

Assessing the Value of Risk Predictions by Using Risk Stratification Tables

Holly Janes, PhD; Margaret S. Pepe, PhD; and Wen Gu, MS

The recent epidemiologic and clinical literature is filled with studies evaluating statistical models for predicting disease or some other adverse event. Risk stratification tables are a new way to evaluate the benefit of adding a new risk marker to a risk prediction model that includes an established set of markers. This approach involves cross-tabulating risk predictions from models with and without the new marker. In this article, the authors use examples to show how risk stratification tables can be used to compare 3 important measures of model performance between the models with and those without the new marker: the extent to which the risks calculated from the models reflect the actual fraction of persons in the population with events (calibration); the proportions in which the pop-

ulation is stratified into clinically relevant risk categories (stratification capacity); and the extent to which participants with events are assigned to high-risk categories and those without events are assigned to low-risk categories (classification accuracy). They detail common misinterpretations and misuses of the risk stratification method and conclude that the information that can be extracted from risk stratification tables is an enormous improvement over commonly reported measures of risk prediction model performance (for example, c-statistics and Hosmer–Lemeshow tests) because it describes the value of the models for guiding medical decisions.

Ann Intern Med. 2008;149:751-760.

www.annals.org

For author affiliations, see end of text.

The recent epidemiologic and clinical literature is filled with studies evaluating statistical models that predict risk for disease or some other adverse event (1–5). Because risk prediction models are intended to help patients and clinicians make decisions, evaluation of these models requires methods that differ from those used to assess models describing disease etiology. This is because the characteristics of the models are less important than their value for guiding decisions.

Cook and colleagues (1, 6) recently proposed a new approach to evaluate risk prediction models: 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 it has therefore been widely adopted in the literature (2–4). We examine the risk stratification approach in detail in this article, 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 and colleagues (2) to illustrate the concepts.

BACKGROUND

A risk prediction marker is any measure that is used to predict a person's risk for an event. It may be a quantitative measure, such as high-density lipoprotein cholesterol level, 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 for disease. But a large, significant association does not assure that the marker has value in predicting risk for many people.

A risk prediction model is a statistical model that combines information from several markers. Common types include logistic regression models, Cox proportional hazard models, and classification trees. Each type of model produces a predicted risk for each person by 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 persons or disease screening for subpopulations identified as high risk, or to select persons 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 fraction of persons in the population with actual events (its calibration); the proportions in which the population is stratified into clinically relevant risk categories (its stratification capacity); and the extent to which participants with events are assigned to high-risk categories and those without events are assigned to low-risk categories (its classification accuracy).

Risk prediction models are commonly evaluated by using the receiver-operating characteristic (ROC) curve (4,

See also:

Print

Key Summary Points 752

Web-Only

Appendix

Conversion of graphics into slides

Key Summary Points

Risk prediction models are statistical models used to predict the probability of an outcome on the basis of the values of 1 or more risk factors (markers).

The accuracy of the model's predictions is typically summarized with statistics that describe the model's discrimination and calibration.

Risk stratification tables are a more informative way to assess and compare the models.

The tables illustrate the distribution of predictions across risk categories.

That illustration allows users to assess 3 key measures of the models' value for guiding medical decisions: the models' calibration, ability to stratify people into clinically relevant risk categories, and accuracy at classifying patients into higher- and lower-risk categories. This information is contained in the margins of the risk stratification table rather than in its cells.

The tables should only be used to compare risk prediction models when one of the models contains all of the markers that are contained in the other (nested models); they should not be used to compare models with different sets of markers (nonnested models).

The table predictions require corrections when case-control data are used.

7), which is a standard tool for evaluating the discriminatory accuracy of diagnostic or screening markers. This curve shows the true-positive rate plotted against the false-positive rate for rules that classify persons by using risk thresholds that vary over all possible values. Receiver-operating characteristic curves are generally not helpful for evaluating risk prediction models because they do not provide information about the actual risks that the models predict or about the proportion of participants who have high or low risk values. Moreover, when comparing ROC curves for 2 risk prediction models, the models are aligned according to their false-positive rates (that is, different risk thresholds are applied to the 2 models to achieve the same false-positive rate). This is clearly inappropriate. In addition, the area under the ROC curve or *c*-statistic, a commonly reported summary measure that can be interpreted as the probability that the predicted risk for a participant with an event is higher than that for a participant without an event, has little direct clinical relevance. Clinicians are never asked to compare risks for a pair of patients—one who will eventually have the event and one who will not.

Neither the ROC curve nor the *c*-statistic relates to the practical task of predicting risks for clinical decision making.

Cook and colleagues (1, 6) propose using risk stratification tables to evaluate the incremental value of a new marker, or the benefit of adding a new marker (for example, C-reactive protein), to an established set of risk predictors (for example, Framingham risk predictors, such as age, diabetes, cholesterol level, smoking, and low-density lipoprotein cholesterol levels). 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, stratification capacity, and classification accuracy can be derived from a risk stratification table and used to assess the added value of a marker for clinical and health care policy decisions.

EXAMPLE

Tice and colleagues (2) published a study that builds and evaluates a model for predicting breast cancer risk by using data from 1 095 484 women in a prospective cohort and incidence data from the Surveillance, Epidemiology, and End Results database. Age, race or ethnicity, family history, and history of breast biopsy were used to model risk with 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 age 65 years and 2.2 for women age 65 years or older. This suggests that breast density is strongly associated with disease risk—that is, that 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, which requires knowledge of the frequency distribution of breast density in the population.

To evaluate the added value of breast density, Tice and colleagues defined 5-year breast cancer risk categories as low (<1%), low to intermediate (1% to 1.66%), intermediate to high (1.67% to 2.5%), and high (>2.5%). The 1.67% cutoff for intermediate risk was presumably chosen on the basis of recommendations by the American Society of Clinical Oncology (8) and the Canadian Task Force on Preventive Health Care (9) to counsel women with 5-year risks greater than this threshold about considering tamoxifen for breast cancer prevention. Tice and colleagues used a risk stratification table (Table 1) 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

Table 1. Five-Year Risks for Breast Cancer as Predicted by Models That Do and Do Not Include Breast Density*

5-Year Risk from Model without Breast Density	5-Year Risk from Model with Breast Density				Total
	0% to <1%	1% to 1.66%	1.67% to 2.5%	>2.5%	
0% to <1%					
Women, <i>n</i>	176 831	38 500	71	0	215 402
Events, <i>n</i>	1161	415	0	0	1576
Nonevents, <i>n</i>	175 670	38 085	71	0	213 826
Proportion of women with events	0.7	1.1	0.0	–	0.7
1% to 1.66%					
Women, <i>n</i>	64 297	99 456	37 149	1025	201 927
Events, <i>n</i>	526	1328	754	32	2640
Nonevents, <i>n</i>	63 771	98 128	36 395	993	199 287
Proportion of women with events	0.8	1.3	2.0	3.1	1.3
1.67% to 2.5%					
Women, <i>n</i>	8741	45 478	71 309	23 267	148 795
Events, <i>n</i>	74	609	1419	621	2723
Nonevents, <i>n</i>	8667	44 869	69 890	22 646	146 072
Proportion of women with events	0.9	1.3	2.0	2.7	1.8
>2.5%					
Women, <i>n</i>	90	2672	15 891	44 452	63 105
Events, <i>n</i>	0	38	340	1467	1845
Nonevents, <i>n</i>	90	2634	15 551	42 985	61 260
Proportion of women with events	0.0	1.4	2.1	3.3	2.9
Total					
Women, <i>n</i>	249 959	186 106	124 420	68 744	629 229
Events, <i>n</i>	1761	2390	2513	2120	8784
Nonevents, <i>n</i>	248 198	183 716	121 907	66 624	620 445
Proportion of women with events	0.7	1.3	2.0	3.1	1.4

* Modified from Table 6 of Tice and colleagues (2).

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 persons in the population with those same predictor values who experience the event (10). With many predictors, and especially with continuous predictors, we cannot evaluate calibration at each possible predictor value because there are too few participants with exactly those values. Instead, the standard approach is to place persons within categories of predicted risk and to compare the category values with the observed event rates for participants in each category.

The calibration of the risk prediction models for breast cancer 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 are 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 less than 1% is 0.7%, which falls within the range of the risk category; the same is true for the risk category of 1% to 1.66% (proportion of events, 1.3%), 1.67% to 2.5% (proportion of events, 1.8%), and greater than 2.5% (proportion of events, 2.9%). Thus, the model without breast density seems to be well calibrated. Similarly, when participants are categorized according to their risk as calculated by the

model that includes breast density, the proportion of events in each category falls within the range of risk in the column label (for example, for the risk category of 0% to <1%, the proportion of events is 0.7%; for 1% to 1.66%, the proportion is 1.3%). Therefore, the model with breast density is also well calibrated.

The Hosmer–Lemeshow test (11), a common test of model calibration, is 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 (for example, *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** to be a useful adjunct to the *P* value from the Hosmer–Lemeshow test because it makes the calibration assessment more concrete.

Event rates in the inner cells of the risk stratification table do not provide useful information about calibration, despite suggestions to the contrary (12, 13). Consider, for example, the cells in the second row of **Table 1** labeled “1% to 1.66%,” in which the proportion of events ranges from 0.8% in the “0% to <1%” column to 3.1% in the “>2.5%” column. All but 1 of these event rates fall outside the range of risk in the row labels. Remember that participants 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 (i.e., in the “>2.5%” column) presumably have higher breast density values than those in the row as a whole ($n = 201\,927$). However, the model that does not include breast density is not designed to capture the higher risk (3.1%) in this subgroup because breast density is not included in the model. The model without breast density is still well calibrated, however, because of the good agreement between the row label (1% to 1.66%) and the event rate in the margin (1.3%).

Nevertheless, the event rates in the inner cells of the risk stratification table are 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 way to express this gradient of risk is to use 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 a hazard ratio. As noted, Tice and colleagues (2) provided a measure of risk gradient when they reported hazard ratios for breast density that adjusted for baseline risk factors, estimating that, among women younger than age 65 years, the rate of breast cancer is 4.2 times higher for women with breasts that are extremely dense than for those whose breasts are almost entirely fat, whereas for women age 65 years or older, the rate is 2.2 times higher.

Risk-Stratification Capacity

Having established that a model can reliably be used to calculate the chance of an event (that is, it 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 that, if based on relatively uninformative markers, will predict risks close to the average for all participants and 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 participants at the extremes of the risk distribution, in which 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, in which there are still uncertainties 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 629 229) at highest risk

(>2.5% row); and 56% in the 2 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 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 any further 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 [2]), 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 persons 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 participants from the 1.67% to 2.5% risk category to the highest-risk category must be paired with the fact that 15 891 of those at highest risk are moved down a risk group. The net changes in risk category allocation are of most interest (15) and 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. **Table 2** shows the possible risk assignments separately for participants who do and those who do not have events. The benefit of measuring the marker in the population can be characterized by 2 proportions: the proportion of participants with subsequent events who are identified as high risk and the proportion of participants without events who are identified as low risk. The former group can potentially receive an intervention that may prevent their events, whereas the latter can avoid unnecessary interventions. The cost of measuring the marker can also be characterized by 2 proportions, the proportion of participants without subsequent events who are classified as high risk and the proportion of participants with events who are identified as low risk. The former group may be subjected to unnecessary

Table 2. Possible Risk Stratifications for Participants Who Do and Do Not Eventually Have Events*

Participants	Low Risk	High Risk	Total
Nonevents	a	b	r_1
Events	c	d	r_2

* The number of participants in each cell is represented by a symbol. The true-positive rate is d/r_2 ; the false-positive rate is b/r_1 .

medical interventions and burden the medical system, whereas the latter group will not receive the interventions they need. Because the proportions for any given group (events or nonevents) sum to 1, the benefit and cost can be summarized by just 2 numbers. The benefit is represented by the proportion of participants with subsequent events who are classified as high risk according to the model (16). This is simply the true-positive rate or sensitivity associated with classifying in terms of the high-risk threshold. The cost is represented by the proportion of participants without subsequent events who are designated as high risk; this is the false-positive rate or 1 minus specificity (16). We stress, however, that these differ from the true- and false-positive rates that make up the ROC curve because they use a specific, clinically meaningful threshold for high risk. Public health practitioners use information about true- and false-positive rates to determine the value of the model for guiding decisions in clinical practice. Clinicians and patients may also find this information helpful when deciding whether to measure the markers (or have them measured).

The true- and false-positive rates for the breast cancer risk prediction models can be estimated by using information displayed in the margins of the event and nonevent rows of the risk stratification table. From **Table 1**, we calculate that by using the model with breast density and a high-risk threshold of 1.67%, 53% of participants $((2513 + 2120)/8784 \times 100)$ 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 (2). However, this benefit comes at a cost of falsely identifying 30% of participants $((121\,907 + 66\,624)/620\,445 \times 100)$ who remain free of breast cancer as high risk. These participants may be needlessly sent for additional screening and perhaps even be given unnecessary medications, causing undue stress and burden to 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 2 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 (2–4). That is, instead of cross-tabulating risk stratifications of 2 models, in which 1 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 (nonnested models).

We caution against this practice. Cross-tabulation of nonnested models gives information only about the extent of correlation between the risks calculated from the 2 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 2 nonnested logistic regression models for predicting cardiovascular disease risk, one including systolic blood pressure only and the other including total cholesterol level only, on the basis of simulated data. The 2 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 (**Table 3A**), the model with systolic blood pressure reclassifies a substantial proportion (69%) of participants compared with the model with total cholesterol level. In contrast, when the markers are simulated to be highly correlated (**Table 3B**), there is relatively little overall reclassification (24%). Therefore, the amount of reclassification does not represent differences in model performance, remembering that the markers are simulated to have the same performance. The large amount of reclassification in **Table 3A**, which reflects the low correlation between the markers, suggests that the information in the 2 models might be usefully combined to predict risk. However, a more informative way to evaluate this combination would be to compare the composite model, including both systolic blood pressure and total cholesterol level, 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 nonnested risk prediction models is that the proportion of events displayed in the table's inner cells may be misleading. Because the cells contain participants 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 systolic blood pressure and total cholesterol levels into a single model, but again, a more informative way to assess the combination of markers is by using nested models, such as **Table 4**.

Risk Stratification with Case–Control Data

Risk stratification tables have also been used to evaluate risk prediction models fit by using case–control data (4). It is well known that absolute risk estimates from case–control samples are biased, as the risk for an event is artificially fixed in the data by choosing the number of case patients and control participants in the design (17–19). **Table 5** illustrates this point. We again use the simulated

Table 3. Simulated Data for 5000 Case Patients and 45 000 Control Participants*

Risk from Model with Total Cholesterol Level	Risk from Model with SBP				Total	Reclassified, %
	<5%	5% to <10%	10% to 15%	>15%		
A. Uncorrelated markers						
<5%						
Participants, <i>n</i>	7724	5447	2676	3375	19 222	60
Proportion of sample with CVD events	0.7	2.3	3.1	8.2	2.8	–
5% to <10%						
Participants, <i>n</i>	5190	3798	1941	2657	13 586	72
Proportion of sample with CVD events	1.9	5.3	8.4	19.2	7.2	–
10% to 15%						
Participants, <i>n</i>	2616	1900	1097	1448	7057	84
Proportion of sample with CVD events	3.1	8.8	15.0	29.3	11.9	–
>15%						
Participants, <i>n</i>	3361	2618	1506	2650	10 135	74
Proportion of sample with CVD events	8.4	19.3	32.0	52.2	26.2	–
Total						
Participants, <i>n</i>	18 887	13 763	7220	10 130	50 000	69
Proportion of sample with CVD events	2.7	7.3	12.3	25.6	10.0	–
B. Highly correlated markers†						
<5%						
Participants, <i>n</i>	16 893	2325	23	0	19 241	12
Proportion of sample with CVD events	2.5	4.9	0	–	2.8	–
5% to <10%						
Participants, <i>n</i>	2304	9039	2061	167	13 571	33
Proportion of sample with CVD events	5.4	7.0	9.4	11.4	7.1	–
10% to 15%						
Participants, <i>n</i>	21	2007	3502	1462	6992	50
Proportion of sample with CVD events	0	9.0	12.6	15.6	12.1	–
>15%						
Participants, <i>n</i>	0	179	1530	8487	10 196	17
Proportion of sample with CVD events	–	8.9	15.2	28.3	26.0	–
Total						
Participants, <i>n</i>	19 218	13 550	7116	10 116	50 000	24
Proportion of sample with CVD events	2.8	7.0	12.2	26.2	10.0	–

CVD = cardiovascular disease; SBP = systolic blood pressure.

* Two markers, total cholesterol level and SBP, have the same performance individually. One risk prediction model includes total cholesterol level only and the other includes SBP only.

† Correlation coefficient, 0.95.

data of Table 4 and compare the performance of 2 logistic regression models for cardiovascular disease, one with total cholesterol level only and the other with systolic blood pressure and total cholesterol level. Table 5A uses data randomly sampled from the population with an event rate of 10%. Table 5B uses data sampled under a case-control design with a 1:1 case-control ratio or an event rate of 50%. The risks estimated by using the case-control data are higher than those in the population sample; the population is shifted to higher risk categories in Table 5B compared with Table 5A. This is to be expected, because 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 data set representative of that in the general population. For example, by using a risk threshold of 15%, the model with systolic blood pressure and total cholesterol level detects 97.4% of events (the true-positive rate) and 70.2% of nonevents (the false-positive rate); this is in contrast with the same risk threshold detecting 65.2% of events and 15.0% of nonevents in the

general population. The 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) (Appendix, available at www.annals.org). Table 5C shows the results of this correction. The distributions of risk among participants with and without events are now very similar to those observed in the population shown in Table 5A. Among participants with events, the correct population proportions are in each risk category. The same is true for participants without events. In other words, the true- and false-positive rates are correctly estimating the population true- and false-positive rates shown in Table 5A.

However, even with the correction to the model-predicted risks, the model's capacity to stratify the population into clinically relevant risk categories in the case-control sample is not representative of its capacity to do so in the general population. In Table 5C, the proportions of participants in the various risk categories are very different

from the sample in part A. This is because there are more events 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 participants with and without events and combining these estimates by using the event rate in the population (20), as shown in the **Appendix** (available at www.annals.org). By using this approach with the case–control data of **Table 5C**, we calculate that the model with total cholesterol level classifies 37.1%, 28.0%, 14.3%, and 20.6% of participants into each of the 4 risk categories, whereas for the model with total cholesterol level, the values are 52.3%, 18.5%, 9.2%, and 20.1%. Both sets of values agree closely with those estimated by using the population sample, shown in **Table 5A**.

When participants with and without events 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 (21).

These strategies can be used for case–control data to evaluate 2 of the criteria pertaining to model performance: capacity for risk stratification and classification accuracy. For the third criterion, calibration, 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 subpopulation. 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. Our simulated example (**Table 4**) shows that systolic blood pressure may be useful in predicting cardiovascular disease risk among those at 10% to 15% risk according to total cholesterol level alone, as it puts 1947 of 7057 (28%) of these participants in the highest risk category and 2082 of 7057 (30%) in the lowest risk category. Cook and colleagues (1) used this approach to evaluate the added benefit of C-reactive protein level for predicting cardiovascular disease risk. Focusing on the middle rows of the risk stratification table, they concluded that C-reactive protein level may be useful in subpopulations at intermediate risk according to standard predictors (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 participants destined to have (or not to have) an event. A standard result from decision theory shows that, given specified cost (C) and benefit (B) values, the optimal threshold (that is, the one for which the expected benefit is above 0) is $1/(1 + B/C)$ (22, 23). Presumably this kind of reasoning underlies the choice of thresholds in practice. For example, in the breast cancer setting, solving $1/(1 + B/C) = 0.0167$ yields a benefit-to-cost ratio of 59. Put another way, the working risk threshold of 1.67% for initiating preventive treatment of breast cancer corresponds with the assumption that the benefit of correctly identifying a woman who develops breast cancer is 59 times greater

Table 4. Simulated Data for 5000 Case Patients and 45 000 Control Participants*

Risk from Model with Total Cholesterol Level	Risk from Model with SBP and Total Cholesterol Level				Total	Reclassified, %
	<5%	5% to <10%	10% to 15%	>15%		
<5%						
Participants, <i>n</i>	16 381	2009	477	355	19 222	15
Proportion of sample with CVD events	1.6	6.7	13.0	21.1	2.8	–
5% to <10%						
Participants, <i>n</i>	7139	3420	1446	1581	13 586	75
Proportion of sample with CVD events	2.4	7.1	12.7	23.5	7.2	–
10% to 15%						
Participants, <i>n</i>	2082	1828	1200	1947	7057	83
Proportion of sample with CVD events	2.5	7.2	12.2	26.0	11.9	–
>15%						
Participants, <i>n</i>	982	1668	1381	6104	10 135	40
Proportion of sample with CVD events	4.2	7.3	13.5	37.8	26.2	–
Total						
Participants, <i>n</i>	26 584	8925	4504	9987	50 000	46
Proportion of sample with CVD events	2.0	7.1	12.8	32.6	10.0	–

CVD = cardiovascular disease; SBP = systolic blood pressure.

* The same data as in **Table 3A**, in which the 2 markers are uncorrelated. One risk prediction model includes total cholesterol level only, whereas the other model includes both SBP and total cholesterol level.

Table 5. Simulated Data for 2 Uncorrelated Markers, SBP and Total Cholesterol Level, as Predictors of Cardiovascular Disease Risk*

Sample	Risk Category				Total
	<5%	5% to <10%	10% to 15%	>15%	
A. Random sample from population					
Total cholesterol level					
Participants, <i>n</i> (%)	19 222 (38.44)	13 586 (27.17)	7057 (14.11)	10 135 (20.27)	50 000 (100)
Events, <i>n</i>	539	971	836	2654	5000
Nonevents, <i>n</i>	18 683	12 615	6221	7481	45 000
Proportion of sample with 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 level					
Participants, <i>n</i> (%)	26 584 (53.17)	8925 (17.85)	4504 (9.01)	9987 (19.97)	50 000 (100)
Events, <i>n</i>	532	632	577	3259	5000
Nonevents, <i>n</i>	26 052	8923	3927	6728	45 000
Proportion of sample with 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					
Total cholesterol level					
Participants, <i>n</i> (%)	181 (0.36)	939 (1.88)	1828 (3.66)	47 052 (94.10)	50 000 (100)
Events, <i>n</i>	9	67	226	24 698	25 000
Nonevents, <i>n</i>	172	872	1602	22 354	25 000
Proportion of sample with events	5.0	7.1	12.4	52.5	50.0
TPR at threshold, %†	100	100	99.7	98.8	–
FPR at threshold, %‡	100	99.3	95.8	89.4	–
SBP and total cholesterol level					
Participants, <i>n</i> (%)	2137 (4.27)	3041 (6.08)	2918 (5.84)	41 094 (83.81)	50 000 (100)
Events, <i>n</i>	60	220	373	24 347	25 000
Nonevents, <i>n</i>	2077	2821	2545	17 557	25 000
Proportion of sample with 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	–
C. Case-control sample, risks corrected for case-control sampling					
Total cholesterol level					
Participants, <i>n</i> (%)	12 646 (25.29)	12 311 (24.62)	7890 (15.96)	17 063 (34.13)	50 000 (100)
Events, <i>n</i>	2625	5093	4521	12 761	25 000
Nonevents, <i>n</i>	10 021	7218	3459	4302	25 000
Proportion of sample with 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	–
SBP and total cholesterol level					
Participants, <i>n</i> (%)	16 896 (33.79)	8145 (16.29)	5022 (10.04)	19 937 (39.87)	50 000 (100)
Events, <i>n</i>	2676	3388	2772	16 164	25 000
Nonevents, <i>n</i>	14 220	4757	2250	3773	25 000
Proportion of sample with 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	–

FPR = false-positive result; SBP = systolic blood pressure; TPR = true-positive result.

* One risk prediction model includes total cholesterol level only, whereas the other model includes both SBP and total cholesterol level. Part A shows the same data as in Table 4. This is a random sample from the population; the event rate is 10%. In part B, a case-control sample is drawn from the population; the sample event rate is 50%. In part C, the estimated risks from the case-control data are corrected for the case-control sampling.

† TPR is the number of events in this category and higher risk categories divided by total number of events.

‡ FPR is the number of nonevents in this category and higher risk categories divided by total number of nonevents.

than the cost of falsely designating as high risk a woman who will not develop breast cancer. 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, 24).

It is important that uncertainty in the estimates of model performance be acknowledged. This means reporting CIs for all parameters, including the proportions of participants stratified into each risk category, overall and separately for participants with and without events. There are 2 sources of uncertainty: uncertainty in the predicted risks due to estimating the coefficients in the risk predic-

tion model, and 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 CIs (10). Using the data from **Table 5A** as an example, the estimated proportions of participants risk-stratified with corresponding bootstrapped 95% CIs are 38.4% (CI, 36.8% to 40.2%), 27.2% (CI, 26.2% to 28.1%), 14.1% (CI, 13.5% to 15.5%), and 20.3% (CI, 19.3% to 21.1%) for the model with total cholesterol level and 53.2% (CI, 51.8% to 54.4%), 17.9% (CI, 17.2% to 18.6%), 9.0% (CI, 8.6% to 9.4%), and 20.0% (CI, 19.3% to 20.6%) for the model with systolic blood pressure and total cholesterol level. The CIs around the true- and false-positive rates are similarly narrow in this large data set.

Finally, we note that the same data should not be used to fit and evaluate the risk prediction model. To avoid overoptimism associated with fitting and evaluating the model on the same data, a training and test data split or internal cross-validation should be used (10). The model may also perform differently in another population because of different risk (that is, different model coefficients) or because of different distribution of the model's predictors (25–27).

CONCLUSION

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.

From the Fred Hutchinson Cancer Research Center and the University of Washington, Seattle, Washington.

Acknowledgment: The authors thank Patrick Bossuyt for very helpful discussions of the material.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Holly Janes, PhD, Fred Hutchinson Cancer Research Center; 1100 Fairview Avenue North M2-C200, Seattle, WA 98109; e-mail, hjanes@scharp.org.

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

References

1. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann Intern Med.* 2006;145:21–9.
2. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann Intern Med.* 2008;148:337–47. [PMID: 18316752]
3. 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. *Ann Intern Med.* 2007;147:821–8. [PMID: 18087052]
4. van der Steeg WA, Boekholdt SM, Stein EA, El-Harchaoui K, Stroes ES, Sandhu MS, et al. Role of the apolipoprotein B-apolipoprotein A-I ratio in cardiovascular risk assessment: a case-control analysis in EPIC-Norfolk. *Ann Intern Med.* 2007;146:640–8. [PMID: 17470832]
5. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, et al. A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study. *Ann Intern Med.* 2008;148:102–10. [PMID: 18195335]
6. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation.* 2007;115:928–35. [PMID: 17309939]
7. Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med.* 2006;355:2631–9. [PMID: 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. *J Clin Oncol.* 1999;17:1939–55. [PMID: 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. *Can Med Assoc J.* 2001;164:1681–90.
10. Harrell FE. *Regression Modelling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York: Springer-Verlag; 2001.
11. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* New York: Wiley; 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 M. J. Pencina et al., *Statistics in Medicine* (DOI: 10.1002/sim.2929). *Stat Med.* 2008;27:191–5. [PMID: 17671959]
13. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem.* 2008;54:17–23. [PMID: 18024533]
14. Pencina MJ, Van Rossum TS, D'Agostino RB Sr. Algorithms for assessing cardiovascular risk in women [Letter]. *JAMA.* 2007;298:175–6; author reply 177–8. [PMID: 17622595]
15. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Van Rossum TS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med.* 2008;27:157–72; discussion 207–12. [PMID: 17569110]
16. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, et al. Integrating the predictiveness of a marker with its performance as a classifier. *Am J Epidemiol.* 2008;167:362–8. [PMID: 17982157]
17. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika.* 1979;66:403–11.
18. Benichou J, Gail MH. Methods of inference for estimates of absolute risk derived from population-based case-control studies. *Biometrics.* 1995;51:182–94. [PMID: 7766773]
19. Langholz B, Borgan O. Estimation of absolute risk from nested case-control data. *Biometrics.* 1997;53:767–74. [PMID: 9192463]
20. Huang Y, Pepe MS. A parametric ROC model based approach for evaluating the predictiveness of continuous markers in case-control studies. *Collection of Biostatistics Research Archive.* Accessed at www.bepress.com/uwbiostat/paper318 on 29 September 2008.
21. Huang Y, Pepe MS. Semiparametric methods for evaluating the covariate-specific predictiveness of continuous markers in matched case-control studies. *Collection of Biostatistics Research Archive.* Accessed at www.bepress.com/uwbiostat/paper329 on 29 September 2008.
22. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis.

N Engl J Med. 1975;293:229-34. [PMID: 1143303]

23. **Pepe MS**. Invited discussion of "The skill plot: a graphical technique for evaluating diagnostic tests". *Biometrics*. 2008;64:256-8.

24. **Huang Y, Sullivan Pepe M, Feng Z**. Evaluating the predictiveness of a continuous marker. *Biometrics*. 2007;63:1181-8. [PMID: 17489968]

25. **Copas JB**. Regression, prediction, and shrinkage. *Journal of the Royal Statis-*

tical Society, Series B. 1983;45:311-54.

26. **Chatfield C**. Model uncertainty, data mining, and statistical inference. *Journal of the Royal Statistical Society, Series A*. 1995;158:419-66.

27. **Harrell FE Jr, Lee KL, Mark DB**. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361-87. [PMID: 8668867]

Current Author Addresses: Drs. Janes and Pepe: Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109. Dr. Gu: Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195.

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):

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

in which $\text{logit } (a) = \log (a/[1 - a])$, 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 nonevent, and cc denotes restriction to 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 follows:

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

The second line follows because the distribution of risk among participants with events is the same in the case-control sample as in the population. Similarly, among participants without events, the risk is the same in the case-control sample and population.