Research Article

Statistics in Medicine

Received 19 December 2011, Accepted 11 December 2012 Published online 7 January 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5727

Testing for improvement in prediction model performance

Margaret Sullivan Pepe,[a*†] Kathleen F. Kerr,[b] Gary Longton[a] and Zheyu Wang[b]

Authors have proposed new methodology in recent years for evaluating the improvement in prediction performance gained by adding a new predictor, Y, to a risk model containing a set of baseline predictors, X, for a binary outcome D. We prove theoretically that null hypotheses concerning no improvement in performance are equivalent to the simple null hypothesis that Y is not a risk factor when controlling for X, $H_0 : P(D = 1|X, Y) = P(D = 1|X)$. Therefore, testing for improvement in prediction performance is redundant if Y has already been shown to be a risk factor. We also investigate properties of tests through simulation studies, focusing on the change in the area under the ROC curve (AUC). An unexpected finding is that standard testing procedures that do not adjust for variability in estimated regression coefficients are extremely conservative. This may explain why the AUC is widely considered insensitive to improvements in prediction performance and suggests that the problem of insensitivity has to do with use of invalid procedures for inference rather than with the measure itself. To avoid redundant testing and use of potentially problematic methods for inference, we recommend that hypothesis testing for no improvement be limited to evaluation of Y as a risk factor, for which methods are well developed and widely available. Analyses of measures of prediction performance should focus on estimation rather than on testing for no improvement in performance.

Keywords: biomarker; logistic regression; receiver operating characteristic curve; risk factors; risk reclassification

1. Introduction

Prediction modeling has long been a mainstay of statistical practice. The field has been re-energized recently because of the promise of highly predictive biomarkers identified through imaging and molecular biotechnologies. Accordingly, there has been renewed interest in methods for evaluating the performance of prediction models. In particular, statisticians have been examining methods for evaluating improvement in performance that is gained by adding a novel marker to a baseline set of predictors.

For example, novel markers for predicting risk of breast cancer beyond traditional factors in the Gail model [1,2] include breast density [3,4] and genetic polymorphisms [5–7]. For cardiovascular outcomes, authors have performed numerous studies in recent years to evaluate candidate markers for their capacities to improve upon factors in the standard Framingham risk score [8]. Tzoulaki *et al.* [9] recently performed a meta-analysis of 79 such published studies.

A typical approach to analysis is to first determine the statistical significance of an observed association between the novel marker, Y, and the outcome, D, controlling for the baseline predictors that we denote by X. The p-value is usually derived from regression modeling techniques. If the contribution of Y to the risk model is found to be statistically significant, then the second step in the typical approach is to test a null hypothesis about improvement in prediction performance for the model that includes

[a]*Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A.*
[b]*Department of Biostatistics, University of Washington, F-600 Health Sciences Building, Campus Mail Stop 357232, Seattle, WA 98195, U.S.A.*
**Correspondence to: Margaret Sullivan Pepe, Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500, Seattle, WA 98105, U.S.A.*
[†]*E-mail: mspepe@u.washington.edu*

Y in addition to X compared with the baseline model that includes only X. The most popular statistic for testing improvement in prediction performance is the change in the area under the receiver operating characteristic (ROC) curve [9]. Authors also use alternate measures, including risk redistribution metrics [10, 11] and risk reclassification metrics [12–16].

In this paper, we question the strategy of testing the null hypothesis about no improvement in prediction performance after testing the statistical significance of Y in the risk model. Our main theoretical result is that the null hypotheses are equivalent. This implies that if Y is shown to be a risk factor, the prediction performance of the model that includes Y cannot be the same as the performance of the baseline model, and there is no point to a second, redundant hypothesis test.

In Section 2, we prove our main result that the null hypothesis about Y as a risk factor can be expressed equivalently as a variety of null hypotheses about the improvement in performance of the expanded model compared with the baseline model. In Sections 3 and 4, we consider the choice of methodology for testing the common null hypothesis. We recommend use of standard statistics derived from regression modeling of the risk as a function of X and Y. This recommendation is based partly on the superior power achieved with likelihood-based tests and also on the new finding corroborated by other recent reports in the literature [14, 17, 18], that is, standard hypothesis testing methods to compare performances of nested models appear to be invalid. We emphasize that estimation of the increment in prediction performance is more important than testing the null hypothesis of no improvement. We discuss the results in Section 5 in the context of a real dataset concerning risk of renal artery stenosis as a function of baseline predictors and a biomarker, serum creatinine.

2. Equivalent null hypotheses

Suppose that the outcome is binary, $D = 1$ for cases or $D = 0$ for controls, which could represent occurrence of an event within a specified time period, say breast cancer within 5 years. Let $\text{risk}(X) = P(D = 1|X)$ and $\text{risk}(X, Y) = P(D = 1|X, Y)$ be the baseline and enhanced model risk functions, respectively, which we assume have absolutely continuous distributions. To evaluate the incremental value of Y for prediction over use of X alone, the first step is often to test the null hypothesis

$$H_0 \ : \ \text{risk}(X, Y) = \text{risk}(X). \tag{1}$$

We use subscripts (X, Y) and X to indicate entities relating to use of $\text{risk}(X, Y)$ and $\text{risk}(X)$, respectively. For example, $\text{ROC}_{(X,Y)}$ is the ROC curve for $\text{risk}(X, Y)$, whereas ROC_X is the ROC curve for $\text{risk}(X)$. The ROC curve for W is a plot of $P(W > w|D = 1)$ versus $P(W > w|D = 0)$, and it is a classic plot for displaying discrimination achieved with a variable W [19, Chapter 4]. To test if discrimination provided by $\text{risk}(X, Y)$ is better than that provided by $\text{risk}(X)$, one could test

$$H_0 \ : \ \text{ROC}_{(X,Y)}(\cdot) = \text{ROC}_X(\cdot). \tag{2}$$

In ROC analysis, the area under the ROC curve (AUC) is typically used as the basis of a test statistic. Then, the null hypothesis is more specifically stated as

$$H_0 \ : \ \text{AUC}_{(X,Y)} = \text{AUC}_X. \tag{3}$$

In the ROC framework, another approach is to assess if, conditional on X, the ROC curve for Y is equal to the null ROC curve [20]. This is particularly relevant when controls are matched by design to cases on X [21]. The corresponding null hypothesis is

$$H_0 \ : \ \text{ROC}_{Y|X}(f) = f, f \in (0, 1) \quad \forall \quad X. \tag{4}$$

Several authors have proposed alternatives to ROC analysis for comparing nested prediction models. The predictiveness curve displays the distribution of risk as the risk quantiles [10, 22, 23]. We write the cumulative distribution of risk as $F_{(X,Y)}(p) = P(\text{risk}(X, Y) \leqslant p)$ and $F_X(p) = P(\text{risk}(X) \leqslant p)$. One can test if the risk distributions based on X or on (X, Y) are different by testing the null hypothesis

$$H_0 : F_{(X,Y)}(\cdot) = F_X(\cdot). \tag{5}$$

Another view is to consider the risk distributions in the case population (denoted with superscript D) and in the control population (superscript $\bar{D}$), separately. We could test

$$H_0 : F^D_{(X,Y)}(\cdot) = F^D_X(\cdot) \text{ and } F^{\bar{D}}_{(X,Y)}(\cdot) = F^{\bar{D}}_X(\cdot) \tag{6}$$

The integrated discrimination improvement statistic is a summary measure based on the difference in average risks between cases and controls, also known as the mean risk difference $\text{MRD} = E(\text{risk}(\cdot)|D = 1) - E(\text{risk}(\cdot)|D = 0)$. The MRD has many interpretations, for example, as the proportion of explained variation, as an R^2 statistic, as Yates slope, and as an average Youden's index [14, 24, 25]. Pencina and others [12] define the integrated discrimination improvement (IDI) as $\text{IDI} = \text{MRD}_{(X,Y)} - \text{MRD}_X$ and propose testing $H_0 : \text{IDI} = 0$. That is, they propose testing

$$H_0 : \text{MRD}_{(X,Y)} = \text{MRD}_X. \tag{7}$$

Another interesting summary of the difference between the case and control risk distributions concerns proportions with risk above the average population risk, $\rho = P(D = 1)$. The above average risk difference is $\text{AARD} = P(\text{risk}(\cdot) > \rho|D = 1) - P(\text{risk}(\cdot) > \rho|D = 0)$. Like the MRD, the AARD has multiple interpretations and relates to existing measures of prediction performance. The AARD is the continuous net reclassification index (NRI (>0), defined in the following text) [13] for comparing a risk model with the null model that has no predictors in which all subjects are assigned risk $P(D = 1) = \rho$. The AARD is also equal to the two-category NRI for comparing a model with the null model when the two risk categories are defined as low risk $\equiv$ 'risk $\leqslant \rho$' and high risk $\equiv$ 'risk $> \rho$'. The AARD can also be considered as a measure relating to the risk distribution in the population, F in Equation (5). In particular, Bura and Gastwirth [26] defined the total gain statistic as the area between the predictiveness curve for risk$(\cdot)$ and the horizontal line at ρ, which is the predictiveness curve for the null model. Gu and Pepe [25] showed that the standardized total gain, total gain/$2\rho(1-\rho)$, is AARD. One can compare the performance of two risk models by evaluating the AARD values and testing the null hypothesis

$$H_0 : \text{AARD}_{(X,Y)} = \text{AARD}_X \tag{8}$$

Authors have also used the medical decision making framework to compare risk models. Vickers and Elkin [27] suggested use of decision curves that plot the net benefit, $\text{NB}(t) \equiv \rho P(\text{risk}(\cdot) > t|D = 1) - (1-\rho)\frac{t}{1-t}P(\text{risk}(\cdot) > t|D = 0)$ against t, the risk threshold. One could envision testing the equality of decision curves

$$H_0 : \text{NB}_{(X,Y)}(\cdot) = \text{NB}_X(\cdot) \tag{9}$$

to compare performance of a model that includes Y with one that does not. Baker [28, 29] suggests standardizing the net benefit by the maximum possible benefit resulting in a relative utility measure. Testing equality of relative utility curves is the same as testing equality of decision curves in (9).

Risk reclassification methodology is yet another approach to comparing risk models. In this framework, for each individual indexed by ι, $\text{risk}(X_\iota, Y_\iota)$ is compared directly with $\text{risk}(X_\iota)$. The NRI statistic is a risk reclassification measure that has gained tremendous popularity since its introduction by Pencina and colleagues in 2008 [12]. The continuous NRI [13] is defined as

$$\text{NRI}(>0) = 2\{P[\text{risk}(X, Y) > \text{risk}(X)|D = 1] - P[\text{risk}(X, Y) > \text{risk}(X)|D = 0]\}$$

The final null hypothesis that we consider testing is

$$H_0 : \text{NRI}(>0) = 0 \tag{10}$$

Our key result is that all of the null hypotheses in Equations (1) through (10) are equivalent. The only condition required for this equivalence is that the distributions of $\text{risk}(X)$ and $\text{risk}(X, Y)$ are absolutely continuous. Before stating the main theorem, we state a result used to prove the theorem, a result that is important in its own right.

Result 1

$$\text{ROC}_{(X,Y)}(f) \geqslant \text{ROC}_X(f) \quad \forall \quad f \in (0, 1)$$

Proof

A fundamental result from decision theory is that decision rules of the form 'risk$(X, Y) > c$' have the best operating characteristics in the sense that when c is chosen to yield a false-positive rate f, $f = P(r(X, Y) > c(f)|D = 0)$, the corresponding true-positive rate $t = P(r(X, Y) > c(f)|D = 1)$

cannot be exceeded by the true positive rate of any other decision rule based on (X, Y) that has the same false-positive rate f [30]. This result follows from that in [31] and is discussed in detail in [32].

It follows that the ROC curve for risk(X, Y) is at least as high at all points as the ROC curve for any other function of (X, Y). In particular, the ROC curve for the function risk(X) cannot exceed the ROC curve for risk(X, Y), namely, $\mathrm{ROC}_{(X,Y)}(\cdot)$, at any point. □

Theorem 1
The following null hypotheses are equivalent

$$\begin{aligned}
& H_0^1 : \mathrm{risk}(X,Y) = \mathrm{risk}(X) \text{ with probability 1} \\
\Leftrightarrow\ & H_0^2 : \mathrm{AUC}_{(X,Y)} = \mathrm{AUC}_X \\
\Leftrightarrow\ & H_0^3 : \mathrm{ROC}_{(X,Y)}(f) = \mathrm{ROC}_X(f) && \forall \quad f \in (0,1) \\
\Leftrightarrow\ & H_0^4 : \mathrm{ROC}_{Y|X}(f) = f && \forall \quad f \in (0,1) \\
\Leftrightarrow\ & H_0^5 : F_{(X,Y)}(p) = F_X(p) && \forall \quad p \in (0,1) \\
\Leftrightarrow\ & H_0^6 : F^D_{(X,Y)}(p) = F^D_X(p) \text{ and } F^{\bar{D}}_{(X,Y)}(p) = F^{\bar{D}}_X(p) && \forall \quad p \in (0,1) \\
\Leftrightarrow\ & H_0^7 : \mathrm{MRD}_{(X,Y)} = \mathrm{MRD}_X \qquad \text{i.e., IDI} = 0 \\
\Leftrightarrow\ & H_0^8 : \mathrm{AARD}_{(X,Y)} = \mathrm{AARD}_X \\
\Leftrightarrow\ & H_0^9 : \mathrm{NB}_{(X,Y)}(t) = \mathrm{NB}_X(t) && \forall \quad t \in (0,1) \\
\Leftrightarrow\ & H_0^{10} : \mathrm{NRI}(>0) = 0
\end{aligned}$$

Proof
That H_0^1 implies each of $H_0^2 - H_0^{10}$ is obvious. Therefore, we focus on showing that each of $H_0^2 - H_0^{10}$ implies H_0^1. We start with H_0^7 and work in reverse order through $H_0^6, H_0^5, \ldots, H_0^2$. Then, we show $H_0^8, H_0^9, H_0^{10} \Rightarrow H_0^1$.

(i) H_0^7 implies H_0^1
Pepe *et al.* [24] write

$$\mathrm{MRD}_{(X,Y)} - \mathrm{MRD}_X = \{\mathrm{var}(\mathrm{risk}(X,Y)) - \mathrm{var}(\mathrm{risk}(X))\}/P(D=1)P(D=0),$$

and because $E(\mathrm{risk}(X,Y)) = E(\mathrm{risk}(X)) = \mathrm{Prob}(D=1)$, it follows that

$$\begin{aligned}
\mathrm{var}(\mathrm{risk}(X,Y)) - \mathrm{var}(\mathrm{risk}(X)) &= E(\mathrm{risk}(X,Y))^2 - E(\mathrm{risk}(X))^2 \\
&= E(\mathrm{risk}(X,Y))^2 - 2E(\mathrm{risk}(X))^2 + E(\mathrm{risk}(X))^2.
\end{aligned}$$

Because $E\{\mathrm{risk}(X,Y)|X\} = \mathrm{risk}(X)$, we have $E(\mathrm{risk}(X)\mathrm{risk}(X,Y)) = E(\mathrm{risk}(X)E(\mathrm{risk}(X,Y)|X)) = E(\mathrm{risk}(X))^2$. Therefore,

$$\begin{aligned}
\mathrm{var}(\mathrm{risk}(X,Y)) - \mathrm{var}(\mathrm{risk}(X)) &= E(\mathrm{risk}(X,Y))^2 - 2E(\mathrm{risk}(X)\mathrm{risk}(X,Y)) + E(\mathrm{risk}(X))^2 \\
&= E\{\mathrm{risk}(X,Y) - \mathrm{risk}(X)\}^2
\end{aligned}$$

Therefore, if $\mathrm{MRD}_{(X,Y)} - \mathrm{MRD}_X = 0$, it follows that $E\{\mathrm{risk}(X,Y) - \mathrm{risk}(X)\}^2 = 0$ and so $\mathrm{risk}(X,Y) = \mathrm{risk}(X)$ with probability 1. That is, H_0^1 follows.

(ii) H_0^6 implies H_0^1
Equality of the case-specific distributions implies that the case-specific means are equal: $E(\mathrm{risk}(X,Y)|D=1) = E(\mathrm{risk}(X)|D=1)$. Similarly, $E(\mathrm{risk}(X,Y)|D=0) = E(\mathrm{risk}(X)|D=0)$. Therefore, H_0^6 implies H_0^7, which we have shown implies H_0^1.

(iii) H_0^5 implies H_0^1
The case-specific distribution of risk can be derived from the population distribution of risk using Bayes' theorem [33].

$$\begin{aligned}
P(\mathrm{risk}(\cdot) = r|D=1) &= \frac{P(D=1|\mathrm{risk}(\cdot)=r)P(\mathrm{risk}(\cdot)=r)}{P(D=1)} \\
&= rP(\mathrm{risk}(\cdot)=r)/P(D=1)
\end{aligned}$$

A similar argument applies to the control-specific distributions. Therefore, equality of population risk distributions in H_0^5 implies equality of case-specific and control-specific risk distributions in H_0^6, which in turn implies H_0^1.

(iv) H_0^4 implies H_0^1

H_0^4 states that, conditional on X, the distributions of Y in the case and control populations are equal:

$$P(Y|D=1,X) = P(Y|D=0,X) = P(Y|X)$$

Using Bayes' theorem, it follows that

$$\frac{P(D=1|Y,X)P(Y|X)}{P(D=1|X)} = P(Y|X)$$

and so $P(D=1|Y,X) = P(D=1|X)$. That is, H_0^1 holds.

(v) H_0^3 implies H_0^1

Huang and Pepe [34] derived the one–one mathematical relationship between the ROC curve for risk$(\cdot)$ and the predictiveness curve, which characterizes the risk distribution. Therefore, equality of ROC curves for risk(X,Y) and risk(X) implies equality of the risk distributions, H_0^5, which in turn implies H_0^1.

(vi) H_0^2 implies H_0^1

We now show that equality of AUCs for risk(X,Y) and risk(X) implies equality of the ROC curves, that is, $H_0^2 \Rightarrow H_0^3$, from which H_0^1 follows. This follows from Result 1 that states $\mathrm{ROC}_{(X,Y)}(\cdot) \geqslant \mathrm{ROC}_X(\cdot)$. If the areas under $\mathrm{ROC}_{(X,Y)}(\cdot)$ and $\mathrm{ROC}_X(\cdot)$ are equal, Result 1 implies that the functions must be equal at all points. That is, H_0^3 must hold.

(vii) H_0^8 implies H_0^1

In the Appendix, Theorem A.1 considers the entity $\mathrm{ROC}_{(X,Y)}\left(t^{\rho}_{(X,Y)}\right) - t^{\rho}_{(X,Y)}$ where $t^{\rho}_{(X,Y)} \equiv P(\mathrm{risk}(X,Y) > \rho|D=0)$. But, by definition of $t^{\rho}_{(X,Y)}$ and the ROC curve, we recognize $\mathrm{ROC}_{(X,Y)}\left(t^{\rho}_{(X,Y)}\right) = P(\mathrm{risk}(X,Y) > \rho|D=1)$. Therefore, Theorem A.1 states that if $P(\mathrm{risk}(X,Y) > \rho|D=1) - P(\mathrm{risk}(X,Y) > \rho|D=0) = P(\mathrm{risk}(X) > \rho|D=1) - P(\mathrm{risk}(X) > \rho|D=0)$ it follows that $\mathrm{ROC}_{(X,Y)}(t) = \mathrm{ROC}_X(t) \quad \forall \quad t$. That is, H_0^8 implies H_0^3, which in turn implies H_0^1.

(viii) H_0^9 implies H_0^1

If $\mathrm{NB}_{(X,Y)}(t) = \mathrm{NB}_X(t) \quad \forall \quad t$, then in particular, we have equality at $t=\rho$: $\mathrm{NB}_{(X,Y)}(\rho) = \mathrm{NB}_X(\rho)$. Recall that $\mathrm{NB}(t)$ is defined as

$$\mathrm{NB}(t) = \rho P(\mathrm{risk} > \rho|D=1) - (1-\rho)\frac{t}{1-t}P(\mathrm{risk} > \rho|D=0)$$

so at $t=\rho$, we have

$$\mathrm{NB}(\rho) = \rho \mathrm{AARD}.$$

Therefore, H_0^9 implies H_0^8, which in turn implies H_0^1.

(ix) H_0^{10} implies H_0^1

We show below that $P(\mathrm{risk}(Y) > \rho|D=1) \geqslant P(\mathrm{risk}(Y) > \rho|D=0)$. An analogous proof that conditions each component on X implies that

$$0 \leqslant P(\mathrm{risk}(X,Y) > \mathrm{risk}(X)|D=1,X) - P(\mathrm{risk}(X,Y) > \mathrm{risk}(X)|D=0,X).$$

But

$$\begin{aligned}\mathrm{NRI}(>0) &= 2\{P(\mathrm{risk}(X,Y) > \mathrm{risk}(X)|D=1) - P(\mathrm{risk}(X,Y) > \mathrm{risk}(X)|D=0)\} \\ &= 2E\{P(\mathrm{risk}(X,Y) > \mathrm{risk}(X)|D=1,X) - P(\mathrm{risk}(X,Y) > \mathrm{risk}(X)|D=0,X)\}\end{aligned}$$

So, if $\mathrm{NRI}(>0) = 0$, it follows that for all X with probability 1, we have

$$P(\mathrm{risk}(X,Y) > \mathrm{risk}(X)|D=1,X) - P(\mathrm{risk}(X,Y) > \mathrm{risk}(X)|D=0,X) = 0$$

The corollary to Theorem A.1 in the Appendix then implies that the ROC curve for Y conditional on X is the null ROC curve. That is, for all X with probability 1, $\mathrm{ROC}_{Y|X}(f) = f \quad \forall \quad f$. In other words, H_0^4 holds, which in turn implies H_0^1.
To complete the proof, we need to prove our assertion that $P(\mathrm{risk}(Y) > \rho | D = 1) \geqslant P(\mathrm{risk}(Y) > \rho | D = 0)$. Using Bayes' theorem, this can be restated as

$$\frac{P(D = 1|\mathrm{risk}(Y) > \rho)}{P(D = 1)} \geqslant \frac{P(D = 0|\mathrm{risk}(Y) > \rho)}{P(D = 0)}$$

$$\frac{a}{b} \geqslant \frac{1 - a}{1 - b}$$

But this holds because we have $a \geqslant b$, implying that $1 - a \leqslant 1 - b$, from which it follows that $a/b \geqslant 1 \geqslant (1 - a)/(1 - b)$.

□

Theorem 1 is a mathematical result involving the functions $\mathrm{risk}(X, Y)$ and $\mathrm{risk}(X)$ and performance measures that are functionals of them. No modeling of the risk functions is assumed. No data sampling is involved in Theorem 1. In the next sections, we consider practical implications of Theorem 1 for data analysis in which models for $\mathrm{risk}(X, Y)$ and $\mathrm{risk}(X)$ may be fit to data.

3. Properties of hypothesis testing procedures

The equivalence of the various null hypotheses in Theorem 1 should not be confused with the equivalence of different hypothesis tests. Two tests can have the same null hypothesis but still be different tests and give different results on a dataset because they are based on different test statistics with different statistical properties. However, it does not make sense to test the same null hypothesis twice — a single test should be chosen. How does one choose the statistical test for the null hypothesis of no incremental predictive value?

There are many possible choices. Here, we focus on the choice between a test for the coefficient for Y in a regression model of the risk function $\mathrm{risk}(X, Y)$ and the change in the AUC for the ROC curves associated with estimated risk functions, $\mathrm{risk}(X)$ and $\mathrm{risk}(X, Y)$. To make the discussion concrete, we consider the likelihood ratio test for β_Y, which is the coefficient for Y in a model for $\mathrm{risk}(X, Y)$, and a test based on the difference $\Delta\widehat{\mathrm{AUC}} = \widehat{\mathrm{AUC}}_{(X,Y)} - \widehat{\mathrm{AUC}}_X$, where $\widehat{\mathrm{AUC}}$ is calculated with the empirical distributions of the fitted values for the risk function in subjects with $D = 1$ and $D = 0$.

3.1. *Testing the regression coefficient has highest power*

When the data are independent and identically distributed observations, the likelihood ratio test is asymptotically the most powerful test for testing $H_0^1 - H_0^{10}$, and so, at least in this classic setting, the test based on β_Y is to be preferred. We see the power advantage demonstrated in the second row of the simulation results in Table I where the procedure based on $\Delta\widehat{\mathrm{AUC}}$ is fixed to have size equal to the nominal level of 0.05. It is also instructive to consider the special case where there are no baseline covariates. In that setting, $\Delta\widehat{\mathrm{AUC}}$ is equivalent to the nonparametric two-sample Wilcoxon statistic, whereas $\hat{\beta}_Y$ from a linear logistic risk model is asymptotically equivalent to the difference in means and so is equivalent to a two-sample Z-statistic. The Z-test is well known to have superior performance compared with the Wilcoxon test for normal data. That is, testing using $\hat{\beta}_Y$ is well known to be superior to testing using $\Delta\widehat{\mathrm{AUC}}$ for normally distributed data and no baseline covariates.

3.2. *Standard tests of performance measures may not be valid*

From a practical point of view, there are additional issues that make the likelihood ratio test more desirable than the $\Delta\widehat{\mathrm{AUC}}$ test. In particular, procedures for fitting risk regression models and for testing coefficients in regression models are highly developed. In contrast, surprisingly little work has been carried out regarding properties of tests that are based on estimates of performance improvement measures. The typical approach to testing with $\Delta\widehat{\mathrm{AUC}}$ uses the fitted values for $\mathrm{risk}(X, Y)$ and $\mathrm{risk}(X)$ as data

Table I. Performance of two-sided nominal 0.05 level tests.

	Size ($\beta_Y = 0$)		Power ($\beta_Y = 0.37$)		Power ($\beta_Y = 0.74$)	
Test statistic	$n_0 = n_{\bar{D}} = 50$	$n_D = 100, n_{\bar{D}} = 900$	$n_D = n_{\bar{D}} = 50$	$n_D = 100, n_{\bar{D}} = 900$	$n_D = n_{\bar{D}} = 50$	$n_D = 100, n_{\bar{D}} = 900$
$LR(\beta_Y)$††	0.051	0.052	0.415	0.909	0.933	1.000
$\Delta\widehat{\text{AUC}}$ size fixed†	0.050	0.050	0.256	0.799	0.775	1.000
$\Delta\widehat{\text{AUC}}$ se standard	0.000	0.002	0.039	0.280	0.356	0.988
$\Delta\widehat{\text{AUC}}$ bootstrap standard	0.000	0.002	0.047	0.291	0.365	0.988
$\Delta\widehat{\text{AUC}}$ bootstrap adjusted	0.012	0.014	0.183	0.666	0.692	0.999

Tests are based on the likelihood ratio for β_Y, the regression coefficient for Y in the risk model $\text{logit risk}(X, Y) = \beta_0 + \beta_X X + \beta_Y Y$ and on $\Delta\widehat{AUC} = \widehat{AUC}_{(X,Y)} - \widehat{AUC}_X$. Tests based on $\Delta\widehat{AUC}$ were 'adjusted' if regression coefficients were estimated in each bootstrap resampled dataset and 'standard' if bootstrap resampling (bootstrap) or DeLong standard error (se) calculation used fitted values $\widehat{\text{risk}}(X, Y)$ and $\widehat{\text{risk}}(X)$ derived from the original dataset. Data were simulated with $X, Y \sim N(0, 1)$ in controls, $X \sim N(0.74, 1)$ in cases, $Y \sim N(0, 1)$ in cases under the null, and $Y \sim N(0.37, 1)$ or $Y \sim N(0.74, 1)$ in cases under the alternative. There were 1000 simulations for each scenario and 1000 bootstrap samples per analysis.

†The rejection thresholds for this test were chosen using the null distribution calculated from 50,000 simulated datasets. In practice, the null distribution is unknown so this test cannot be applied.

††Results are shown for the likelihood test. Almost identical results were obtained with the Wald test, which is asymptotically equivalent to the likelihood ratio test.

inputs to a test of equal AUCs for two diagnostic tests such as the DeLong test [35] or the resampling-based test [36]. The fact that the coefficients in the fitted values are estimated from the data is ignored in these testing procedures.

We used simulation studies to investigate the properties of these tests in a simple scenario. We generated data for X and Y as independent and normally distributed with standard deviation 1 in cases ($D = 1$) and controls ($D = 0$). The mean of X was 0.74 in cases and 0 in controls yielding an AUC of 0.7 for the baseline risk model. The mean of Y was 0 in cases and in controls under scenarios simulating the null setting for evaluating size, whereas the means were 0.37 or 0.74 in cases and 0 in controls under scenarios simulating the alternative setting for evaluating power. We see from the third and fourth rows in Table I that standard tests ignoring sampling variability in the estimated risk regression coefficients are extremely conservative. Both the DeLong test [35] that uses the normal approximation with a standard error formula and the test using percentiles of the bootstrap distribution [36] have size less than 0.005 with sample sizes as large as 100 cases and 900 controls. The problem is due to estimating the coefficients in the nested models because the same tests comparing X alone to another independent marker with equal performance were not conservative with comparable sample sizes (data not shown). Demler *et al.* [18] recently provided some theoretical arguments for poor performance of the DeLong test. Tests using $\Delta\widehat{\text{AUC}}$ standardized by the bootstrap variance had similar properties (data not shown).

We implemented an alternative version of the $\Delta\widehat{\text{AUC}}$ test in the hope that acknowledging sampling variability in the estimated regression coefficients would lead to a test with correct size. This approach used the bootstrap, rejecting the null hypothesis if the 2.5th percentile of the bootstrap distribution exceeded 0. In these simulations, we resampled observations from the original dataset, fit the risk models, and calculated $\Delta\widehat{\text{AUC}}$ for each resampled dataset. We provide the results in line 5 of Table I as $\Delta\widehat{\text{AUC}}$ adjusted. These tests were less conservative than procedures not adjusting for variability in regression coefficients but remained conservative nevertheless. Therefore, simply acknowledging variability of the estimated regression coefficients in the bootstrap procedure does not rectify the problem. Rather, it appears that estimating the regression coefficients leads to non-normality of the distribution of $\Delta\widehat{\text{AUC}}$ under the null, which in turn invalidates use of the bootstrap procedure. Authors reported previously this non-normality for the $\Delta\widehat{\text{AUC}}$ statistic [18]. Kerr *et al.* [37] found a similar phenomenon for the IDI statistic under the null.

We conclude that all currently available procedures for testing incremental value based on $\Delta\widehat{\text{AUC}}$ in the full dataset are unacceptably conservative in the classic scenarios we studied. From Table I, we observe that as a consequence, they have extremely low power compared with the likelihood ratio test for β_Y.

Note that a simple split sampling approach can yield valid inference about ΔAUC. In particular, if the dataset is split into a training set where the risk models are fit and a test set where fitted values are calculated, standard methods for inference about ΔAUC can be applied in the test set using those fitted values because they are fixed functions of the test set observations. However, the power of this split sample ΔAUC approach is compromised by the reduced sample size of the test set relative to the full dataset. The power will be much lower than that of a test based on β_Y that uses the entire dataset.

4. Recommendations for practice

When risk functions fit a dataset reasonably well, Result 1 and Theorem 1 hold approximately in the data. There are at least two important implications of these results and of the simulation results described earlier for practical data analysis.

(i) If parametric risk models are employed and if there are no overfitting issues that could invalidate inference, one should use likelihood-based methods for testing if coefficients associated with Y are zero in order to test the null hypothesis of no performance improvement. The rationale is that likelihood-based procedures have optimal power and that highly developed, well-performing methods for inference are available.

(ii) If risk models are fit using alternatives to parametric models, Theorem 1 implies that methods to test $H_0^1 : \text{risk}(X, Y) = \text{risk}(X)$ can be based on performance measures mentioned in Theorem 1, such as $\text{AUC}_{(X,Y)} - \text{AUC}_X$ or IDI or $\text{AARD}_{(X,Y)} - \text{AARD}_X$ or NRI (>0). However, we need to develop valid methods for inference about these performance measures under the null hypothesis before we can implement such hypothesis testing procedures.

The estimated risk model, $\widehat{\text{risk}}(X)$, is considered to fit the data well if, within subgroups, S_X defined by X, the frequency of events, $\hat{P}(D=1|X \in S_X)$, is approximately equal to the average estimated risk:

$$\hat{P}(D_i = 1|X_i \in S_X) \approx \text{ Average } \left(\widehat{\text{risk}}(X_i)|X_i \in S_X\right).$$

This is also called good calibration of the model $\widehat{\text{risk}}(X)$. Similarly, the fitted model for $\text{risk}(X,Y)$ is considered well calibrated if

$$\hat{P}(D_i = 1|(X_i, Y_i) \in S_{(X,Y)}) \approx \text{ Average } \left(\widehat{\text{risk}}(X_i, Y_i)|(X_i, Y_i) \in S_{(X,Y)}\right)$$

where $S_{(X,Y)}$ denotes a subset of observations defined by X and Y. It is our opinion that risk models should be shown to be well calibrated in the dataset before proceeding to evaluate their prediction performances. Poorly calibrated models, by definition, are known to misrepresent the risk functions in the population and are therefore inappropriate as risk calculators for use in practice. Consequently, the prediction performance of poorly calibrated models is not important. There is a large literature on methods to estimate well-fitting models and to evaluate calibration. See, for example, textbooks by Harrell [38] and Steyerberg [39]. The focus of this paper is on settings where the dimensions of X and Y are low. Ensuring good calibration is tenable when X and Y are low dimensional.

Overfitting is a concern primarily when the number of predictors is large. Overfitting yields estimated risks that are likely to be biased toward values more extreme than the true risks. We can use techniques such as penalized likelihood to address this issue. Another approach is to use a subset of the data to reduce the dimensions of X and Y to one before fitting and evaluating risk models for X and (X,Y) in the remaining data. We demonstrate this approach in the context of an example in Section 5.

In summary, if one employs parametric risk models, our recommendation is to ensure the use of well-calibrated models and to base hypothesis testing on β_Y rather than on ΔAUC. Current procedures based on ΔAUC do not have correct size. Kerr *et al.* [37] found similar problems for the IDI statistic under the null. It is possible that new approaches to testing based on ΔAUC could be developed to properly account for estimation of the risk values and thereby yield appropriately sized tests. However, even if such procedures were developed, we have argued and observed in Table I (line 2) that tests based on β_Y are still likely to be more powerful, at least when likelihood-based procedures are used to estimate parameters in the risk models. Therefore, testing based on β_Y would still be the better choice.

More important than testing if there is *any* increment in prediction performance is estimating the size of the gain in performance. The sizes of the regression coefficients for Y and X in $\text{risk}(X,Y)$ are not sufficient because prediction performance depends on the population distribution of the predictors (X,Y) in addition to the conditional probability function $P(D=1|X,Y) = risk(X,Y)$. We described a variety of measures to quantify the prediction performance of a risk model in Section 2, and a comparison of the measures calculated with $\text{risk}(X)$ and $\text{risk}(X,Y)$ constitutes the corresponding increment in performance due to Y. The field of risk prediction however has not yet settled debates about which are the best measures for quantifying performance increment, and we do not debate this question further here. Our recommendation is to focus on estimating a compelling measure of increment in prediction performance. Any testing should be limited to testing whether Y is a risk factor when controlling for X in a regression model.

We note that one further implication of Theorem 1 is that if Y is found to be statistically significant as a risk factor at the α level, then 0 should be excluded from confidence intervals for changes in AUC, MRD, and AARD statistics and for the NRI (>0) statistic. The effect on confidence interval coverage probabilities caused by excluding 0 when H_0 is rejected is a subject for future study. Nevertheless, the practice seems intuitively reasonable and provides internal consistency for data analysis results.

5. Application to a renal artery stenosis dataset

Diagnosis of stenosis in the renal artery involves a risky surgical procedure and is only undertaken for patients deemed likely to have a positive finding. The risk of having renal artery stenosis is estimated from clinical data in order to guide decisions about undergoing invasive surgery for definitive diagnostic procedures. Janssens and others [40] reported data for 426 patients who were surgically assessed for renal stenosis. We consider the improvement in prediction performance that is gained by adding serum creatinine to the baseline predictors.

We randomly chose one third of the observations ($n = 142$; 33 of the 98 cases and 109 of the 328 controls) as a training set to generate a baseline risk predictor X that is a combination of the candidate clinical variables. Using linear logistic regression, we found that age, body mass index (BMI), and abdominal bruit (bruit) were highly significantly associated with renal stenosis but that gender, hypertension, and vascular stenosis were not. We refit the model including only age (in years), BMI (kg/m^2), and bruit (yes = 1, no = 0) to derive the linear combination

$$X = 0.87 \times \text{ age } - 0.27 \text{ BMI } + 1.39 \times \text{ bruit } - 1.99.$$

We then evaluated the performances of risk models based on this one-dimensional predictor X and on the combination of X and $Y = \log$ (serum creatinine) using the remaining two thirds of the data ($n = 284$), the evaluation dataset.

We fit linear logistic models in the evaluation dataset:

$$\text{Baseline: } \operatorname{logit} P(D = 1|X) = \alpha_0 + \alpha_1 X$$
$$\text{Enhanced: } \operatorname{logit} P(D = 1|X, Y) = \beta_0 + \beta_1 X + \beta_2 Y.$$

The predictiveness curves in Figure 1 show that these models are well calibrated to the evaluation study cohort because observed event rates in each decile of modeled risk (shown as open circles) are approximately equal to the average modeled risks (shown as the points on the solid predictiveness curves). Hosmer–Lemeshow goodness-of-fit statistics [10, 41, 42] do not provide any evidence against the null of good calibration (p-values of 0.42 for the baseline model, risk(X), and 0.51 for the enhanced model, risk(X, Y)).

As an aside, we also calculated risk estimates in the evaluation dataset using the model for risk(X) originally fit in the one-third dataset that generated X. The dashed predictiveness curve in Figure 1 shows that the original model does not fit the evaluation dataset. There is evidence that the original model suffers from overfitting because the risk quantiles are more extreme than the observed event rates within each decile of X. One can think of the model for risk(X) fit in the evaluation dataset as a recalibrated version of the original model fit in the one-third dataset. Authors have previously described this approach. See, for example, [43]. Recalibrating the risk model is a step missing from the usual split sample approach that simply evaluates the performance of the original model fit in the training dataset.

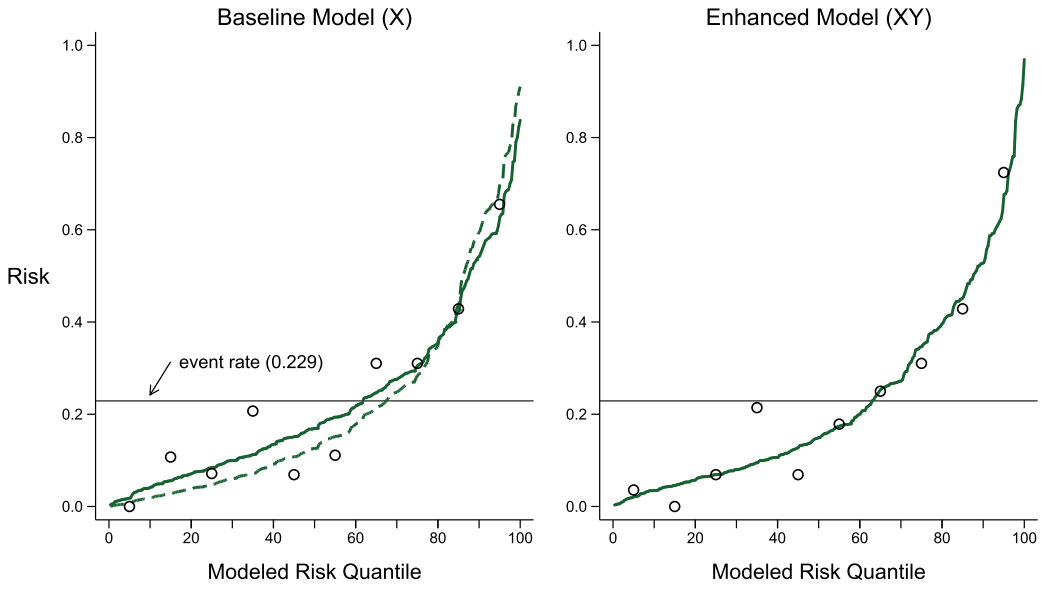

Figure 1. Predictiveness curves to assess calibration of baseline and enhanced risk models for renal artery stenosis in the evaluation dataset ($n = 284$). Shown are the modeled risk quantiles (as curves) and the observed event rates within each decile of modeled risk (as open circles). Hosmer–Lemeshow statistics corresponding to the plots have p-values equal to 0.43 (baseline model) and 0.51 (enhanced model). The quantiles of the risk function fitted in the one-third dataset used to generate X is also shown as the dashed curve and appears steeper than the model recalibrated in the evaluation dataset.

If that model is over fit, however, we have argued that its prediction performance is not of interest as the risk estimates derived from it are biased toward extreme values. We recommend recalibration before evaluating prediction performance in the evaluation set.

Consider now the comparison between the risk functions risk(X) and risk(X, Y) in the evaluation dataset. The likelihood ratio test for $H_0 : \beta_2 = 0$ is highly significant with $p = 0.0003$ (Table II). According to Theorem 1, we can conclude that prediction model performance is improved by addition of Y to the model. Nevertheless, we implemented tests on the basis of ΔAUC as well to compare inference. The test for equality of AUCs is also significant but with much weaker p-value, $p = 0.034$, using the DeLong variance formula and $p = 0.026$ using percentiles of the bootstrap distribution. Recall that these tests do not acknowledge variability in the estimated regression coefficients $\left(\widehat{\beta}_1, \widehat{\beta}_2\right)$ and are extremely conservative. Bootstrapping that incorporated refitting the risk model in each resampled dataset is also not valid according to our simulations. It yielded a p-value $= 0.07$, which was even more conservative in contrast to results of our simulation study. In accordance with our recommendation in Section 4, the test based on β_2 yielded the strongest evidence against H_0 that prediction performance is not improved by including serum creatinine as a predictor.

We repeated the analysis using a weaker marker, Y^*, for illustration. Here, $Y^* = Y + \varepsilon$, where ε is a standard normal random variable, adding noise to Y. In this analysis, the coefficient for Y^* is highly statistically significant ($p = 0.006$ with the likelihood ratio test, Table II), whereas the standard tests based on ΔAUC are not ($p = 0.21$ using either the DeLong variance formula or the percentiles of the bootstrap distribution). The bootstrap-adjusted ΔAUC test that refits the models in each bootstrap sample is not significant either, $p = 0.28$. Again, this supports our recommendation for testing the null hypothesis of no performance improvement on the basis of the regression coefficient for Y in the enhanced risk model, risk(X, Y).

We provide estimates of prediction performance in Table III for the baseline and enhanced risk models. We calculated confidence intervals using 2.5th and 97.5th percentiles of bootstrap distributions with models refit in each bootstrapped dataset. We estimated that the AUC increased from 0.78 to 0.81 with addition of serum creatinine. We also considered a point on the ROC curve. In particular, setting the risk threshold so that 80% of the cases are sent for the invasive diagnostic renal arteriography, we find that the proportion of controls who unnecessarily undergo the procedure, denoted by $\text{ROC}^{-1}(0.8)$ in Table III, decreases from 0.36 to 0.33. Note that Pfeiffer and Gail [44] recommend calculating the percent needed to follow (PNF) that is a simple function of $\text{ROC}^{-1}(f) : \text{PNF}(f) = \rho f + (1 - \rho)\text{ROC}^{-1}(f)$. Therefore, the PNF decreased from 0.46 to 0.44. The IDI statistic is the change in the MRD statistic and is calculated as 0.05, whereas the conceptually similar change in the AARD is 0.014. The continuous-NRI statistic is NRI$(> 0) = 0.45$. Note that the NRI is measured on a scale from 0 to 2, unlike most other

Table II. Logistic regression models for risk of renal artery stenosis fit to data for 284 patients.

	Intercept	X	Y or Y^*
Baseline model (X)			
Coefficient	−0.12	0.76	—
se	0.21	0.12	—
p-value	0.56	<0.001	—
Enhanced model (X, Y)			
Coefficient	−0.32	0.67	0.62
se	0.22	0.12	0.18
p-value	0.15	<0.001	<0.001
Enhanced model (X, Y^*)			
Coefficient	−0.19	0.73	0.31
se	0.21	0.12	0.12
p-value	0.38	<0.001	0.006

The addition of $Y = \log(\text{serum creatinine})$ to a model including the baseline covariate X is assessed. Also shown are results for a model including $Y^* = Y + \varepsilon$, where $\varepsilon \sim N(0, 1)$ random variable. Log odds ratios are displayed along with standard errors and p-values calculated with likelihood ratio tests.

Table III. Performance of baseline and enhanced models for prediction of renal artery stenosis and performance improvement with 95% confidence interval calculated with 1000 bootstrap samples.

Performance measure		Baseline model X	Enhanced model (X, Y)	Performance improvement†
ROC area	AUC	0.78	0.81	0.03 (0.01, 0.07)
FPR at TPR = 0.8	$ROC^{-1}(0.8)$	0.36	0.33	−0.03 (−0.26, 0.07)
Mean risk difference	MRD	0.20	0.25	0.05* (0.011, 0.116)
Above average risk difference	AARD**	0.43	0.47	0.014(0, 0.11)
Continuous NRI	NRI (>0)	—	—	0.45 (0.17, 0.79)
Net benefit at 0.25	NB (0.25)	8.1%	9.3%	1.2% (−1.6%, 3.4%)

†Performance improvement is defined as the difference between the measure for the enhanced model and that for the baseline model for all measures except for the NRI.
*Also known as the IDI statistic.
**Also known as the total gain statistic.

measures that are restricted to (0,1). We calculated the net benefit using a risk threshold of 0.25. This threshold implicitly assumes that the net benefit of diagnosis for a subject with renal artery stenosis is three times the net cost of the diagnostic procedures for a subject without stenosis because the cost–benefit ratio = risk threshold/(1-risk threshold) [27]. The maximum possible benefit of a risk model in this population would be that associated with diagnosing all 67 (24%) subjects who have renal stenosis and not sending any controls for the diagnostic procedure. We calculate that the net benefit is 35.4% of maximum with use of the baseline model and 40.5% of maximum with use of the model that includes serum creatinine. We see that 95% confidence intervals for most measures of improvement in performance exclude the null value of 0. We modified the bootstrap confidence interval for the change in AARD to exclude values $\leqslant 0$ because the null hypothesis $H_0 : \beta_2 = 0$ was rejected, implying that the null hypothesis $H_0 : \Delta\text{AARD} = 0$ is also rejected.

6. Discussion

The main result of this paper is that the common practice of performing separate hypothesis tests, for the coefficient of Y in the risk prediction model and for the change in performance of the model, is literally testing the same null hypothesis twice. Vickers *et al.* [17] make a heuristic argument for this point. We have proven the result with formal mathematical theory. Testing the same null hypothesis in multiple ways is poor statistical practice and should be replaced with a more thoughtful strategy for analysis that employs a single test of the null. Arguments in favor of basing the single test on the regression coefficient for Y in a risk model include the following: (i) such tests are most powerful asymptotically; and (ii) techniques are well developed and widely available for performing such tests. This strategy relies on employing risk models that are well calibrated in the data. We have argued that good calibration is a crucial aspect of risk model assessment. If necessary, models should be recalibrated to the population of interest prior to assessing model performance. We demonstrated this in the renal artery stenosis dataset.

The split sample recalibration approach that we employed fit a risk model to the baseline predictors $U = (U_1, \ldots, U_K)$ to derive a univariate score that we denoted by $X = \sum \gamma_k U_k$. We then fit and evaluated models for risk(X) and risk(X, Y) in the evaluation dataset. Note that the model risk$(X, Y) = P(D = 1|X, Y)$ should not be interpreted as a model for risk$(U, Y) = P(D = 1|U_1, \ldots, U_K, Y)$. The former fixes the combination of $\{U_1, \ldots, U_K\}$ considered in joint modeling with Y and therefore may be less predictive than $r(U, Y)$ that allows the contributions of $\{U_1, \ldots, U_K\}$ each to vary in the presence of Y. That is, our strategy that evaluates risk(X, Y) rather than risk(U, Y) might lead to a somewhat pessimistic assessment of the incremental value of Y over the baseline prediction model risk$(X) = P(D = 1|X) = P(D = 1|U_1, \ldots, U_K)$. The advantage of our approach is that it offers a way around calibration and performance assessment problems associated with overfitting, but admittedly, a potentially pessimistic evaluation of incremental value is a possible downside of this approach.

After testing if there is any improvement in prediction performance, the next task is to estimate the extent of improvement achieved. How to quantify the improvement in performance is a topic of much debate in the literature. A multitude of metrics exist, including ΔAUC, ΔMRD, ΔAARD, approaches based on risk reclassification tables [13, 15, 16], approaches based on the Lorenz curve

[44], and approaches based on medical decision making [7, 27, 29, 45]. This paper does not seek to provide guidance on the choice of measure, but we emphasize that estimating the magnitude of improvement is far more important than testing for any improvement. Moreover, if hypothesis testing based on performance measures is employed, it should be with regard to a null hypothesis concerning *minimal* improvement, H_0 : performance improvement $\leqslant$ minimal, rather than *any* improvement, H_0 : performance improvement $= 0$. The exercise of setting standards for minimal improvement may have the added benefit of helping us to choose a clinically relevant measure of performance improvement.

Appendix A

We use the following notation

$$t^{\rho}_{X} \equiv 1 - F^{\bar{D}}_{X}(\rho) = P(\text{risk}(X) > \rho | D = 0)$$
$$t^{\rho}_{(X,Y)} \equiv 1 - F^{\bar{D}}_{(X,Y)}(\rho) = P(\text{risk}(X,Y) > \rho | D = 0)$$

We also assume that the distributions of $\text{risk}(X,Y)$ and $\text{risk}(X)$ are absolutely continuous. This implies that their ROC curves have second derivatives.

Theorem A.1

$$\begin{aligned} & \text{ROC}_{(X,Y)}\left(t^{\rho}_{(X,Y)}\right) - t^{\rho}_{(X,Y)} = \text{ROC}_{X}\left(t^{\rho}_{X}\right) - t^{\rho}_{X} \\ \Rightarrow \quad & \text{ROC}_{(X,Y)}(t) = \text{ROC}_{X}(t) \quad \forall \quad t \\ & \text{and } t^{\rho}_{(X,Y)} = t^{\rho}_{X} \end{aligned} \tag{A.1}$$

Proof
For $W = \text{risk}(X)$ or $W = \text{risk}(X,Y)$, it is well known that $\text{ROC}_{W}(t) - t$ is a concave function [19, p. 71]. Therefore, $\text{ROC}_{(X,Y)}(t) - t$ has a unique maximizer. Moreover, the maximizer occurs when $\text{ROC}'_{(X,Y)}(t) = 1$. Arguments in the following proof of Corollary A.1 show that this implies $\text{ROC}_{(X,Y)}(t) - t$ is maximized at $t^{\rho}_{(X,Y)}$.

Because $\text{ROC}_{(X,Y)}(t) \geqslant \text{ROC}_{X}(t) \quad \forall \quad t$, we have

$$\text{ROC}_{(X,Y)}\left(t^{\rho}_{X}\right) - t^{\rho}_{X} \geqslant \text{ROC}_{X}\left(t^{\rho}_{X}\right) - t^{\rho}_{X}$$

and Equation (A.1) implies therefore that

$$\text{ROC}_{(X,Y)}\left(t^{\rho}_{X}\right) - t^{\rho}_{X} \geqslant \text{ROC}_{(X,Y)}\left(t^{\rho}_{(X,Y)}\right) - t^{\rho}_{(X,Y)}.$$

It follows that $t^{\rho}_{X} = t^{\rho}_{(X,Y)}$ because, as noted earlier, $\text{ROC}_{(X,Y)}(t) - t$ has a unique maximizer at $t^{\rho}_{(X,Y)}$. This also implies by Equation (A.1) that $\text{ROC}_{(X,Y)}(t^{\rho}) = \text{ROC}_{X}(t^{\rho})$ where we now use the notation t^{ρ} for the common value of $t^{\rho}_{(X,Y)}$ and t^{ρ}_{X}.

Next, we show that $\text{ROC}'_{X}(t) \leqslant \text{ROC}'_{(X,Y)}(t)$ when $t < t^{\rho}$. To show this, we suppose that $\text{ROC}'_{X}(t) > \text{ROC}'_{(X,Y)}(t)$ for some $t < t^{\rho}$ and show a contradiction by constructing decision rules based on (X,Y) that have an ROC curve exceeding $\text{ROC}_{(X,Y)}$ on a subinterval of $(0, t^{\rho})$. If $\text{ROC}'_{X}(t) > \text{ROC}'_{(X,Y)}(t)$ at some point t, by continuity of ROC'_{X} and $\text{ROC}'_{(X,Y)}$, we have $\text{ROC}'_{X}(t) > \text{ROC}'_{(X,Y)}(t)$ on an interval $(a,b) \subset (0, t^{\rho})$. Let r^{a} denote the risk threshold corresponding to the false-positive rate a and consider the family of decision rules that classify positive if $\left\{\text{'risk}(X,Y) > r^{a}_{(X,Y)}\text{'} \text{ or } \left[\text{'risk}(X,Y) < r^{a}_{(X,Y)} \text{ and risk}(X) < r^{a}_{X} \text{ and risk}(X) > k\right]\text{' for } k > r^{b}_{X}\right\}$. These decision rules vary in K. They have an ROC curve equal to $\text{ROC}_{(X,Y)}(t)$ at $t = a$ and ROC derivative equal to ROC'_{X}, which we assume is higher than $\text{ROC}'_{(X,Y)}$ over (a,b). Therefore, this ROC curve exceeds $\text{ROC}_{(X,Y)}$ over (a,b). But this is impossible because the Neyman–Pearson lemma implies that $\text{ROC}_{(X,Y)}(t)$ is optimal at all t. In particular, $\text{ROC}_{(X,Y)}(t) \geqslant \text{ROC}_{X}(t)$ at all t. Therefore, we cannot have $\text{ROC}'_{X}(t) > \text{ROC}'_{(X,Y)}(t)$ for any $t < t^{\rho}$.

Recall from the previous discussion that

$$0 = \mathrm{ROC}_{(X,Y)}(t^\rho) - \mathrm{ROC}_X(t^\rho) = \int_0^{t^\rho} \left(\mathrm{ROC}'_{(X,Y)}(t) - \mathrm{ROC}'_X(t)\right) \mathrm{d}t.$$

But having shown that the integrand is ≥ 0, we must conclude that the integrand is 0,

$$\mathrm{ROC}'_{(X,Y)}(t) = \mathrm{ROC}'_X(t) \quad \forall \quad t < t^\rho.$$

Moreover, equality of $\mathrm{ROC}_{(X,Y)}(t)$ and $\mathrm{ROC}_X(t)$ at $t = 0$ and at $t = t^\rho$ implies

$$\mathrm{ROC}_{(X,Y)}(t) = \mathrm{ROC}_X(t) \quad \forall \quad t < t^\rho.$$

Similar arguments show that $\mathrm{ROC}_{(X,Y)}(t) = \mathrm{ROC}_X(t) \quad \forall \quad t > t^\rho$. □

Corollary A.1
Let $\mathrm{ROC}_\omega(\cdot)$ be the ROC curve for the risk function $\mathrm{risk}(\omega) = P(D = 1|\omega)$. We show that

$$\begin{aligned} & \mathrm{ROC}_\omega\left(t^\rho_\omega\right) = t^\rho_\omega \\ \Rightarrow \quad & \mathrm{ROC}_\omega(t) = t \quad \forall \quad t \in (0,1) \end{aligned} \tag{A.2}$$

Proof
$\mathrm{ROC}_\omega(t) - t$ is maximized at the point where $\mathrm{ROC}'_\omega(t) = 1$. Bayes' theorem implies that

$$\mathrm{logit}P(D = 1|\mathrm{risk}(\omega) = r) = \mathrm{logit}\rho + \log \mathrm{ROC}'\left(t^r_\omega\right)$$

where $t^r_\omega = P(\mathrm{risk}(\omega) > r|D = 0)$. When $\mathrm{ROC}'\left(t^r_\omega\right) = 1$, therefore, $P(D = 1|\mathrm{risk}(\omega) = r) = \rho$. That is, the point that maximizes $\mathrm{ROC}_\omega(t) - t$ is t^ρ_ω. We write

$$\sup|\mathrm{ROC}_\omega(t) - t| = \mathrm{ROC}_\omega\left(t^\rho_\omega\right) - t^\rho_\omega \tag{A.3}$$

but (A.2) then implies that $\sup|\mathrm{ROC}_\omega(t) - t| = 0$. In other words, (A.2) implies $\mathrm{ROC}_\omega(t) = t \quad \forall \quad t \in (0, 1)$. Note that Equation (A.3) also follows from the fact that both sides of (A.3) were shown to equal the standardized total gain statistic (see Equations (6) and (7) of [25]). □

Acknowledgement

Grants from the National Institutes of Health (GM54438 and CA86368) supported Margaret S. Pepe.

References

1. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* 1989; **81**:1879–1886. DOI: 10.1093/jnci/81.24.1879.
2. Gail MH, Costantino JP. Validating and improving models for projecting the absolute risk of breast cancer. *Journal of the National Cancer Institute* 2001; **93**:334–335. DOI: 10.1093/jnci/93.5.334.
3. Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, Tice JA, Buist DS, Geller BM, Rosenberg R, Yankaskas BC, Kerlikowske K. Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute* 2006; **98**:1204–1214. DOI: 10.1093/jnci/djj331.
4. Chen J, Pee D, Ayyagari R, Graubard B, Schairer C, Byrne C, Benichou J, Gail MH. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *Journal of the National Cancer Institute* 2006; **98**:1215–1226. DOI: 10.1093/jnci/djj332.
5. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, Thun MJ, Cox DG, Hankinson SE, Kraft P, Rosner B, Berg CD, Brinton LA, Lissowska J, Sherman ME, Chlebowski R, Kooperberg C, Jackson RD, Buckman DW, Hui P, Pfeiffer R, Jacobs KB, Thomas GD, Hoover RN, Gail MH, Chanock SJ, Hunter DJ. Performance of common genetic variants in breast-cancer risk models. *New England Journal of Medicine* 2010; **362**:986–993. DOI: 10.1056/NEJMoa0907727.
6. Gail MH. Probability discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *Journal of the National Cancer Institute* 2008; **100**:1037–1041. DOI: 10.1093/jnci/djn180.
7. Gail MH. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. *Journal of the National Cancer Institute* 2009; **101**:959–963. DOI: 10.1093/jnci/djp130.
8. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998; **97**:1837–1847. DOI: 10.1161/01.CIR.97.18.1837.

9. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *Journal of the American Medical Association* 2009; **302**:2345–2352. DOI: 10.1001/jama.2009.1757.
10. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology* 2008; **167**:362–368. DOI: 10.1093/aje/kwm305.
11. Pepe MS, Gu JW, Morris DE. The potential of genes and other markers to inform about risk. *Cancer Epidemiology Biomarkers and Prevention* 2010; **3**:655–665. DOI: 10.1158/1055-9965.EPI-09-0510.
12. Pencina MJ, D'Agostino RB, Sr, D'Agostino RB, Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* 2008; **27**:157–172. DOI: 10.1002/sim.2929.
13. Pencina MJ, D'Agostino RB, Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine* 2010; **30**:11–21. DOI: 10.1002/sim.4085.
14. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. *Biometrical Journal* 2011; **53**:237–258. DOI: 10.1002/bimj.201000078.
15. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Annals of Internal Medicine* 2009; **150**:795–802. DOI: 10.1059/0003-4819-150-11-200906020-00007.
16. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. *American Journal of Epidemiology* 2011; **173**:1327–1335. DOI: 10.1093/aje/kwr013.
17. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC Medical Research Methodology* 2011; **11**:13. DOI: 10.1186/1471-2288-11-13.
18. Demler OV, Pencina MJ, D'Agostino RB, Sr. Misuse of DeLong test to compare AUCs for nested models. *Statistics in Medicine* 2012; **31**(23):2577–2587. DOI: 10.1002/sim.5328.
19. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
20. Janes H, Pepe MS. Adjusting for covariate effects on classification accuracy using the covariate adjusted ROC curve. *Biometrika* 2009; **96**:371–382. DOI: 10.1093/biomet/asp002.
21. Janes H, Pepe MS. Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. *Biometrics* 2008; **64**:1–9. DOI: 10.1111/j.1541-0420.2007.00823.x.
22. Huang Y, Pepe MS, Feng Z. Evaluating the predictiveness of a continuous marker. *Biometrics* 2007; **63**:1181–1188. DOI: 10.1111/j.1541-0420.2007.00814.x.
23. Stern RH. Evaluating new cardiovascular risk factors for risk stratification. *Journal of Clinical Hypertension* 2008; **10**:485–488. DOI: 10.1111/j.1751-7176.2008.07814.x.
24. Pepe MS, Feng Z, Gu JW. Invited commentary on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond'. *Statistics in Medicine* 2008; **27**:173–181. DOI: 10.1002/sim.2991.
25. Gu J W, Pepe MS. Measures to summarize and compare the predictive capacity of markers. *International Journal of Biostatistics* 2009; **5**(1):article 27. DOI: 10.2202/1557-4679.1188.
26. Bura E, Gastwirth JL. The binary regression quantile plot: assessing the importance of predictors in binary regression visually. *Biometrical Journal* 2001; **43**:5–21. DOI: 10.1002/1521-4036(200102)43:1<5::AID-BIMJ5>3.0.CO;2-6.
27. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 2006; **26**:565–574. DOI: 10.1177/0272989X06295361.
28. Baker SG. Putting risk prediction in perspective: relative utility curves. *Journal of the National Cancer Institute* 2009; **101**:1538–1542. DOI: 10.1093/jnci/djp353.
29. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society Series B* 2009; **172**:729–748. DOI: 10.1111/j.1467-985X.2009.00592.x.
30. Green DM, Swets JA. *Signal Detection Theory and Psychophysics*. Wiley: New York, 1966.
31. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society of London, Series A* 1933; **231**:289–337. DOI: 10.1098/rsta.1933.0009.
32. McIntosh MS, Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics* 2002; **58**:657–664. DOI: 10.1111/j.0006-341X.2002.00657.x.
33. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics* 2005; **6**:227–239. DOI: 10.1093/biostatistics/kxi005.
34. Huang Y, Pepe MS. A parametric ROC model based approach for evaluating the predictiveness of continuous markers in case-control studies. *Biometrics* 2009; **65**:1133–1144. DOI: 10.1111/j.1541-0420.2009.01201.x.
35. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**:837–845. DOI: 10.1111/j.0006-341X.2002.00657.x.
36. Pepe M, Longton G, Janes H. Estimation and comparison of receiver operating characteristic curves. *Stata Journal* 2009; **9**:1–16. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2774909/pdf/nihms90148.pdf.
37. Kerr KF, McClelland RL, Brown ER, Lumley T. Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *American Journal of Epidemiology* 2011; **174**:364–374. DOI: 10.1093/aje/kwr086.
38. Harrell F. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Verlag: New York, 2001.
39. Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer: New York, 2010.
40. Janssens AC, Deng Y, Borsboom GJ, Eijkemans MJ, Habbema JD, Steyerberg EW. A new logistic regression approach for the evaluation of diagnostic test results. *Medical Decision Making* 2005; **25**:168–177. DOI: 10.1177/0272989X05275154.
41. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow SA. Comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* 1997; **16**:965–980. DOI: 10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O.
42. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. Wiley: New York, 2000.

43. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine* 2000; **19**:3401–3415. DOI: 10.1002/1097-0258(20001230)19:24<3401::AID-SIM554>3.0.CO;2-2.
44. Pfeiffer RM, Gail MH. Two criteria for evaluating risk prediction models. *Biometrics* 2011; **67**:1057–1065. DOI: 10.1111/j.1541-0420.2010.01523.x.
45. Rapsomaniki E, White IR, Wood AM, Thompson SG, Emerging Risk Factors Collaboration. A framework for quantifying net benefits of alternative prognostic models. *Statistics in Medicine* 2012; **31**:114–130. DOI: 10.1002/sim.4362.