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

Year 2015 *Paper* 30

Inferential Methods to Assess the Difference in the Area Under the Curve From Nested Binary Regression Models

Glenn Heller* Venkatraman E. Seshan†
Chaya S. Moskowitz‡ Mithat Gonen**

*Memorial Sloan Kettering, hellerg@mskcc.org
†Memorial Sloan-Kettering Cancer Center, seshanv@mskcc.org
‡Memorial Sloan-Kettering Cancer Center, moskowc1@mskcc.org
**Memorial Sloan-Kettering Cancer Center, gonenm@mskcc.org

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

http://biostats.bepress.com/mskccbiostat/paper30

Copyright ©2015 by the authors.

Inferential Methods to Assess the Difference in the Area Under the Curve From Nested Binary Regression Models

Glenn Heller, Venkatraman E. Seshan, Chaya S. Moskowitz, and Mithat Gonen

Abstract

The area under the curve (AUC) is the most common statistical approach to evaluate the discriminatory power of a set of factors in a binary regression model. A nested model framework is used to ascertain whether the AUC increases when new factors enter the model. Two statistical tests are proposed for the difference in the AUC parameters from these nested models. The asymptotic null distributions for the two test statistics are derived from the scenarios: (A) the difference in the AUC parameters is zero and the new factors are not associated with the binary outcome, (B) the difference in the AUC parameters is less than a strictly positive value. A confidence interval for the difference in AUC parameters is developed. Simulations are generated to determine the finite sample operating characteristics of the tests and a pancreatic cancer data example is used to illustrate this approach.

Inferential Methods to Assess the Difference in the Area Under the Curve From Nested Binary Regression Models

Glenn Heller, Venkatraman E. Seshan, Chaya S. Moskowitz, Mithat Gönen

Department of Epidemiology and Biostatistics

Memorial Sloan Kettering Cancer Center

485 Lexington Ave. New York, NY 10017

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Abstract

The area under the curve (AUC) is the most common statistical approach to evaluate the discriminatory power of a set of factors in a binary regression model. A nested model framework is used to ascertain whether the AUC increases when new factors enter the model. Two statistical tests are proposed for the difference in the AUC parameters from these nested models. The asymptotic null distributions for the two test statistics are derived from the scenarios: (A) the difference in the AUC parameters is zero and the new factors are not associated with the binary outcome, (B) the difference in the AUC parameters is less than a strictly positive value. A confidence interval for the difference in AUC parameters is developed. Simulations are generated to determine the finite sample operating characteristics of the tests and a pancreatic cancer data example is used to illustrate this approach.

KEY WORDS: Area under the receiver operating characteristic curve; Incremental value; Maximum rank correlation; Nested models; Risk classification model

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

1. Introduction

Receiver operating characteristic (ROC) curves and the areas under the ROC curves (AUCs) are popular tools for assessing how well biomarkers and clinical risk prediction models distinguish between patients with and without a health outcome of interest. Historically, in cases where a new biomarker panel was developed and interest lies in evaluating its ability to add information beyond that provided by established risk factors, a two-step approach was taken. First, analysts would fit a regression model containing both the established factors and the new biomarkers and test whether the association between the outcome and new markers was statistically significant. Secondly, the linear predictor function from this model would be used to construct an AUC. This AUC would be compared to the AUC from a model containing only the established risk factors. This comparison typically involved testing whether the difference in the two AUCs was statistically significantly different from zero.

Recent work has pointed out that this approach is problematic for at least two reasons. First, when evaluating incremental value as we have described, the AUCs arise from nested regression models. The current convention is to test the difference in the AUCs with the DeLong test (DeLong, DeLong, and Clarke-Pearson 1988). In the context of AUCs that are derived from nested regression models, Seshan, Gönen, and Begg (2013) and Vickers, Cronin, and Begg (2011) have illustrated through simulation that the distributional assumptions of the DeLong test are violated resulting in a biased test statistic. Second, Pepe et al. (2013) demonstrate that the null hypothesis of no association between the new biomarkers and the outcome when established risk factors are included in the model is equivalent to the null hypothesis that the AUCs

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

from the two models are equal, and consequently, testing both is superfluous. The conclusions from these papers all coalesce to the same recommendation: when testing for whether a new set of biomarkers add any incremental value, only one statistical test should be done and the preferable one is a test of whether the regression coefficient from a binary regression model is significantly different from zero. This can be done with either a Wald, score, or likelihood ratio test.

These parametric association test statistics are more sensitive than the nonparametric difference in AUC statistic. Specifically, high odds ratios and small p-values corresponding to new markers in a classification model can produce only modest increments in the observed difference in AUCs. Such seemingly incongruous results may lead to dissonance when explaining the results to a collaborator not sufficiently versed in statistical inference. As Pepe et al. (2013) emphasize, the equivalence of two null hypotheses does not imply that the two corresponding statistical tests are the same. If the AUCs from the nested models are the primary focus of the study, then a direct method for testing this difference would provide a coherent analysis. The first part of this work derives a test of equality based on the difference in AUCs from nested models.

The second part of this work derives the distribution theory needed to accurately apply hypothesis testing and confidence interval construction for a nonzero difference in population AUCs. Rigorous evaluation of a new biomarker panel, particularly in a prospective study, necessitates that some thought be given to the minimally acceptable degree of incremental value provided by the panel. A decision as to whether the biomarkers are clinically useful need not be based on a statistical test of whether there is evidence of any incremental value, but on whether the magnitude of additional

COBRA
BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

information is sufficiently large to consider the biomarker panel promising and either worthy to study further or to recommend for use in practice. While both Pepe et al. (2013) and Seshan, Gönen, and Begg (2013) emphasize this point, neither they, nor anyone else as far as we are aware, offer guidance on how to formally test for a minimally acceptable degree of incremental value.

Although alternative model performance metrics have their merits, the AUC still remains one of the most often used measures of medical test performance. It is ubiquitous in clinical, bioinformatic, and radiology journals, and many researchers are familiar with it. Having a way to test for a minimal change in AUCs could thus be useful in multiple contexts. Furthermore, this familiarity may facilitate clinicians abilities to judge what constitutes a clinically meaningful difference. In addition to the development of a test under a non-zero null, the methodology developed enables an asymptotic confidence interval for the difference in the population AUCs; a useful inferential approach that we could not find in the literature.

2. The Difference in AUCs with Nested Models

A generalized binary regression model

$$\Pr(Y = 1|\boldsymbol{X}) = G(\boldsymbol{\beta}_0^T \boldsymbol{X})$$

is used create risk scores $\boldsymbol{\beta}^T \boldsymbol{X}$ that predict a binary classifier Y, with outcomes referred to as response ($Y = 1$) and nonresponse ($Y = 0$). In this model, the monotone link function G is unknown, making the parameter vector $\boldsymbol{\beta}$ identifiable up to a scale factor. To establish scale normalization, the first parameter component is set equal to 1 and is expressed as $\boldsymbol{\beta} = (1, \boldsymbol{\eta}^T)^T$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

The model based performance in terms of classification is evaluated using the area under the receiver operating characteristic curve (AUC). The area under the curve is defined as

$$\Pr(\boldsymbol{\beta}^T \boldsymbol{X}_1 > \boldsymbol{\beta}^T \boldsymbol{X}_2 | Y_1 = 1, Y_2 = 0),$$

which represents the probability that a responder's risk score is greater than a non-responder's risk score.

Often a new set of markers are under consideration to improve risk classification. A direct approach for this assessment is to test whether the new risk factors in tandem with existing markers increase the area under the curve relative to the AUC derived solely from the established factors. This evaluation is based on the difference in AUCs from the nested models

$$\Pr(Y = 1 | \boldsymbol{X}, \boldsymbol{Z}) = G(\boldsymbol{\beta}_0^T \boldsymbol{X} + \boldsymbol{\gamma}_0^T \boldsymbol{Z})$$

$$\Pr(Y = 1 | \boldsymbol{X}) = G(\boldsymbol{\beta}^{0T} \boldsymbol{X}),$$

where the existing markers are denoted by the p-dimensional covariate vector $\boldsymbol{X}$ and the new markers are represented by the q-dimensional covariate vector $\boldsymbol{Z}$. The estimated area under the curve for the nested models are:

$$\tilde{A}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = (n_0 n_1)^{-1} \sum_i \sum_j I[y_i > y_j] I[\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_{ij} + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z}_{ij} > 0]$$

$$\tilde{A}_n(\hat{\boldsymbol{\beta}^0}, 0) = (n_0 n_1)^{-1} \sum_i \sum_j I[y_i > y_j] I[\hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_{ij} > 0]$$

where the notation $\boldsymbol{x}_{ij}$ is used to represent the pairwise difference $\boldsymbol{x}_i - \boldsymbol{x}_j$ and $n_k = \sum_i I[y_i = k]$. Note that due to the identifiability constraint, $\hat{\boldsymbol{\beta}} = (1, \hat{\boldsymbol{\eta}}^T)^T$, $\hat{\boldsymbol{\beta}}^0 =$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

$(1, \hat{\boldsymbol{\eta}^0}^T)^T$ and the corresponding parameters are denoted by $\boldsymbol{\beta}_0 = (1, \boldsymbol{\eta}_0^T)^T$, $\boldsymbol{\beta}^0 = (1, \boldsymbol{\eta}^{0T})^T$.

The parameter estimates from these nested models are computed using the maximum rank correlation (MRC) procedure (Han 1987). The MRC is a rank based estimation procedure that maximizes the AUC. For the full model, the MRC estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ are computed as

$$\arg\max_{(\boldsymbol{\eta},\boldsymbol{\gamma})} \ (n_0 n_1)^{-1} \sum_i \sum_j I[y_i > y_j] I[\boldsymbol{\beta}^T \boldsymbol{x}_i + \boldsymbol{\gamma}^T \boldsymbol{z}_i > \boldsymbol{\beta}^T \boldsymbol{x}_j + \boldsymbol{\gamma}^T \boldsymbol{z}_j].$$

Sherman (1993) demonstrated that $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}})$ and $\hat{\boldsymbol{\eta}}^0$ are asymptotically normal and are consistent estimates of $(\boldsymbol{\eta}_0, \boldsymbol{\gamma}_0)$ and $\boldsymbol{\eta}^0$.

3. Hypothesis Testing

To test the hypothesis that the new markers improve the AUC, we denote the limiting values of the estimated AUC from the reduced model and full model as $\alpha(\boldsymbol{\beta}^0, 0)$ and $\alpha(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$, respectively. Han (1987) demonstrates that these limiting forms represent the maximum population AUCs when the markers are combined linearly.

The hypothesis test may be characterized as

$H_0 : \alpha(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) - \alpha(\boldsymbol{\beta}^0, 0) \leq \delta$

$H_a : \alpha(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) - \alpha(\boldsymbol{\beta}^0, 0) > \delta$

A standard approach to derive a testing procedure is to find an asymptotic reference distribution for the difference in nested AUCs via a Taylor series expansion around the true parameter vectors. This expansion, however, requires differentiation with respect to the parameters $(\boldsymbol{\eta}, \boldsymbol{\gamma})$, which is problematic due to the discontinuity

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

induced by the indicator function in the AUC statistic. As a result, the expansions utilized in this paper use a smooth version of $\tilde{A}_n$ based on the asymptotic approximation

$$I[\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \boldsymbol{\gamma}^T \boldsymbol{z}_{ij} > 0] \approx \Phi\left(\frac{\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \boldsymbol{\gamma}^T \boldsymbol{z}_{ij}}{h_n}\right)$$

where Φ is the standard normal distribution function and h_n is a bandwidth that goes to 0 as the sample size n gets large (Horowitz 1992). The smoothed empirical AUCs are written as

$$A_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = (n_0 n_1)^{-1} \sum_i \sum_j I[y_i > y_j] \Phi\left(\frac{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_{ij} + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z}_{ij}}{h_n}\right)$$

$$A_n(\hat{\boldsymbol{\beta}^0}, 0) = (n_0 n_1)^{-1} \sum_i \sum_j I[y_i > y_j] \Phi\left(\frac{\hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_{ij}}{h_n}\right).$$

Ma and Huang (2007) demonstrate the asymptotic normality of the parameter estimates from the smoothed AUC and the uniform consistency of the smoothed AUCs to the maximum population AUCs. As a result, the smoothed versions of the MRC based AUC estimates are used to derive the null asymptotic reference distribution.

To determine the distribution of the test statistic, there are two null scenarios for the threshold that are considered separately

$A:\ \delta = 0,\ \boldsymbol{\gamma}_0 = 0\ (\boldsymbol{\beta}_0 = \boldsymbol{\beta}^0)$

$B:\ \delta > 0,\ \boldsymbol{\gamma}_0 \neq 0$

For the null in scenario A, the set of new factors are not associated with the response, and as a result, the limiting AUCs are equal (Pepe et al. 2013). For the null hypothesis in scenario B, the new factors are associated with response, but the difference in the limiting AUCs is not larger than an apriori determined value (δ).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

3.1. Scenario A: $\delta = 0,\ \boldsymbol{\gamma}_0 = 0\ (\boldsymbol{\beta}_0 = \boldsymbol{\beta}^0)$

The most common approach for testing scenario A is to apply the asymptotic normal U-statistic theory to the studentized difference in empirical AUCs (Delong, DeLong, and Clarke-Pearson 1988). As shown below, root-n normality is not the correct null reference distribution for the difference in AUCs from nested models. Seshan, Gönen, and Begg (2013) recognized the inaccuracy of the normal reference distribution and developed a resampling approach to attain an approximate distribution for the difference in nested AUCs. They illustrated that the estimated risk scores, derived from a logistic regression model, oriented the difference in AUCs in a positive direction. In addition, they noted that the variance-covariance matrix for the AUCs under the null is singular, further distorting this application. They addressed these issues by constructing a projection-permuation reference distribution and demonstrated its operating characteristics through simulation.

We reexamine the asymptotic null distribution theory. The theorem below provides the distribution for the difference in nested AUCs when the new factors are not associated with response. The proof of this theorem is found in the appendix.

Theorem 1: The difference in nested AUCs under scenario A may be asymptotically represented as

$$2n[A_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - A_n(\hat{\boldsymbol{\beta}^0}, 0)] = \sum_{j=1}^{q} \lambda_j \chi_j^2 + o_p(1),$$

where $\{\chi_j^2\}$ are independent chi-square random variables each with one degree of freedom, $\{\lambda_j\}$ are the eigenvalues of the product matrix $-V_{\boldsymbol{\gamma}}[D^{\gamma\gamma}]^{-1}$, where both V and D are derived from the full model, $V_{\boldsymbol{\gamma}}$ is the asymptotic variance of the MRC

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

estimate $\hat{\boldsymbol{\gamma}}$, D is the second derivative matrix of A_n, and its partitioned form along with its inverse are represented as

$$D = \begin{bmatrix} D_{\boldsymbol{\eta\eta}} & D_{\boldsymbol{\eta\gamma}} \\ D_{\boldsymbol{\gamma\eta}} & D_{\boldsymbol{\gamma\gamma}} \end{bmatrix} \qquad D^{-1} = \begin{bmatrix} D^{\boldsymbol{\eta\eta}} & D^{\boldsymbol{\eta\gamma}} \\ D^{\boldsymbol{\gamma\eta}} & D^{\boldsymbol{\gamma\gamma}} \end{bmatrix}$$

Comment 1: Although the distribution of a weighted sum of independent chi-square random variables does not have a closed form, the distribution can be approximated by generating q independent squared standard normal random variables $\{Z_j^2\}$, computing the linear combination $\sum \lambda_j Z_j^2$, and repeating a large number of times.

Comment 2: Vuong (1989) and Fine (2002) present this distributional result for the likelihood ratio statistic from mispecified nested (semi)parametric models. Further, the result is a generalization of the asymptotic distribution theory for the likelihood ratio statistic. If $A_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ and $A_n(\hat{\boldsymbol{\beta}}^0, 0)$ were replaced by the loglikelihoods from the full and constrained parametric regression models, then D is the negative information matrix and from standard likelihood theory $[-D^{\boldsymbol{\gamma\gamma}}]^{-1}$ approximates $V_{\boldsymbol{\gamma}}$. It follows that the q eigenvalues of $-V_{\boldsymbol{\gamma}}[D^{\boldsymbol{\gamma\gamma}}]^{-1}$ are each equal to 1, and the result reduces to $\sum_{j=1}^{q} \chi_j^2 + o_p(1)$; the standard result that the likelihood ratio test statistic is a chi-square with q degrees of freedom.

Comment 3: Seshan, Gönen, and Begg (2013) used maximum likelihood from a logistic model to estimate the regression coefficients for the AUC calculations. Their results indicated that a nontrivial percentage of the simulations produced a negative difference in the nested AUCs, which was difficult to interpret. The MRC coefficient estimates, derived through maximization of the AUCs from the constrained and unconstrained models, result in a non-negative difference in AUCs up to the limitations

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

of the algorithmic maximization search.

Comment 4: The first derivative of the AUC, when evaluated at the MRC parameter estimate, is equal to zero. As a result, the quadratic is the lowest order nonzero term in the asymptotic expansion of the difference in AUCs. This simplifies the derivation of the null asymptotic distribution.

3.2. Scenario B: $\delta > 0,\ \boldsymbol{\gamma}_0 \neq 0$

We obtain the asymptotic distribution under a null that indicates that a new set of factors are associated with response after controlling for the established risk factors, but the parameter AUCs in the nested models do not differ by more than δ. In deciding what constitutes a relevant increase in the model AUC, the analyst will often follow practical and empirical considerations that depend upon the particular application. As has been noted previously, putting the AUC increase in a clinical context has been challenging, but experience with this measure has enabled investigators to gauge improvement (Kerr, Bansal, and Pepe 2012). Less well appreciated is that the magnitude of the improvement is a function of the baseline model AUC. This point was made by Pencina et al. (2012), and suggests that a calibrated determination, as a function of the baseline model AUC, be used for testing an improvement in nested AUCs. For example, a large δ may be useful when testing for an improvement over a relatively weak baseline model AUC, whereas a small δ may be justified when testing for an improvement over a stronger baseline model AUC.

The theorem below provides the asymptotic distributional framework for hypothesis testing and confidence interval estimation for δ.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

THEOREM 2: The difference in nested AUCs under scenario B may be asymptotically represented as

$$n^{1/2}[A_n(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) - A_n(\hat{\boldsymbol{\beta}}^0,0) - \delta] =$$

$$n^{1/2}\left[(n_0n_1)^{-1}\sum_i\sum_j I[y_i > y_j]\left\{\Phi\left(\frac{\boldsymbol{\beta}_0^T\boldsymbol{x}_{ij} + \boldsymbol{\gamma}_0^T\boldsymbol{z}_{ij}}{h_n}\right) - \Phi\left(\frac{\boldsymbol{\beta}^{0T}\boldsymbol{x}_{ij}}{h_n}\right) - \delta\right\}\right] + o_p(1)$$

The asymptotic expression is simply the zero order term in the asymptotic expansion. This asymptotic approximation is a two-sample U-statistic of degree 2 with no estimated parameters. It follows from U-statistic theory that under the δ null, the difference in AUCs is asymptotically normal with mean 0. The variance estimate from this U-statistic is provided in the appendix. Interestingly, the studentized statistic is the DeLong statistic. In contrast, as shown in the previous section, the asymptotic normal distribution is incorrectly applied to the DeLong statistic under scenario A.

The simulation results in Section 5 demonstrate that for a sample size as large as 500, this asymptotic normal test is conservative under scenario B. An explanation for this lack of accuracy is illustrated in Figure 1a, which is a plot of the difference in the AUCs $[\hat{\delta} = A_n(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) - A_n(\hat{\boldsymbol{\beta}}^0,0)]$ and its estimated asymptotic variance $[\hat{V}]$. The points are the realizations of a simulation where $\delta = 0.01$, the baseline AUC is 0.70, and the sample size within each replication is 500. The graph indicates a linear relationship between the estimate and its variance. To remove this mean-variance relationship, an Anscombe variance stabilizing reparameterization $g(\delta) = \sqrt{\delta + \frac{3}{8n}}$ is used to provide greater accuracy for the normal approximation. The transformed estimate for testing the difference in AUCs and its estimated asymptotic variance are

$$\hat{\tau} = \sqrt{\hat{\delta} + \frac{3}{8n}} \qquad \widehat{\text{var}}(\hat{\tau}) = \frac{\hat{V}}{4(\hat{\delta} + \frac{3}{8n})}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

Stemming from comment 3 in Section 3, estimating the regression parameters by maximizing the AUCs in the reduced and full models, leads to a nonnegative $\hat{\delta}$, and removes a barrier to applying the square root transformation. Figure 1b depicts the variance stabilization after the Anscombe transformation was applied.

4. Confidence Intervals

In addition to providing more accurate level tests, the normalizing transformation enables the construction of a confidence interval for the difference in the AUC parameters, $\delta = \alpha(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) - \alpha(\boldsymbol{\beta}^0, 0)$. The 95% confidence interval is obtained by using the variance stabilizing transformation $\tau = \sqrt{\delta + \frac{3}{8n}}$ and selecting the set of values not in the critical region of the asymptotic normal test

$$\left\{ \tau : \left| \frac{\hat{\tau} - \tau}{\sqrt{\text{var}(\hat{\tau})}} \right| < 1.96 \right\}$$

A back transformation of the upper and lower 95% confidence limits for τ leads to an asymptotic confidence interval for δ.

$$\Pr\left[\left\{ \hat{\tau} - 1.96\sqrt{\text{var}(\hat{\tau})} \right\}^2 - \frac{3}{8n} < \delta < \left\{ \hat{\tau} + 1.96\sqrt{\text{var}(\hat{\tau})} \right\}^2 - \frac{3}{8n} \right] \approx 0.95.$$

5. Simulations

A simulation study is performed to examine the validity the proposed test. A bivariate normal equal correlation model with correlation parameters $\{0, 0.5\}$ and $\Pr(Y = 1) = 0.5$ were used to generate the simulation data. Five hundred observations per replicate and and 5000 replicates were run for each simulation. The range of population AUCs examined was $(0.55 - 0.85)$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

The choice of bandwidth used for smoothing the AUC is flexible, since the only asymptotic constraint is that it goes to zero as the sample size gets large. For scenario A, the second derivative matrix D, derived from the smoothed AUC, is a function of the normal density ϕ. Guidance from kernel density estimation led to the bandwidth $h_n = \hat{\omega} n^{-1/5}$, where ω^2 is the variance of $\boldsymbol{\beta}^T \boldsymbol{x} + \boldsymbol{\gamma}^T \boldsymbol{z}$. For scenario B, the test statistic is based on the normal distribution function Φ, but none of its derivatives. Since the stability of its derivative ϕ does not play a role, a tighter bandwidth $h_n = \hat{\omega} n^{-1/2}$ was chosen for these simulations.

Scenario A size and power calculations are presented in Tables 1 and 2. For Table 1, the new factors are not associated with response ($\boldsymbol{\gamma}_0 = 0$) and in Table 2, the difference δ varies with the underlying baseline population AUC. Scenario B size results, with the null difference in AUC parameters equal to δ, are given in Table 3.

For scenario A, the asymptotic reference distribution, based on a linear combination of chi-square random variables, results in an accurate size test except when the AUC is near the 0.50 boundary. The results in Table 1 also confirm the validity of the Wald test under this scenario. The power results in Table 2 illustrate that the parametric Wald test is more sensitive than the nonparametric difference in AUC test, but that the difference in power is not substantial.

The size results for scenario B are displayed in Table 3. The difference in AUCs test (DIFF), based on the studentized asymptotic normal test, is conservative, but improving as δ increases. To remove the mean-variance relationship in the studentized test, the variance stabilizing transform is applied and it is verified that the variance stabilized difference in AUCs (DIFFvst) does generate a valid test, but has increasing size as the AUC gets closer to the 0.50 boundary. The Wald test is inappropriate in

COBRA
BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

this scenario.

6. Application to pancreatic cancer

Intraductal papillary mucinous neoplasms (IPMN) are cystic lesions of the pancreas and present with difficult treatment decisions. Surgical removal is difficult and morbid. It is essential if the lesions are high-risk (defined as malignant or high-grade) but also a potential for harm to the patient for low-risk lesions (low-grade or benign). Unfortunately lesion risk (malignancy and grade) can only be evaluated pathologically, leaving the clinician to use alternative clinical markers of risk such as main duct involvement. It is widely accepted that lesions involving the main pancreatic duct are at higher risk of being malignant and current guidelines of the International Association of Pancreatology recommend resection of all main-duct lesions (Tanaka et al. 2012). Using the data which supported these guidelines one can infer that 40 percent of patients with main duct IPMN will undergo resection to remove low-risk lesions. Therefore the search for markers that improve our ability to select patients for resection continues. Lesion size and presence of a solid component on imaging are recently reported to be predictors of high-risk lesions (Correa-Gallego et al. 2013) although they are not yet incorporated into the international gudielines. In this analysis we evaluate whether a novel marker, recent weight loss, provides incremental improvement in risk classification, when used in conjunction with main duct involvement, lesion size and the presence of a solid component in imaging.

Two hundred and six patients at Memorial Sloan Kettering who were candidates for surgical removal of IPMNs were evaluated. The Wald statistic, derived from a

CORRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

logistic regression analysis, indicated that recent weight loss is positively associated with high vs. low risk lesions ($p = 0.007$) in the presence of a solid component on imaging, main duct involvement, and lesions size. The maximum rank correlation AUC estimates from models without and with the weight loss factor were 0.794 and 0.809, respectively. Thus, although the Wald statistic indicates that weight loss is associated with resection, it is unclear whether its inclusion is sufficiently helpful in terms of risk classification.

We examined the importance of weight loss, first in scenario A, confirming the logistic analysis that weight loss is associated with high-risk lesions. The observed difference in model AUCs was 0.015 and the test that the added factor increased the population AUC generated a p-value equal to 0.007. Given that the population AUCs from the nested models have a non-zero difference, we next examined scenario B and tested whether this difference was greater than 0.01. We choose 0.01 because both lesion size and the presence of a solid component on imaging displayed improvement over main duct involvement by more than 0.01 on the AUC scale. The results using the variance stabilizing transformation, generated a studentized test statistic that resulted in a p-value equal to 0.652, indicating that adding recent weight loss to the existing factors did not improve surgical risk classification via the AUC metric by more than 0.01. The 95% confidence interval for the difference in AUCs was $(-0.001, 0.053)$. Thus, weight loss does not provide sufficient additional information for incorporation into the current surgical risk classification algorithm.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

7. Discussion

http://biostats.bepress.com/mskccbiostat/paper30

The complexity of human disease and response to treatment can only be captured by the use of multiple clinical features and biomarkers. While most clinical features that are in use for predictive purposes are well-established, new biomarkers (including genomic and proteomic ones) are rapidly being introduced into clinical research. These novel markers are useful to the extent that they improve our ability to prognosticate and predict response to therapy over and beyond what we can currently do using clinical features and established biomarkers. This requires the development of a statistical model that includes both established and novel markers, and using this model to test the added predictive value of the novel components. This is typically done comparing the AUCs from the full (the model containing all variables) and reduced (the model excluding the novel variables) resulting in nested models.

The current recommendation to establish an increase in the AUC for nested models is to perform a likelihood ratio or Wald test on the additional factors. While this is a valid test it does not directly address the aim of the AUC analysis. The direct method is to measure the difference in AUCs from the nested models. This approach is analogous to using the F statistic for prediction in linear regression rather than the likelihood ratio test to examine the predictive importance of a subset of factors. In this article we provide the asymptotic theory nececesary for the statistical comparison of two AUCs resulting from nested models. In addition we provide a method to construct an asymptotically valid confidence interval for the difference in AUCs filling another gap in the methodology.

As prediction becomes more important in medical research and practice, metrics other than AUC have been introduced (Pencina et al. 2008). It is noted that the methodological framework, including the smoothing approximation for indicator

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

functions and the distribution theory for nested models, is sufficiently general to be applied to assess the added value of new markers to other measures of discrimination, such as: sensitivity, specificity, net benefit, net reclassification improvement, and integrated discriminant improvement. The application of the proposed methodology to these statistics will be explored in future work.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

Table 1: Size simulations for scenario A ($\boldsymbol{\gamma}_0 = 0$)

AUCf	AUCr	ρ	LCCS	WALD	ρ	LCCS	WALD
0.55	0.55	0	0.1040	0.0428	0.5	0.1056	0.0450
0.60	0.60	0	0.0504	0.0602	0.5	0.0506	0.0612
0.65	0.65	0	0.0454	0.0500	0.5	0.0460	0.0512
0.70	0.70	0	0.0500	0.0526	0.5	0.0500	0.0516
0.75	0.75	0	0.0460	0.0456	0.5	0.0460	0.0456
0.80	0.80	0	0.0474	0.0482	0.5	0.0472	0.0478
0.85	0.85	0	0.0554	0.0490	0.5	0.0554	0.0500

Table 2: Power simulations for scenario A ($\boldsymbol{\gamma}_0 \neq 0,\ \delta = \{0.005, 0.01, 0.02\}$)

δ	AUCr	ρ	LCCS	WALD	ρ	LCCS	WALD
0.02	0.55	0	0.3674	0.5140	0.5	0.3528	0.4992
0.02	0.60	0	0.6434	0.7428	0.5	0.6536	0.7500
0.01	0.65	0	0.5600	0.6182	0.5	0.5634	0.6224
0.01	0.70	0	0.6914	0.7318	0.5	0.6838	0.7234
0.01	0.75	0	0.8142	0.8398	0.5	0.8210	0.8446
0.005	0.80	0	0.6302	0.6486	0.5	0.6416	0.6584
0.005	0.85	0	0.7590	0.7566	0.5	0.7594	0.7580

AUCf = Area under the curve for full model with covariates (X, Z)

AUCr = Area under the curve for reduced model with covariate X

δ = AUCf - AUCr

ρ = Correlation between the covariates (X, Z)

LCCS = linear combination of chi-square random variables

Wald = Wald statistic

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 3: Size simulation results for scenario B ($\boldsymbol{\gamma}_0 \neq 0, \quad \delta = \{0.005, 0.01, 0.02\}$)

δ	AUCr	ρ	DIFF	DIFFvst	WALD	ρ	DIFF	DIFFvst	WALD
0.02	0.55	0	0.0378	0.0864	0.5280	0.5	0.0350	0.0776	0.4876
0.02	0.60	0	0.0286	0.0668	0.7428	0.5	0.0300	0.0670	0.7600
0.01	0.65	0	0.0238	0.0620	0.6182	0.5	0.0196	0.0582	0.6216
0.01	0.70	0	0.0230	0.0568	0.7318	0.5	0.0210	0.0558	0.7392
0.01	0.75	0	0.0238	0.0538	0.8398	0.5	0.0218	0.0524	0.8478
0.005	0.80	0	0.0168	0.0502	0.6486	0.5	0.0162	0.0472	0.6530
0.005	0.85	0	0.0220	0.0534	0.7566	0.5	0.0192	0.0522	0.7614

AUCf = Area under the curve for full model with covariates (X, Z)

AUCr = Area under the curve for reduced model with covariate X

δ = AUCf - AUCr

ρ = Correlation between the covariates (X, Z)

DIFF = Difference in AUC test

DIFFvst = Difference in AUC test with variance stabilizing transformation

Wald = Wald statistic

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

Appendix

The following notation and regularity conditions are used in this appendix.

Notation:

$$\boldsymbol{\beta}^T = (1, \eta_1, \ldots, \eta_{p-1}), \quad \boldsymbol{\gamma}^T = (\gamma_1, \ldots, \gamma_q), \quad \boldsymbol{\theta} = (\boldsymbol{\eta}^T, \boldsymbol{\gamma}^T)^T$$

$$A_n(\boldsymbol{\theta}) = (n_0 n_1)^{-1} \sum_i \sum_j I[y_i > y_j] \Phi\left(\frac{\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \boldsymbol{\gamma}^T \boldsymbol{z}_{ij}}{h_n}\right)$$

The second derivative matrix of $A_n(\boldsymbol{\theta})$ and its inverse are partitioned as

$$D(\boldsymbol{\theta}) = \begin{bmatrix} D_{\boldsymbol{\eta}\boldsymbol{\eta}} & D_{\boldsymbol{\eta}\boldsymbol{\gamma}} \\ D_{\boldsymbol{\gamma}\boldsymbol{\eta}} & D_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \end{bmatrix} \qquad D^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} D^{\boldsymbol{\eta}\boldsymbol{\eta}} & D^{\boldsymbol{\eta}\boldsymbol{\gamma}} \\ D^{\boldsymbol{\gamma}\boldsymbol{\eta}} & D^{\boldsymbol{\gamma}\boldsymbol{\gamma}} \end{bmatrix} \qquad \text{where } D_{\boldsymbol{\eta}\boldsymbol{\gamma}} = \frac{\partial^2 A_n(\boldsymbol{\theta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\gamma}}$$

Regularity conditions:

1. $\boldsymbol{\theta} \in \Theta$ a compact subspace of $\mathcal{R}^{p-1+q}$.
2. The domain of $(\boldsymbol{x}, \boldsymbol{z})$ is not contained in a linear subspace of $\mathcal{R}^{p+q}$.
3. The density of x_1 conditional on all other covariates is everywhere positive.

The null asymptotic distribution of the difference in AUCs: Scenario A

A three term expansion of $A_n(\boldsymbol{\theta}_0)$ around $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}})$ is,

$$A_n(\hat{\boldsymbol{\theta}}) - \left\{ A_n(\hat{\boldsymbol{\theta}}) + 0 + \frac{1}{2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T D(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) \right\},$$

where the first order term is zero since $\hat{\boldsymbol{\theta}}$ is obtained through maximization of $A_n(\boldsymbol{\theta})$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

A similar argument produces a three term expansion of $A_n(\boldsymbol{\theta}^0)$ around $\hat{\boldsymbol{\theta}}^0 = (\hat{\boldsymbol{\eta}}^0, 0)$,

$$A_n(\hat{\boldsymbol{\theta}}^0) - \left\{ A_n(\hat{\boldsymbol{\theta}}^0) + 0 + \frac{1}{2}(\boldsymbol{\eta}^0 - \hat{\boldsymbol{\eta}}^0)^T D_{\boldsymbol{\eta}\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}^0)(\boldsymbol{\eta}^0 - \hat{\boldsymbol{\eta}}^0) \right\}.$$

Therefore, the test statistic $2n[A_n(\hat{\boldsymbol{\theta}}) - A_n(\hat{\boldsymbol{\theta}}^0)]$ is asymptotically approximated by

$$n(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T \left[-D(\hat{\boldsymbol{\theta}}) \right] (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) - n(\boldsymbol{\eta}^0 - \hat{\boldsymbol{\eta}}^0)^T \left[-D_{\boldsymbol{\eta}\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}^0) \right] (\boldsymbol{\eta}^0 - \hat{\boldsymbol{\eta}}^0) + o_p(1).$$

Further simplification may be achieved by relating the unrestricted and the restricted estimates $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\eta}}^0$ under the null (Cox and Hinkley 1974),

$$(\boldsymbol{\eta}^0 - \hat{\boldsymbol{\eta}}^0) = (\boldsymbol{\eta}_0 - \hat{\boldsymbol{\eta}}) + D^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}^0) D_{\boldsymbol{\eta}\boldsymbol{\gamma}}(\hat{\boldsymbol{\theta}}^0)(\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\gamma}}) + o_p(n^{-1/2}).$$

Thus, the test statistic under the null may be asymptotically approximated by

$$2n[A_n(\hat{\boldsymbol{\theta}}) - A_n(\hat{\boldsymbol{\theta}}^0)] = n(\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\gamma}})^T [-D^{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\hat{\boldsymbol{\theta}})]^{-1} (\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\gamma}}) + o_p(1).$$

The quadratic on the right hand side is asymptotically a weighted sum of independent chi-square random variables, each with one degree of freedom (Johnson and Kotz 1970).

Therefore under scenario A, a test for the difference in nested AUCs may be based on the null reference distribution

$$2n[A_n(\hat{\boldsymbol{\theta}}) - A_n(\hat{\boldsymbol{\theta}}^0)] \approx \sum_{j=1}^{q} \lambda_j \chi^2_j.$$

where $\{\lambda_j\}$ are the eigenvalues of the product matrix $-V_{\boldsymbol{\gamma}}[D^{\boldsymbol{\gamma}\boldsymbol{\gamma}}]^{-1}$ and $V_{\boldsymbol{\gamma}}$ is the asymptotic variance of $\hat{\boldsymbol{\gamma}}$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

The null asymptotic distribution of the difference in AUCs: Scenario B

The test statistic and its asymptotic distribution are derived under a null that indicates that the new set of factors are associated with response, but the AUCs do not differ by more than δ.

Consider the first order asymptotic approximation

$$n^{1/2}[A_n(\hat{\boldsymbol{\theta}}) - A_n(\hat{\boldsymbol{\theta}}^0) - \delta] = n^{1/2}[A_n(\boldsymbol{\theta}_0) - A_n(\boldsymbol{\theta}^0) - \delta] +$$

$$\left[\frac{\partial A_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]^T n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \left[\frac{\partial A_n(\boldsymbol{\eta}, 0)}{\partial \boldsymbol{\eta}}\bigg|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^0}\right]^T n^{1/2}(\hat{\boldsymbol{\eta}}^0 - \boldsymbol{\eta}^0) + o_p(1).$$

Again, since $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\eta}}^0$ maximize their respective smooth AUCs,

and it follows that

$$n^{1/2}[A_n(\hat{\boldsymbol{\theta}}) - A_n(\hat{\boldsymbol{\theta}}^0) - \delta] = n^{1/2}[A_n(\boldsymbol{\theta}_0) - A_n(\boldsymbol{\theta}^0) - \delta] + o_p(1).$$

Therefore,

$$n^{1/2}[A_n(\boldsymbol{\theta}_0) - A_n(\boldsymbol{\theta}^0) - \delta] =$$

$$n^{1/2}\left[(n_0 n_1)^{-1}\sum_i\sum_j I[y_i > y_j]\left\{\Phi\left(\frac{\boldsymbol{\beta}_0^T\boldsymbol{x}_{ij} + \boldsymbol{\gamma}_0^T\boldsymbol{z}_{ij}}{h_n}\right) - \Phi\left(\frac{\boldsymbol{\beta}^{0T}\boldsymbol{x}_{ij}}{h_n}\right) - \delta\right\}\right]$$

is a two-sample U-statistic of degree 2 (with no estimated parameters) and a test for the difference in nested AUCs under scenario B is based on a normal mean 0 null reference distribution. The variance from this U-statistic is

$$V = \frac{n}{n_0}\sigma_1^2 + \frac{n}{n_1}\sigma_2^2,$$

which may be estimated with the following components

$$\hat{\sigma}_1^2 = [n_0 n_1(n_0 - 1)]^{-1}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1,k\neq j}^n I[y_i = 1]I[y_j = 0]I[y_k = 0](e_{ij} - \delta)(e_{ik} - \delta)$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

$$\hat{\sigma}_2^2 = [n_0 n_1 (n_1 - 1)]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1, k \neq j}^{n} I[y_i = 1] I[y_j = 0] I[y_k = 1] (e_{ij} - \delta)(e_{kj} - \delta)$$

$$\text{and } e_{ij} = \Phi \left[\frac{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_{ij} + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z}_{ij}}{h_n} > 0 \right] - \Phi \left[\frac{\hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_{ij}}{h_n} > 0 \right].$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

References

Correa-Gallego, C., Do, R., Lafemina, J., Gonen, M., D'Angelica, M. I., DeMatteo, R. P., Fong, Y., Kingham, T. P., Brennan, M. F., Jarnagin, W. R., Allen, P.J. (2013), "Predicting dysplasia and invasive carcinoma in intraductal papillary mucinous neoplasms of the pancreas: development of a preoperative nomogram," *Annals of Surgical Oncology*, 4348-4355.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, New York, NY: Chapman and Hall.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988), "Comparing areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, 44, 837-845.

Fine, J. P. (2002), "Comparing nonnested Cox models," *Biometrika*, 89, 635-647.

Han, A. (1987), "Nonparametric analysis of a generalized regression model," *Journal of Econometrics*, 35, 303-316.

Horowitz, J. L. (1992), "A smoothed maximum score estimator for the binary response model," *Econometrica*, 60, 505-531.

Johnson, N. L. and Kotz, S. (1970), *Distributions in Statistics: Continuous Univariate Distributions - 2*, New York, NY: John Wiley and Sons.

Kerr, K. F., Bansal, A., and Pepe, M. S. (2012), "Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context," *American Journal of Epidemiology*, 176, 482-487.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Ma, S., and Huang, J. (2007), "Combining multiple markers for classification using ROC," *Biometrics*, 63, 751-757.

Pencina M. J., D'Agostino, R. B. Sr, D'Agostino, R. B. Jr, Ramachandran, R. S. (2008), "Evaluating the added predicitve ability of a new marker: From area under the ROC curve to reclassification and beyond," *Statistics in Medicine*, 27, 157-172.

Pencina, M. J., D'Agostino, R. B. Sr, Pencina, K. M., Janssens, C. J. W., and Greenland, P. (2012), "Interpreting incremental value of markers added to risk prediction models," *American Journal of Epidemiology*, 176, 473-481.

Pepe, M. S., Janes, H., and Li, C. I. "Net risk reclassification p values: valid or misleading?" *Journal of the National Cancer Institute*, 106,

Pepe, M. S., Kerr, K. F., Longton, G., and Wang, Z. (2013), "Testing for improvement in prediction model performance," *Statistics in Medicine*, 32, 1467-1482.

Seshan, V. E., Gonen, M., and Begg, C. B. (2013), "Comparing ROC curves derived from regression models," *Statistics in Medicine*, 32, 1483-1493.

Sherman, R. P. (1993), "The limiting distribution of the maximum rank correlation estimator," *Econometrica*, 61, 123-137.

Tanaka, M., Fernandez-de Castillo, C., Adsay, V., Chari, S., Falconi, M., Jang, J. Y., Kimura, W., Levy, P., Pitman, M. B., Schmidt, C. M., Shimizu, M., Wolfgang, C. L., Yamaguchi, K., Yamao, K. (2012), " International consensus guideline 2012 for the mangement of IPMN and MCN of the pancreas," *Pancreatology*, 12, 183-197.

Vickers, A. J., Cronin, A. M., and Begg, C.B. (2011), "One statistical test is suf-

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mskccbiostat/paper30

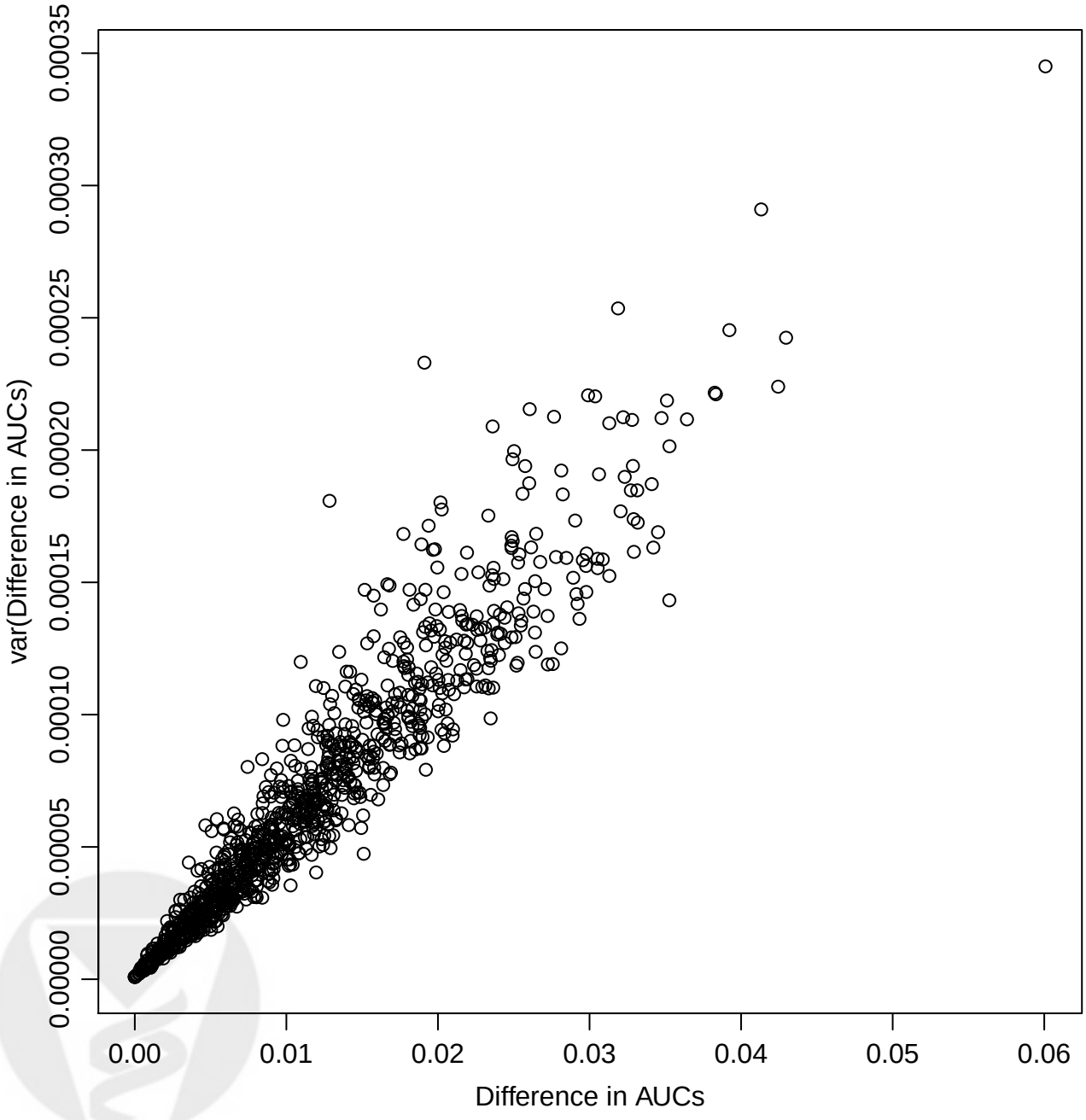
ficient for assessing new predictive markers," *BMC Medical Research Methodology*, 11, 13.

Vuong, Q. H. (1989), "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, 57, 307-333.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

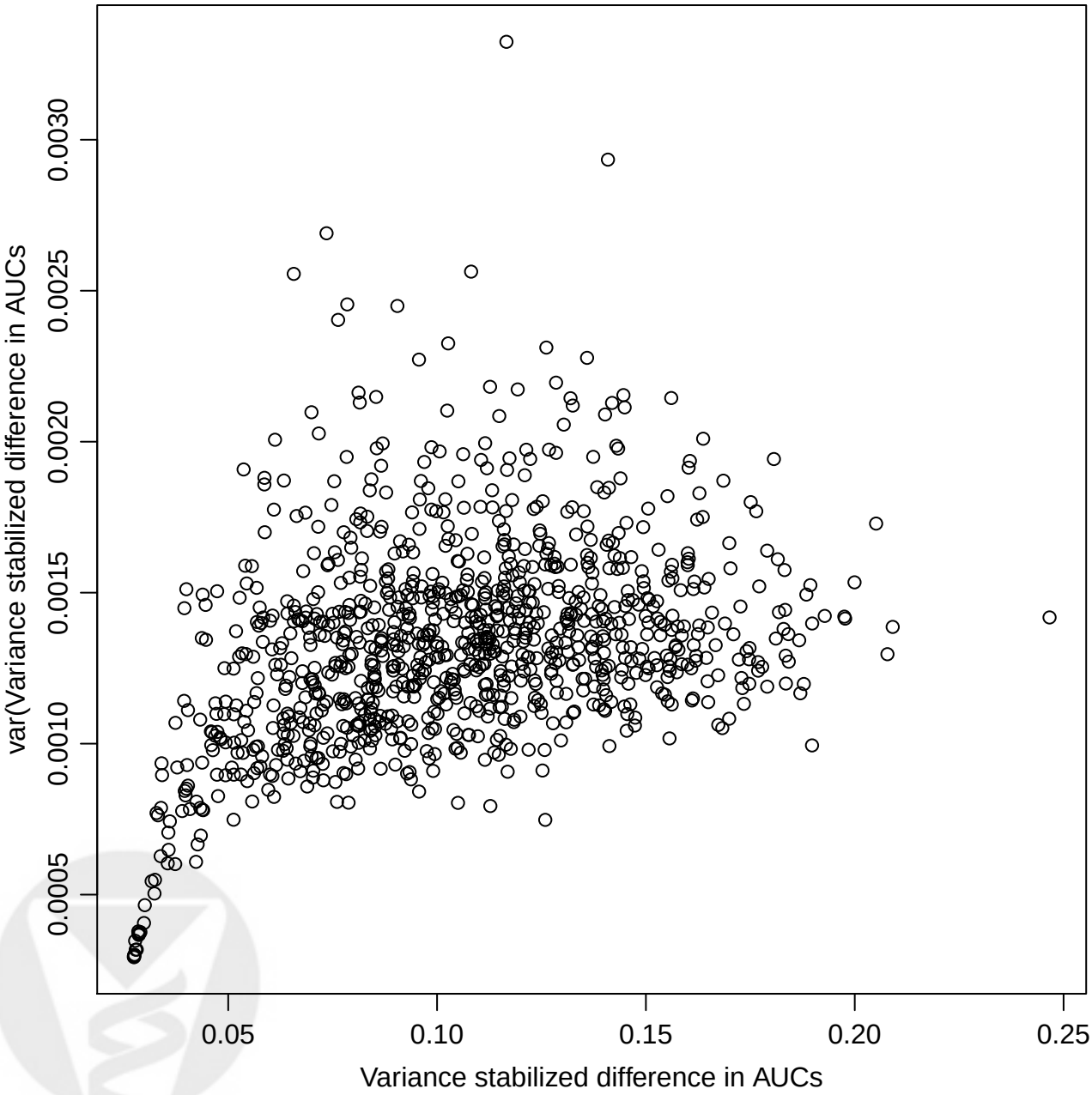
Hosted by The Berkeley Electronic Press

Figure 1a



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Figure 1b



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press