10-17-2012

Methods for Evaluating Prediction Performance of Biomarkers and Tests

Margaret Pepe
University of Washington, Fred Hutch Cancer Research Center, mspepe@u.washington.edu

Holly Janes
Fred Hutchinson Cancer Research Center, hjanes@scharp.org

Methods for Evaluating Prediction Performance of Biomarkers and Tests

Margaret Pepe and Holly Janes

1 Introduction

1.1 Background

Predicting an individual's risk of a bad outcome is a key component of medical decision making. For example, the Framingham Risk Calculator (www.framinghamheartstudy.org) provides 10 year risks of cardiovascular event outcomes such as coronary heart disease and myocardial infarction events as functions of risk factors [5]. Risk factors for hard coronary heart disease (HCHD), defined as myocardial infarction or coronary death, include age, smoking status, treatment for hypertension and levels of total cholesterol, high density lipoproteins and systolic blood pressure. If the 10-year risk exceeds 20%, longterm treatment with cholesterol lowering drug therapy is recommended. Another risk calculator routinely used in clinical practice is the Breast Cancer Risk Assessment Tool (BCRAT)[30]. The predicted outcome may be a future event such as a cardiovascular event for the Framingham Risk Calculator or breast cancer diagnosis for BCRAT. However, a current condition can also constitute the predicted outcome. For example, presence of acute kidney injury is the outcome predicted in Parikh et al. [15] and presence of critical illness requiring hospitalization is the outcome predicted by Seymour et al. [25]. New molecular biology techniques for measuring biomarkers and new imaging technologies portend the availability of excellent predictors of risk in the future. Moreover, easy dissemination of risk calculators over the internet will increase the impact of risk prediction models on clinical practice.

Margaret Pepe
Fred Hutchinson Cancer Research Center and University of Washington, Seattle, Washington, USA
e-mail: mspepe@u.washington.edu

Holly Janes
Fred Hutchinson Cancer Research Center, Seattle, Washington, USA e-mail: hjanes@scharp.org

In this chapter we discuss methods for evaluating risk prediction models. Since the goal is to use risk calculators in medical decision making, our perspective is that the crucial evaluations are about determining whether or not good decisions can be made with use of the risk prediction model. For much of the chapter we define a good decision rule as one that recommends treatment for people who would get the bad outcome in the absence of treatment (called *cases* here) and does not recommend treatment for those who would not have a bad outcome (called *controls* here). The rationale is that the cases could possibly benefit from treatment while the controls would not benefit but would be subjected to toxicities, expenses and other costs associated with treatment. Therefore, we consider that a prediction model is good if it leads to a large proportion of cases being classified into a treatment category and a large proportion of controls into a no-treatment category. However, prediction models are not just classifiers. A prediction model is an algorithm that people use to calculate their risk and as such it has real meaning and interpretation for individuals. This must be accounted for in evaluating a prediction model and sets it apart from evaluations of other classifiers such as diagnostic tests where a numerical score itself may not have meaning.

1.2 Notation and Assumptions

We write D for the outcome of interest and assume that it is binary $D = 1$ for a case and $D = 0$ for a control. If the outcome is an event occurring within a specific time period, for example a cardiovascular event within 10 years, the cases may be called *events* and the controls may be called *nonevents*. The prevalence or event rate in the population is denoted by ρ:

$$\rho = P(D = 1).$$

The predictors are denoted by X and Y, both of which may be multidimensional. In Section 2, we consider a single risk model and use X for the predictors in the model. In Section 3, we consider two nested risk models, one with the baseline predictors denoted by X and one expanded model that includes the predictors Y in addition to X.

To focus the presentation we assume that in the absence of predictor information subjects do not receive treatment. The purpose of the risk model is to identify subjects for treatment. One may be interested in the opposite scenario in some settings. That is, standard of practice may be to receive treatment and the purpose of the model is to identify subjects at low risk who may forego treatment. This setting can be dealt with using methods analogous to those we describe here and is mentioned later, but for the most part, and to keep the discussion focused, we consider the default no-treatment scenario.

We assume that data are available for a cohort of N independent untreated subjects. We write the data as $\{(D_i, Y_i, X_i);\ i = 1, \ldots, N\}$. Most of this chapter concerns conceptual formulations of measures to quantify and compare the prediction perfor-

mance of risk models. As such, sampling variability and statistical inference is not a major focus, at least in sections 2 and 3. In other words we assume N is very large.

1.3 Illustrative Data

For illustrative purposes we use a simulated data set. The simulated data is available on the DABS website (http://labs.fhcrc.org/pepe/dabs/index.html) and was previously used in a publication [18]. It reflects the prevalence and risk ranges that have been reported for cohort studies of cardiovascular disease. The data are comprised of 10,000 observations of which 1017 are case subjects and 8983 are control subjects. Predictors X and Y are one-dimensional and continuous. In practice the predictors X and/or Y may be scores derived from multiple risk factors or biomedical measurements. In Section 2 we focus on the predictor X only, while in Section 3 we consider X and Y together as predictors. Table 1 shows fitted logistic regression models for these data.

Table 1 Estimated coefficients for baseline and enhanced risk models.

	Baseline Model		**Enhanced Model**	
Factor	Coefficient	SE	Coefficient	SE
Intercept	-3.67	0.07	-4.23	0.09
X	1.72	0.05	1.77	0.05
Y		1.01	0.05	

1.4 Chapter Outline

In section 2 we discuss the concept of risk and validity of a risk model. The performance of a risk model is discussed in Section 3. A plethora of performance measures are used to assess prediction performance and we describe the main ones. Insights and relationships amongst the measures are provided. In section 4 we consider the comparison of two risk prediction models focusing especially on the comparison of two nested models. This is a somewhat controversial area of statistical methodology where public debate and discourse is needed. We hope that this chapter will add constructively to the discourse.

2 Validity of the Risk Calculator

2.1 What is risk(X)?

The function $\text{risk}(x) = P(D = 1|X = x)$ is the *frequency* of events among subjects with predictor values $X = x$. It is important to remember that statistical analysis delivers information about population level entities, such as averages and frequencies and distributions. We emphasize this point because the term 'individual level risk' is often used in this era of 'personalized medicine'. But the risk value calculated from the risk calculator for subject j with predictors X_j, $\text{risk}(X_j)$, is not the probability of a random event for that subject. Rather it is the frequency of events in the group of subjects with the same predictors as him.

To make the distinction concrete, suppose that $\text{risk}(X_j) = 0.20$ and consider the (large) group of subjects with predictors X_j. The following scenarios are all consistent with $\text{risk}(X_j) = 0.20$: (i) 20% of the subjects are destined to have the event with probability 1 while 80% are destined not to have the event; (ii) for each subject i there is a stochastic mechanism giving rise to an event $D = 1$ with individual level probability $\pi_i = 0.20$; (iii) 10% of subjects are destined to have the event ($\pi_i = 1.00$ for them), 50% are destined not to have the event ($\pi_i = 0.00$ for them) and for 40% of subjects the outcome is stochastic with $\pi_i = 0.25$. Gail and Pfeiffer [6] discuss the distinction between the unobservable individual level probabilities denoted by π_i and $\text{risk}(x) = P(D = 1|X = x)$. They note the equality: $\text{risk}(x) = E\{\pi_i|X_i = x\}$. A simulated example that illustrates the distinction can be found in Pepe [18].

Unless repeated observations of the outcome were available for an individual, one cannot make inference about individual level risks. In this sense, the individual level risks, i.e. the π_i's, are not observable. It is not clear that they are even well defined when a subject can only experience one event. We regard the concept of individual level risk π as a useless distraction. Individualized risk will not be discussed further in this chapter.

Risk is a function of the predictors modeled. It is important to remember that an individual with two sets of predictors X_i and Y_i can have at least three risk values, $\text{risk}(X_i) = P(D = 1|X = X_i)$, $\text{risk}(Y_i) = P(D = 1|Y = Y_i)$ and $\text{risk}(X_i, Y_i) = P(D = 1|X = X_i, Y = Y_i)$. Each is his 'true risk'. Each is a frequency of events but calculated amongst different groups of subjects: those with $X = X_i$, those with $Y = Y_i$ and those with $X = X_i$ and $Y = Y_i$, respectively.

2.2 The Meaning of Calibration

The traditional definition of calibration is that a well calibrated risk calculator $\text{risk}^*(\cdot)$, is one for which the frequency of events among subjects with $X = x$ is equal to $\text{risk}^*(x)$: $P(D = 1|X = x) = \text{risk}^*(x)$. When X is multidimensional it can be difficult to assess calibration defined in this strong sense. A weaker definition of cal-

ibration is typically used in practice: the criterion is that $P(D=1|\text{risk}^*(X)=r)\approx r$ for all r. In words, if the frequency of events is r among subjects whose calculated risks are equal to r, then the model risk$^*(\quad)$ is considered well calibrated in the weak sense. This level of validity seems like a minimal requirement to justify use of the risk model in practice.

We note that strong calibration implies weak calibration because under strong calibration, where $P(D=1|X=x)=\text{risk}^*(x)$ we have $P(D=1|\text{risk}^*(X)=r)=E\{P(D=1|X=x)|\text{risk}^*(X)=r\}=E\{\text{risk}^*(X)|\text{risk}^*(X)=r\}=r$. However, weak calibration does not imply strong calibration and must not be interpreted as such.

Calibration is an attribute that may not transport from one population to another. For example, if there are predictors (known or unknown) that are not included in the model and that have different distributions in two populations, the true risk models in the two populations will likely be different, $P(D=1|X,\text{ population A})\neq P(D=1|X,\text{ population B})$. Strong calibration may not transport for this reason. However, if all relevant predictors are included in the model, strong calibration will not be affected by a change in the distribution of predictors. On the other hand, a model that is weakly calibrated but not strongly calibrated may not transport its weak calibration to other populations where distributions of modeled covariates differ.

2.3 Assessing Calibration

To evaluate calibration one compares the observed event rates within subgroups of subjects defined by the modeled predictors to the average modeled risk values among those subjects. The subgroups are usually selected as having modeled risk values in a narrow range. A visual aid to this comparison is the predictiveness curve

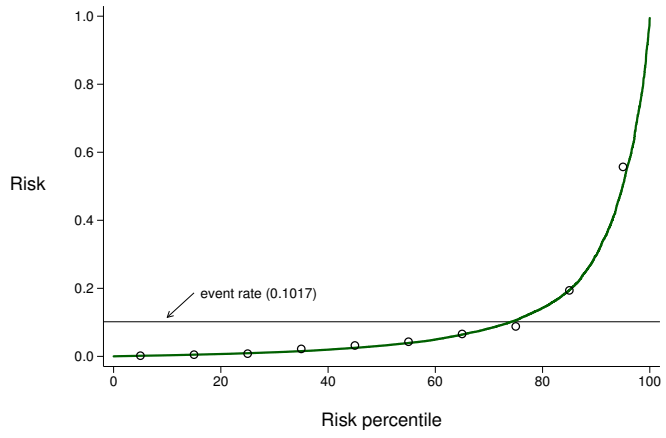

Fig. 1 Visual assessment of calibration with the predictiveness curve.

[20]. The plot orders subjects from lowest to highest modeled risk, plotting the risk versus the corresponding percentile. The x-axis, allows one to identify subgroups similar in regards to modeled risk. For example, groups may be defined by decile of modeled risk. The observed event rate in each subgroup is superimposed on the plot as shown in Figure 1. If the circles follow the predictiveness curve, we conclude that modeled risks are close to observed risks and the model is well calibrated (in the weak sense) in the dataset. The predictiveness curve in Figure 1 shows that the fitted model is extremely well calibrated. An alternative but related visual display

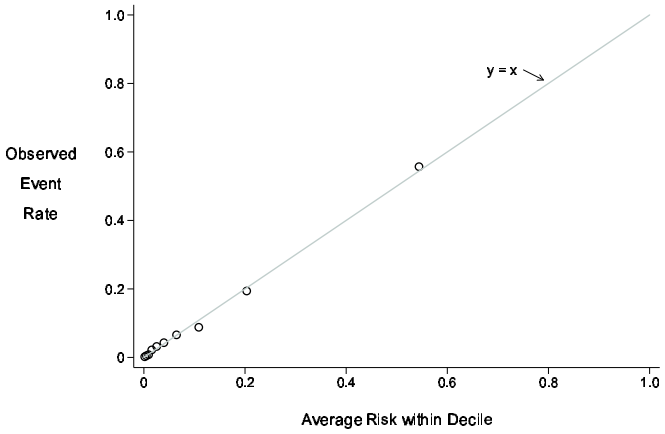

Fig. 2 Calibration with the calibration plot.

is the calibration plot (Figure 2) [25]. This also uses intervals of modeled risk (e.g. deciles) but plots observed event rates versus average modeled risk producing points that should lie along the 45° line if the model is well calibrated. A disadvantage of the calibration plot versus the predictiveness curve is that points can be more clumped. Moreover the variability in modeled risk within an interval risk category is not evident from the calibration plot. If substantial variation exists in a category one might choose to use smaller subdivisions of that category in comparing observed event rates with modeled rates.

The Hosmer-Lemeshow test is often reported as a test for calibration [14]. It also uses subsets defined by modeled risk, typically deciles, and calculates

$$H \equiv \sum_{k=1}^{10} N_k \frac{(0_k - \overline{\hat{r}_k}(X))^2}{\overline{\hat{r}_k}(X)(1-\overline{\hat{r}_k}(X))}$$

where N_k = the number of subjects in the k^{th} category, 0_k = the observed event rate and $\overline{\hat{r}_k}(X)$ is the average estimated modeled risk. Under the null hypothesis of good model calibration, the statistic H has a chi-squared distribution. The statistic corre-

sponds closely with the predictiveness and calibration plots in that it compares 0_k and $\overline{\hat{r}_k}$. However, the statistic has been criticized for many reasons including that it is highly dependent on sample size (almost certainly significant if N is large enough and non-significant if N is small enough). In of itself, it does not convey the extent to which the model is well calibrated to the data. But it can serve as a descriptive adjunct to the visual display of calibration manifested in the predictiveness or calibration plots. Although deciles of risk are typically used for the predictiveness plot, calibration plot and Hosmer-Lemeshow statistic, as mentioned above there is no reason that other subgroupings couldn't be used.

2.4 Achieving a Well Calibrated Model

This chapter does not cover procedures to fit regression models. We refer to textbooks that cover the topic in depth [8, 26]. When only a few predictors are included, the task of fitting a model that is well calibrated to the data and not over-fit is relatively straightforward. Although sampling variability in the fitted model remains, assuming good study design practices have been followed and in the absence of additional data, one will propose the fitted model for use in practice. The next task will be to evaluate its performance for prediction.

Sometimes an externally fitted model will be proposed for validation on a new dataset. If the model is not found to be well calibrated on the new dataset, it must be regarded as not having been validated. Nevertheless, some investigators proceed to evaluate its classification performance. In our opinion this is inappropriate. Individuals will want to use the prediction model not just as an aid in classification but to calculate their risk as a function of predictors. A poorly calibrated model is known to be invalid for this purpose and should be abandoned.

A better strategy perhaps is to derive a revised risk prediction model with the new dataset. This may be done by using the original modeled risk value as a sole predictor and fitting a model with that single predictor to the data. This is called *recalibration*. Because only one predictor is involved it should be easy to arrive at a revised model that is well calibrated to the new dataset and therefore worthy of evaluation for its predictive performance.

Another strategy is to begin anew and fit a model with each predictor in the original model included as a candidate predictor for the new model. If many predictors are involved, issues pertaining to overfitting arise and one will need to use techniques such as 'shrinkage' in order to arrive at a believable model. Fruitfully applying such techniques requires considerable skill and experience. An advantage of starting over with the new dataset is that a combination of predictors that is closer to optimal for application in the target population may be arrived at. Recalibrating an existing model is easer and subject to less sampling variability, but one maintains the same predictor combination of the original model that may be suboptimal if it was derived from a population that is not the one of interest.

Why might an externally derived model fail to be well calibrated? Certainly if it was derived in a different population with different predictor effects (or distributions, see 2.2) it may not be valid for use in the target population. Another common cause is overfitting the model in the original population.

For the remainder of this chapter we assume that a model that is well calibrated is to be evaluated for its performance.

3 Measuring Prediction Performance of a Single Model

3.1 Context

The focus of this section is on describing conceptual approaches to evaluating a risk prediction model. We suppose that we have an extremely large population and the true risk function, $\text{risk}(X) = P(D = 1|X)$, is available. How do we measure the performance of this risk function for use in the population?

It is important to remember that the purpose of calculating $\text{risk}(X)$ is to affect medical decisions. Recall that we assume that in the absence of knowledge of $\text{risk}(X)$ no treatment will be offered, but that if $\text{risk}(X)$ is found to be large enough, treatment will be offered. Implicitly we assume that treatment must have associated with it some costs, e.g. toxicity, monetary costs, inconvenience. Otherwise all subjects, regardless of their risk value, could be treated even if the benefit was minimal.

3.2 Case and Control Risk Distributions

All metrics to gauge the performance of a risk model are derived from the distribution of $\text{risk}(X)$ in cases and in controls. Having a visual display of the distributions is often helpful. Although probability density functions (Figure 3) give a sense for the separation between case and control distributions, cumulative distributions, or $1 - cdfs$ denoted by HR_D and $\text{HR}_{\bar{D}}$ in Figure 4 are more useful because they explicitly show the proportions of cases and controls above any threshold value used to define 'high risk.' Since decisions to opt for treatment will be based on 'high risk' designation, it is of interest and can be seen directly from the cdfs how many cases and controls are recommended for treatment using the risk model.

Gail and Pfeiffer [6] make the point that the cumulative distribution function(cdf) of risk in the population as a whole, cases and controls together, is sufficient to calculate performance measures. This is true because the case distribution and the control distribution can both be calculated from the overall population distribution of risk. Using $f(\text{risk}(X) = r)$ to denote probability density for $\text{risk}(X)$ at r, this follows because

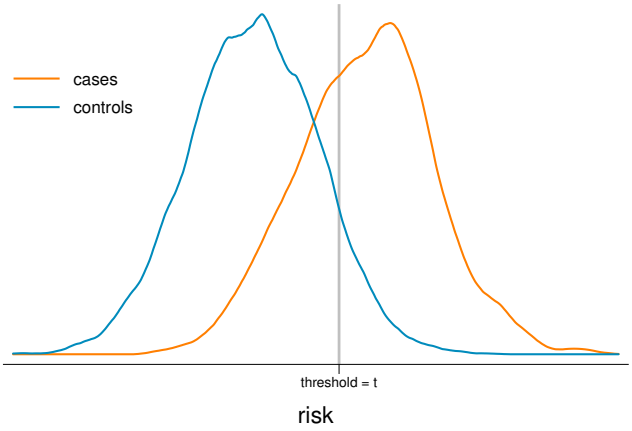

Fig. 3 Case and Control risk distributions on the logit scale.

$$f(\text{risk}(X)=r|D=1)=\frac{P(D=1|\text{risk}(X)=r)}{P(D=1)}f(\text{risk}(X)=r)$$
$$=\frac{rf(\text{risk}(X)=r)}{P(D=1)}$$
$$f(\text{risk}(X)=r|D=0)=\frac{(1-r)f(\text{risk}(X)=r)}{P(D=0)}$$

The overall cdf of risk(X) in the population as a whole is shown by the *predictiveness curve* that displays quantiles of risk in the population (Figure 5) but we find the case and control specific cdfs to be a more informative display.

The *receiver operating characteristic* (ROC) curve, that plots $\text{HR}_D(r)=P(\text{risk}(X)>r|D=1)$ versus $\text{HR}_{\bar{D}}(r)\equiv P(\text{risk}(X)>r|D=0)$ (Figure 6) is another popular visual display to assess performance. When the case and control distributions are well separated, the ROC curve

$$\text{ROC}(f)=\text{HR}_D(\text{HR}_{\bar{D}}^{-1}(f))$$

lies close to the upper left hand corner of the $[0,1]\times[0,1]$ quadrant. Huang and Pepe [10]show that the ROC curve and prevalence, $\rho=P[D=1]$, together can be used to calculate the predictiveness curve. And, since the predictiveness curve can be used to calculate the case and control distributions of risk it follows that $(\text{ROC}(f), f\in(0,1)\,;\,\rho)$ contain all the information available in the case and control distributions of risk(X). However, the risk thresholds r corresponding to the points on the ROC curve, $(\text{HR}_{\bar{D}}(r),\text{HR}_D(r))$, are not visible from the ROC curve detracting from its interpretation. When plotting a single ROC curve it is possible to add the risk thresholds to the axes (Figure 6). However, with two or more ROC curves this is not possible. Since ROC curves do not align models according to the same risk

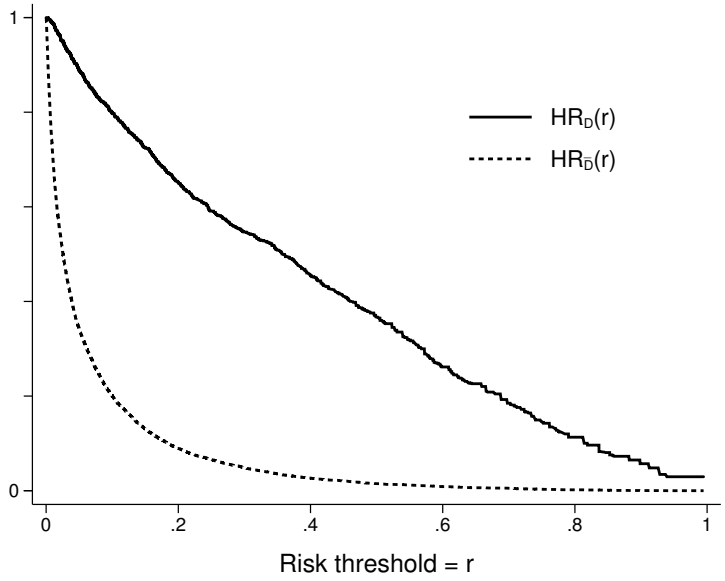

Fig. 4 Case and control distributions of risk shown with $1-cdf$ functions, $\mathrm{HR}_C(t) = P(\mathrm{risk}(X) > t|D=1)$ and $\mathrm{HR}_N(t) = P(\mathrm{risk}(X) > t|D=0)$

thresholds and do not display risk thresholds they are less useful for evaluating prediction models than they are for evaluating diagnostic tests whose numeric scales are often irrelevant in data displays.

3.3 Risk thresholds

How should one choose the risk threshold for designating a patient as at sufficiently high risk to warrant treatment? Intuitively the costs and benefits associated with treatment dictate the choice. If the treatment is very costly in terms of toxicities, monetary expense or inconvenience, a high threshold may be warranted. If the treatment is very likely to be effective at preventing a bad outcome, this might lower the threshold for treatment. In the extreme, an ineffective or prohibitively costly treatment dictates use of a risk threshold close to 1 corresponding to few people being treated. At the other extreme, a highly effective inexpensive treatment with few toxicities dictates that many people should be treated, i.e., use of a low risk threshold.

An explicit relationship is given in the next result between the risk threshold for treatment, r_H, and the net costs and benefits of treatment. Write the *net benefit of treatment* to a subject who would have an event in the absence of treatment as B. For example, if treatment reduces the risk of an event by 50%, the benefit might be $0.5\times$

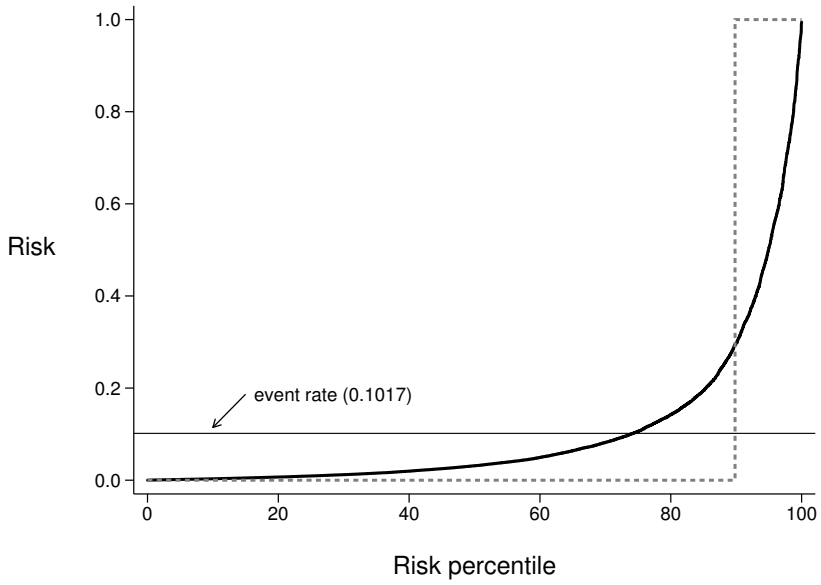

Fig. 5 The predictiveness curve that shows the quantiles of risk(X) in the population. The prediction curve for the ideal model where all cases have risk $= 1$ and all controls have risk $= 0$, is shown as a dashed step function.

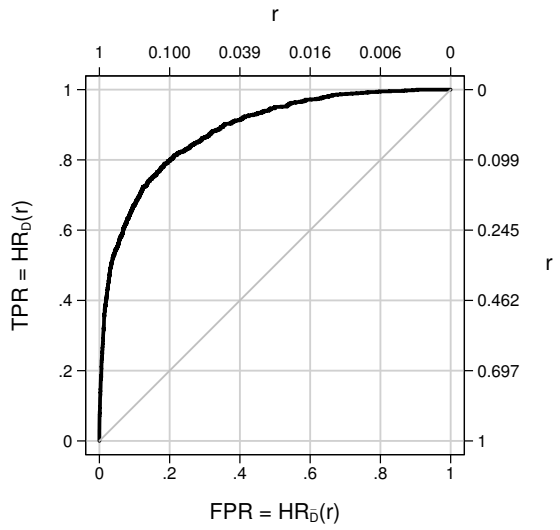

Fig. 6 ROC plot with risk thresholds *r* shown on top and right axes.

{value of an event} and the net benefit is the benefit less the costs associated with treatment for a subject who would otherwise have an event. Note that these costs must be put on the same scale as the benefit, i.e. {value of an event} in our example. A subject who would not have an event in the absence of treatment suffers only costs and no benefit from being treated. We use Cost to denote the corresponding cost for such a subject.

Result 1

The risk threshold for treatment that should be employed to ensure subjects with risk values risk(X) benefit on average is

$$r_H = \text{Cost}/(\text{Cost}+B).$$

The threshold does not depend on the model or on the predictors in the model.

Proof. The expected net benefit for subjects with risk$(X) = r$ is

$$B\,P(D=1|\text{risk}(X)=r) - \text{Cost}\,P(D=0|\text{risk}(X)=r) = B\,r\, - \text{Cost}(1-r)$$

which is positive if

$$\frac{r}{1-r} > \frac{\text{Cost}}{B}$$

In other words, to ensure a positive average benefit for subjects with risk values r they should opt for treatment if $r/(1-r) > \text{Cost}/B$ and should not opt for treatment if $r/(1-r) < \text{Cost}/B$. That is the treatment risk threshold is $\text{Cost}/(\text{Cost}+B)$. □

The result agrees with the intuition that high costs and/or low benefits should be correspond to high values of the risk threshold for opting for treatment.

Example Suppose treatment reduces the risk of breast cancer within 5 years by 50% but it can cause other bad outcomes such as other cancers, cardiovascular events and hip fractures that are considered equally bad. Assume other bad outcomes, O, occur with a frequency of x in the absence of treatment but with a frequency of z in the presence of treatment regardless of whether $D = 1$ or 0. Let $D(0), D(1), O(0), O(1)$ denote the potential outcomes with and without treatment. A subject who would not develop breast cancer absent treatment suffers the increased risk of other bad events; his/her net cost of treatment is

$$\begin{aligned}
&P(D=1 \text{ or } O=1|T=1, D(0)=0) - P(D=1 \text{ or } O=1|T=0, D(0)=0)\\
&= P(O=1|T=1, D(0)=0) - P(O=1|T=0, D(0)=0)\\
&= P(O=1|T=1) - P(O=1|T=0) = z-x.
\end{aligned}$$

The net benefit of treating a subject who would get breast cancer absent treatment is the reduction in his/her event probability,

$$
\begin{aligned}
&P(D=1 \text{ or } O=1|T=0,D(0)=1) - P(D=1 \text{ or } O=1|T=1,D(0)=1) \\
&= 1 - [P(D=1|T=0,D(0)=1) + P(D=0 \text{ and } O=1|T=1,D(0)=1)] \\
&= 1 - [0.5 + (1-0.5)P(O=1|T=1)] = 0.5 + 0.5z
\end{aligned}
$$

The treatment risk threshold is therefore $\frac{z-x}{z-x+0.5(1+z)}$.

In practice it is often difficult to specify costs and benefits associated with treatment. It can be especially difficult to specify them on a common scale when they are qualitatively different entities. On the other hand a treatment threshold for risk is often easier to specify. For example, the ATP guidelines recommend that subjects with risks above 20% consider longterm treatment with cholesterol lowering therapy to reduce risk of cardiovascular events. Individuals make decisions such as whether or not to have genetic testing of their fetus based on their risk of having a child with genetic abnormalities. Their chosen risk threshold is often derived intuitively from their knowledge of the qualitative costs and benefits of amniocentesis. We similarly make decisions about procuring insurance using our tolerance for risk. Result 1 tells us the explicit relationship between the risk threshold and the perceived cost-benefit ratio. For example, by choosing a risk threshold equal to 20% for cholesterol lowering therapy, we are implicitly stating that the net benefit of therapy for a would-be case is 4 times the net cost of therapy for a would-be control because $\frac{r_H}{1-r_H} = .2/(1-.2) = 1/4$.

3.4 Summary Statistics When a Risk Threshold is Available

In this section we consider settings where a risk threshold, r_H, exists that defines high risk status with possible recommendation for treatments or, perhaps, for entry into a clinical trial. The context then is essentially reduced to a binary classification rule, high risk or not high risk, and measures commonly used to summarize performance of binary classifiers are appropriate. We already defined the proportions of cases and controls classified as high risk as

$$
\begin{aligned}
\mathrm{HR}_D(r_H) &= P(\mathrm{risk}(X) > r_H|D=1) \\
\mathrm{HR}_{\bar{D}}(r_H) &= P(\mathrm{risk}(X) > r_H|D=0).
\end{aligned}
$$

A perfect model classifies all cases and no controls as high risk, $\mathrm{HR}_D(r_H) = 1$ and $\mathrm{HR}_{\bar{D}}(r_H) = 0$. A good model classifies a large proportion of cases as high risk and a low proportion of controls as high risk. Terms commonly used for $\mathrm{HR}_D(r_H)$ are true positive rate and sensitivity while false positive rate and 1-specificity are terms used for $\mathrm{HR}_{\bar{D}}(r_H)$.

The $\mathrm{HR}_D(r_H)$ is a good attribute of a prediction model while $\mathrm{HR}_{\bar{D}}(r_H)$ is a negative attribute. The expected population net benefit of using the model with risk threshold r_H, $\mathrm{NB}(r_H)$, combines the two into an overall population measure that balances the positive and negative attributes:

$$\begin{aligned}\mathrm{NB}(r_H) &= P(\mathrm{risk}(X) > r_H)\{BP(D=1|\mathrm{risk}(X) > r_H) - \mathrm{Cost}\, P(D=0|\mathrm{risk}(X) > r_H)\} \\ &= B\, P(\mathrm{risk}(X) > r_H|D=1)P(D=1) - \mathrm{Cost}\, P(\mathrm{risk}(X) > r_H|D=0)\, P(D=0) \\ &= B\mathrm{HR}_D(r_H)\rho - \mathrm{Cost}\mathrm{HR}_{\bar{D}}(r_H)(1-\rho).\end{aligned}$$

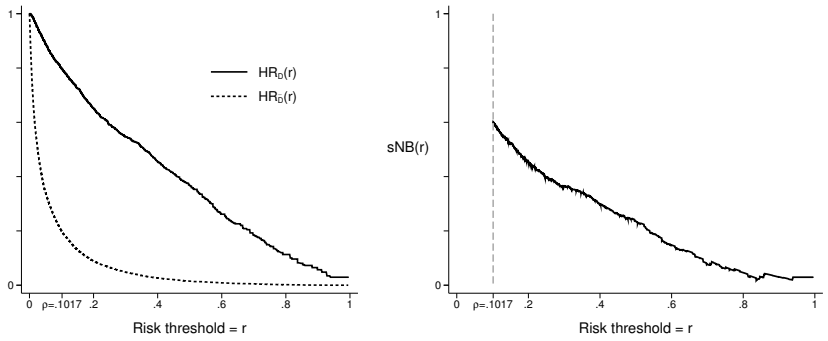

Fig. 7 Proprtions of cases and controls above risk threshold r and corresponding standardized net benefit.

Observe that $\mathrm{NB}(r_H)$ is an expectation over the entire population and assumes treatment is offered if $\mathrm{risk}(X) > r_H$. In contrast the expected net benefit in the proof of result 1 concerns only subjects with $\mathrm{risk}(X) = r$ and considers their net benefit if they receive treatment.

Recall that Result 1 tells us that use of the risk threshold r_H implies $\mathrm{Cost}/B = r_H/(1-r_H)$. Substituting into the above gives us an expression for expected net benefit that depends only on model performance parameters $(\mathrm{HR}_D(r_H), \mathrm{HR}_{\bar{D}}(r_H))$ and the constants (ρ, r_H).

$$\mathrm{NB}(r_H) = \{\rho\mathrm{HR}_D(r_H) - \frac{r_H}{1-r_H}(1-\rho)\mathrm{HR}_{\bar{D}}(r_H)\}B$$

Vickers and Elkin [27] propose measuring net benefit in units that assign B a value 1. In those units $\mathrm{NB}(r_H) = \rho\mathrm{HR}_D(r_H) - \frac{r_H}{1-r_H}(1-\rho)\mathrm{HR}_{\bar{D}}(r_H)$

A standardized version of $\mathrm{NB}(r_H)$ is found by dividing $\mathrm{NB}(r_H)$ by the maximum possible value that can be achieved, namely ρB, corresponding to a perfect model with $\mathrm{HR}_D(r_H) = 1$ and $\mathrm{HR}_{\bar{D}}(r_H) = 0$. Baker et al. used the term 'relative utility' but we prefer the more descriptive term standardized net benefit' and use notation that corresponds:

$$sNB(r_H) = NB(r_H)/\max(NB(r_H)) = NB(r_H)/\rho B$$
$$= HR_D(r_H) - \frac{r_H}{(1-r_H)}\frac{(1-\rho)}{\rho}HR_{\bar{D}}(r_H).$$

An advantage of standardizing net benefit is that it no longer depends on the measurement unit B, an entity that is sometimes difficult to digest. $sNB(r_H)$ is a unitless numerical summary in the range (0,1).

Another interpretation for $sNB(r_H)$ is that it discounts the true positive rate $HR_D(r_H)$ using the scaled false positive rate $HR_{\bar{D}}(t)$ to yield a discounted true positive rate. The FPR is scaled so that the units are on the same scale as the TPR. $sNB(r_H)$ can therefore be interpreted as the true positive rate of a prediction model that has no false positives but equal benefit.

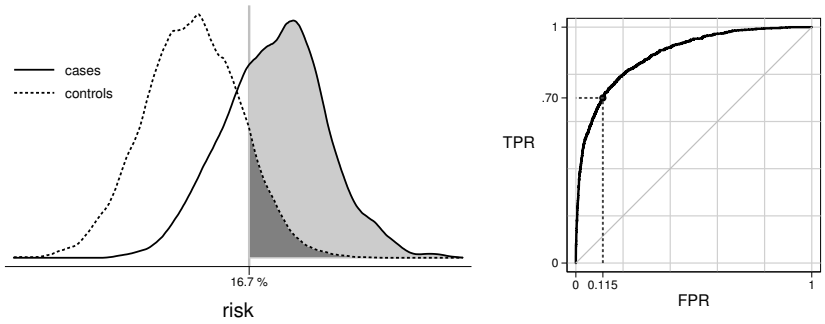

Fig. 8 $ROC^{-1}(0.7) - 0.115$

In our opinion a reasonably complete reporting of a prediction model's performance is given by $HR_D(r_H)$, $HR_{\bar{D}}(r_H)$ and $sNB(r_H)$. One also needs to keep in mind the prevalence, ρ, and the risk threshold, r_H, in order to interpret $sNB(r_H)$ and its components $(HR_D(r_H), HR_{\bar{D}}(r_H))$.

Example

Figure 7 shows $HR_D(r_H)$, $HR_{\bar{D}}(r_H)$ and their weighted average $sNB(r_H)$ for various choices of risk threshold, r_H. For example, at $r_H = 0.2$ we have that 65.2% of cases and 8.9% of controls are classified as high risk. This corresponds to a benefit that is 45.6% of the maximum possible benefit, where the maximum possible benefit would be achieved by classifying all 10.17% cases and no controls as high risk. Alternatively we can consider that the observed true positive rate of 65.2% is discounted to 45.6% by the 8.9% of controls that are designated as high risk.

The plot in Figure 7 allows one to view performance achieved with different choices of risk threshold. Observe that the net benefit curve is plotted only for $r_H > \rho$. This is because the assumed default action is to *not* treat subjects. In the absence of predictors, all subjects are assigned risk values ρ. Therefore a risk value of ρ must correspond to the 'no treatment' rule. To be consistent and rational we still assign

no treatment if risk$(X) < \rho$ when predictors are available. Therefore risk thresholds for treatment below ρ are not relevant. If one were to instead assign treatment as the default decision and use the model for decisions to forego treatment, then the expected net benefit and its standardized version would be calculated differently: $sNB(r_L) = (1 - HR_{\bar{D}}(r_L)) - \frac{\rho}{(1-\rho)} \frac{(1-r_L)}{r_L}(1 - HR_D(r_L))$, where r_L is the low risk threshold below which subjects would *not* receive treatment. Moreover this version of $sNB(r_L)$ would only be calculated for $r_L < \rho$. See Baker [1] for details.

In many circumstances a fixed risk threshold for assigning treatment does not exist. It can be useful to consider a variety of thresholds. Consider that a prediction model is often developed to determine eligibility criteria for a clinical trial of a new treatment. For example, new treatments for acute kidney injury are under development and prediction models are sought to identify high risk subjects for upcoming clinical trials. Potentially toxic or expensive treatments will require higher risk thresholds than less toxic, inexpensive treatments. Having displays that show performance across a range of risk thresholds will allow researchers to entertain use of risk prediction models for designing trials of different types of treatment.

Another important reason to display performance as a function of risk threshold is that it allows individuals with different tolerances to assess the value of ascertaining predictor information for them. If the distribution of risk thresholds among individuals in the population were known, those could be overlaid on the plots. One could summarize the information by integrating over the distribution of risk thresholds: $HR_{\bar{D}} = E(HR_{\bar{D}}(r_H)) =$ the overall proportions of controls that do not receive treatment; $HR_D = E(HR_D(r_H)) =$ the overall proportion of cases who receive treatment; and $E(NB(r_H)) =$ expected net benefit.

Although we emphasize $HR_D(r_H)$, $HR_{\bar{D}}(r_H)$ and $sNB(r_H)$ as the key measures of prediction performance for settings where a risk threshold reduces the model to a binary decision rule, we acknowledge that other measures of performance for binary classifiers could also be reported. Classic measures include: the misclassification rate, $(1 - HR_D(r_H))\rho + HR_{\bar{D}}(r_H)(1-\rho)$; Youden's index, $HR_D(r_H) - HR_{\bar{D}}(r_H)$; and the Brier score, $E(D - \text{risk}(X))^2$. These measures, like $sNB(r_H)$, are functions of $HR_D(r_H)$ and $HR_{\bar{D}}(r_H)$ but seem to lack its compelling interpretation and practical relevance. Therefore we do not endorse them for evaluating risk prediction models.

3.5 Multiple Risk Categories

For prevention of cardiovascular events two possible treatment strategies are recommended. Longterm treatment with cholesterol lowering drugs is recommended for high risk subjects while an inexpensive, non-toxic intervention, namely healthy lifestyle changes, is recommended for subjects at moderately elevated risk. Three risk categories are therefore of interest in this clinical setting: low risk (risk $\leq 5\%$), moderate risk $(5 - 20\%)$ and high risk $(\geq 20\%)$. The parameters HR_D and $HR_{\bar{D}}$ are easily generalized to the setting of multiple risk categories. One reports the pro-

portions of cases in each category and the proportions of controls in each category. These fractions can be read off of the distribution function displays in Figure 7. Values are shown in Table 2.

Table 2 Proportions of subjects in each of 3 risk categories. Risk refers to 10-year probability of a cardiovascular event.

Risk Category	Cases	Controls
low ($< 10\%$)	.112	.657
medium ($10-20\%$)	.236	.253
high ($> 20\%$)	.652	.089

The standardized net benefit function can also be generalized to accommodate more than two categories of risk, but it requires specifying some relative costs and benefits explicitly in addition to the cost-benefit ratios that are implied by the risk threshold values that define the risk categories.

As an example, suppose there are 3 categories of risk with treatment recommendations being none for the low risk category $(\text{risk} \leq r_{\text{low}})$, intermediate treatment for the medium risk category $(r_{\text{low}} \leq \text{risk} < r_{\text{high}})$ and intense treatment for the high risk category $(\text{risk} > r_{\text{high}})$. Suppose that in the absence of a predictive model all subjects are recommended for intermediate treatment. We use the following notation for costs and benefits: B_{high}= net benefit of intense treatment to a subject who would be a case; C_{high}= net cost of intense treatment to a would-be control; B_{low}= net benefit of no treatment to a would be control and C_{low} = net cost of no treatment to a would-be case. All of these quantities are relative to the intermediate treatment and as always, cases (controls) are those who would (would not) get the bad outcome in the absence of treatment. The expected net benefit in the population associated with use of the risk model is:

$$\rho\{-C_{\text{low}}\text{LR}_D + B_{\text{high}}\text{HR}_D\} + (1-\rho)\{B_{\text{low}}\text{LR}_{\bar{D}} - C_{\text{high}}\text{HR}_{\bar{D}}\}$$

where LR and HR are probabilities of being in the low and high risk categories and the subscripts D and $\bar{D}$ indicate cases and controls as usual. Let r_L and r_H denote the risk thresholds that separate low from medium risk and medium from high risk, respectively. The arguments of Result 1 implies that $(r_L/1-r_L) = B_{\text{low}}/C_{\text{low}}$ and $(r_H/1-r_H) = C_{\text{high}}/B_{\text{high}}$. Let $\lambda = B_{\text{high}}/C_{\text{low}}$ be the ratio of the net benefit of intense treatment to the net cost of no treatment for a subject that would be a case. The population net benefit of using the model with these risk categories can then be written as

$$B_{\text{high}}\left\{\rho\{-\frac{1}{\lambda}\text{LR}_D + \text{HR}_D\} + (1-\rho)\{\frac{r_L}{(1-r_L)\lambda}\text{LR}_{\bar{D}} - \frac{r_H}{1-r_H}\text{HR}_{\bar{D}}\right\}.$$

This can be standardized by the maximum possible benefit that is achieved with a perfect prediction model, $B_{\text{high}}\{\rho + (1-\rho)\frac{r_L}{(1-r_L)\lambda}\}$, yielding the standardized net

benefit function

$$sNB(r_M, r_H) = \frac{\rho\{HR_D - LR_D/\lambda\} + (1-\rho)\{\frac{r_L}{(1-r_L)\lambda}LR_{\bar{D}} - \frac{r_H}{(1-r_H)}HR_{\bar{D}}\}}{\rho + (1-\rho)\frac{r_L}{(1-r_L)\lambda}}$$

This expression is a function of the risk thresholds, prevalence and case and control risk distributions and it requires specifying another parameter, namely λ. If we assume the net benefit of cholesterol lowering treatment relative to healthy lifestyle is 10 times as large as the net cost of no treatment relative to healthy lifestyle intervention to a would-be case, then the standardized net benefit associated with the fitted model risk(X) is

$$\frac{0.1017\{0.652 - 0.112/10\} - 0.8983\{\frac{0.05}{0.95}\frac{0.657}{10} - \frac{0.20}{0.80} \times 0.089\}}{0.1017 + 0.8983 \times \frac{0.05}{0.95 \times 10}}$$
$$= 77.1\%$$

This example demonstrates that calculation of net benefit associated with a risk model becomes quite complicated with 3 treatment categories compared with the calculation when only two treatment categories exist.

3.6 Implicit Use of Risk thresholds

In some circumstances a risk threshold for treatment that maximizes expected benefit cannot be adopted. Policy makers may require that alternative criteria are met. For example, Pfeiffer and Gail [23] consider using a risk threshold that determines a proportion v of the population is recommended for treatment: Allocation of financial resources might determine such a policy. The risk threshold is written as

$$r_H(v) \ : \ v = P(\text{risk}(X) > r_H(v)).$$

Having fixed v, they propose the proportion of cases that meet the treatment threshold as a measure of model performance:

$$PCF(v) = P(\text{risk}(X) > r_H(v) | D = 1),$$

with larger values indicating better performance. Observe that in our previous notation we can write

$$PCF(v) = HR_D(r_H(v)).$$

Another policy based criterion might require that a fixed proportion of the cases are recommended for treatment. In this case the treatment threshold is

$$r_H(w) \ : \ w = P(\text{risk}(X) > r_H(w) | D = 1)$$

and the prediction model performance measure proposed by Pfeiffer and Gail is the corresponding proportion of the population needed to follow, i.e., testing positive,

$$\text{PNF}(w) = P(\text{risk}(X) > r_H(w)).$$

Smaller values of $PNF(w)$ are more desirable.

These measures are closely related to the ROC curve that plots the true positive rate $\text{ROC}(f) = P(\text{risk}(X) > r_H | D = 1)$ versus the false positive rate $f = P(\text{risk}(X) > r_H | D = 0)$ for all possible thresholds $r_H \in (0,1)$. In fact a little algebra shows that

$$\text{PNF}(w) = \rho w + (1-\rho) ROC^{-1}(w)$$

and

$$\text{PCF}(v) = \text{ROC}(f(v))$$

where $f(v)$ is found by solving $v = \rho ROC(f) + (1-\rho) f$.

One can also directly use ROC curve points to characterize performance and it follows from arguments above that this approach is essentially equivalent to Pfeiffer and Gail's approach. In ROC analysis one fixes the proportion of cases deemed at high risk, derives the corresponding threshold $r_H(w)$ defined above, and evaluates the corresponding proportion of controls classified as high risk,

$$\text{ROC}^{-1}(w) = P(\text{risk}(X) > r_H(w) | D = 0).$$

Alternatively, one can fix the proportion of controls classified as high risk at f, derive the corresponding threshold

$$r_H(f) \ : \ f = P(\text{risk}(X) > r_H(f) | D = 0)$$

and use as the performance measure the corresponding proportion of cases classified as high risk

$$\text{ROC}(f) = P(\text{risk}(X) > r_H(f) | D = 1).$$

Example

Using our dataset, suppose we require that $w = 70\%$ of cases go forward for treatment. We calculate that the corresponding risk threshold will be $r_H(w) = 0.167$ and that 11.5% of controls will exceed this threshold with use of the model. See Figure 8. Since the prevalence is 10.17%, the overall proportion of the population that will undergo treatment is 17.5%. Using our notation:

$$w = 70\%, \quad r_H(w) = 0.167, \quad \text{ROC}^{-1}(w) = .115, \quad \text{PNF}(w) = .175$$

3.7 Measures Independent of Risk Thresholds

When a risk threshold for decision making is not forthcoming, a descriptive summary of the risk distributions in cases and controls may be of interest. In particular,

one can describe the separation between the case and control distributions of risk. For example, we could report: the average risk in cases, 0.391 in our example; the average risk in controls, 0.069 in our example, and the difference that we write as MRD, the *mean risk difference*:

$$\text{MRD} = E(\text{risk}(X)|D=1) - E(\text{risk}(X)|D=0)$$

which is 0.322 in our example.

The MRD statistic is also called Yates' slope. It is closely related to the integrated discrimination improvement (IDI) statistic proposed by Pencina et al. [16] for comparing nested risk prediction models. Specifically, if we consider the baseline model as the null model without predictors, so all subjects have estimated risk equal to ρ, then the IDI for comparing the model, risk(X), with the null model is the MRD. It has also been shown [19] that the MRD can be interpreted as the proportion of explained variation or coefficient of determination, $R^2 = \text{var}\{E(D|X)\}/\text{var}(D)$.

Another way to summarize the distance between the case and control risk distributions is with the *above average risk difference*, AARD:

$$\text{AARD} \equiv P(\text{risk}(X) > \rho|D=1) - P(\text{risk}(X) > \rho|D=0).$$

Noting that the average risk in the population is ρ, this measure compares the proportion of cases with risks exceeding ρ, $\text{HR}_D(\rho) = 0.797$ in our example, with the corresponding proportion of controls, $\text{HR}_{\bar{D}}(\rho) = 0.198$, in our example, and calculates the difference, AARD$= 0.797 - 0.198 = 0.599$.

The AARD has several additional noteworthy interpretations. First, Youden's index for a dichotomous diagnostic test is defined as the true positive rate minus the false positive rate. We see that AARD is Youden's index for the rule that classifies subjects as positive when risk$(X) > \rho$. Second, we see that AARD$= s\text{NB}(\rho)$, the standardized net benefit defined earlier, for the decision rule that uses ρ as the high risk threshold. Third, the AARD is closely related to the net reclassification index (NRI) that will be defined in Section 4.3.2. The NRI is currently a very popular measure for comparing nested models. It can be shown that for comparing the model with predictors X, *risk*(X), to the null model that assigns all subjects a risk of ρ, AARD$=$ NRI(>0) and AARD $=$ NRI(ρ) where NRI(>0) is known as the continuous NRI and NRI(ρ) is the categorical NRI with two risk categories defined by the risk threshold ρ.

Interestingly, it can be shown that the difference, $\text{HR}_D(r) - \text{HR}_{\bar{D}}(r)$, is maximized at $r = \rho$ (see proof of Theorem A.1 [22]). Therefore, the AARD can also be interpreted as a Kolmogorov-Smirnov distance between the case and control risk distributions. Finally, Huang and Pepe [10] and Gu and Pepe [7] showed that the standardized total gain statistic proposed by Bura and Gastwirth [2] as a measure of predictive capacity of a model risk(X), $sTG = \int |\text{risk}(X) - \rho| dF(X)/2\rho(1-\rho)$ is equal to the AARD.

Another nonparametric measure of distance between the case and control risk distributions is the area under the ROC curve (AUC):

$$\text{AUC} = P(\text{risk}(X_i) \geq \text{risk}(X_j) | D_i = 1, D_j = 0)$$

where X_i and X_j are predictors for randomly drawn independent subjects from the case and control distributions, respectively. This is also known as the Mann-Whitney U-statistic and is calculated as AUC $= 0.884$ for our data. The AUC has a long history of use in evaluating diagnostic tests and other classifiers including risk models. It is still the most popular metric in use. However, its use in evaluating risk models has been criticized [3, 21]. One of the criticisms leveled against the AUC is that the measure has no practical relevance. Certainly this is true. If subjects were presented in case-control pairs to the physician for deciphering which one is the case, the measure $P(\text{risk}(X_i) > \text{risk}(X_j) | D_i = 1, D_j = 0)$ would be of interest. But this is not the usual clinical task. Another criticism of the AUC is that the measure may be dominated by differences in risk distributions that are clinically irrelevant. In particular in a low prevalence or incidence setting, the AUC is dominated by the low end of the risk range where most of the population's risks lie. Yet small differences in the distributions over that range are of no clinical relevance. Consequently the AUC may not be sensitive to differences in risk distributions over more clinically relevant ranges.

These two criticisms, practical irrelevance and insensitivity to clinically important differences in distributions, however, apply not only to the AUC, but also apply to other measures of distance between case and control risk distributions such as the MRD and AARD. We see no particular advantage to MRD or AARD or the related reclassification measures (IDI and NRI) that will be described later. We caution against using any of these measures as a sole focus for evaluating and comparing models. Rather they may be more suitable to assessing if a model is at one extreme or the other in terms of prediction performance. As such, their use may be justified in algorithms to sift through many models in order to select some that have predictive performance worthy of more thorough evaluation, i.e., discovery research.

3.8 Recommendations

For evaluating a single risk prediction model we have the following recommendations:

(i) Assess the model for its calibration in the population of interest and if necessary recalibrate the model.
(ii) Plot the case and control risk distributions, possibly providing a summary index of distance between the distributions as a descriptive adjunct.
(iii) In collaboration with clinical and health policy colleagues, elicit a risk threshold or several thresholds that could be used for making treatment decisions. Evaluate the model in terms of corresponding case and control classification rates and in terms of net benefit of using the model.

4 Comparing Two Risk Models

In this section we consider the comparison of two risk models for their prediction performance. In a nutshell we recommend that (i) each model be evaluated for its calibration; (ii) plots of risk distributions and net benefit be prepared for each model; (iii) having chosen a clinically relevant risk threshold for treatment (or several) compare the corresponding case and control categorized risk distributions for the two models and (iv) compare the corresponding net benefits associated with the two models. An illustrative example is provided in Section 4.1. This approach applies when the two models are nested (where one model has predictors X and the other has additional predictors Y) and when the models are not nested. Several methods have been proposed in recent years for the specific problem of comparing nested models, notably risk reclassification methods. We describe those risk reclassification methods in section 4.3. Finally, we provide a result concerning the equivalence of null hypotheses about improvement in prediction performance gained by adding Y to a set of baseline predictors X and the classic null hypothesis about Y as a risk factor after controlling for X.

4.1 Example

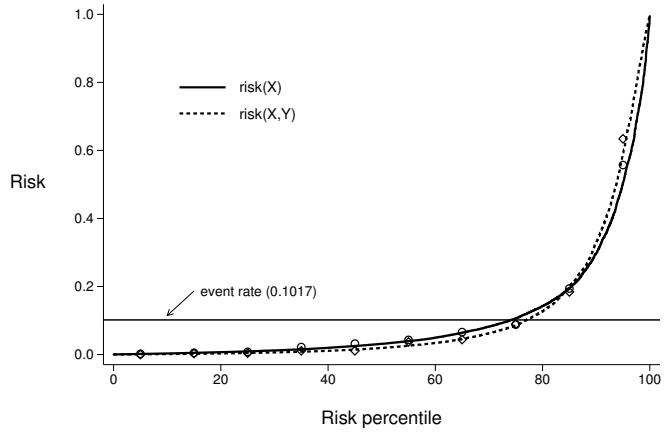

Fig. 9 Event rates for subjects in each decile of modeled risk align well with the risk values for subjects in each decile that are shown with predictiveness curves.

We compare the logistic models, risk(X) and risk(X,Y), fit to our illustrative data, as described in Section 1.3. Predictiveness curves for the fitted models superimposed with observed event rates in each decile of fitted risk are shown in Figure

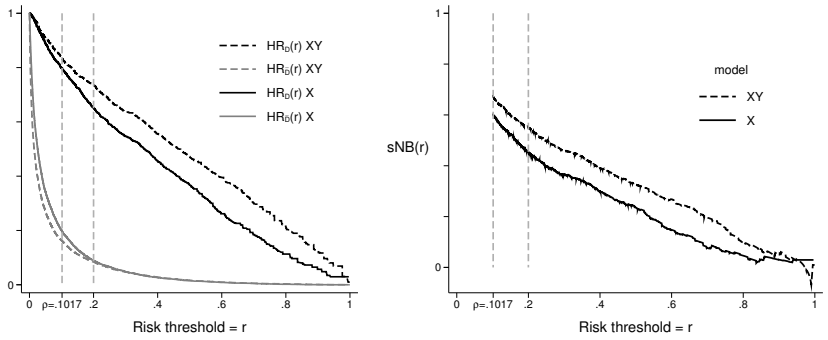

Fig. 10 Plots showing high risk classification for cases (D) and controls $(\bar{D})$ under the baseline model (X) and the expanded model (X,Y). A comparison of standardized net benefit is also shown.

9. Both models are well calibrated. The risk distributions in cases and controls are shown in the left panel of Figure 10 using 1−cumulative distribution functions. That is, for each risk threshold r we show the proportions of cases and controls above that threshold. We see that at all thresholds more cases and fewer controls have risks above the threshold when using the model risk(X,Y) than with model risk(X). Consequently the case and control risk distributions are more separated by the model risk(X,Y) than by the model risk(X). The measures of separation, AUC, MRD, and AARD, are presented in Table 2, and confirm that greater separation is achieved with the model including Y as a predictor.

Table 3

	risk(X)	risk(X,Y)	Difference
AUC	0.884	0.920	0.036
MRD	0.322	0.416	0.094[†]
AARD	0.599	0.673	0.074
$\text{HR}_D(0.20)$	0.652	0.735	0.084
$\text{HR}_{\bar{D}}(0.20)$	0.089	0.084	−0.005
$s\text{NB}(0.20)$	0.455	0.550	0.095
PNF	0.174	0.134	−0.040

[†]difference in MRDs is called the IDI [16]

The right hand side of Figure 10 shows that, regardless of which risk threshold is employed for recommending treatment, the expected benefit is larger when Y is included in the risk model. This is not surprising since the change in the net benefit is a weighted sum of the increase in the value $\text{HR}_D(r)$ plus the decrease in the value of $\text{HR}_{\bar{D}}(r)$, both of which are positive.

$$sNB^{(X,Y)}(r) - sNB^{X}(r) = \{HR_D^{(X,Y)}(r) - HR_D^{X}(r)\} + \frac{(1-\rho)}{\rho}\frac{r}{(1-r)}\{HR_{\bar{D}}^{X}(r) - HR_{\bar{D}}^{(X,Y)}(r)\}$$

Suppose that subjects with risks above 20% are recommended for treatment because the net benefit of treatment to a (would-be) case is considered 4 times the net cost of treatment to a (would be) control. The model that includes Y recommends 8.4% more cases and .5% fewer controls for treatment. Relative to a perfect prediction model, the model with X only achieves 45.5% of maximum benefit while that including Y as well as X achieves 55.0% of maximum benefit. That is, the standardized net benefit or discounted true positive rate is improved by 9.5%.

We also consider performance when certain criteria are set by policy makers. Suppose that a treatment risk threshold will be employed that will guarantee $w = 70\%$ of cases are treated. Using the model with X only will require treating $\text{PNF}(w) = 17.4\%$ of the population (and correspondingly $\text{ROC}^{-1}(w) = 0.114$ of controls) since the largest risk threshold exceeded by 70% of cases is $r_H^X(w) = 0.167$. On the other hand, a higher risk threshold $r_H^{(X,Y)}(w) = 0.231$, can be employed with the model $\text{risk}(X,Y)$ as the case risk distribution is higher. Consequently only $\text{PNF}(w) = 13.4\%$ of the population (and only $\text{ROC}^{-1}(w) = 0.07$ of controls) will be treated if $\text{risk}(X,Y)$ is used for assigning treatment.

In this example, better performance is achieved by including Y in the risk model regardless of how performance is measured.

4.2 Risk Reclassification within Subpopulations Defined by $\text{risk}(X)$

In addition to determining whether or not use of the model $\text{risk}(X,Y)$ is better than use of $\text{risk}(X)$ in the population as a whole, one might ask if the additional information provided by knowledge of Y is useful in subsets of the population. Specifically, one might consider subsets of the population determined to be at low (or high or intermediate) risk according to $\text{risk}(X)$ and evaluate use of the enhanced model $\text{risk}(X,Y)$ in that subpopulation. This is one motivation for constructing the risk reclassification table illustrated in Table 4.

Table 4 shows risk reclassification tables for cases in the 'Events' panel of the table and for controls in the 'Nonevents' panel of the table. In cardiovascular disease the tables are often constructed using 3 categories corresponding to the 3 treatment recommendations. Here we focus on the most important reclassifications to or from the high risk category that associated with cholesterol lowering treatment, and ignore reclassification between the medium and low risk categories that are less consequential. Table 5 summarizes the performance of the enhanced model $\text{risk}(X,Y)$ within each of the subpopulations determined to be at low ($< 5\%$), medium($5-20\%$) and high($> 20\%$) risk according to the baseline model, $\text{risk}(X)$.

In the low risk population, where the event rate $\rho = 1.89\%$, only 3.5% of those cases are reclassified by Y to the high risk category and almost no controls 0.4% are reclassified. The maximum possible benefit is to reclassify all cases (18.9 per 1000

Table 4 Event and Nonevent Risk Reclassification Tables

Events

	risk(X,Y)			
$r(X)$	$< 5\%$	$5-20\%$	$\geq 20\%$	Total
$\leq 5\%$	72	38	4	114 (11.2%)
$5-20\%$	21	105	114	240 (23.6%)
$\geq 20\%$	0	33	630	663 (65.2%)
Total	93	176	748	1017
	(9.1%)	(17.3%)	(73.5%)	

Nonevents

	risk(X,Y)			
$r(X)$	$< 5\%$	$5-20\%$	$\geq 20\%$	Total
$\leq 5\%$	5486	399	21	5906 (65.8%)
$5-20\%$	1015	990	272	2277 (25.3%)
$\geq 20\%$	40	296	464	800 (8.9%)
Total	6541	1685	757	8983
	(72.8%)	(18.8%)	(8.4%)	

Table 5 Performance of $r(X,Y)$ within strata defined by $r(X)$

Population	$\rho(X)$ Event Rate	cases $\text{HR}_D(0.20)$	controls $\text{HR}_{\bar{D}}(0.20)$	% of max benefit $s\text{NB}(0.20)$
low risk $r(X)$	1.89%	0.035	0.004	-1.7%
med risk $r(X)$	9.54%	0.475	0.119	19.3%
high risk $r(X)$	45.32%	0.950	0.580	25.4%

subjects) and no controls but the model reclassifies only 3.5% of the cases (0.66 cases per 1000 subjects). Consequently, the standardized net benefit of using the model is negligible. (The negative value, -1.7%, must be due to sampling variability as the net benefit cannot be negative if the model is correct.) It appears that use of Y in the low risk population is not beneficial.

In the medium risk population, 47.5% of the cases are favorably reclassified to the high risk group while only 11.9% of the controls are. The maximum possible benefit is that achieved if all 95.4/1000 cases were moved to the high risk category without moving any of the 904.6/1000 of the controls. With use of Y it appears that $s\text{NB}(0.2) = 19.3\%$ of this maximum possible benefit is achieved. In other words the benefit reached by measuring Y in the medium risk population is the same as that achieved by a model that moved about 19/1000 cases to treatment with statins without moving any controls into that high risk category.

In the high risk group, benefit can be obtained by moving controls down to a lower risk category but at the possible cost of moving cases down. The reclassification tables show however that with use of Y only 5% of the cases are moved down while 42% of controls move down. The maximum benefit in a population of 1000

subjects would be that achieved by moving none of the 453 cases but all of the 547 controls down. With use of Y, we are able to move 230 controls off treatment at the expense of moving 23 cases off treatment. Is this a net benefit? Arguments similar to those in Section 3.4 can be used to show the standardized net benefit of a rule that denies treatment when the risk $< r_H$ in a population with prevalence ρ is

$$sNB(r_H) = \{1 - HR_{\bar{D}}\} - \frac{\rho}{1-\rho}\frac{1-r_H}{r_H}\{1 - HR_D(r_H)\}.$$

We calculate that $sNB(r_H) = 25.4\%$ of maximum benefit is achieved with use of Y in the population deemed at high risk according to risk(X). This is equivalent to moving $.254 \times 547 = 140$ controls down without moving any cases in a set of 1000 subjects. This benefit seems substantial.

We find risk reclassification tables useful for evaluating an expanded model risk(X,Y) within subpopulations defined by risk levels calculated according to the baseline model risk(X), as illustrated above. One might also choose to plot risk distributions and calculate other statistics for evaluating a risk model within each subpopulation using techniques described in Section 3. However, analyses of risk reclassification tables have been proposed for purposes beyond evaluation use of risk(X,Y) within subpopulations defined by risk(X). In particular, analyses that purport to compare the two models in the entire population have been proposed. In the next section we describe the two main approaches that have been proposed and point out some fundamental weaknesses in them.

4.3 Risk Reclassification to Compare Two Models

4.3.1 The Cook and Ridker Analysis Method

Cook and Ridker [4] combine the event and nonevent reclassification tables in a single table with the elements shown in Table 6.

Table 6 Risk reclassification tables showing numbers of subjects and event rates (%) in each cell.

	risk(X,Y)			
$r(X)$	$\leq 5\%$	$5-20\%$	$> 20\%$	Total
$\leq 5\%$	5558	437	25	6020
	1.30	8.71	16.00	1.89
$5-20\%$	1036	1095	386	2517
	2.03	9.59	29.53	9.54
$> 20\%$	40	329	1094	1463
	0.00	10.03	57.59	45.32
Total	6634	1861	1505	10,000
	1.40	9.46	49.70	10.17

They calculate the following entities that are explained below:

(i) the overall reclassification rate = RC
(ii) the percent correctly reclassified = RC-correct
(iii) the baseline model reclassification calibration statistic: RCC^X and its p-value
(iv) the expanded model reclassification calibration statistic: $\text{RCC}^{(X,Y)}$ and its p-value

The RC is the proportion of subjects in the off-diagonal cells of the table, 22.5% in our example. This is simply a descriptive statistic that is not used to compare models although a small value indicates that treatment recommendations will be changed for few subjects by including Y as a risk predictor in addition to X.

The RC-correct is defined to be the proportion of subjects in off-diagonal cells where the observed event rate is within the risk(X,Y) category label and not within the risk(X) category label. In our data RC-correct= 100%. If the RC-correct value is large, this is taken as evidence that prediction performance with the model that includes Y is better than prediction performance with the reduced model risk(X). However, it has been shown that by definition, when models are well calibrated in the standard sense, RC-correct$\approx$ 100% in large samples. This follows because for observations in an off-diagonal cell where risk$(X) \in A$ and risk$(X,Y) \in B$ the expected event rate

$$\begin{aligned}
&P(D=1|\text{risk}(X) \in A, \text{risk}(X,Y) \in B)\\
=&E(D=1|\text{risk}(X) \in A, \text{risk}(X,Y) \in B)\\
=&E(E(D=1|X,Y)|\text{risk}(X) \in A, \text{risk}(X,Y) \in B)\\
=&E(\text{risk}(X,Y)|\text{risk}(X) \in A, \text{risk}(X,Y) \in B).
\end{aligned}$$

This average risk is in the interval B because for all subjects in the off-diagonal cell, risk$(X,Y) \in B$. Moreover, the average risk is not in the interval A because, being an off-diagonal cell, the interval A lies outside of interval B. It follows that for each off-diagonal cell, in large samples the event rate is in the interval defined by the enhanced model. That is, we expect that RC-correct = 100%. Any deviation from 100% that occurs must be due to sampling variability. There is no point in calculating a statistic that is $\approx$ 100% by definition and it cannot be used to compare the performances of the two models.

The reclassification calibration statistics are:

$$\text{RCC}^X = \sum_{k=1}^{K} \frac{(\widehat{p}_k - ave(\text{risk}(X))_k)^2}{ave(\text{risk}(X))_k(1-ave(\text{risk}(X))_k)/n_k}$$

and

$$\text{RCC}^{(X,Y)} = \sum_{k=1}^{K} \frac{(\widehat{p}_k - ave(\text{risk}(X,Y))_k)^2}{ave(\text{risk}(X,Y))_k(1-ave(\text{risk}(X))_k)/n_k)}$$

where the summation is over the K interior cells of the table with sample sizes $n_k \geq 20$, $ave(\text{risk}(\quad))_k$ is the average of modeled risks for subjects in cell k and $\widehat{p}_k$ is the observed event rate in that cell. Cook and Ridker refer the reclassification

calibration statistics to chi-squared distributions with $K-2$ degrees of freedom for calculating p-values.

The arguments above show that in large samples the observed event rates in off-diagonal cells converge to $ave(\text{risk}(X,Y))_k$ but not to $ave(\text{risk}(X))_k$. Therefore the implicit null hypothesis for the statistic $\text{RCC}^{(X,Y)}$ is satisfied. That is, the enhanced model will not be rejected at a rate above the nominal significance level in large samples. On the other hand the implicit null hypothesis for the baseline model will be rejected in large samples assuming that Y is a risk factor that moves even a small proportion of subjects to off-diagonal cells. In other words, if there are subjects in off-diagonal cells, there is no point in performing the RCC (X,Y) statistical test since the result is predetermined in large samples. If there are no subjects in off-diagonal cells the setting is degenerate and there is obviously no point in performing the test either. We implemented the RCC tests on our illustrative data and in agreement with our arguments above, the baseline model was rejected, $p < .001$, while the enhanced model was not, $p = 0.29$.

In conclusion, we regard the risk-reclassification table proposed by Cook and Ridker [4] as an interesting descriptive device. However, the analysis strategy based on the table is not useful for comparing risk models.

4.3.2 The Net Reclassification Index

Pencina et al. [16, 17] introduced the Net Reclassification Index (NRI) as a measure to compare the prediction performance of nested risk models. The statistic is calculated as the sum of two components, the NRI-event and NRI-nonevent. When risk categories exist, the data are summarized in reclassification tables of the form shown in Table 4 and the corresponding NRI statistics are

$$\begin{aligned}
\text{cat-NRI-event} &= P[\text{risk}_c(X,Y) > \text{risk}_c(X) | D=1] - P[\text{risk}_c(X,Y) < \text{risk}_c(X) | D=1] \\
\text{cat-NRI-nonevent} &= P[\text{risk}_c(X,Y) < \text{risk}_c(X) | D=0] - P[\text{risk}_c(X,Y) > \text{risk}_c(X) | D=0] \\
\text{cat-NRI} &= \text{cat-NRI-event} + \text{cat-NRI-nonevent}
\end{aligned}$$

where $\text{risk}_c(X,Y)$ is the risk category in which the subject's value for $\text{risk}(X,Y)$ falls and $\text{risk}_c(X)$ is that in which his value of $\text{risk}(X)$ falls. In words, cat-NRI-event is the proportion of cases above the diagonal of the event reclassification table minus the proportion below the diagonal. The counterpart, cat-NRI-nonevent is the proportion of controls below the diagonal minus the proportion of controls above the diagonal. Their sum, cat-NRI, takes values between 0 and 2. For our data with 3 risk-categories, cat-NRI-event$= 0.100$, cat-NRI-nonevent=0.073 and cat-NRI=0.174.

The cat-NRI provides a descriptive summary of the reclassification tables. However, it does not seem particularly well suited to the purpose of comparing the prediction performance of the models $\text{risk}(X)$ and $\text{risk}(X,Y)$. In general it does not represent a comparison of the performance of the model $\text{risk}(X,Y)$ with the performance of the model $\text{risk}(X)$. To see this, consider that the performance of the model $\text{risk}(X)$ must be derived from the case and control distributions of $\text{risk}(X)$. These

distributions are contained in the *vertical margins* of the event and nonevent reclassification tables, respectively. Similarly the performance of the model risk(X,Y) must be derived from the *horizontal margins* of the reclassification tables. However, the cat-NRI statistic is a function that depends on the interior cells of the tables, not just their margins. This is illustrated in Table 7 that shows an example where the margins of the reclassification tables are equal but cat-NRI> 0.

Table 7 Example where cat-NRI > 0 but there is no performance improvement.

Events		$r(X,Y)$				Non-Events		$r(X,Y)$			
		low	med	high				low	med	high	
	low	10	10	0	20		low	500	100	0	600
$r(X)$	med	5	20	10	35		med	100	200	0	300
	high	5	5	35	45		high	0	0	100	100
		20	35	45	100			600	300	100	900

That is, in Table 7, prediction performance for the model risk(X,Y) is the same as that for the model risk(X) since the vertical and horizontal margins are equal. However, the cat-NRI statistic $= (20-15)/100 = 0.05 > 0$, indicating that performance has improved.

Although the cat-NRI statistic was originally proposed for use with 3 or more risk categories, it is interesting to consider it in the simpler and more common setting where only two risk categories exist that are separated at a treatment risk threshold r_H. Using the notation in Table 8, it is easy to see that with two categories

$$\begin{aligned}
\text{cat-NRI-event} &= b-c = b+d-(c+d) = \text{HR}_D^{(X,Y)}(r_H) - \text{HR}_D^X(r_H) \\
\text{cat-NRI-nonevent} &= f-g = f+h-(g+h) = \text{HR}_{\bar{D}}^X(r_H) - \text{HR}_{\bar{D}}^{(X,Y)}(r_H) \\
\text{cat-NRI} &= \{\text{HR}_D^{(X,Y)}(r_H) - \text{HR}_D^X(r_H)\} - \{\text{HR}_{\bar{D}}^{(X,Y)}(r_H) - \text{HR}_{\bar{D}}^X(r_H)\}.
\end{aligned}$$

That is, cat-NRI-event is the increase in the proportion of cases classified as high risk and cat-NRI-nonevent is the decrease in the proportion of controls classified as high risk. The simple summation that is cat-NRI however, does not weight the relative contributions appropriately unless $r_H = \rho$. To see this, recall that the change in the standardized net benefit does weight the contributions appropriately and is written as

$$s\text{NB}^{(X,Y)}(r_H) - s\text{NB}^X(r_H) = \{\text{HR}_D^{(X,Y)}(r_H) - \text{HR}_D^X(r_H)\} - \frac{1-\rho}{\rho}\frac{r_H}{(1-r_H)}\{\text{HR}_{\bar{D}}^{(X,Y)}(r_H) - \text{HR}_{\bar{D}}^X(r_H)\}.$$

Only when $r_H = \rho$ does cat-NRI correspond to the appropriately weighted combination, the change in $s\text{NB}(r_H)$. Interestingly a weighted version of the two-category NRI statistic has recently been proposed to correspond with the form of change in $s\text{NB}(r_H)$, by weighting cat-NRI-nonevent by $\frac{(1-\rho)}{\rho}\frac{r_H}{(1-r_H)}$ [17]. However, weighted versions of the cat-NRI statistic have not been proposed to correspond with change in standardized net benefit when more than two categories are involved.

Table 8 Example where cat-NRI > 0 but there is no performance improvement.

Events			Non-Events		
	$r(X,Y)$			$r(X,Y)$	
	low	high		low	high
low	a	b	low	e	f
high	c	d	high	g	h

A continuous version of the NRI statistic has been proposed for use when no clinically relevant risk categories exist:

$$\begin{aligned}
\text{cont-NRI-event} &= P(\text{risk}(X,Y) > \text{risk}(X) | D=1) - P(\text{risk}(X,Y) < \text{risk}(X) | D=1) \\
&= 2P(\text{risk}(X,Y) > \text{risk}(X) | D=1) - 1 \\
\text{cont-NRI-nonevent} &= P(\text{risk}(X,Y) < \text{risk}(X) | D=0) - P(\text{risk}(X,Y) > \text{risk}(X) | D=0) \\
&= 1 - 2\,P(\text{risk}(X,Y) > \text{risk}(X) | D=0) \\
\text{cont-NRI} &= 2\{P(\text{risk}(X,Y) > \text{risk}(X) | D=1) - P(\text{risk}(X,Y) > \text{risk}(X) | D=0)\}.
\end{aligned}$$

The statistic cont-NRI is also denoted by $NRI(>0)$. Again, since it is not a function of the marginal case and control risk distributions, cont-NRI does not seem well suited to quantifying the improvement in prediction performance of the model $\text{risk}(X,Y)$ versus that of the model $\text{risk}(X)$. It is not composed as a difference between an index of the performance of $\text{risk}(X,Y)$ and an index of the performance of $\text{risk}(X)$. But it is an interesting easily understood descriptive statistic about the joint distributions of $(\text{risk}(X), \text{risk}(X,Y))$ based on the comparison of $\text{risk}(X,Y)$ and $\text{risk}(X)$ within individuals. In our data we calculate that for 69.4% of cases their calculated risk increased with addition of Y and for 29.5% of controls their risks increased with addition of Y. Consequently cont-NRI-event$= 0.388$, cont-NRI-nonevent$= 0.411$ and cont-NRI$= 0.799$.

4.4 Hypothesis Testing for Nested Models

When evaluating if a model that includes marker Y in addition to X improves performance over the baseline model that includes X only, it is common practice to do several hypothesis tests. One will typically test the hypothesis $H_0^1 : \text{risk}(X,Y) = \text{risk}(X)$, using, for example, likelihood techniques based on regression models. If H_0^1 is rejected, one may test if the prediction performance of $\text{risk}(X,Y)$ is equal to that of $\text{risk}(X)$ using one or more statistics, such as the difference in the AUCs or the IDI statistic, which is the difference in MRDs, amongst others. The following result indicates that the null hypotheses concerning many measures of performance improvement are identical to $H_0^1 : \text{risk}(X,Y) = \text{risk}(X)$.

The practical implication of this result is that if H_0^1 is rejected, one can conclude that the other hypotheses listed in Result 2 are also rejected. Testing any one of hy-

potheses $H_0^2 - H_0^6$ is equivalent to testing H_0^1. We recommend using well standard methods from regression modeling to test the hypothesis formulated as H_0^1. Corresponding statistical techniques are well developed and they are often efficient. In contrast techniques based on estimates of the performance measures in $H_0^2 - H_0^6$ are likely to be less efficient and have in some cases been shown to have bizarre distributions under the null hypothesis of no change in performance [13]. This is an active area of research.

Result 2

The following conditions are equivalent

$$
\begin{aligned}
H_0^1 &: \text{risk}(X,Y) = \text{risk}(X) \\
H_0^2 &: \text{ROC}^{(X,Y)}(f) = \text{ROC}^X(f) \quad \forall \quad f \\
H_0^3 &: \text{AUC}^{(X,Y)} = \text{AUC}^X \\
H_0^4 &: \text{MRD}^{(X,Y)} = \text{MRD}^X \\
H_0^5 &: \text{AARD}^{(X,Y)} = \text{AARD}^X \\
H_0^6 &: \text{NRI}(>0) = 0
\end{aligned}
$$

■

For a proof of Result 2 and additional related results see Pepe [22].

5 Concluding Remarks

We now recap some of the main points made in this chapter. First, a necessary condition for a useful risk prediction model is that it be well-calibrated. Whereas the scale of a marker used for classifying individuals according to disease status is not in and of itself of interest, risks calculated using a prediction model are used to advise a patient and to make medical decisions. The scale of the risk model predictions is therefore a fundamental aspect of the model's utility. Good calibration is essential.

The distributions of risk predicted by the model, for cases and for controls, are the building blocks for evaluating model performance. Summary measures are functions of these distributions. A variety of summary measures have been described here which rely on specification of a high risk threshold (or multiple risk thresholds) for classifying subjects. Our preferred measures are the proportions of cases and controls classified as high risk (or classified into each risk category) and the standardized net benefit of using the model. Performance measures that do not rely on risk thresholds can be useful for screening many models in order to select a subset for further evaluation.

The choice of the risk threshold(s) should be based on an assessment of costs and benefits associated with a high risk (or each risk category) designation. These costs and benefits are also used in calculating the scaled net benefit of the model.

When comparing models, our recommendation is to compare measures of marginal performance. This is in contrast to basing comparisons on statistics that summarize the cross-classification of the two models. The cross-classification is useful for descriptive analysis but cannot be used as the basis for forming conclusions regarding the relative performance of the two models.

When testing the incremental value of a new marker added to a risk model, standard likelihood methods should be used. Tests based on contrasts of performance measures between the baseline and expanded risk models are testing the same null hypothesis. These tests have, in some cases, been shown to poorly control the type-I error rate and to be less powerful [13, 28, 22].

This chapter has focused on conceptual approaches to evaluation assuming a very large sample size. In practice, where sample sizes are finite, all measures of model performance should be accompanied by confidence intervals to characterize the level of uncertainty. This practice is much more informative than reporting p-values based on hypothesis tests; moreover in some instances as mentioned above, tests based on model performance measures have poor properties. Bootstrapping is a simple and flexible technique that can be used to construct confidence intervals. The bootstrapping should reflect the actual data analysis that was performed; if the risk model was fit using the data (versus using a separate dataset), the model should be re-fit and performance estimated in each bootstrap sample. Bootstrapping is flexible in that unique attributes of the original study design can be accommodated, such as repeated measures or case-control sampling.

When a risk model is fit and evaluated using the same data, the apparent performance will tend to be over-optimistic. This is particularly true if an intensive model-selection procedure was employed. Standard approaches to dealing with this problem include separating the data into "training" and "test" portions, and more efficient methods such as cross-validation [8, 9]. The disadvantage of the latter approach is the requirement for a prespecified and automated model selection procedure. If either approach is used, it should be reflected in the bootstrapping described above. Specifically, in each bootstrap sample the complete model selection procedure should be performed.

In many contexts there are covariates that need to be taken into account when predicting risk and evaluating risk model performance. We distinguish between covariates (Z) that predict risk of bad outcome, eg age, and covariates that modify the performance of a risk calculator ($risk(X)$), e.g., the laboratory in which the biomarker X is assayed. Of course some coviarates, such as disease comorbidities, may have both types of effects. For covariates that predict risk only, the approach is to simply include these as predictors in the risk model, ie to model $P(D=1|X,Z)=risk(X,Z)$. To do this, the covariates Z are treated as additional markers and the methods described in this chapter apply directly. On the other hand, covariates that modify the distribution of $risk(X)$ will have an impact on the performance of the model. Evaluating how the performance of $risk(X)$ varies with Z will generally require modeling X as a function of Z and we refer the reader to Huang [11] for details. Covariates that both predict risk and modify performance can be accommodated using the same

methods where $risk(X,Z)$ is the risk calculator and the joint distribution of (X,Z) is modeled as a function of Z.

Our discussion of the choice of risk threshold(s) assumed implicitly that the cost and benefit of being treated are constant across individuals, and in particular are independent of the marker X. Markers that predict the benefit or cost of treatment have greater potential net benefit [24, 29, 12]. However evaluating these markers requires data from a randomized trial where the marker is measured at baseline, in order to assess the cost and benefit of treatment as a function of X.

References

1. Baker, S.: Putting risk prediction in perspective: relative utility curves. Journal of the National Cancer Institute **101**(22), 1538–1542 (2009)
2. Bura, E., Gastwirth, J.: The binary regression quantile plot: assessing the importance of predictors in binary regression visually. Biometrical Journal **43**(1), 5–21 (2001)
3. Cook, N.: Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation **115**(7), 928–935 (2007)
4. Cook, N., Ridker, P.: Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Annals of internal medicine **150**(11), 795–802 (2009)
5. Gail, M., Costantino, J.: Validating and improving models for projecting the absolute risk of breast cancer. Journal of the National Cancer Institute **93**(5), 334–335 (2001)
6. Gail, M., Pfeiffer, R.: On criteria for evaluating models of absolute risk. Biostatistics **6**(2), 227–239 (2005)
7. Gu, W., Pepe, M.: Measures to summarize and compare the predictive capacity of markers. Int J Biostat **5**(1), Article 27 (2009). DOI 10.2202/1557-4679.1188
8. Harrell, F.: Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer Verlag (2001)
9. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Springer Verlag (2001)
10. Huang, Y., Pepe, M.: A parametric roc model-based approach for evaluating the predictiveness of continuous markers in case–control studies. Biometrics **65**(4), 1133–1144 (2009)
11. Huang, Y., Pepe, M.S.: Semiparametric methods for evaluating the covariate-specific predictiveness of continuous markers in matched case-control studies. Journal of the Royal Statistical Society, Series C (Applied Statistics) **59**(3), 437–456 (2010)
12. Janes, H., Pepe, M.S., Bossuyt, P.M., Barlow, W.E.: Measuring the performance of markers for guiding treatment decisions. Annals of Internal Medicine **154**, 253–259 (2011)
13. Kerr, K.F., McClelland, R.L., Brown, E.R., Lumley, T.: Evaluating the incremental value of new biomarkers with integrated discrimination improvement. American Journal of Epidemiology **174**(3), 364–374 (2011)
14. Lemeshow, S., Hosmer Jr, D.: A review of goodness of fit statistics for use in the development of logistic regression models. American Journal of Epidemiology **115**(1), 92–106 (1982)
15. Parikh, C.R., Devarajan, P., Zappitelli, M., Sint, K., Thiessen-Philbrook, H., Li, S., Kim, R.W., Koyner, J.L., Coca, S.G., Edelstein, C.L., Shlipak, M.G., Garg, A.X., Krawczeski, C.D., , T.R.I.B.E.A.K.I.C.: Postoperative biomarkers predict acute kidney injury and poor outcomes after pediatric cardiac surgery. J Am Soc Nephrol **22**(9), 1737–1747 (2011)
16. Pencina, M., D'Agostino, R., D'Agostino, R., Vasan, R.: Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in Medicine **27**(2), 157–172 (2008)

17. Pencina, M., D'Agostino Sr, R., Steyerberg, E.: Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Statistics in Medicine **30**(1), 11–21 (2011)
18. Pepe, M.: Problems with risk reclassification methods for evaluating prediction models. American Journal of Epidemiology **173**(11), 1327 (2011)
19. Pepe, M., Feng, Z., Gu, J.: 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 (DOI: 10.1002/sim. 2929). Statistics in Medicine **27**(2), 173–181 (2008)
20. Pepe, M., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I., Zheng, Y.: Integrating the predictiveness of a marker with its performance as a classifier. American Journal of Epidemiology **167**(3), 362 (2008)
21. Pepe, M., Janes, H.: Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. Journal of the National Cancer Institute **100**(14), 978–979 (2008)
22. Pepe PhD, M., Kerr, K., Longton, G., Wang, Z.: Testing for improvement in prediction model performance. UW Biostatistics Working Paper Series p. 379 (2011)
23. Pfeiffer, R., Gail, M.: Two criteria for evaluating risk prediction models. Biometrics **67**(3), 1057–1065 (2011)
24. Sargent, D.J., Conley, B.A., Allegra, C., Collette, L.: Clinical trial designs for predictive marker validation in cancer treatment trials. Journal of Clinical Oncology **23**(9), 2020–2027 (2005)
25. Seymour, C.W., Kahn, J.M., Cooke, C.R., Watkins, T.R., Heckbert, S.R., Rea, T.D.: Prediction of critical illness during out-of-hospital emergency care. JAMA **304**(7), 747–754 (2010)
26. Steyerberg, E.: Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (2009)
27. Vickers, A., Elkin, E.: Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making **26**(6), 565 (2006)
28. Vickers, A.J., Cronin, A.M., Begg, C.B.: One statistical test is sufficient for assessing new predictive markers. BMC Medical Research Methodology **11**, 13 (2011)
29. Vickers, A.J., Kattan, M.W., Daniel, S.: Method for evaluating prediction models that apply the results of randomized trials to individual patients. Trials **5**, 8–14 (2007)
30. Wilson, P., D'Agostino, R., Levy, D., Belanger, A., Silbershatz, H., Kannel, W.: Prediction of coronary heart disease using risk factor categories. Circulation **97**(18), 1837–1847 (1998)