Published in final edited form as:

Ann Intern Med. 2008 November 18; 149(10): 751–760.

Assessing the Value of Risk Predictions Using Risk Stratification Tables

Holly Janes¹, Margaret S. Pepe^{1,2}, and Wen Gu¹

¹Division of Public Health Sciences, Fred Hutchinson Cancer Research Center

²Department of Biostatistics, University of Washington

Abstract

In many clinical settings, statistical models are being developed for predicting risk of disease or other adverse event. These models are intended to help patients and physicians make informed decisions. A new approach to assessing the value of adding a new marker to a risk prediction model, called the risk stratification approach, was recently proposed by Cook and colleagues (1,2). This involves cross-tabulating risk predictions on the basis of models with and without the new marker, and has been widely adopted in the literature. We argue that important information with regard to three important model validation criteria can be extracted from risk stratification tables: 1) model fit or calibration; 2) capacity for risk stratification; and 3) accuracy of classifications based on risk. However, we describe how the information contained in the tables must be interpreted carefully, and caution against common misuses of the method. The concepts are illustrated using data from a recently published study of a breast cancer risk prediction model by Tice et al. (3).

Introduction

The recent epidemiologic and clinical literature is filled with studies evaluating statistical models for predicting risk of disease or some other adverse event (e.g. (1,3-6)). These models are intended to help patients and clinicians make decisions. Therefore their evaluation is different from models describing disease etiology. It's not the characteristics of the models themselves that are of interest, but the value of the models for guiding decisions.

Cook and colleagues (1,2) have recently proposed a new approach to evaluating risk prediction models using a risk stratification table. This methodology appropriately focuses on the key purpose of a risk prediction model, which is to classify individuals into clinically relevant risk categories, and has therefore been widely adopted in the literature (see, e.g., (3-5)). In this paper, we examine the risk stratification approach in detail, identifying the relevant information that can be abstracted from a risk stratification table, and cautioning against misuses of the method that frequently occur in practice. We use a recently published study of a breast cancer risk prediction model by Tice et al. (3) to illustrate the concepts.

Background

A risk prediction marker is any measure that is used to predict a person's risk of an event. It may be a quantitative measure such as HDL cholesterol, or a qualitative measure such as family history of disease. Risk predictors are also risk factors, in the sense that they will necessarily be strongly associated with the risk of disease. But a large statistically significant association does not assure that the marker has value in terms of risk prediction for many people.

A risk prediction model is a statistical model that combines information from several markers. Common types of models include logistic regression models, Cox proportional hazards models, and classification trees. Each type of model produces, for each individual, a predicted risk using information in the model. Consider, for example, a model predicting breast cancer risk that includes age as the only predictor. The resulting risk prediction for a woman of a given age is simply the proportion of women her age who develop breast cancer. The woman's predicted risk will change if more information is included in the model. For instance, if family history information is added, her predicted risk will be the proportion of women her age and with her family history who develop breast cancer.

The purpose of a risk prediction model is to accurately stratify individuals into clinically relevant risk categories. This risk information can be used to guide clinical or policy decisions, for example about preventive interventions for individuals, or disease screening for subpopulations identified as high risk, or to select individuals for inclusion in clinical trials. The value of a risk prediction model for guiding these kinds of decisions can be judged by the extent to which the risk calculated from the model reflects the actual fraction of people in the population with events (its calibration); the proportions in which the population is stratified into clinically relevant risk categories (its stratification capacity); and the extent to which subjects with events are assigned to high risk categories and subjects without events are assigned to low risk categories (its classification accuracy).

Risk prediction models are commonly evaluated using receiver operating characteristic (ROC) curves (e.g. (5,7)), which are standard tools for evaluating the discriminatory accuracy of diagnostic or screening markers. The ROC curve shows the true-positive rate versus the false-positive rate for rules that classify individuals using risk thresholds that vary over all possible values. ROC curves are generally not helpful for evaluating risk prediction models because they do not provide information about the actual risks the model predicts, or about the proportions of subjects who have high or low risk values. Moreover, when comparing ROC curves for two risk prediction models, the models are aligned according to their false-positive rates, where different risk thresholds are applied to the two models in order to achieve the same false-positive rate. This is clearly inappropriate. In addition, the area under the ROC curve (AUC or C-statistic), a commonly reported summary measure which can be interpreted as the probability that the predicted risk for a subject with an event is higher than that for a subject without an event, has little direct clinical relevance because clinicians are never asked to compare risks for a pair of subjects, one who will go on to have the event and one who will not. Neither the ROC curve nor the AUC relates to the practical task of predicting risks for clinical decision making.

Cook and colleagues propose using risk stratification tables to evaluate the incremental value of a new marker, or the benefit of adding a new marker (eg C-reactive protein) to an established set of risk predictors (eg Framingham risk predictors such as age, diabetes, cholesterol, smoking, and LDL levels) (1,2). In these stratification tables, risks calculated from models with and without the new marker are cross-tabulated. This approach represents a substantial improvement over the use of ROC methodology because it displays the risks calculated by use of the model and the proportions of individuals in the population who are stratified into the risk groups. We will provide an example of this approach, and show how information about model calibration, capacity for risk stratification, and classification accuracy can be derived from a risk stratification table and used to assess the added value of a marker for clinical and healthcare policy decisions.

Example

Tice and colleagues (3) published a study which builds and evaluates a model for predicting breast cancer risk using data from 1,095,484 women in a prospective cohort and using incidence data from the Surveillance, Epidemiology, and End Results (SEER) database. Age, race/ethnicity, family history, and history of breast biopsy were used to model risk using a Cox proportional hazard model. The study focused on the benefit of adding breast density information to the model. The hazard ratio for breast density in the multivariate model (extremely dense vs. almost entirely fat) was estimated as 4.2 for women younger than 65 years and 2.2 for women over 65 years. This suggests that breast density is strongly associated with disease risk. More specifically, breast cancer rates are higher among women with higher breast density. However it does not describe the value of breast density for helping women make informed clinical decisions. This involves the frequency distribution of breast density in the population as well.

In order to evaluate the added value of breast density, 5-year breast cancer risk categories were defined as: low risk (< 1%), low or intermediate risk (1% - 1.66%), intermediate or high risk (1.67% - 2.49%), and high risk (> 2.5%). The 1.67% cutoff for intermediate risk was presumably chosen based on recommendations by the American Society of Clinical Oncology (8) and Canadian Task Force on Preventive Health Care (9) to counsel women with 5-year risks above this threshold about considering tamoxifen for breast cancer prevention. A risk stratification table, reproduced below as Table 1, was used to compare risk prediction models with and without breast density.

Calibration

Assessing model calibration is an important first step in evaluating any risk prediction model. Good calibration is essential; it means that the model-predicted probability of an event for a person with specified predictor values is the same as or very close to the proportion of all people in the population with those same predictor values that experience the event (10 [pp. 60-3]). With many predictors and especially with continuous predictors, we cannot evaluate calibration at each possible value of the predictors since there are too few subjects with exactly those values. Instead the standard approach is to place individuals within categories of predicted risk, and to compare the category values with the observed event rates for subjects in each category.

The calibration of the breast cancer risk prediction models can be assessed by comparing the proportions of events in the margins of Table 1 with the corresponding row and column labels. For the model without breast density the proportions of observed events within each risk category appear in the far right "Total" column, and they generally agree with the row labels. That is, the proportion of observed events within the risk category of 0% to 1% is 0.7%, which falls within the risk category range; the same is true for the risk category of 1% to 1.66% (proportion of events 1.3%), 1.67% to 2.5% (1.8%), and > 2.5% (2.9%). Thus, the model without breast density appears to be well-calibrated. Similarly when subjects are categorized according to their risk calculated using the model that includes breast density, in each category the proportion of events falls within the range of risk in the column label (for the risk category 0% to 1% the proportion of events is 0.7%; for 1% to 1.66% the proportion is 1.3%, and so on). Therefore the model with breast density is also well-calibrated.

The Hosmer-Lemeshow test (11) is a common test of model calibration based on this notion. The test sums the squared differences between observed event rates in the row (or column) margins and average predicted risks for people in each category, typically using 10 categories. Low p-values (eg $p < 0.05$) indicate that observed and predicted risks are significantly different, implying poor calibration. We find the display of event rates in the

margins of Table 1 a useful adjunct to the p-value from the Hosmer-Lemeshow test because the calibration assessment is made more concrete.

Event rates in the inner cells of the risk stratification table do not provide useful information about calibration, despite suggestions otherwise (12,13). Consider for example the cells in the second row of Table 1 labeled “1% to 1.66%,” where the proportion of events range from 0.8% in the “0 to 1%” column to 3.1% in the “> 2.5%” column. All but one of these event rates fall outside the range of risk in the row labels. Remember that subjects are selected into each cell on the basis of breast density as well as other baseline factors. For example, the 1025 women in the last cell of the row (in the > 2.5% column) presumably have high breast density values compared with the row as a whole (N = 201,927). But the model without breast density is not geared towards capturing the higher risk (3.1%) in this subgroup since breast density is not included in the model. The model without breast density is still well-calibrated, however, because there is good agreement between the row label (1% to 1.66%) and the event rate in the margin (1.3%).

The event rates in the inner cells of the risk stratification table are nevertheless informative in one respect. Event rates in the second row of Table 1 increase from left to right, suggesting that risk increases with breast density. However the standard and more straightforward approach to expressing this gradient of risk is with use of the coefficient of breast density in the risk prediction model that includes breast density and baseline predictors. This coefficient can be transformed into a commonly used measure of association such as an odds ratio or hazard ratio. We noted previously that Tice et al. (3) provided hazard ratios for breast density that adjusted for baseline risk factors, estimating that among women younger than 65 the rate of breast cancer is 4.2 times higher for women with breasts that are extremely dense versus almost entirely fat, while for women over 65 the rate is 2.2 times higher.

Risk Stratification

Having established that a model can reliably be used to calculate the chance of an event (i.e. is well-calibrated), it is appropriate to consider the model’s value in terms of its capacity to stratify the population into clinically relevant risk categories. After all, it is possible to have a perfectly calibrated model but if the markers are weak it will predict risks close to the average for all subjects, and therefore be useless for clinical decision making.

A model’s capacity for risk stratification can be described by the proportions in which the population is allocated into clinically relevant risk categories. A better model places more subjects at the extremes of the risk distribution where there are clear implications for future actions. A perfect model would assign the entire population to the very highest or very lowest risk categories, and leave no one in the middle categories where there is still uncertainty about the appropriate course of action. A useless model resulting from uninformative markers would assign the same risk to the entire population, that value being the overall event rate or prevalence.

The capacity for risk stratification of the model without breast density can be calculated from numbers in the final column of Table 1. We see that it puts 34% of women (215,402 of 629,299) at lowest risk (0 to 1% row); 10% of women (63,105 of 629,299) at highest risk (> 2.5% row); and leaves 56% in the two middle categories. The corresponding values for the model with breast density can be calculated from the Table’s bottom row. This model puts 40% of women (249,959 of 629,229) at lowest risk (0 to 1% column); 11% of women (68,744 of 629,229) at highest risk (> 2.5% column); and leaves 49% in the middle categories. Therefore adding breast density to the model has a small benefit in terms of moving an additional 7% of people to the most determinant, highest and lowest risk

categories. The value of this movement, and therefore the added value of adding breast density, depends on the cost of ascertaining breast density, which is presumably small because it is routinely assessed with standard mammography (3), and the benefits of moving individuals to the highest and lowest risk categories.

Once again the risk stratification table must be interpreted carefully. It may be tempting to focus on movements of individuals between risk categories, but the cells in the table cannot be interpreted in isolation (14). For example in Table 1, the fact that adding breast density to the model moves 23,267 subjects from the 1.67% to 2.5% risk category to the highest risk group must be paired with the fact that 15,891 of those at highest risk are moved down a risk group. It is the net changes in risk category allocation that are of most interest (15), and these are displayed in the margins rather than in the cells of the table.

Classification Accuracy

The third perspective from which to evaluate a risk prediction model is its classification accuracy. High and low risk designations based on the model are typically associated with medical decisions about interventions. We refer to Table 2 which shows the possible risk assignments separately for subjects who do and do not go on to have events. The benefit of measuring the marker in the population can be characterized by two proportions, the proportion of subjects with subsequent events who are identified as high risk and the proportion of subjects without events who are identified as low risk. The former group can potentially receive an intervention that may prevent their events, while the latter avoid unnecessary interventions. The cost of measuring the marker can also be characterized by two proportions, the proportion of subjects without subsequent events who are classified as high risk and the proportion of subjects with events who are identified as low risk. The former group may suffer unnecessary medical interventions and place a burden on the medical system, while the latter group will not receive interventions they need. Because the proportions for any given group (events or non-events) sum to one, the benefit and cost can be summarized by just two numbers. The benefit is represented by the proportion of subjects with subsequent events who are classified as high risk according to the model (16). This is simply the true-positive rate (TPR) or sensitivity associated with classifying using the high risk threshold. The cost is represented by the proportion of subjects without subsequent events who are designated as high risk; this is the false-positive rate (FPR) or $1 - \text{specificity}$ (16). We stress however that these are different from the true- and false-positive rates that make up the ROC curve, as they use a specific, clinically meaningful threshold for high risk. Public health practitioners use information about the TPR and FPR to determine the value of the model for guiding decisions in clinical practice. Individual clinicians and patients may also find this information helpful when deciding whether or not to measure the markers (or have them measured).

The TPR and FPR for the breast cancer risk prediction models can be estimated using information displayed in the margins of the event and non-event rows of the risk stratification table. From Table 1, we calculate that using the model with breast density and

a high risk threshold of 1.67%, $\frac{2513+2120}{8784} \times 100 = 53\%$ of subjects who develop breast cancer within 5 years are identified as high risk. These individuals can presumably benefit from additional screening and chemoprevention with tamoxifen or raloxifene (3). However

this benefit comes at a cost of falsely identifying $\frac{121907+66624}{620445} \times 100 = 30\%$ of subjects who remain breast-cancer-free as high risk. These subjects may be needlessly sent for additional screening and perhaps even be placed on unnecessary medications, causing undue

stress and burdening the medical system. The performance of the model when considered in this light depends on the relative importance one places on these benefits and costs.

Misuses of the Risk Stratification Method

We next caution against two misuses of the risk stratification method.

Risk Stratification Tables for Head-To-Head Model Comparisons

Risk stratification tables were originally proposed for evaluating the value of adding a new marker to a set of baseline predictors, but they have recently been applied to head-to-head comparisons of risk models (3-5). That is, instead of cross-tabulating risk categories of two models where one includes baseline predictor variables and the other includes baseline predictor variables plus the new marker (nested models), risk stratification tables have been used to contrast models that include entirely different sets of markers or predictors (non-nested models).

We caution against this practice. Cross-tabulation of non-nested models gives information only about the extent of correlation between the risks calculated from the two models. Large amounts of reclassification suggest low correlation, and low amounts suggest high correlation. However correlation provides no information about differences in model performance. Table 3 compares two non-nested logistic regression models for predicting CVD risk, one including SBP only and the other including total cholesterol only, based on simulated data. The markers have been simulated to be normally distributed with the same mean and variance, so they have the same performance.

When the markers are simulated to be uncorrelated (panel A), the model with SBP reclassifies a substantial proportion (69%) of subjects compared to the model with total cholesterol. In contrast, when the markers are simulated to be highly correlated (panel B), there is relatively little overall reclassification (24%). Therefore the amount of reclassification does not represent differences in model performance, remembering that by design the markers are simulated to have the same performance. The large amount of reclassification in panel A, which reflects the low correlation between the markers, suggests that the information in the two models might be usefully combined to predict risk. But a more informative way to evaluate this combination would be to compare the composite model, including both SBP and total cholesterol, with each of the individual models as shown in Table 4. This is the original nested model setting.

A second problem with using risk stratification tables to evaluate non-nested risk prediction models is that the proportion of events displayed in the table's inner cells may be misleading. Because the cells contain subjects selected on the basis of factors in *both* models, there is no reason to expect that the proportion of events in any cell will fall within the ranges in either the row or column labels. Instead, as before, calibration can be evaluated by observing whether the proportion of events in the Table margins (the "Total" column or row) falls within the risk ranges in the corresponding labels. Cell event rates that fall outside of these ranges give some information about the value of combining SBP and total cholesterol into a single model, but again, a more informative way to assess the combination of markers is by using nested models, as in Table 4.

Risk Stratification with Case-Control Data

Risk stratification tables have also been used to evaluate risk prediction models fit using case-control data (eg (5)). It is well known that absolute risk estimates from case-control samples are biased, as the risk of an event is artificially fixed in the data by choosing the number of cases and controls in the design (17-20). This is illustrated in Table 5, which

shows the relevant information extracted from the margins of Table 4. We again use the simulated data and compare the performance of two logistic regression models for CVD, one with total cholesterol only and the other with SBP and total cholesterol. Panel A uses data sampled randomly from the population with an event rate of 10%. Panel B uses data sampled under a case-control design with a 1:1 case-control ratio, or an event rate of 50%. Observe that the risks estimated using the case-control data are higher than in the population sample; the population is shifted to higher risk categories in panel B as compared to panel A. This is to be expected, since the event rate is much higher in the case-control sample. The risks estimated from the case-control data are not representative of risks in the general population. Nor is classification accuracy in the case-control dataset representative of the general population. For example, using a risk threshold of 15%, the model with SBP and total cholesterol detects 97.4% of events (the TPR) and 70.2% of non-events (the FPR); this is in contrast with detection of 65.2% of events and 15.0% non-events in the general population using the same risk threshold. This difference is due to the artificially high risks in the case-control sample.

When logistic regression is used for risk modeling, as in this example, the intercept of the model can be adjusted to correct for the bias in the risk estimates due to case-control sampling (17) (see appendix). Panel C of Table 5 shows the results of this correction: the risks are now very similar to those observed in the population shown in Panel A. Observe also that among subjects with events the correct population proportions are in each risk category, and similarly for subjects without events. In other words, the TPR and FPR are correctly estimating the population TPR and FPR shown in Panel A.

However even with the correction to the model-predicted risks, the models' capacity to stratify the population into clinically relevant risk categories in the case-control sample is not representative of that in the general population. Observe in Panel C of Table 5 that the proportions of subjects in the various risk categories are very different than in the population sample in Panel A. This is a consequence of the fact that there are more cases in the case-control sample than in the general population. The distribution of risk can be corrected for the case-control sampling by calculating the distribution of risk separately for cases and controls, and combining these estimates using the event rate in the population (21), as shown in the appendix. Using this approach with the case-control data of Table 5B we calculate that the model with total cholesterol classifies 37.1%, 28.0%, 14.3%, and 20.6% of subjects into the four risk categories, while for the model with total cholesterol the values are 52.3%, 18.5%, 9.2%, and 20.1%. Both sets of values agree closely with those estimated using the population sample, shown in Panel A of Table 5.

Note that when cases and controls are matched with respect to factors known to predict risk, the risk estimates and distribution of risk need to be corrected in a more complex fashion (see (22) for details).

Using the above strategies for case-control data one can evaluate two of the criteria pertaining to model performance: 2) capacity for risk stratification, and 3) classification accuracy. For the third criterion, calibration, we note that the Hosmer-Lemeshow test is a valid test of model calibration even with case-control data. If a model is well-calibrated in the case-control sample, it is well-calibrated in the population.

Additional Issues in Evaluating Risk Prediction Models

In some applications one might want to measure a new marker only in a sub-population. The value of the marker for this subgroup can be evaluated by restricting the summaries we have described to the subpopulation of interest. One subpopulation of interest may be those at a specified risk according to the baseline model. We see in our simulated example (Table 4)

that SBP may be useful in predicting CVD risk among those at 10% to 15% risk according to total cholesterol alone, as it classifies $1947/7057 = 28\%$ of these subjects in the highest risk category and $2082/7057 = 30\%$ in the lowest risk category. Cook and colleagues (1) used this approach to evaluate the added benefit of C-reactive protein for predicting CVD risk. Focusing on the middle rows of the risk stratification table, they concluded that C-reactive protein may be useful in subpopulations at intermediate risk according to standard predictors, constituting 12% of the population. However the margins of the table indicated no major benefit for the population as a whole.

Careful thought should be put into choosing thresholds for defining risk categories. One approach is to specify benefits (and costs) associated with high risk designations for subjects destined to have an event (and not to have an event). A standard result from decision theory shows that given specified cost (C) and benefit (B) values, the optimal threshold, i.e. the one

for which the expected benefit is above 0, is $\frac{1}{1+B/C}$ (23,24). Presumably this kind of reasoning underlies the choice of thresholds in practice. For example, in the breast cancer

setting, solving $\frac{1}{1+B/C} = 0.0167$ yields a benefit-to-cost ratio of 59. Put another way, the working risk threshold of 1.67% for initiating preventive treatment for breast cancer corresponds with assuming that the benefit of correctly identifying a woman who develops breast cancer is 59 times greater than the cost of falsely designating a woman who will not develop breast cancer as high-risk. Viewing the thresholds in this manner may be helpful in some contexts. We note that predictiveness curves avoid the issue of choosing thresholds entirely; the same information contained in the margins of the risk stratification table is shown for all possible thresholds (16,25).

It is important that uncertainty in the estimates of model performance be acknowledged. This means reporting confidence intervals for all parameters, including the proportions of subjects stratified into each risk category, overall and separately for subjects with and without events. There are two sources of uncertainty. The first is uncertainty in the predicted risks, due to estimating the coefficients in the risk prediction model. The second is variability in the distribution of the predictors in the population. With small sample sizes or in subpopulations with small sample sizes, uncertainty can be substantial. We recommend bootstrapping both the model fit and model evaluation to capture both sources of variability in the confidence intervals (10 [p. 93]). Using the data from Table 5A as an example, the estimated proportions of subjects risk stratified with corresponding bootstrapped 95% confidence intervals are: 38.4% (36.8% to 40.2%), 27.17% (26.2% to 28.1%), 14.1% (13.5% to 15.5%), and 20.3% (19.3%, 21.1%) for the model with total cholesterol, and 53.2% (51.8% to 54.4%), 17.9% (17.2%, 18.6%), 9.0% (8.6%, 9.4%), and 20.0% (19.3%, 20.6%) for the model with SBP and total cholesterol. The confidence intervals around the TPRs and FPRs are similarly narrow in this large dataset.

Finally, we note that the same data should not be used to fit and evaluate the risk prediction model. In order to avoid overoptimism associated with fitting and evaluating the model on the same data, a training/test data split or internal cross-validation should be used (10 [pp. 93-4]). The model may also perform differently in another population, because the risk may be different (i.e. different model coefficients) and/or because the distribution of the predictors in the model may be different (26-28).

Conclusions

In summary, when used appropriately, a risk stratification table can be used to gauge the value of adding a marker to a risk prediction model. The key attributes of a model are its

calibration, its capacity to stratify the population into clinically relevant risk categories, and its classification accuracy, all of which can be extracted from the margins of the risk stratification table. These summaries represent an enormous improvement over the commonly reported C-statistic and Hosmer-Lemeshow test, because they describe the risks that are estimated by the model, the reliability of these risks, the distribution of the risks in the population, and the benefit and cost of classifying individuals using clinically relevant risk thresholds. These are key elements for understanding the value of the model for guiding medical decisions.

Appendix

The risk in the population can be calculated from the case-control data if an external measure of the population event rate is available. The relationship between the population risk (r) and the risk in the case-control sample (r_{cc}) is (14):

$$\text{logit } r = \text{logit } r_{cc} - \text{logit}(\rho) + \text{logit}(\rho_{cc}),$$

where $\text{logit}(x) = \log(x/(1-x))$, ρ is the population event rate, and ρ_{cc} is the event rate in the case-control sample.

The distribution of risk in the population can be estimated from case-control data in the following way (18). Let $D = 1$ denote an event, $D = 0$ a non-event, and cc a quantity in the case-control sample. The proportion of the sample in a specified risk range, say risk in (r_0, r_1) , can be calculated as

$$\begin{aligned} P(\text{risk in } (r_0, r_1)) &= P(\text{risk in } (r_0, r_1)|D=1)\rho + P(\text{risk in } (r_0, r_1)|D=0)(1-\rho) \\ &= P(\text{risk in } (r_0, r_1)|D=1, cc)\rho_{cc} + P(\text{risk in } (r_0, r_1)|D=0, cc)(1-\rho_{cc}), \end{aligned}$$

where the second line follows because the distribution of risk in cases and controls separately is the same in the case-control sample and general population.

References

1. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Annals of Internal Medicine*. 2006; 145:21–9. [PubMed: 16818925]
2. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007; 115:928–35. [PubMed: 17309939]
3. 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. *Annals of Internal Medicine*. 2008; 148:337–47. [PubMed: 18316752]
4. Lauer MS, Pothier CE, Magid DJ, Smith SS, Kattan MW. An externally validated model for predicting long-term survival after exercise treadmill testing in patients with suspected coronary artery disease and a normal electrocardiogram. *Annals of Internal Medicine*. 2007; 147:821–8. [PubMed: 18087052]
5. van der Steeg WA, Boekholdt M, Stein EA, El-archaoui K, Stroes ESG, Sandhu MS, Wareham NJ, Jukema JW, Luben R, Zwinderman AH, Kastelein JJP, Khaw K-T. Role of the apolipoprotein B – apolipoprotein A-I ratio in cardiovascular risk assessment: A case-control analysis in EPIC-Norfolk. *Annals of Internal Medicine*. 2007; 146:640–8. [PubMed: 17470832]
6. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, D'Agostino RB, Kannel WB, Vasan RS. A risk score for predicting near-term incidence of hypertension: The Framingham heart study. *Annals of Internal Medicine*. 2008; 148:102–10. [PubMed: 18195335]

7. Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, Benjamin EJ, D'Agostino RB, Vasan RS. Multiple biomarkers for the prediction of first major cardiovascular events and death. *New England Journal of Medicine*. 2006; 355:2631–39. [PubMed: 17182988]
8. Chlebowski RT, Collyar DE, Somerfield MR, Pfister DG. American Society of Clinical Oncology technology assessment on breast cancer risk reduction strategies: Tamoxifen and raloxifene. *Journal of Clinical Oncology*. 1999; 17:1939–55. [PubMed: 10561236]
9. Levine M, Moutquin JM, Walton R, Feightner J. Chemoprevention of breast cancer. A joint guideline from the Canadian Task Force on preventive health care and the Canadian Breast Cancer Initiative's steering committee on clinical practice guidelines for the care and treatment of breast cancer. *Canadian Medical Association Journal*. 2001; 164:1681–90. [PubMed: 11450210]
10. Harrell, FE. *Regression Modelling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer-Verlag; New York: 2001.
11. Hosmer, DW.; Lemeshow, S. *Applied Logistic Regression*. John Wiley and Sons; New York: 1989. Section 5.2.2
12. Cook NR. Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by MJ Pencina et al. *Statistics in Medicine*. 2008; 27:191–5. [PubMed: 17671959]
13. Cook NR. Statistical evaluation of prognostic versus diagnostic models: Beyond the ROC curve. *Clinical Chemistry*. 2008; 54:17–23. [PubMed: 18024533]
14. Pencina MJ, Vasan RS, D'Agostino RB. Algorithms for assessing cardiovascular risk in women. *Journal of the American Medical Association*. 2007; 298:176–7.
15. 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(2):157–72. [PubMed: 17569110]
16. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*. 2008; 167(3):362–8. [PubMed: 17982157]
17. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–11.
18. Borgan O, Langholz B. Non-parametric estimation of relative mortality from nested case-control studies. *Biometrics*. 1993; 49:593–602. [PubMed: 8369390]
19. Benichou J, Gail MH. Methods of inference for estimates of absolute risk derived from population-based case-control studies. *Biometrics*. 1995; 51:182–94. [PubMed: 7766773]
20. Langholz B, Borgan O. Methods of inference for estimates of absolute risk derived from population-based case-control studies. *Biometrics*. 1997; 53:767–74. [PubMed: 9192463]
21. Huang, Y.; Pepe, MS. A Parametric ROC Model Based Approach for Evaluating the Predictiveness of Continuous Markers in Case-control Studies. UW Biostatistics Working Paper Series. Workig Paper 318. <http://www.bepress.com/uwbiostat/paper318>
22. Huang, Y.; Pepe, MS. Semiparametric methods for evaluating the covariate-specific predictiveness of continuous markers in matched case-control studies. UW Biostatistics Working Paper Series. Working Paper 329. <http://www.bepress.com/uwbiostat/paper329>
23. Pauker SG, Kassirer JP. Therapeutic decision making: A cost-benefit analysis. *N Engl J Med*. 1975; 293(5):229–34. [PubMed: 1143303]
24. Pepe MS. Invited discussion of "The skill plot: a graphical technique for evaluating diagnostic tests". *Biometrics*. 2008; 64:256–258.
25. Huang Y, Pepe MS, Feng Z. Evaluating the predictiveness of a continuous marker. *Biometrics*. 2007; 63:1181–8. [PubMed: 17489968]
26. Copas JB. Regression, prediction, and shrinkage. *Journal of the Royal Statistical Society, Series B*. 1983; 45:311–54.
27. Chatfield C. Model uncertainty, data mining, and statistical inference. *Journal of the Royal Statistical Society, Series A*. 1995; 158(part 3):419–66.

28. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*. 1996; 15:361–87. [PubMed: 8668867]

Modified from Table 6 of Tice et al. (3). 5-year risks of breast cancer as predicted by models that do and do not include breast density. In each cell, the number of subjects (N), number of events and non-events, and proportion of subjects with events are shown.

Table 1

5-year risk from model without breast density						
5-year risk from model without breast density		0 to 1%	1% to 1.66%	1.67% to 2.5%	> 2.5%	Total
0 to 1%	N	176,831	38,500	71	0	215,402
	Events	1,161	415	0	0	1,576
	Non-events	175,670	38,085	71	0	213,826
	Prop. Events (%)	0.7	1.1	0.0	--	0.7
1% to 1.66%	N	64,297	99,456	37,149	1,025	201,927
	Events	526	1,328	754	32	2,640
	Non-events	63,771	98,128	36,395	993	199,287
	Prop. Events (%)	0.8	1.3	2.0	3.1	1.3
1.67% to 2.5%	N	8,741	45,478	71,309	23,267	148,795
	Events	74	609	1,419	621	2,723
	Non-events	8,667	44,869	69,890	22,646	146,072
	Prop. Events (%)	0.9	1.3	2.0	2.7	1.8
> 2.5%	N	90	2,672	15,891	44,452	63,105
	Events	0	38	340	1,467	1,845
	Non-events	90	2,634	15,551	42,985	61,260
	Prop. Events (%)	0.0	1.4	2.1	3.3	2.9
Total	N	249,959	186,106	124,420	68,744	629,229
	Events	1,761	2,390	2,513	2,120	8,784
	Non-events	248,198	183,716	121,907	66,624	620,445
	Prop. Events (%)	0.7	1.3	2.0	3.1	1.4

Table 2

Possible risk stratifications for subjects who do and do not go on to experience events. The number of subjects in each cell is represented by a symbol. The true-and false-positive rates (TPR and FPR) of the high risk classifications are shown.

	Low risk	High risk	Total
Non-events	a	b	r_1
Events	c	d	r_2

$$\text{TPR} = d/r_2$$

$$\text{FPR} = b/r_1$$

Simulated data for 5000 cases and 45000 controls. Two markers, total cholesterol and systolic blood pressure (SBP) have the same performance individually. One risk prediction model includes total cholesterol only and the other includes SBP only. In Panel A the two markers are uncorrelated. In Panel B the markers are highly correlated with a correlation coefficient 0.95. Each cell includes the number of subjects (N) and proportion of subjects with CVD events.

Table 3

A)						
Risk from model with SBP						
Risk from model with total cholesterol	< 5%	5% - 10%	10%- 15%	> 15%	Total	% Reclassified
< 5%	N 7724	5447	2676	3375	19222	60
	Prop. Events (%) 0.7	2.3	3.1	8.2	2.8	
5% – 10%	N 5190	3798	1941	2657	13586	72
	Prop. Events (%) 1.9	5.3	8.4	19.2	7.2	
10 – 15%	N 2616	1900	1097	1448	7057	84
	Prop. Events (%) 3.1	8.8	15.0	29.3	11.9	
> 15%	N 3361	2618	1506	2650	10135	74
	Prop. Events (%) 8.4	19.3	32.0	52.2	26.2	
Total	N 18887	13763	7220	10130	50000	69
	Prop. Events (%) 2.7	7.3	12.3	25.6	10.0	
B)						
Risk from model with SBP						
Risk from model with total cholesterol	< 5%	5% - 10%	10%- 15%	> 15%	Total	% Reclassified
< 5%	N 16893	2325	23	0	19241	12
	Prop. Events (%) 2.5	4.9	0	–	2.8	
5% – 10%	N 2304	9039	2061	167	13571	33
	Prop. Events (%) 5.4	7.0	9.4	11.4	7.1	
10 – 15%	N 21	2007	3502	1462	6992	50
	Prop. Events (%) 0	9.0	12.6	15.6	12.1	

A)							
Risk from model with SBP							
Risk from model with total cholesterol		< 5%	5% - 10%	10%- 15%	> 15%	Total	% Reclassified
> 15%	N	0	179	1530	8487	10196	17
	Prop. Events (%)	--	8.9	15.2	28.3	26.0	
Total	N	19218	13550	7116	10116	50000	24
	Prop. Events (%)	2.8	7.0	12.2	26.2	10.0	

Simulated data for 5000 cases and 45000 controls. The same data as in Panel A of Table 3, where the two markers are uncorrelated. One risk prediction model includes total cholesterol only while the other model includes both SBP and total cholesterol. Each cell includes the number of subjects (N) and proportion of subjects with CVD events.

Table 4

Risk from model with SBP and total cholesterol							
Risk from model with total cholesterol	< 5%	5% - 10%	10%-15%	> 15%	Total	% Reclassified	
< 5%	N	16381	2009	477	355	19222	15
	Prop. Events (%)	1.6	6.7	13.0	21.1	2.8	
5% - 10%	N	7139	3420	1446	1581	13586	75
	Prop. Events (%)	2.4	7.1	12.7	23.5	7.2	
10 - 15%	N	2082	1828	1200	1947	7057	83
	Prop. Events (%)	2.5	7.2	12.2	26.0	11.9	
> 15%	N	982	1668	1381	6104	10135	40
	Prop. Events (%)	4.2	7.3	13.5	37.8	26.2	
Total	N	26584	8925	4504	9987	50000	46
	Prop. Events (%)	2.0	7.1	12.8	32.6	10.0	

Table 5

Simulated data for two uncorrelated markers, SBP and total cholesterol, as predictors of CVD risk. One risk prediction model includes total cholesterol only while the other model includes both SBP and total cholesterol. In each risk category, the number of subjects (N), number of events and non-events, proportion of subjects who have events, and TPR and FPR using the low end of the risk category as the threshold are shown. Panel A shows the same data as in Table 4. This is a random sample from the population; the event rate is 10%. In Panel B, a case-control sample is drawn from the population; the sample event rate is 50%. In Panel C the estimated risks from the case-control data are corrected for the case-control sampling.

A) Random Sample from Population		Risk Category				
		< 5%	5% to 10%	10% to 15%	> 15%	Total
Total cholesterol	N (%)	19222 (38.44)	13586 (27.17)	7057 (14.11)	10135 (20.27)	50000 (100)
	Events	539	971	836	2654	5000
	Non-events	18683	12615	6221	7481	45000
	Prop. Events (%)	2.8	7.2	11.9	26.2	10.0
	TPR at threshold*	100	89.2	69.8	53.1	--
	FPR at threshold ⁺	100	58.5	30.5	16.6	--
SBP and total cholesterol	N (%)	26584 (53.17)	8925 (17.85)	4504 (9.01)	9987 (19.97)	50000 (100)
	Events	532	632	577	3259	5000
	Non-events	26052	8923	3927	6728	45000
	Prop. Events (%)	2.1	7.1	12.8	32.6	10.0
	TPR at threshold*	100	89.4	76.7	65.2	--
	FPR at threshold ⁺	100	42.1	23.7	15.0	--
B) Case-Control Sample		Risk Category				
		< 5%	5% to 10%	10% to 15%	> 15%	Total
Total cholesterol	N (%)	181 (0.36)	939 (1.88)	1828 (3.66)	47052 (94.10)	50000 (100)
	Events	9	67	226	24698	25000
	Non-events	172	872	1602	22354	25000
	Prop. Events (%)	5.0	7.1	12.4	52.5	50.0
	TPR at threshold*	100	100	99.7	98.8	--

A) Random Sample from Population		Risk Category				
		< 5%	5% to 10%	10% to 15%	> 15%	Total
	FPR at threshold ⁺	100	99.3	95.8	89.4	--
<hr/>						
SBP and total cholesterol	N (%)	2137 (4.27)	3041 (6.08)	2918 (5.84)	41094 (83.81)	50000 (100)
	Events	60	220	373	24347	25000
	Non-events	2077	2821	2545	17557	25000
	Prop. Events (%)	2.8	7.2	12.8	58.1	50.0
	TPR at threshold*	100	99.8	98.9	97.4	--
	FPR at threshold ⁺	100	91.7	80.4	70.2	--
<hr/>						
C) Case-Control Sample, Risks Corrected for Case-Control Sampling		Risk Category				
		< 5%	5% to 10%	10% to 15%	> 15%	Total
<hr/>						
Total cholesterol	N (%)	12646 (25.29)	12311 (24.62)	7890 (15.96)	17063 (34.13)	50000 (100)
	Events	2625	5093	4521	12761	25000
	Non-events	10021	7218	3459	4302	25000
	Prop. Events (%)	20.8	41.4	56.7	74.8	50.0
	TPR at threshold*	100	89.5	69.1	51.0	--
	FPR at threshold ⁺	100	59.9	31.0	17.2	--
<hr/>						
SBP and total cholesterol	N (%)	16896 (33.79)	8145 (16.29)	5022 (10.04)	19937 (39.87)	50000 (100)
	Events	2676	3388	2772	16164	25000
	Non-events	14220	4757	2250	3773	25000
	Prop. Events (%)	15.8	41.6	55.2	81.1	50.0
	TPR at threshold*	100	89.3	75.7	64.7	--
	FPR at threshold ⁺	100	43.1	24.1	15.1	--

* TPR = no. events in this category and above / total no. of events

⁺ FPR = no. non-events in this category and above / total no. of non-events