



UW Biostatistics Working Paper Series

3-28-2012

Testing for improvement in prediction model performance

Margaret S. Pepe PhD

Fred Hutchinson Cancer Research Center, nnoble@fhcrc.org

Kathleen F. Kerr

University of Washington, katiek@u.washington.edu

Gary M. Longton

Fred Hutchinson Cancer Research Center, glongton@fhcrc.org

Zheyu Wang

Fred Hutchinson Cancer Research Center, wangzy@u.washington.edu

Suggested Citation

Pepe, Margaret S. PhD; Kerr, Kathleen F.; Longton, Gary M.; and Wang, Zheyu, "Testing for improvement in prediction model performance" (March 2012). *UW Biostatistics Working Paper Series*. Working Paper 379.
<http://biostats.bepress.com/uwbiostat/paper379>

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

Copyright © 2011 by the authors

Testing for improvement in prediction model performance

Margaret Sullivan Pepe^{a*}, Kathleen F. Kerr^b, Gary Longton^a, Zheyu Wang^b

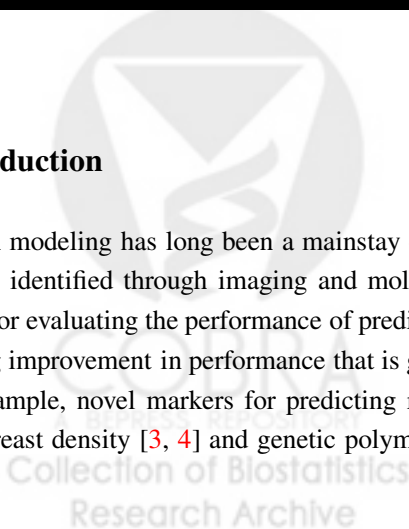
New methodology has been proposed in recent years for evaluating the improvement in prediction performance gained by adding a new predictor, Y , to a risk model containing a set of baseline predictors, X , for a binary outcome D . We prove theoretically that null hypotheses concerning no improvement in performance are equivalent to the simple null hypothesis that the coefficient for Y is zero in the risk model, $P(D = 1|X, Y)$. Therefore, testing for improvement in prediction performance is redundant if Y has already been shown to be a risk factor. We 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. Copyright © 2011 John Wiley & Sons, Ltd.

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 due to the promising 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, 6, 7]. For cardiovascular outcomes, numerous studies have



^aBiostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

^bDepartment of Biostatistics, University of Washington, F-600 Health Sciences Building, Campus Mail Stop 357232, Seattle, WA 98195, USA

* Correspondence to: Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500, Seattle, WA 98105, USA
E-mail: mspepe@u.washington.edu

been performed 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, the second step in the typical approach is to test a null hypothesis about improvement in prediction performance for the model that includes 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]. Alternate measures are also used, including risk redistribution metrics [10, 11] and risk reclassification metrics [12, 13, 14, 15, 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 Section 3 we discuss 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, but also on the new finding corroborated by other recent reports in the literature [14, 17], that standard ROC methods for nested models appear to be excessively conservative. We emphasize that estimation of the increment in prediction performance is more important than testing the null hypothesis of no improvement. The results are discussed in Section 4 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. To evaluate the incremental value of Y for prediction over use of X alone, the first step is often to test the null hypothesis

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

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 $\text{risk}(X, Y)$ while 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 is a classic plot for displaying discrimination achieved with a variable W [18] (Chapter 4). To test if discrimination provided by $\text{risk}(X, Y)$ is better than that provided by $\text{risk}(X)$, one could test

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

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

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

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 [19]. This is particularly relevant when controls are matched by design to cases on X [20]. The corresponding null hypothesis is

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

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, 21, 22]. We write the cumulative distribution of risk as $F_{(X,Y)}(p) = P(\text{risk}(X, Y) \leq p)$ and $F_X(p) = P(\text{risk}(X) \leq p)$. One can test if the risk distributions based on X or on (X, Y) are different by testing the null hypothesis

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

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

$$H_0 : F_{(X,Y)}^D(\cdot) = F_X^D(\cdot) \text{ and } F_{(X,Y)}^{\bar{D}}(\cdot) = F_X^{\bar{D}}(\cdot) \quad (6)$$

The integrated discrimination improvement statistic is a summary measure based on the difference in average risks between cases and controls, $MRD = E(\text{risk}(\cdot)|D = 1) - E(\text{risk}(\cdot)|D = 0)$. 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 [15, 23, 24]. Pencina and others [12] define the integrated discrimination improvement (IDI) as $IDI = MRD_{(X,Y)} - MRD_X$ and propose testing $H_0 : IDI = 0$. That is, they propose testing

$$H_0 : MRD_{(X,Y)} = MRD_X. \quad (7)$$

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 $AARD = P(\text{risk}(\cdot) > \rho|D = 1) - P(\text{risk}(\cdot) > \rho|D = 0)$. 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 below) statistic [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 $\leq \rho$ ' 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 [25] defined the total gain statistic as the area between the predictiveness curve for $\text{risk}(\cdot)$ and the horizontal line at ρ , which is the predictiveness curve for the null model. Gu and Pepe [24] showed that the standardized total gain, $\text{total gain}/2\rho(1 - \rho)$, is $\rho AARD$. One can compare the performance of two risk models by evaluating the AARD values and testing the null hypothesis

$$H_0 : AARD_{(X,Y)} = AARD_X \quad (8)$$

The medical decision making framework has also been used to compare risk models. Vickers and Elkin [26] 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

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

to compare performance of a model that includes Y with one that does not. Baker [27, 28] 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 i , $\text{risk}(X_i, Y_i)$ is compared directly with $\text{risk}(X_i)$. 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.

Theorem 1

The following null hypotheses are equivalent

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

Proof

That H_0^1 implies each of $H_0^2 - H_0^{10}$ is obvious. Therefore we focus on showing that each of $H_0^2 - H_0^{10}$ imply H_0^1 . We start with H_0^7 and work in reverse order through $H_0^6, H_0^5, \dots, H_0^2$. 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, Feng and Gu [23] write

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

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

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

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

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

Therefore if $\text{MRD}_{(X, Y)} - \text{MRD}_X = 0$ it follows that $E\{\text{risk}(X, Y) - \text{risk}(X)\}^2 = 0$ and so $\text{risk}(X, Y) = \text{risk}(X)$ 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: $E(\text{risk}(X, Y)|D = 1) = E(\text{risk}(X)|D = 1)$. Similarly $E(\text{risk}(X, Y)|D = 0) = E(\text{risk}(X)|D = 0)$. 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 [29].

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

A similar argument applies to the control specific distributions. Therefore equality of population risk distributions in H_0^5 implies equality of case 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 [30] derived the one-one mathematical relationship between the ROC curve for $\text{risk}(\cdot)$ and the predictiveness curve which characterizes the risk distribution. Therefore equality of ROC curves for $\text{risk}(X, Y)$ and $\text{risk}(X)$ 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 $\text{risk}(X, Y)$ and $\text{risk}(X)$ implies equality of the ROC curves, i.e. $H_0^2 \Rightarrow H_0^3$, from which H_0^1 follows. A fundamental result from decision theory is that decision rules of the form ' $\text{risk}(X, Y) > c$ ' 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 another decision rule based on (X, Y) . This result follows from Neyman-Pearson [31] and is discussed in detail in McIntosh and Pepe [32].

It follows that the ROC curve for $\text{risk}(X, Y)$ is at least as high at all points than the ROC curve for any other function of (X, Y) . In particular, the ROC curve for the function $\text{risk}(X)$ cannot exceed $\text{ROC}_{(X, Y)}(\cdot)$ at any point. Therefore, if the areas under $\text{ROC}_{(X, Y)}(\cdot)$ and $\text{ROC}_X(\cdot)$ are equal, 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 $\text{ROC}_{(X, Y)}(t_{(X, Y)}^\rho) - t_{(X, Y)}^\rho$ where $t_{(X, Y)}^\rho \equiv P(\text{risk}(X, Y) > \rho|D = 0)$. But, by definition of $t_{(X, Y)}^\rho$ and the ROC curve, we recognize $\text{ROC}_{(X, Y)}(t_{(X, Y)}^\rho) = P(\text{risk}(X, Y) > \rho|D = 1)$. Therefore Theorem A.1 states that if $P(\text{risk}(X, Y) > \rho|D = 1) - P(\text{risk}(X, Y) > \rho|D = 0) = P(\text{risk}(X) > \rho|D = 1) - P(\text{risk}(X) > \rho|D = 0)$ it follows that $\text{ROC}_{(X, Y)}(t) = \text{ROC}_X(t) \quad \forall t$. 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 $\text{NB}_{(X, Y)}(t) = \text{NB}_X(t) \quad \forall t$, then in particular we have equality at $t = \rho$: $\text{NB}_{(X, Y)}(\rho) = \text{NB}_X(\rho)$. Recall that $\text{NB}(t)$ is defined as

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

so at $t = \rho$ we have

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

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 $P(\text{risk}(Y) > \rho|D = 1) \geq P(\text{risk}(Y) > \rho|D = 0)$. The analogous statement when conditioning on X is that

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

But

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

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

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

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, $\text{ROC}_{Y|X}(f) = f \quad \forall f$.

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 $P(\text{risk}(Y) > \rho | D = 1) \geq P(\text{risk}(Y) > \rho | D = 0)$. Using Bayes' theorem this can be restated as

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

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

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

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

3. Recommendations for Hypothesis Testing

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 value?

There are many possible choices, but we focus here on the choice between a test for the coefficient for Y in a regression model of the risk function $\text{risk}(X, Y)$ and the change in the AUC for the ROC curves associated with estimated risk functions, $\text{risk}(X)$ and $\text{risk}(X, Y)$. To make the discussion concrete we consider the Wald test based on $\hat{\beta}_Y$ where β_Y is the coefficient for Y in a model for $\text{risk}(X, Y)$ and a test based on the difference $\Delta \widehat{\text{AUC}} = \widehat{\text{AUC}}_{(X, Y)} - \widehat{\text{AUC}}_X$ where $\widehat{\text{AUC}}$ 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 identically distributed observations, the Wald 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 $\hat{\beta}_Y$ is to be preferred. We see the power advantage demonstrated in the second row of the simulation results in Table 1 where the procedure based on $\Delta \widehat{\text{AUC}}$ 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 $\Delta \widehat{\text{AUC}}$ is equivalent to the nonparametric two-sample Wilcoxon statistic while $\hat{\beta}_Y$ 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-statistic is well known to have superior performance compared with the Wilcoxon statistic for normal data. That is, testing using $\hat{\beta}_Y$ is well known to be superior to testing using $\Delta \widehat{\text{AUC}}$ 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 Wald test more desirable than the $\Delta \widehat{\text{AUC}}$ 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 done regarding inference for the ROC performance of nested

models. The typical approach to testing with $\widehat{\Delta\text{AUC}}$ uses the fitted values for $\text{risk}(X, Y)$ and $\text{risk}(X)$ as data inputs to a test of equal AUCs for two diagnostic tests such as the DeLong test [33] or the resampling based test [34]. 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, while 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 1 that standard tests ignoring sampling variability in the estimated risk regression coefficients are extremely conservative. Both the DeLong test [33] that uses the normal approximation and a standard error formula and the test using percentiles of the bootstrap distribution [34] have size less than .005 with sample sizes as large as 100 cases and 900 controls. The conservatism is due to estimating the coefficients in the nested models since the same tests comparing X alone to another independent marker with equal performance were not conservative with comparable sample sizes (data not shown).

We implemented an alternative version of the $\widehat{\Delta\text{AUC}}$ 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. We resampled observations from the original dataset, fit the risk models, and calculated $\widehat{\Delta\text{AUC}}$ for each resampled dataset. Tests using percentiles of the bootstrap distribution (results shown in line 5 of Table 1 as $\widehat{\Delta\text{AUC}}$ -adjusted) or the standard deviation of the bootstrap distribution (results not shown) remained conservative, though they were less conservative than procedures not adjusting for variability in regression coefficients. We conclude that all currently available procedures for testing incremental value based on $\widehat{\Delta\text{AUC}}$ are unacceptably conservative in the classic scenarios we studied. From Table 1 we observe that as a consequence they have extremely low power compared with the Wald test for β_Y .

3.3. What if models are not well calibrated?

The results of Theorem 1 have implications for the analysis of data when the functional form of $\text{risk}(X, Y)$ is approximately correct. That is, for testing the null hypothesis of no improvement in prediction performance the theorem implies that we can test if regression coefficients for Y are zero, but this requires that the regression model for $\text{risk}(X, Y)$ is correctly specified at least under the null. If the form of the regression model for $\text{risk}(X, Y)$ under the null is in doubt, i.e., if $\text{risk}(X)$ is misspecified, tests based on the regression model may not be valid.

With misspecified model forms one might still use the risk modeling exercise to generate combination scores, $\widehat{\text{risk}}^*(X, Y)$ and $\widehat{\text{risk}}^*(X)$. Under these circumstances we believe that a case can be made for considering tests of increment in performance using statistics based on performance measures such as ΔAUC instead of tests based on the regression coefficient for Y . Note that with this approach one should derive the combination scores in a training set and test for equality of performance in an independent test dataset since our simulation results (Table 1) and others [14, 17] show that current testing procedures are not valid if both steps are undertaken in the same dataset.

It is our opinion however that poorly calibrated models are not acceptable in the context of risk prediction research. It is unethical to apply a model known to be poorly calibrated when calculating a person's risk. The assessment of calibration must be a priority and should precede assessment of prediction performance. A minimum requirement is weak calibration defined by :

$$\widehat{P}(D = 1 | \text{risk}(\cdot) = r) \approx r$$

so that of persons whose risks are calculated as r , approximately a proportion r are cases. If a model is poorly calibrated in the weak sense, it is simply not valid for the population and its performance characteristics are of less interest. One can and should at least recalibrate the models to the study population. The performance characteristics of the recalibrated models in the study population will then be of interest and can be assessed.

3.4. Recommendation

Our recommendation is to ensure the use of well calibrated risk models and to base hypothesis testing on $\widehat{\beta}_Y$ rather than on $\widehat{\Delta AUC}$. Procedures based on $\widehat{\Delta AUC}$ do not have correct size. Kerr et al [35] found similar problems for the IDI statistic under the null. It is possible that new approaches to testing based on $\widehat{\Delta AUC}$ could be developed to properly account for sampling variability in the fitted risk values and thereby yield appropriately sized tests. However, even if such procedures were developed, we have argued and observed in Table 1 (line 2) that tests based on $\widehat{\beta}_Y$ 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 $\widehat{\beta}_Y$ 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 $P(D = 1|X, Y) = \text{risk}(X, Y)$. A variety of measures to quantify the prediction performance of a risk model were described 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 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.

4. 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. Data for 426 patients who were surgically assessed for renal stenosis were reported by Janssens and others [36]. 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 ($n = 142$) 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

$$X = 0.93 \times \text{age} - 0.24 \text{ BMI} + 1.58 \times \text{bruit}.$$

We then evaluated the performances of risk models based on X and on the combination of X and $Y = \log(\text{serum creatinine})$ using the remaining two thirds of the data ($n = 284$).

Linear logistic models were fit:

$$\text{Baseline: } \text{logit}P(D = 1|X) = \alpha_0 + \alpha_1 X$$

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

Figure 1 shows that these models are well calibrated to the study cohort since observed event rates in each decile of modeled risk (shown as open circles) are approximately equal to the modeled risks (shown as the points on the risk quantile curves). Hosmer-Lemeshow goodness of fit statistics [10, 38] do not provide any evidence against the null of good calibration (p -values of 0.39 for the baseline model, $\text{risk}(X)$, and 0.98 for the enhanced model, $\text{risk}(X, Y)$).

The Wald test for $H_0 : \beta_2 = 0$ is highly significant with $p < 0.001$ (Table 2). According to Theorem 1 we can conclude that prediction model performance is improved by addition of Y to the model. Nevertheless we implemented tests based on ΔAUC as well to compare inference. The test for equality of AUCs is also significant but with much weaker p -value, $p = 0.014$, using the DeLong variance formula and $p = 0.012$ using percentiles of the bootstrap distribution. Recall that these tests do not acknowledge variability in the estimated regression coefficients $(\hat{\beta}_1, \hat{\beta}_2)$ and are extremely conservative. Bootstrapping that incorporated refitting the risk model in each resampled dataset yielded a stronger p -value ($p < 0.001$). However, in accordance with our recommendation in Section 3, the test based on $\hat{\beta}_2$ yielded the strongest evidence that prediction performance is 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 ($p = 0.009$, Table 2) while the standard tests based on ΔAUC are not ($p = 0.12$ using the DeLong variance formula and $p = 0.13$ using percentiles of the bootstrap distribution). The bootstrapped adjusted ΔAUC test that refits the models is significant, $p = 0.02$, but not as convincing as the test for Y^* in the risk model ($p = 0.009$). Again, this is consistent with our simulation results and theoretical expectations and 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, $\text{risk}(X, Y)$.

Estimates of prediction performance are shown in Table 3 for the baseline and enhanced risk models. Confidence intervals were calculated using 2.5th and 97.5th percentiles of bootstrap distributions with models refit in each bootstrapped dataset. We estimated that the area under the ROC curve increased from 0.78 to 0.82 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 3, decreases from 0.44 to 0.31. Note that Pfeiffer and Gail [37] 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 0.55 to 0.45. The IDI statistic is the change in the MRD statistic and is calculated as 0.06 while the conceptually similar change in the AARD is 0.03. The continuous-NRI statistic is $\text{NRI}(> 0) = 0.54$. Note that the NRI is measured on a scale from 0 to 2, unlike most other 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 3 times the net cost of the diagnostic procedures for a subject without stenosis since the cost-benefit ratio = risk threshold/(1-risk threshold) [26]. 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 27.1% of maximum with use of the baseline model and 36.4% of maximum with use of the model that includes serum creatinine. We see that 95% confidence intervals for some but not all measures of improvement in performance exclude the null value of 0. However the single test of $H_0 : \beta_2 = 0$ is sufficient to conclude that performance is improved.

5. 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: (i) that such tests are most powerful asymptotically; and (ii) that techniques are well developed and widely available for performing such tests. This strategy relies on employing risk models that have approximately correct forms. 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. After recalibrating the models, testing based on regression coefficients is the most powerful and reliable approach to detecting if there is *any* improvement in predictor performance by adding a predictor to a baseline model.

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 [37] and approaches based on medical decision making [7, 26, 28, 39]. This paper does not seek to provide guidance on the choice of measure, but we do emphasize that estimation of the improvement gained is crucial. 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 \leq 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.

Acknowledgments

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

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, 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**:1037–41, 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 J. W., Morris D.E. 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. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
19. 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.
20. 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.
21. 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.
22. 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.
23. 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.
24. Gu JW, 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.
25. 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.
26. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 2006; **26**:565–574.
27. 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
28. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. *Jornal of the Royal Statistical Society Series B* 2009; **172**:729–748. doi: 10.1111/j.1467-985X.2009.00592.x
29. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics* 2005; **6**:227–239, DOI: 10.1093/biostatistics/kxi005.
30. 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.
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.
32. McIntosh MS, Pepe MS. Combining several screening tests: Optimality of the risk score. *Biometrics* 2002; **58**:657–64.
33. 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.
34. Pepe M, Longton G, Janes H. Estimation and comparison of receiver operating characteristic curves. *Stata Journal* 2009; **9**:1–16.
35. 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
36. 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.
37. 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.
38. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. Wiley: New York, 2000.
39. 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* 2011 Early view, DOI: 10.1002/sim.4362

Table 1. Performance of two-sided nominal 0.05 level tests. Tests are based on $\hat{\beta}_Y$, the estimated regression coefficient for Y in the risk model $\text{logit risk}(X, Y) = \beta_0 + \beta_X X + \beta_Y Y$ and on $\Delta\widehat{\text{AUC}} = \widehat{\text{AUC}}_{(X, Y)} - \widehat{\text{AUC}}_X$. Tests based on $\Delta\widehat{\text{AUC}}$ were: ‘adjusted’ if regression coefficients were estimated from each bootstrap resampled dataset; and ‘standard’ if bootstrap resampling (bootstrap) or DeLong standard error (se) calculation used $\widehat{\text{risk}}(X, Y)$ 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 $Y \sim N(0.37, 1)$ in cases under the alternative. 1000 simulations for each scenario and 1000 bootstrap samples per analysis.

Test Statistic	Size ($\beta_Y = 0$)		Power ($\beta_Y = 0.37$)		Power ($\beta_Y = 0.74$)	
	$n_0 = n_D = 50$	$n_D = 100, n_D = 900$	$n_D = n_D = 50$	$n_D = 100, n_D = 900$	$n_D = n_D = 50$	$n_D = 100, n_D = 900$
$\hat{\beta}_Y$	0.048	0.051	0.388	0.908	0.928	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

[†]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.

Table 2. Logistic regression models for risk of renal artery stenosis fit to data for 284 patients. The addition of $Y = \log(\text{serum creatinine})$ to a model including the baseline covariate $X = 0.93 \text{ age} - 0.24 \text{ BMI} + 1.58 \text{ bruit}$ 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.

	Intercept	X	Y or Y^*
Baseline Model (X)			
coefficient	0.03	0.74	—
se	0.22	0.11	—
p -value	0.88	<0.001	—
Enhanced Model (X, Y)			
coefficient	-0.20	0.66	0.70
se	0.24	0.12	0.19
p -value	0.39	<0.001	<0.001
Enhanced Model (X, Y^*)			
coefficient	-0.05	0.71	0.34
se	0.23	0.12	0.13
p -value	0.81	<0.001	0.009



Table 3. 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	0.82	0.04 (0.01,0.08)
FPR at TPR=0.8	$\text{ROC}^{-1}(0.8)$	0.44	0.31	-0.12 (-0.22,0.06)
Mean Risk Difference	MRD	0.20	0.26	0.06* (0.015,0.13)
Above Average Risk Difference	AARD	0.43	0.46	0.03** (-0.05,0.13)
Continuous NRI	NRI (> 0)	—	—	0.54 (0.23,0.84)
Net Benefit at 0.25	NB (0.25)	6.4%	8.6%	2.2% (-0.6%,5.1%)

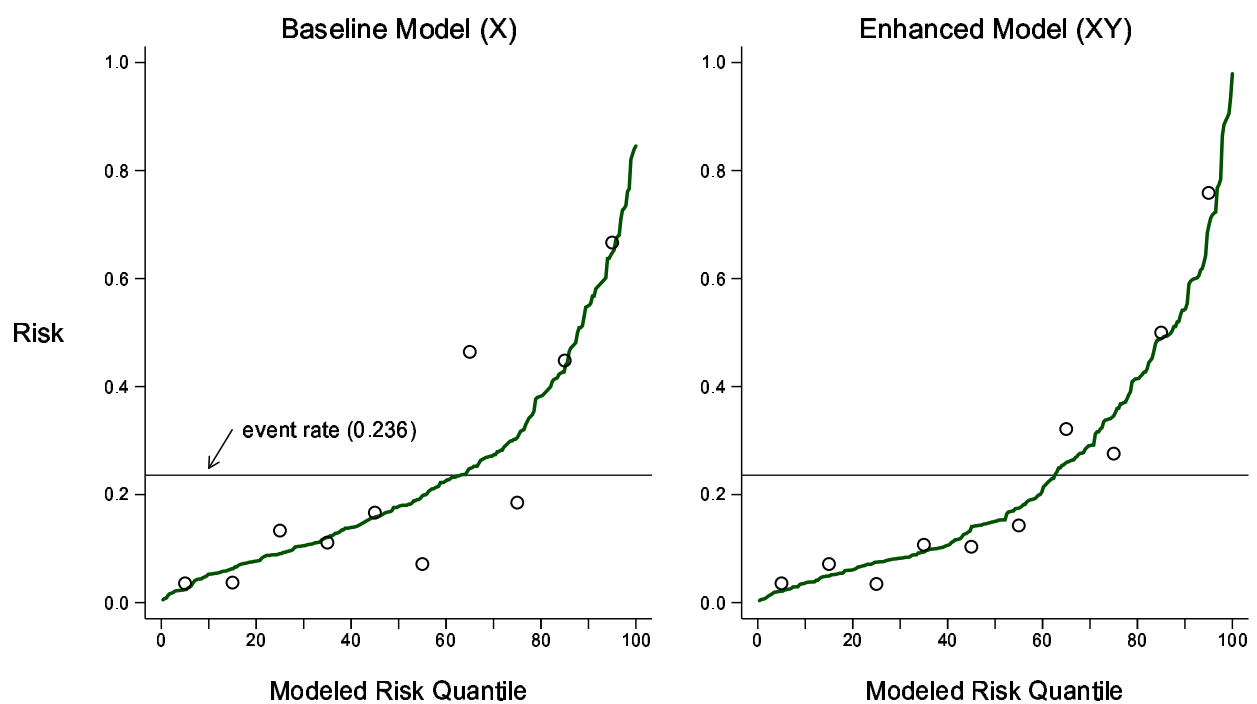
[†] Performance improvement is defined as the difference between the measure for the enhanced model and that for the baseline model for all measures except for the NRI.

* Also known as the IDI statistic.

** Also known as the Total Gain statistic.



Figure 1. Predictiveness curves to assess calibration of baseline and enhanced risk models for renal artery stenosis. 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 0.39 (baseline model) and 0.98 (enhanced model).



Appendix

We use the following notation

$$t_X^\rho \equiv 1 - F_X^{\bar{D}}(\rho) = P(\text{risk}(X) > \rho | D = 0)$$

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

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

$$\begin{aligned} \text{ROC}_{(X,Y)}(t_{(X,Y)}^\rho) - t_{(X,Y)}^\rho &= \text{ROC}_X(t_X^\rho) - t_X^\rho & (\text{A.1}) \\ \Leftrightarrow \text{ROC}_{(X,Y)}(t) &= \text{ROC}_X(t) \quad \forall t \end{aligned}$$

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 (Pepe 2003, page 71 [18]). Therefore $\text{ROC}_{(X,Y)}(t) - t$ has a unique maximizer. Moreover, the maximizer occurs when $\text{ROC}'_{(X,Y)}(t) = 1$. Arguments below in the proof of Corollary A.1 show that this implies $\text{ROC}_{(X,Y)}(t) - t$ is maximized at $t_{(X,Y)}^\rho$.

Since $\text{ROC}_{(X,Y)}(t) \geq \text{ROC}_X(t) \quad \forall t$, we have

$$\text{ROC}_{(X,Y)}(t_X^\rho) - t_X^\rho \geq \text{ROC}_X(t_X^\rho) - t_X^\rho$$

and equation (A.1) implies therefore that

$$\text{ROC}_{(X,Y)}(t_X^\rho) - t_X^\rho \geq \text{ROC}_{(X,Y)}(t_{(X,Y)}^\rho) - t_{(X,Y)}^\rho.$$

It follows that $t_X^\rho = t_{(X,Y)}^\rho$ because, as noted above, $\text{ROC}_{(X,Y)}(t) - t$ has a unique maximizer at $t_{(X,Y)}^\rho$. 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 $t_{(X,Y)}^\rho$ and t_X^ρ .

Next we show that $\text{ROC}'_X(t) \leq \text{ROC}'_{(X,Y)}(t)$ 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 construct decision rules based on (X, Y) with 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 and consider the family of decision rules that classify positive if $\{ \text{risk}(X, Y) > r^a_{(X,Y)} \}$ or $[\text{risk}(X, Y) < r^a_{(X,Y)} \text{ and } \text{risk}(X) < r^a_X \text{ and } \text{risk}(X) > k \text{ for } k > r^b_X]$. These decision rules have an ROC curve equal to $\text{ROC}_{(X,Y)}^{(t)}$ at $t = a$ and with derivative 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 $\text{ROC}_{(X,Y)}(t) \geq \text{ROC}_X(t)$ at all t .

Recall from above that

$$0 = \text{ROC}_{(X,Y)}(t^\rho) - \text{ROC}_X(t^\rho) = \int_0^{t^\rho} (\text{ROC}'_{(X,Y)}(t) - \text{ROC}'_X(t)) dt. \quad \text{Hosted by The Berkeley Electronic Press}$$

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

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

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

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

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

■

Corollary A.1

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

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

Proof

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

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

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

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

but (A.2) then implies that $\sup|\text{ROC}_\omega(t) - t| = 0$. In other words $\text{ROC}_\omega(t) = t \quad \forall \quad t \in (0, 1)$. Note that equation (A.3) also follows from the fact that both sides of (A.3) were show to equal the standardized total gain statistic (see equations (6) and (7) of Gu and Pepe [24]).

■

