.0	10.0-10.0	0
	Events, <i>n</i>	0	0	0	0	400	1,600	0	2,000		
	Outcome Prevalence	0	0	0	0	98	621	0	719		
5/33.0	Women, <i>n</i>	-	-	-	-	24.4	38.8	-	35.9	21.4-38.8	5.8
	Events, <i>n</i>	0	0	0	0	0	1,600	6,400	8,000		
	Outcome Prevalence	0	0	0	0	0	621	4,326	4,947		
All	Women, <i>n</i>	51,200	25,600	3,200	10,000	400	3,200	6,400	100,000		
	Events, <i>n</i>	1,434	1,638	260	1,000	98	1,242	4,326	10,000		
	Outcome Prevalence	2.8	6.4	8.2	10.0	24.4	38.8	67.6	10.0	2.8-67.6	-

^aall outcome prevalences are multiplied by 100

^bSee Appendix I for description of populations and models

^cBiased risk obtained by calibrating model to covariate distribution of Population A (see text)

^dstandard deviation, multiplied by 100

^eNet Reclassification Index, multiplied by 100

^fincrease of cross-classified outcome prevalence per unit increase in Model 2 risk

Table III

Estimated Ten-year Prevalence^a of ER+ Breast Cancer among Post-menopausal Women^b, by Risk Decile of Model 1 (without Estradiol) and Model 2 (with Estradiol)

Model 1 Risk Decile		Model 2 Risk Decile										SD ^c	NRI ^d	Slope
		1	2	3	4	5	6	7	8	9	10	All		
1	Women (%)	8.289	1.745	-	-	-	-	-	-	-	-	10.034		
	Outcome Prev	0.663	0.998	-	-	-	-	-	-	-	-	0.721	0.127	13.46 0.00
2	Women (%)	1.650	5.655	2.365	0.365	-	-	-	-	-	-	10.035		
	Outcome Prev	0.731	0.805	2.039	2.933	-	-	-	-	-	-	1.161	0.631	47.29 0.12
3	Women (%)	-	2.470	4.582	2.456	0.526	-	-	-	-	-	10.035		
	Outcome Prev	-	0.488	1.345	1.690	2.292	-	-	-	-	-	1.268	0.501	30.68 0.16
4	Women (%)	-	0.346	2.490	4.253	2.442	0.505	-	-	-	-	10.036		
	Outcome Prev	-	0.387	0.968	1.323	2.358	3.714	-	-	-	-	1.575	0.739	31.91 0.26
5	Women (%)	-	-	0.483	2.467	4.225	2.447	0.424	-	-	-	10.047		
	Outcome Prev	-	-	1.385	0.706	1.490	2.737	6.313	-	-	-	1.800	1.190	35.82 1.09
6	Women (%)	-	-	-	0.518	2.456	4.391	2.345	0.326	-	-	10.035		
	Outcome Prev	-	-	-	1.034	1.145	1.525	3.085	5.343	-	-	1.895	0.976	52.16 2.06
7	Women (%)	-	-	-	-	0.420	2.378	4.723	2.334	0.181	-	10.036		
	Outcome Prev	-	-	-	-	0.638	1.239	2.155	2.639	6.662	-	2.069	0.841	8.14 1.67
8	Women (%)	-	-	-	-	-	0.321	2.357	4.797	2.197	-	9.673		
	Outcome Prev	-	-	-	-	-	1.668	1.761	2.429	4.877	-	2.797	1.163	34.81 1.04
9	Women (%)	-	-	-	-	-	-	0.185	2.161	6.203	1.486	10.035		
	Outcome Prev	-	-	-	-	-	-	2.903	1.549	3.023	5.860	3.123	1.284	22.68 0.45
10	Women (%)	-	-	-	-	-	-	-	-	1.531	8.503	10.034		
	Outcome Prev	-	-	-	-	-	-	-	-	1.924	4.962	4.498	1.092	18.29 0.09
All	Women (%)	9.939	10.217	9.920	10.060	10.069	10.042	10.033	9.619	10.113	9.989	100.000		
	Outcome Prev	0.674	0.747	1.418	1.305	1.623	1.867	2.470	2.381	3.325	5.095	20.908		44.07

^a all prevalences are multiplied by 100

^b adapted from Table 4 of Rosner et al (2008)

^c standard deviation, multiplied by 100