Memorial Sloan-Kettering Cancer Center
Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology
& Biostatistics Working Paper Series

Year 2011 *Paper* 20

Comparing ROC Curves Derived From Regression Models

Venkatraman E. Seshan* Mithat Gonen†
Colin B. Begg‡

*Memorial Sloan-Kettering Cancer Center, seshanv@mskcc.org
†Memorial Sloan-Kettering Cancer Center, gonenm@mskcc.org
‡Memorial Sloan-Kettering Cancer Center, beggc@mskcc.org

http://biostats.bepress.com/mskccbiostat/paper20

Comparing ROC Curves Derived From Regression Models

Venkatraman E. Seshan, Mithat Gonen, and Colin B. Begg

Abstract

In constructing predictive models, investigators frequently assess the incremental value of a predictive marker by comparing the ROC curve generated from the predictive model including the new marker with the ROC curve from the model excluding the new marker. Many commentators have noticed empirically that a test of the two ROC areas often produces a non-significant result when a corresponding Wald test from the underlying regression model is significant. A recent article showed using simulations that the widely-used ROC area test [1] produces exceptionally conservative test size and extremely low power [2]. In this article we show why the ROC area test is invalid in this context. We demonstrate how a valid test of the ROC areas can be constructed that has comparable statistical properties to the Wald test. We conclude that using the Wald test to assess the incremental contribution of a marker remains the best strategy. We also examine the use of derived markers from non-nested models and the use of validation samples. We show that comparing ROC areas is invalid in these contexts as well.

Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Comparing ROC Curves Derived From Regression Models

V. E. Seshan, M. Gönen*, C. B. Begg

In constructing predictive models, investigators frequently assess the incremental value of a predictive marker by comparing the ROC curve generated from the predictive model including the new marker with the ROC curve from the model excluding the new marker. Many commentators have noticed empirically that a test of the two ROC areas [1] often produces a non-significant result when a corresponding Wald test from the underlying regression model is significant. A recent article showed using simulations that the widely-used ROC area test produces exceptionally conservative test size and extremely low power [2]. In this article we show why the ROC area test is invalid in this context. We demonstrate how a valid test of the ROC areas can be constructed that has comparable statistical properties to the Wald test. We conclude that using the Wald test to assess the incremental contribution of a marker remains the best strategy for nested models. We also examine the use of derived markers from non-nested models and the use of validation samples. We show that comparing ROC areas is invalid in these contexts as well.

Keywords: receiver operating characteristic curve, biomarker, predictive model, area under the ROC curve, logistic regression, predictive accuracy, discrimination

1. Introduction

Receiver operating characteristic (ROC) curves provide a standard way of evaluating the ability of a continuous marker to predict a binary outcome. The area under the ROC curve (AUC) is a frequently used summary measure of diagnostic/predictive accuracy. Comparison of two or more ROC curves is usually based on a comparison of the area measures. The standard method comparing AUCs is a non-parametric test [1], hereafter referred to as the "AUC test," although a method developed earlier is also used widely [3]. The AUC test uses the fact that the AUC is a U-statistic and incorporates the dependencies caused by the fact that the markers are usually generated in the same patients, and are thus "paired."

Although the AUC test was originally developed in the context of comparing distinct diagnostic tests or markers, it has increasingly been adopted for use in evaluating the incremental effect of an additional marker in predicting a binary event via a regression model. Indeed authors of several methodological articles on predictive modeling have advocated

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, 10065, USA
** Correspondence to:* gonenm@mskcc.org

Contract/grant sponsor: National Cancer Institute Award CA 136783

the use of ROC curves for this purpose though these groups have generally not advocated statistical testing of the ROC curves specifically [4, 5, 6, 7, 8]. In this setting investigators typically use the fitted values from the regression model to construct an ROC curve and to compare this with the ROC curve derived similarly from the fitted values from the regression excluding the new marker. However, a recent article provided several examples of the use of this strategy in the literature, and demonstrated using simulations that the AUC test has exceptionally conservative test size in this setting, and much lower power than the Wald test of the new marker in the underlying regression model [2] .

We show here that in the context of comparing AUCs from fitted values of two nested regression models the AUC test is invalid. That is the nominal reference distribution does not approximate the distribution of the test statistic under the null hypothesis of no difference between the models. Note that we use the term "compare" to refer to a significance test and not the evaluation of of the incremental predictive ability by various summary measures developed for this purpose, see, for example, Pencina et. al. [9] and the accompanying discussion articles. We show that comparing models using a formal test of ROC areas is valid only if the reference distribution is constructed in recognition of the induced correlations of the predictors from different patients from the fitted models, and also by recognizing an analytical artifact that is invariably adopted in this setting. We also develop and present a procedure that produces the correct reference distribution and remains valid in this context. It turns out, however, that the operating characteristics of the proposed procedure are indistiguishable from those of the Wald test and there seems to be no particular advantage in its use. We use the Wald test as a benchmark in recognition of the well-known result that Wald test is asymptotically equivalent to the likelihood ratio test and the score test [10, Chapter 9]. We also consider the related problems of comparing derived predictors from non-nested regression models and performing the comparison in vaildation samples. We show that the AUC test is invalid in these cases as well.

2. AUC Test for Nested Binary Regression Models

2.1. A Review of the AUC Test

Consider first the comparison of the predictive accuracy of two independently generated predictive markers, denoted W_{1i} and W_{2i} for cases $i = 1, ..., n$. We are interested in predicting a binary outcome Y_i, where $Y_i = 1$ or 0. The estimate of the AUC for marker k can be written as a U-statistic:

$$\hat{A}_k = \frac{1}{n_0 n_1} \sum_{i=1, Y_i=1}^{n} \sum_{j=1, Y_j=0}^{n} I(W_{ki} > W_{kj}) + \frac{1}{2} I(W_{ki} = W_{kj}) \quad (1)$$

where $n_0 = \sum_{j=1}^{n} I(Y_j = 0)$, $n_1 = \sum_{i=1}^{n} I(Y_i = 1)$ and $I(.)$ is the indicator function. The AUC is equivalent to the Mann-Whitney estimate of the probability that a randomly selected marker with a positive outcome is greater than a randomly selected marker with a negative outcome.

It is important to note that (1) assumes that high marker values are more indicative of positive outcomes, $Y_i = 1$, than low marker values. This assumption, often relegated to small print or overlooked, is highly consequential for our purposes. We will call it *known directionality*. Known directionality accompanies most single-marker analyses but this is not the case for markers derived as predictions from multivariable regression models.

An estimate of the difference between A_2 and A_1 is given by $\hat{\delta} = \hat{A}_2 - \hat{A}_1$. DeLong et. al. [1] derived a consistent estimate for the variance $\hat{V}$=Var($\hat{\delta}$) and proposed the test statistic $T = \hat{\delta}/\sqrt{\hat{V}}$ which has an asymptotic standard normal distribution under the null hypothesis that $\delta = 0$. We heretofore refer to this as the AUC test.

2.2. AUC test is invalid with nested binary regression models

The original derivation of the AUC test assumes that the two markers are to be compared head-to-head [1]. If the goal is to evaluate the incremental value of a marker in the presence of another marker then the AUC test cannot be used directly. Instead one needs to create a "composite" marker that captures the combined effect of the two markers, and then compare this with the first marker. This aggregation of predictive information is usually accomplished using regression. For example in the setting of logistic regression we would compare the first marker $W_1 = \{W_{1i}\}$ with a composite marker derived from the risk predictors from a logistic regression of $Y = \{Y_i\}$ on W_1 and $W_2 = \{W_{2i}\}$. Frequently there are other variables (Z) in the regression and so the comparison is between two composite predictors, derived from the following two models:

$$\mathrm{M}_1: \quad \mathrm{logit}(Y_i) = \quad \beta_0 + \beta_1 W_{1i} + \theta' Z_i \tag{2}$$

$$\mathrm{M}_2: \quad \mathrm{logit}(Y_i) = \quad \beta_0 + \beta_1 W_{1i} + \beta_2 W_{2i} + \theta' Z_i. \tag{3}$$

In this context we are interested in testing the null hypothesis that $\beta_2 = 0$. One can then form the linear predictors using the MLEs of the parameters

$$W^*_{1i} \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 W_{1i} + \hat{\theta}' z_i \tag{4}$$

$$W^*_{2i} \quad = \quad \tilde{\beta}_0 + \tilde{\beta}_1 W_{1i} + \tilde{\beta}_2 W_{2i} + \tilde{\theta}' z_i \tag{5}$$

Let $\hat{A}^*_1$ and $\hat{A}^*_2$ denote the AUCs estimated from (1) using W^*_1 and W^*_2 in place of W_1 and W_2. Also let $\hat{\delta}^* = \hat{A}^*_2 - \hat{A}^*_1$ and let T^* denote the test statistic corresponding to $\hat{\delta}^*$ calculated in the manner outlined in Section 2.1. We heretofore refer to T^* as the AUC test statistic. As reasonable and straightforward as it seems, the comparison using the AUC test in this manner is not valid for two reasons.

The first reason is that the variance estimate $\hat{V}^*$ is based on the assumption that the observations from the patients are mutually independent, i.e. $(W^*_{1i}, W^*_{2i}) \perp (W^*_{1j}, W^*_{2j})$ for all $i \neq j$. With W^*_{1i} and W^*_{2i} defined as in (4-5) this assumption is clearly violated. In fact, typically (W^*_{1i}, W^*_{2i}) and (W^*_{1j}, W^*_{2j}) are strongly correlated, as we demonstrate later in Section 4.

The second reason concerns the construction of $\hat{A}^*_k$ as defined in (1). From the perspective of predictive accuracy of an individual marker it should make a difference whether W_{1i} is ranked from the smallest to the largest or from the largest to the smallest. We know whether a high value of a diagnostic test should be associated with increased risk of disease. That is, we know a priori how to order W with respect to Y. If we define $U^*_{ki} = -W^*_{ki}$, $\forall_i$, and the AUC for U^*_{ki} to be A^{*-}_k then it is easily shown that $A^*_k + A^{*-}_k = 1$. With known directionality, an AUC estimate less than 0.5 is admissible, though it would be recognized that the decrement from 0.5 is likely to be due to random variation. But in the context of a regression model it is not possible to invoke known directionality. In this context the model is constrained to choose the ordering that leads to an increase in the area estimate. That is, for M_2 for example, if W_2 is observed to be positively associated with Y_i after adjusting for W_1 and Z then the sign of $\tilde{\beta}_2$ will usually be positive. If, on the other hand W_2 is observed to be negatively associated with Y then the sign of $\tilde{\beta}_2$ will usually be negative. Either way, the net effect will be to increment the AUC estimate upwards. This is especially problematic when testing the null hypothesis that $\beta_2 = 0$. The reference distribution for the AUC test is constructed under the assumption that half the time the results should lead to a decrement in AUC, but in fact this rarely happens. This creates a bias in T^* such that it no longer has zero mean under the null hypothesis. However, as the true value of β_2 increases the probability of observing a negative residual association of W_2 and Y by chance becomes less likely.

3. A Valid Procedure Based on Projection and Permutation to test $H : \delta = 0$

The previous section makes clear that the problems with the use of the AUC test stem from the artifacts related to the use of regression in conjunction with the estimation of the AUCs. In this section we pursue a modification to the derivation of the reference distribution for the AUC test to demonstrate that it is possible to construct a valid AUC test in this context. We construct an orthogonal decomposition of W_2:

$$W_2 = W_2^p + W_2^c = PW_2 + (I - P)W_2,$$

where $P = (X'X)^{-1}X$ and $X = (1\ W_1\ Z)$. That is, W_2^p is the projection of W_2 on to the vector space spanned by $(1\ W_1\ Z)$ and W_2^c is the orthogonal complement of W_2^p. By definition W_2^c is uncorrelated with $(W_1^*\ Z)$ and hence all the information in W_2 that is incremental to $(W_1\ Z)$ must be contained in W_2^c. Under the null hypothesis W_2^c forms an exchangeable sequence. This suggests that permuting W_2^c and fitting the same logistic regression models as in Section 2 will generate a realization of the data generating mechanism under the null hypothesis. We have examined this conjecture in Section 4 using a reference distribution for T^* in which the test statistic is calculated after repeated permutations of W_2^c.

To summarize, the Projection-Permutation reference distribution is constructed as follows:

1. Compute the projection matrix $P = (X'X)^{-1}X$ where $X = (1W_1Z)$.
2. Compute $W_2^p = PW_2$ and $W_2^c = (I - P)W_2$
3. Obtain a permutation of the vector $\{W_{2i}^c\}$, call it $W_{2,perm}^c$
4. Construct $W_{2,perm} = W_2^p + W_{2,perm}^c$
5. Fit models M_1 and M_2 replacing W_2 with $W_{2,perm}$ from Step 4, and compute the area test statistic T^*
6. Repeat 3-5 B times
7. Construct the reference distribution from the B values of the test statistic computed in step 6.

4. Comparison of the Tests in Nested Models

The extent and magnitude of the problem explained in Section 2 has been investigated in detail by Vickers et. al. [2] who performed simulations to show that the use of the AUC test in nested regression contexts is problematic under several scenarios. In the following we have reproduced simulations constructed in the same way as in [2] with the objective of establishing the validity of the permutation procedure outlined in Section 3. Details of the data generation are provided in [2], but briefly the simulations are constructed as follows. The outcomes $\{Y_i\}$ are generated as Bernoulli random variables with probability 0.5. We note that Vickers et al [2] examined additional input probabilities and found extensive bias regardless of this choice. Pairs of marker values $\{W_{1i}, W_{2i}\}$ are generated as bivariate standard normal variates with correlation ρ, conditional on $\{Y_i\}$. The mean is $(0, 0)$ when $Y_i = 0$ and (μ_1, μ_2) when $Y_i = 1$. Two logistic regressions are then performed: $\text{logit}(Y_i) = \beta_0 + \beta_1 W_{1i}$ and $\text{logit}(Y_i) = \beta_0 + \beta_1 W_{1i} + \beta_2 W_{2i}$, and pairs of predictors $\{W_{1i}^*, W_{2i}^*\}$ generated as in Section 2.2. These are analyzed using the Wald test (for testing $\beta_2 = 0$), the AUC test based in the statistic T^* [1], and the Projection-Permutation test of the AUCs described in Section 3.

Results are displayed in Table 1 for different combinations of ρ, μ_1, μ_2, and for two different sample sizes. Note that μ_2 represents the unconditional predictive strength of W_2, with the null hypothesis equivalently represented by $\mu_2 = 0$ and $\beta_2 = 0$. The parameter μ_1 represents the underlying predictive strength of W_1. These results indicate clearly that the AUC test is extremely insensitive with exceptionally conservative test size and low power. Under the null hypothesis, i.e. when $\mu_2 = 0$, the AUC test is significant typically only once in 5000 simulations (at the nominal 5% significance level) whereas the Wald test has approximately the correct size. As μ_2 moves away from 0, the AUC test is substantially inferior

Table 1. Size ($\mu_2 = 0$) and power ($\mu_2 > 0$) of the tests for n=250 and 500.

		n=250				n=500			
	μ_1	0		0.3		0		0.3	
μ_2	ρ	0	0.5	0	0.5	0	0.5	0	0.5
0	Wald Test	0.04	0.05	0.06	0.05	0.05	0.05	0.06	0.05
	Standard AUC Test	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	Projection AUC Test	0.04	0.05	0.06	0.04	0.05	0.05	0.06	0.05
0.1	Wald Test	0.12	0.14	0.12	0.16	0.19	0.24	0.18	0.25
	Standard AUC Test	0.01	0.02	0.00	0.02	0.02	0.04	0.01	0.01
	Projection AUC Test	0.11	0.15	0.11	0.15	0.17	0.23	0.18	0.22
0.2	Wald Test	0.36	0.44	0.36	0.43	0.59	0.75	0.60	0.70
	Standard AUC Test	0.08	0.11	0.04	0.06	0.18	0.29	0.11	0.17
	Projection AUC Test	0.36	0.42	0.32	0.41	0.57	0.72	0.53	0.68
0.3	Wald Test	0.66	0.76	0.67	0.75	0.93	0.96	0.91	0.96
	Standard AUC Test	0.22	0.33	0.13	0.22	0.56	0.73	0.39	0.54
	Projection AUC Test	0.63	0.75	0.65	0.72	0.91	0.95	0.88	0.95

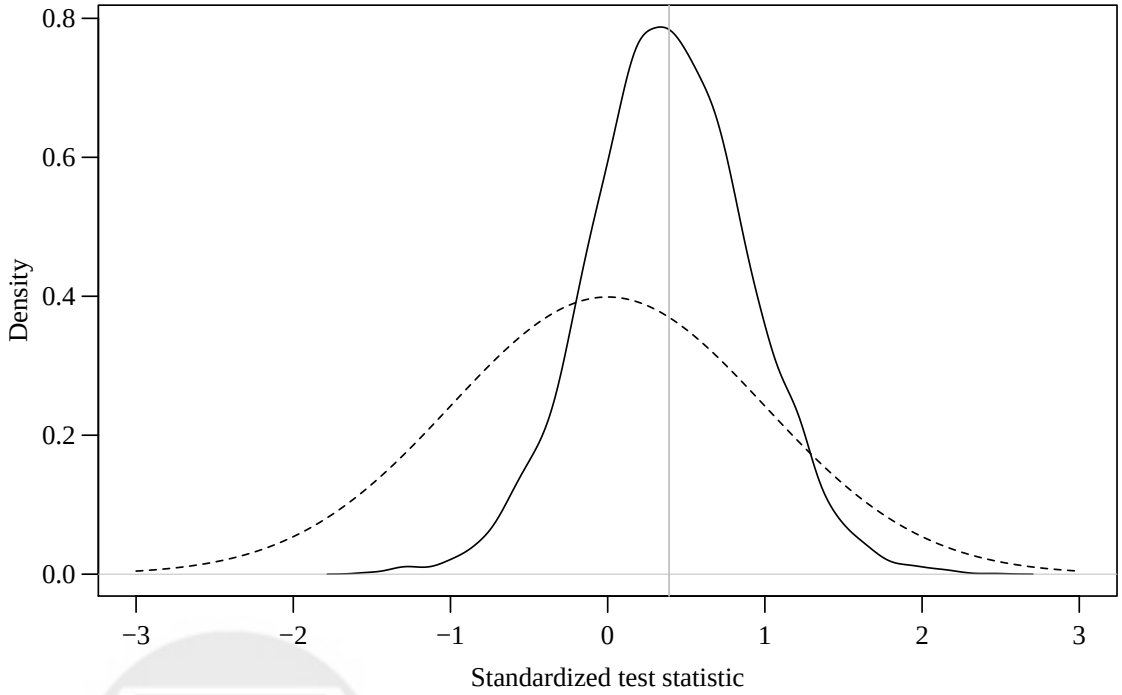

Figure 1. Distribution of the AUC test statistic under the null hypothesis. The solid curve depicts the density of the observed test statistic from 5000 simulations and the dashed curve is a standard normal, the presumed asymptotic distriution of the test statistic under the null hypothesis. Data are generated using $\mu_1 = 0.3$, $\mu_2 = 0$, $\rho = 0$ and $n = 500$.

to the Wald test in terms of power, due to the same factors that make it extremely conservative in the case of $\mu_2 = 0$. We also see that doubling the sample size from 250 to 500 does not remedy the problem. This is expected because the bias involved in the estimation of the mean and the variance of the AUC test statistic does not diminish with increasing sample sizes. On the other hand the Projection-Permutation test has the correct size and comparable power to the Wald test. Poor performance of the AUC test is largely unaffected by the degree of correlation between the markers (represented by ρ).

To better illustrate the biases in using the AUC test we display in Figure 1 results from a specific run of simulations ($\mu_1 = 0.3$, $\mu_2 = 0$ and $\rho = 0$ with $n = 500$). The horizontal axis is the standardized AUC test statistic T^* which should have zero mean and unit variance. The solid line is the kernel density estimate of the observed test statistic over 5000 simulations; it has mean 0.353 and variance 0.232. The dotted line is a standard normal density which is the reference

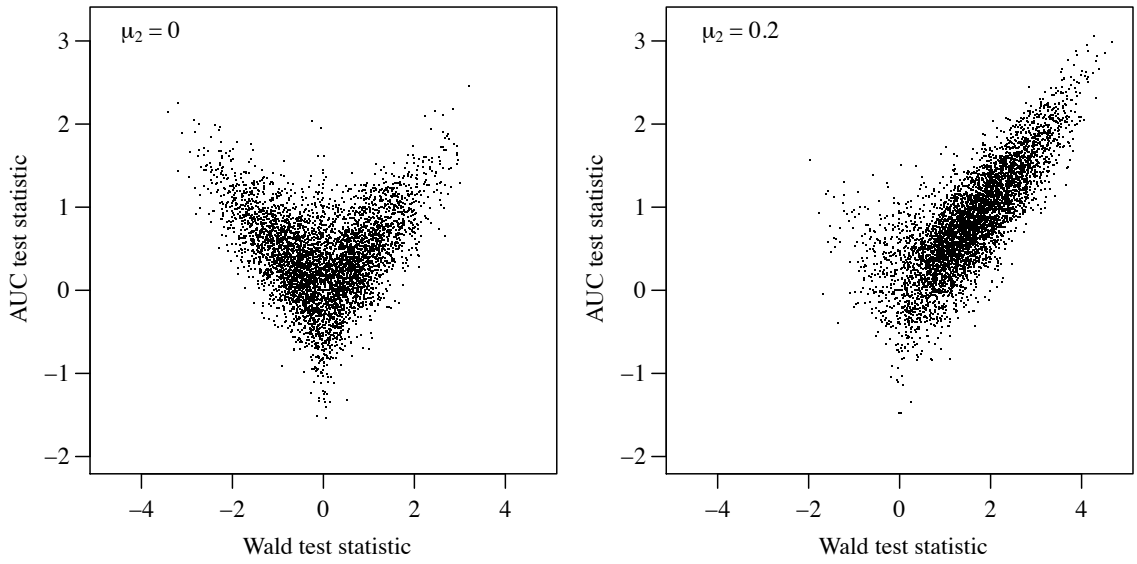

Figure 2. Wald statistic and the AUC test statistic under the null (left panel) and alternative (right panel) hypotheses. Both graphics are generated using 5000 draws from M_2 with $\mu_1 = 0.3$ and $n = 500$ for both panels, $\mu_2 = 0$ for the left panel and 0.2 for the right panel ($\rho = 0$ for both panels).

distribution of the AUC test statistic calculated in the conventional way [1]. Clearly, the difference between the two densities is substantial, both with respect to mean and variance.

The occasional negative values of the AUC test statistic in Figure 2 reveal another source of discrepancy between the AUC and Wald tests. This is due to dissonance between the maximum likelihood and AUC estimates. Since parameter estimates are based on maximizing the likelihood, there will be some data sets where the residual association between W_2^* and Y is positive yet the corresponding the AUC estimate is less than 0.5. In other words, the parameter values that maximize the likelihood result in a positive Wald test statistic but a decrement in the AUC. In our simulations under the null we observed a negative AUC test statistic approximately 20% of the time. This phenomenon becomes less common when W_2 has incremental information, since the estimated coefficient is not only positive but also distant from 0 in most circumstances.

Figure 3 illustrates the validity of the Projection-Permutation test graphically for $n = 250$ and $\rho = 0.5$. The empirical density of the difference in areas ($\hat{\delta}^*$) under the null hypothesis (i.e., when $\mu_2 = 0$) is given by the black curve and the density of the reference distribution of the Projection-Permutation test is given by the red curve. The two are almost exactly the same, establishing the validity of the test. Further, the blue curve depicting the reference distribution under the alternative hypothesis, i.e. when $\mu_2 = 0.3$ is virtually identical to the black and red curves, showing that the null distribution is computed correctly when the data are generated under the alternative hypothesis. The green curve represents the distribution of $\hat{\delta}^*$ under this alternative.

It is instructive to examine graphically the way the bias in the mean of the AUC test statistic operates. Figure 2 plots the standardized Wald test statistic against the AUC test statistic for two scenarios: no incremental information in W_2 (left panel) and strong incremental information in W_2 (right panel). In both cases the statistics should exhibit strong positive correlation. For the left panel, since there is no incremental information, the residual association between W_2 and Y is negative approximately half of the time. For the right panel, as indicated in Section 2.2, the resulting MLE of $\hat{\beta}_2$ will typically be positive in these cases indicating a positive impact on prediction, and the AUC statistic is correspondingly positive. Hence the V-shaped pattern on the left and this V-shape is the source of the bias described in Section 2.2. The magnitude of the problem becomes smaller as the signal increases; the figure on the right exhibits the positive correlation between the two tests that we would expect. This is because it is increasingly unlikely that the residual association between W_2 and Y is negative in these circumstances.

Figure 2 explains the source of the discrepancy with respect to the location of the two densities in Figure 1. An

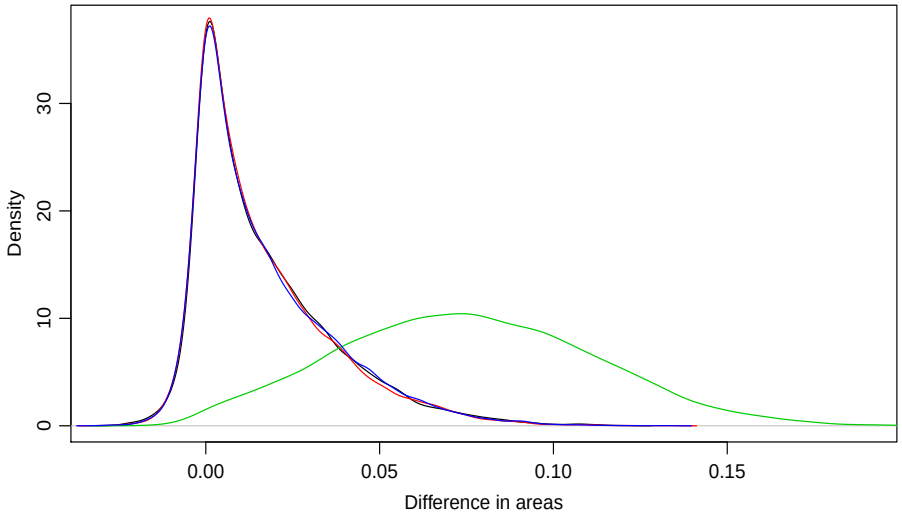

Figure 3. Distribution of $\hat{\delta}$ for $\rho = 0.5$. The black ($\mu_2 = 0$) and green ($\mu_2 = 0.3$) curves are estimated from the data over 10000 simulations. The red ($\mu_2 = 0$) and blue ($\mu_2 = 0.3$) curves are estimated from the reference distribution of the Projection-Permutation test.

explanation of the scale discrepancy is elusive graphically but possible by examining the following between subject correlations: $\rho_1^* = \text{Cor}(W_{1i}^*, W_{1j}^*)$, $\rho_2^* = \text{Cor}(W_{2i}^*, W_{2j}^*)$ and $\rho_{12}^* = \text{Cor}(W_{1i}^*, W_{2j}^*)$ all of which are induced by the derived nature of W^*s. The variance of the asymptotic reference distribution in Figure 1 is computed under the assumption that $\rho_1^* = \rho_2^* = \rho_{12}^* = 0$. Estimates of these correlations obtained from our simulations indicate that these are consistently, and frequently strongly, positive, making clear the source of scale discrepancy. For example for the configurations used in Figures 1 and 2 with $\mu_1 = 0.3$, $\mu_2 = 0$ and $\rho = 0$, we obtained $\hat{\rho}_1^* = 0.50$, $\hat{\rho}_2^* = 0.41$ and $\hat{\rho}_{12}^* = 0.35$.

Construction of Figure 3 deserves some explanation. For each simulated data set $\hat{\delta}^*$ is computed and the values over 10000 simulations are used to construct the density estimates depicted by the black (null) and the green (alternative) curves. Each simulated data set is also used to construct the reference distribution of the Projection-Permutation test as described in Section 3. To obtain a single reference distribution from these 10000 reference distributions we randomly sampled one permutation for each data set and then used those samples to construct the reference distributions shown in the figure.

5. Performance of the Area Test in Non-Nested Models and Validation Samples

5.1. Non-Nested Models

Our primary objective is to compare the incremental value of a new marker which inherently gives rise to the nested regression model. It is, however, logical to ask if the AUC test is valid if the comparison is between the distinct incremental contributions of two different markers. That is we wish to compare non-nested models. In this case the models M_1 and M_2 (2-3) are replaced with

$$\begin{aligned} M_1: \quad \text{logit}(Y_i) &= \beta_0 + \beta_1 W_{1i} + \beta_2 W_{2i} \\ M_2: \quad \text{logit}(Y_i) &= \beta_0 + \beta_1 W_{1i} + \beta_3 W_{3i} \end{aligned}$$

Table 2. Size of the AUC test in non-nested models for n=250 and 500.

	n=250				n=500			
μ_1	0		0.3		0		0.3	
ρ	0	0.5	0	0.5	0	0.5	0	0.5
$\mu_2 = \mu_3$								
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.01	0.01	0.00	0.01	0.02	0.01	0.01	0.00
0.2	0.03	0.02	0.01	0.02	0.05	0.04	0.03	0.04
0.3	0.04	0.04	0.03	0.04	0.06	0.06	0.04	0.03

and the linear predictors (4-5) now become

$$\begin{aligned} W^*_{1i} &= \hat{\beta}_0 + \hat{\beta}_1 W_{1i} + \hat{\beta}_2 W_{2i} \\ W^*_{2i} &= \tilde{\beta}_0 + \tilde{\beta}_1 W_{1i} + \tilde{\beta}_3 W_{3i} \end{aligned}$$

As before, the AUC test is used to compare W^*_1 and W^*_2.

Table 2 reports the results of simulations conducted under this scenario. Here $\mu_2 = E(W_2)$, $\mu_3 = E(W_3)$ and $\rho = \text{Cor}(W_2, W_3)$. It is evident that the test size remains extremely conservative, especially when the diagnostic value of each of the markers being compared is zero (i.e. $\mu_2 = \mu_3 = 0$). This appears to be due to the known directionality issue as well as the correlation between $\{(W^*_{1i}, W^*_{2i})\}$ and $\{(W^*_{1j}, W^*_{2j})\}$ described in Section 2. As the common signal of the two markers being compared strengthens the size approaches the nominal level. Overall, however, it is clear that the AUC test is not valid for comparing markers derived from non-nested regression models.

5.2. Validation Samples

The scenarios we have considered so far have been limited to the case where estimation of regression parameters and comparison of the ROC curves were performed on the same data set. It is not uncommon for marker studies to employ validation samples where coefficients are estimated in a training set and derived predictors are constructed and compared only on a test set. Using independent validation samples is considered to be the gold standard method for marker studies since it can eliminate optimistic bias. Intuition suggests that it should be possible to apply this logic also to formal comparisons based on the AUC test.

To study the characteristics of the AUC test in this scenario we conducted a set of simulations in which the data are generated in exactly the same way as in Section 4. Each simulated data set is then split into training and test sets. Two logistic regressions are estimated using only the training set: $\text{logit}(Y_i) = \hat{\beta}_0 + \hat{\beta}_1 W_{1i}$ and $\text{logit}(Y_i) = \tilde{\beta}_0 + \tilde{\beta}_1 W_{1i} + \tilde{\beta}_2 W_{2i}$. Following this, pairs of predictors $\{W^*_{1i}, W^*_{2i}\}$ are calculated using solely the test data but with $\hat{\beta}$s and $\tilde{\beta}$s obtained from the training set. The data are analyzed only using the the AUC test since the Wald test and the Projection-Permutation test are not applicable in validation samples.

Results are reported in Table 3. These show that the size of the test is close to the nominal level when $\mu_1 = \mu_2 = 0$ but seems to display erratic behavior as μ_1 increases. It is easily shown that $(W^*_{1i}, W^*_{2i}) \perp (W^*_{1j}, W^*_{2j})$ for all $i \neq j$ in this setting, and so dependence between the observations is not the problem. The problem is that the manner in which the derived markers are calculated corrupts the null hypothesis being tested that the AUCs are equivalent. To see this we recognize that the AUC test is a rank test and that the ranks are invariant under a location and scale (up to sign) shift. Consequently a comparison of $\{W^*_{1i}\}$ versus $\{W^*_{2i}\}$ is equivalent to a comparison of $\{W^\dagger_{1i}\}$ versus $\{W^\dagger_{2i}\}$ where $W^\dagger_{1i} = W_{1i}$ and $W^\dagger_{2i} = W_{1i} + \tilde{\beta}_2 W_{2i}/\tilde{\beta}_1$. Since $E(\tilde{\beta}_2) = 0$ under the null we are in effect comparing a single marker $\{W_{1i}\}$ with the same marker with additional noise. Added noise inevitably reduces the AUC and so the AUC corresponding to $\{W^\dagger_{1i}\}$ is necessarily larger than the AUC corresponding to $\{W^\dagger_{2i}\}$. However, this decrement is zero when $\mu_1 = \mu_2 = 0$

Table 3. Size of the AUC test with validation samples

		μ_1	0		0.3	
N_{Train}	N_{Test}	μ_2	$\rho = 0$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$
250	250	0	0.04	0.05	0.07	0.06
500	500	0	0.05	0.05	0.09	0.07
10000	250	0	0.05	0.04	0.03	0.03
10000	500	0	0.05	0.05	0.03	0.02

since both AUCs are 0.5 in this scenario. Thus, even though the test sizes in Table 3 are fairly close to the nominal level we conclude that a test constructed in this fashion is fundamentally invalid.

6. Discussion

In this article we have provided an explanation to the baffling observation that use of the AUC test to compare nested binary regression models is invalid [2]. We found that the validity problems of the AUC test in this context are due to two principal reasons. The first reason is that the test is based on the assumption that the data from individual subjects (W_{1i}, W_{2i}) and (W_{1j}, W_{2j}) are mutually independent. This is grossly violated when using predictors from a regression model. This leads to an incorrect variance estimate of the test statistic as described in Section 4 and illustrated in Figure 1. The second major problem is that in its proper construction the AUC test is fundamentally one-sided in that we know in advance the anticipated directionality of the relationship between predictor and outcome, allowing the possibility of negative effects by chance. In the regression context the methodology does not distinguish any such "known directionality" of the markers and instead constructs predictors to maximize the likelihood. In most cases this results in an optimized ROC curve. Consequently both ROC curves are optimized in the same direction. Both of these phenomena lead to bias in the test statistic. The effect of these two factors is to greatly reduce the sensitivity of the AUC test.

We have also established that using an independent validation sample does not lead to a valid AUC test. This is due to the fact that the null hypothesis of no incremental value does not correspond to the null hypothesis of equal areas under the ROC curves, except for the largely concocted case of no discriminatory power in either of the markers. We also found that the AUC test is invalid when the comparison is between two predictors drawn from non-nested regression models.

Since ROC curves are widely used for assessing the discriminatory ability of predictive models, comparing the ROC curves derived from predictive models is commonplace. In the first four months of 2011 alone, we easily identified seven articles in clinical journals that used the AUC test to compare nested logistic regression models [11, 12, 13, 14, 15, 16, 17] which speaks to the prevalence of the problem in applications of biostatistics. A recent feature in PROC LOGISTIC of SAS (ROCCONTRAST statement in version 9.2) enables users to specify nested logistic regression models, estimate their ROC curves and compare them using the AUC test. Availability of this feature in one of the most commonly used statistical packages is likely to increase the use of this invalid procedure.

We have shown that it is possible to construct a valid reference distribution for the AUC test using permutation. In so doing we have shown that the problem is not due to the AUC test statistic but is instead a consequence of the fact that the standard asymptotic reference distribution is inappropriate in the context of modelled predictors. Nevertheless comparison of two nested models using the AUC test statistic and its valid reference distribution (from Section 3) seems unncessary since its operating characteristics are very similar to those of the Wald test, which is widely available in standard statistical software.

In all of our simulations we generated marker values and other covariates from the multivariate normal distribution. This represents a framework in which the logistic regression is fully valid. Given that the derived AUC test is grossly invalid in these circumstances in which the data generation is a perfect fit for the assumed model, we consider it unnecessary

to investigate alternative sampling models for which observed biases may be caused either by inappropriate modeling assumptions or the phenomena we have described.

The performance of the AUC test has perplexed other investigators, including Demler et. al. [18] who assumed multivariate normality of the markers and employed linear discriminant analysis to construct the risk prediction tool. While these authors also seem to be motivated with the underperformance of the AUC test, they show that the AUCs of M_1 and M_2 are the same if and only if $\alpha_2 = 0$, where α_2 is the coefficient of the second marker from a linear discriminant analysis. They show that the F-test for testing $\alpha_2 = 0$ has the correct size for comparing the AUCs. These authors did not use the empirical estimate of the AUC nor did they consider the AUC test of DeLong et. al. [1], the most commonly used method of comparing the AUCs. Our results specifically explain the poor performance observed in Vickers et. al. [2] by showing that the AUC test is biased and its variance is incorrect in this specific setting. We have also demonstrated that a test comparing the AUCs using an appropriate reference distribution has very similar properties to those of the Wald test under general conditions that do not require distributional assumptions of the markers.

Finally, we clarify that the tests we have investigated are designed to test whether or not a new marker has any incremental value in predicting the outcome. Even if the marker is found to have significant incremental value, it is important to gauge the magnitude of the incremental information to determine if the marker has pratical clinical utility. ROC curves and the change in ROC area in particular have often been used for this purpose, although various other measures and approaches have been proposed [9, 19].

References

1. 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**(3):837–845.
2. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC Medical Research Methodology* 2011; **11**.
3. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; **148**(3):839–843.
4. Harrell FE, Califf RM, Pryor DB. Evaluating the yield of medical tests. *Journal of the American Medical Association* 1982; **247**(18):2543–2546.
5. Kattan MW. Judging new markers by their ability to improve predective accuracy. *Journal of the National Cancer Institute* 2003; **95**(9):634–635.
6. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* 2004; **159**(9):882–890.
7. Cook NR. Statistical evaluation of prognostic versus diagnostic models: Beyond the roc curve. *Clinical Chemistry* 2008; **54**(1):17–23.
8. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MSV, Go AS, Harrell FE, Howard BV, Howard VJ, *et al.*. Criteria for evaluation of novel markers of cardiovascular risk: A scientific statement from the american heart association. *Circulation* 2009; **119**(17):2408–2416.
9. Pencina MJ, D'Agostino RBD, S VR. Evaluating the added predictive ability of a new marker: From the area under the roc curve and beyond. *Statistics in Medicine* 2008; **27**(2):157–172.
10. Cox DR, Hinkley DV. *Theoretical Statistics*. Chapman and Hall, 1974.
11. Kwon S, Kim Y, Shim J, Sung J, Han M, Kang D, Kim JY, Choi B, Chang HJ. Coronary artery calcium scoring does not add prognostic value to standard 64-section ct angiography protocol in low-risk patients suspected of having coronary artery disease. *Radiology* 2011; **259**(1):92–99.
12. Berg K, Stenseth R, Pleym H, Wahba A, Videm V. Mortality risk prediction in cardiac surgery: Comparing a novel model with the euroscore. *Acta Anaesthesiologica Scandinavica* 2011; **55**(3):313–321.
13. Resnic F, Normand SL, Piemonte T, Shubrooks S, Zelevinsky K, Lovett A, Ho K. Improvement in mortality risk prediction after percutaneous coronary intervention through the addition of a compassionate use variable to the national cardiovascular data registry cathpci dataset: A study from the massachusetts angioplasty registry. *Journal of the American College of Cardiology* 2011; **57**(8):904–911.
14. Roe C, Fagan A, Williams M, Ghoshal N, Aeschleman M, Grant E, Marcus D, Mintun M, Holtzman D, Morris J. Improving csf biomarker accuracy in predicting prevalent and incident alzheimer disease. *Neurology* 2011; **76**(6):501–510.
15. Thuret R, Sun M, Abdollah F, Budaus L, Lughezzani G, Liberman D, Morgan M, Johal R, Jeldres C, Latour M, *et al.*. Tumor grade improves the prognostic ability of american joint committee on cancer stage in patients with penile carcinoma. *Journal of Urology* 2011; **185**(2):501–507.
16. Hammill B, Curtis L, Fonarow G, Heidenreich P, Yancy C, Peterson E, Hernandez A. Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. *Circulation: Cardiovascular Quality and Outcomes* 2011; **4**(1):60–67.

17. Liang Y, Ankerst D, Ketchum N, Ercole B, Shah G, Shaughnessy Jr J, Leach R, Thompson I. Prospective evaluation of operating characteristics of prostate cancer detection biomarkers. *Journal of Urology* 2011; **185**(1):104–110.
18. Demler OV, Pencina MJ, D'Agostino RB. Equivalence of improvement in area under roc curve and linear discriminant analysis coefficient under assumption of normality. *Statistics in Medicine* 2011; **30**:1410–1408, doi:10.1002/sim.4196.
19. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann Intern Med* 2009; **150**(11):795–802.