Biostatistics (2017) **18**, 2, *pp.* 260–274
doi:10.1093/biostatistics/kxw045
Advance Access publication on September 21, 2016

Inference for the difference in the area under the ROC curve derived 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 Avenue, New York, NY 10017, USA

hellerg@mskcc.org

Summary

The area under the curve (AUC) statistic is a common measure of model performance in a binary regression model. Nested models are used to ascertain whether the AUC statistic increases when new factors enter the model. The regression coefficient estimates used in the AUC statistics are computed using the maximum rank correlation methodology. Typically, inference for the difference in AUC statistics from nested models is derived under asymptotic normality. In this work, it is demonstrated that the asymptotic normality is true only when at least one of the new factors is associated with the binary outcome. When none of the new factors are associated with the binary outcome, the asymptotic distribution for the difference in AUC statistics is a linear combination of chi-square random variables. Further, when at least one new factor is associated with the outcome and the population difference is small, a variance stabilizing reparameterization improves the asymptotic normality of the AUC difference statistic. A confidence interval using this reparameterization is developed and simulations are generated to determine their coverage properties. The derived confidence interval provides information on the magnitude of the added value of new factors and enables investigators to weigh the size of the improvement against potential costs associated with the new factors. A pancreatic cancer data example is used to illustrate this approach.

Keywords: Area under the receiver operating characteristic curve; Confidence interval; Incremental value; Maximum rank correlation; Nested models; Risk classification model.

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 three-step approach was taken. First, analysts would fit a binary regression model containing both the established factors and the new biomarkers and test whether the association between the outcome and the new markers was statistically significant. If the test of association was significant, using for example a Wald or likelihood ratio test, then the linear predictor function from this model would be used to compute the area under the curve (AUC). Second, an additional statistical test would be carried out

*To whom correspondence should be addressed.

to compare the difference in the AUC for this model and the AUC from a model containing only the established risk factors. Third, if this direct test of AUC equality was significant, a confidence interval was constructed to determine the magnitude of this difference.

In recent years, this multi-step approach has come under criticism. Pepe *and others* (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 from the two models are equal. Thus, it is redundant to perform both the association test and the difference in AUC test. Further, Vickers *and others* (2011); Seshan *and others* (2013) and Pepe *and others* (2013) have illustrated through simulation that the null asymptotic normal distribution assumption for the difference in AUC test does not provide accurate operating characteristics. As a result of these findings, it is recommended that only the test of association be used to infer if the difference in AUCs has improved as a result of the inclusion of new markers.

However, tests of association are not sufficient for understanding the magnitude of the population AUC increase. The new markers may be costly or require an invasive procedure to obtain, and their introduction into a clinical risk prediction model may be justified only if the AUC improvement is meaningful. Conversely, new markers that are not costly and demonstrate no harm to the patient, may have a lower threshold of AUC increase for acceptance. A point estimate for the population difference in AUCs along with a confidence interval for this population difference often provides this important additional information. An interval where the lower confidence bound is close to zero may indicate that the additional factors provide little benefit for use in a clinical decision algorithm. To date, methodology to construct accurate confidence intervals for the difference in AUCs from nested models is incomplete. This work fills the gaps in the AUC methodology by developing a proper null asymptotic distribution for the difference in AUCs and an accurate confidence interval for the population difference when the new markers are associated with the binary outcome.

The outline of the paper is as follows. In Section 2, the nested binary regression models are defined and maximum rank correlation (MRC) methodology is used to estimate the AUC. In Section 3, the asymptotic distribution for the difference in AUCs from nested models is developed. The asymptotic distribution is differentially determined based on whether any of the new factors are associated with the clinical outcome. A confidence interval, derived from a reparameterized population difference in AUC, is proposed in Section 4 and its coverage properties are estimated in Section 5 through simulation. A pancreatic cancer data example is used to illustrate the methodology in Section 6 and a discussion follows in Section 7.

2. The difference in AUCs with nested models

A binary regression model

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

is used to 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, G is a monotone link function. Common link functions for a binary outcome include the logit and the probit.

The model based performance in terms of classification is evaluated using the area under the receiver operating characteristic curve (AUC). The AUC 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 nonresponder's risk score.

Often a new set of markers are under consideration to improve risk classification. This evaluation is based on the difference in AUCs from the nested models

$$\Pr(Y = 1|\boldsymbol{X}, \boldsymbol{Z}) = G(\boldsymbol{\beta}_0^T \boldsymbol{X} + \boldsymbol{\gamma}_0^T \boldsymbol{Z}),$$
$$\Pr(Y = 1|\boldsymbol{X}) = G(\boldsymbol{\beta}^{0^T} \boldsymbol{X}),$$

where the existing markers are denoted by the p-dimensional covariate vector $\boldsymbol{X}$, the new markers are represented by the q-dimensional covariate vector $\boldsymbol{Z}$, and $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\beta}^0)$ represent the true parameter values from the respective models. The estimated AUC 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 $\boldsymbol{x}_i - \boldsymbol{x}_j$, $n_k = \sum_i I[y_i = k]$, and $\hat{\boldsymbol{\beta}}^0$ denotes the $\boldsymbol{\beta}$ parameter estimate when $\boldsymbol{\gamma}$ is set to 0. The difference in the estimated AUCs, derived from the nested models, is written as

$$\tilde{\delta} = \tilde{A}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - \tilde{A}_n(\hat{\boldsymbol{\beta}}^0, 0).$$

Note that the statistic of interest is a function of estimated regression coefficients.

The regression parameter estimates from these nested models are computed using the MRC procedure (Han, 1987). The use of MRC estimates rather than the more commonly applied logistic or probit maximum likelihood estimates results in a simplification in the asymptotic distribution theory, which will be explained further in Section 3, comment 3. 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

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

These estimates are scale invariant (Han, 1987), which creates an identifiability problem for the parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. To resolve the identifiability, the first component of $\boldsymbol{\beta}$ is set to one, and hence $\hat{\boldsymbol{\beta}} = (1, \hat{\boldsymbol{\eta}}^T)^T$, $\hat{\boldsymbol{\beta}}^0 = (1, \hat{\boldsymbol{\eta}}^{0^T})^T$ and the corresponding parameters are denoted by $\boldsymbol{\beta}_0 = (1, \boldsymbol{\eta}_0^T)^T$, $\boldsymbol{\beta}^0 = (1, \boldsymbol{\eta}^{0^T})^T$. Sherman (1993) proves 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. ASYMPTOTIC DISTRIBUTION THEORY

We denote the limiting values of the estimated AUC from the full model and reduced model as $\alpha(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ and $\alpha(\boldsymbol{\beta}^0, 0)$, respectively. Han (1987) demonstrates that these limiting forms represent the maximum population AUCs when the markers are combined linearly. The difference in the limiting AUCs is denoted by

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

and asymptotic distribution theory is derived for inference on this parameter.

A standard approach to derive the asymptotic distribution for a statistic with estimated parameters is via a Taylor series expansion around the true parameter vectors. This expansion, however, requires differentiation with respect to the unknown parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, which is problematic due to the discontinuity 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

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

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

The asymptotic normality of the smoothed AUC parameter estimates and the uniform consistency of the smoothed AUCs are derived in Ma and Huang (2007). As a result, the smoothed versions of the AUC estimates are used to derive the asymptotic distribution of

$$\hat{\delta} = A_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - A_n(\hat{\boldsymbol{\beta}}^0, 0).$$

The asymptotic distribution is derived under two separate conditions: (i) no new factors are associated with the outcome ($\boldsymbol{\gamma}_0 = 0$) and (ii) at least one new factor is associated with the outcome ($\boldsymbol{\gamma}_0 \neq 0$).

3.1. *New factors provide no added value -* $\boldsymbol{\gamma}_0 = 0$ $(\boldsymbol{\beta}_0 = \boldsymbol{\beta}^0)$

The new set of factors are not associated with the clinical outcome, and as a result, the limiting AUCs are equal (Pepe *and others*, 2013). The derived distribution of the difference in the AUC statistic under this condition is useful for deriving a direct test of equality. An approach commonly used to test for the equality of population AUCs from nested models is to apply an asymptotic normal reference distribution to the studentized difference in empirical AUCs (DeLong *and others*, 1988). However, root-n normality is not the correct null reference distribution for this difference. The theorem below provides the asymptotic 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 Assume the following standard conditions for MRC estimation (Han, 1987):

(1) $(\eta, \gamma) \in \Theta$ 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 the first component of $\boldsymbol{x}$ conditional on all other covariates is everywhere positive.

When the new factors are not associated with the response ($\gamma_0 = 0$), as $n \to \infty$,

$$\Pr\left(2n[A_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - A_n(\hat{\boldsymbol{\beta}}^0, 0)] \leq u\right) = \Pr\left(\sum_{j=1}^{q} \lambda_j \chi_j^2 \leq u\right),$$

where $\{\chi_j^2\}$ are independent chi-square random variables each with one degree of freedom and $\{\lambda_j\}$ are the eigenvalues of the product matrix $-V_{\gamma}[D^{\gamma\gamma}]^{-1}$. The matrix V_{γ} is the asymptotic variance of the MRC estimate $\hat{\boldsymbol{\gamma}}$ and D is the second derivative matrix of $A_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$. The partitioned forms of D and its inverse are represented as

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

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: The result in Theorem 1 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 $-D^{\gamma\gamma}$ approximates V_{γ}. It follows that the q eigenvalues of $-V_{\gamma}[D^{\gamma\gamma}]^{-1}$ are each equal to 1, and the result reduces to the standard result that the likelihood ratio test statistic is a chi-square with q degrees of freedom. In addition, Vuong (1989) and Fine (2002) present similar results to Theorem 1 for the likelihood ratio statistic from misspecified nested parametric and semiparametric models.

Comment 3: The first derivative of the AUC, when evaluated at the MRC parameter estimate, is equal to zero. Thus, as a result of using MRC estimates, the quadratic term is the lowest order nonzero term in the asymptotic expansion of the difference in AUCs. Hence, the intrinsic MRC estimates produce a straightforward asymptotic distribution for the difference in AUC statistics. In contrast, if the link function G were specified and the maximum likelihood estimates were used to estimate $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, then the linear and quadratic terms in the Taylor series expansion are nonzero. As a result, maximum likelihood estimation significantly complicates the asymptotic distribution.

Comment 4: Seshan *and others* (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 of the algorithmic maximization search.

3.2. *New factors provide added value*—$\boldsymbol{\gamma}_0 \neq 0$

THEOREM 2 When at least one of the new set of factors is associated with response after controlling for the established risk factors, the difference in nested AUCs is asymptotically represented as

$$\begin{aligned} &n^{1/2}[A_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - A_n(\hat{\boldsymbol{\beta}}^0, 0) - \delta] \\ &\quad = n^{1/2}\left[(n_0 n_1)^{-1} \sum_i \sum_j I[y_i > y_j] \left\{ \Phi\left(\frac{\boldsymbol{\beta}_0^T \boldsymbol{x}_{ij} + \boldsymbol{\gamma}_0^T \boldsymbol{z}_{ij}}{h_n}\right) - \Phi\left(\frac{\boldsymbol{\beta}^{0^T} \boldsymbol{x}_{ij}}{h_n}\right) - \delta \right\}\right] + o_p(1). \end{aligned}$$

The asymptotic expression is the zero order term in the asymptotic expansion and is a two-sample U-statistic of degree 2 with no estimated parameters. It follows from asymptotic U-statistic theory that this expression is asymptotically normal with mean 0. The asymptotic variance estimate from this U-statistic is provided in the Appendix. We again note that if the maximum likelihood estimation for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ were used rather than MRC estimation, the linear term in the asymptotic expansion would be nonzero and would need to be incorporated into the asymptotic variance calculation.

Although asymptotic normality is obtained when $\boldsymbol{\gamma}_0 \neq 0$, statistics derived from nested models (such as the likelihood ratio statistic) tend to be positively skewed with finite samples. For the difference in AUC statistic, Figure 1(a) depicts a plot of this difference $[\hat{\delta} = A_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - A_n(\hat{\boldsymbol{\beta}}^0, 0)]$ and its estimated asymptotic variance $[\hat{V}]$. The points are the realizations of a simulation where the true difference $\delta = 0.01$, the true baseline AUC is 0.70, and the sample size within each replication is 500. The graph indicates a strong linear relationship between the estimated difference in AUCs and its asymptotic variance, indicating that the normal approximation is inaccurate. To remove this mean-variance linear relationship, a square root reparameterization $g(\delta) = \sqrt{\delta}$ is applied. The transformed estimate and its asymptotic variance are

$$\hat{\tau} = \sqrt{\hat{\delta}} \qquad \widehat{\text{var}}(\hat{\tau}) = \frac{\hat{V}}{4\hat{\delta}}.$$

Stemming from comment 4, estimating the regression parameters by maximizing the AUCs in the reduced and full models leads to a non-negative $\hat{\delta}$ and removes a barrier to applying the square root transformation. Figure 1(b) demonstrates the variance stabilization after the square root transformation was applied, suggesting improved accuracy for the normal approximation. Subsequently, we will explore the use of this transformation for the development of accurate confidence intervals.

Finally we note that the asymptotic distribution theory in this section, including the square root transformation, can be applied to develop an accurate test under the null $\delta = \delta_0$, with $\delta_0 \neq 0$. The Wald test for $\boldsymbol{\gamma}_0$ is inappropriate for the nonzero null, since the mapping $f(\delta_0, \boldsymbol{\beta}_0, \boldsymbol{\beta}^0) = \boldsymbol{\gamma}_0$ for the inverse of the limiting difference in AUCs is unknown and not 1-1.

4. Confidence intervals

An interval estimate for the magnitude of the improvement in the AUC due to the inclusion of new factors is important. A confidence interval enables the investigator to weigh this improvement relative to the potential costs in obtaining new markers. An asymptotic 95% confidence interval, derived directly from Theorem 2, for the population difference in AUCs is

$$\text{DIFF} = \left(\hat{\delta} - 1.96\sqrt{\text{var}(\hat{\delta})},\ \hat{\delta} + 1.96\sqrt{\text{var}(\hat{\delta})}\right).$$

A variance stabilizing square root transformation should provide a more accurate asymptotic confidence interval for the difference in the AUC parameters. The 95% confidence interval is obtained by using the reparameterization $\tau = \sqrt{\delta}$, as described above, 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 the confidence interval for δ

$$\text{DIFFvst} = \left(\left\{\hat{\tau} - 1.96\sqrt{\text{var}(\hat{\tau})}\right\}^2,\ \left\{\hat{\tau} + 1.96\sqrt{\text{var}(\hat{\tau})}\right\}^2\right).$$

If $\hat{\tau} - 1.96\sqrt{\text{var}(\hat{\tau})}$ is negative, then the lower confidence bound is set to zero.

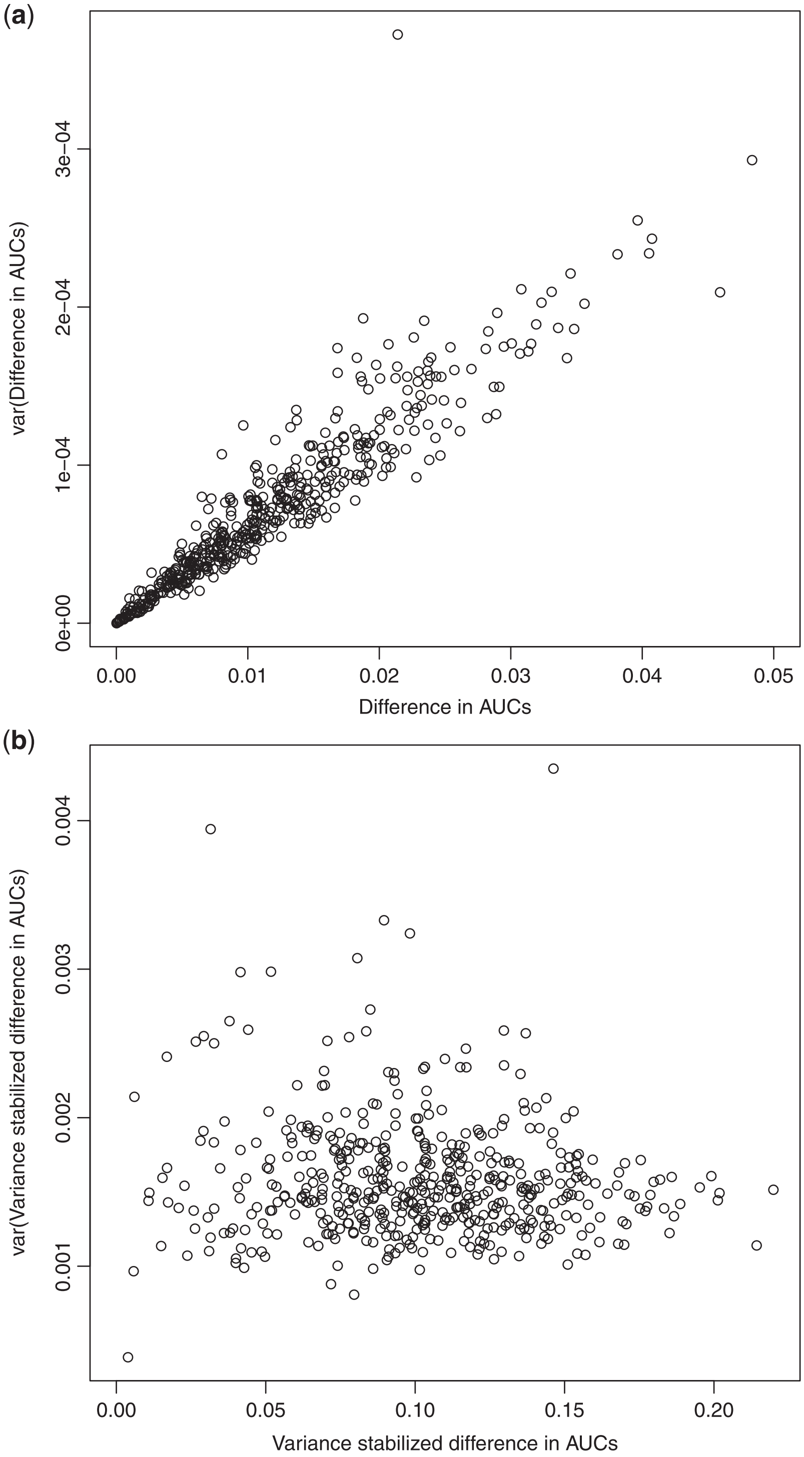

Fig. 1. (a) Difference in AUCs. (b) Variance stabilized difference in AUCs.

5. Simulations

A simulation study was performed to assess the operating characteristics of the direct test of equality of AUCs from nested models and coverage properties of the confidence interval for the difference in AUCs parameter. A binormal logistic risk model was generated with covariate correlation parameters $\{0, 0.5\}$ and $\Pr(Y = 1) = 0.5$. To evaluate the robustness of the proposed method, a probit model with the same covariate structure was generated. Five hundred observations per replicate and 5000 replicates were run for each simulation. The range of population AUCs examined was (0.55–0.85).

For the test statistic, the normal density smooth ϕ was used to compute the second derivative matrix D. Guidance from kernel density estimation led to the bandwidth $h_n = \hat{\omega} n^{-0.20}$, where ω^2 is the variance of $\boldsymbol{\beta}^T \boldsymbol{x} + \boldsymbol{\gamma}^T \boldsymbol{z}$ (Simonoff, 1996). However, when the AUC was near the 0.5 boundary, there were cases when D was not negative definite. For those cases, $\hat{\omega} n^{-\phi}$ $(0.05 < \phi < 0.50)$ were evaluated and the exponent closest to 0.20 (if one existed) which produced a negative definite D was chosen. If a bandwidth could not be found that enabled D to be negative definite, then that replication used $\phi = 0.20$. For confidence interval estimation, the normal distribution function Φ was used to estimate the smooth AUCs and using the kernel smoothing literature for distribution functions, $h_n = \hat{\omega} n^{-0.333}$ was chosen (Lloyd, 1998). The choice of bandwidth used for smoothing in both cases is flexible, since the only asymptotic constraint is that it goes to zero as the sample size gets large.

Tables 1 and 2 compare the size and power estimates for the AUCs test of equality to the Wald test of association for the new factors. The results in Table 1 demonstrate that the difference in AUCs test statistic, based on a linear combination of chi-square random variables as the asymptotic null reference distribution, is a valid test under the null. The results 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 efficient than the nonparametric difference in AUC test.

Table 1. *Size simulations* ($\boldsymbol{\gamma}_0 = 0$). *All entries multiplied by 100*

			Logistic		Probit	
AUCf	AUCr	ρ	LCCS	WALD	LCCS	WALD
0.55	0.55	0	5.30	4.42	5.34	5.42
		0.5	5.34	4.50	5.40	5.34
0.60	0.60	0	4.38	6.02	3.92	5.84
		0.5	4.42	5.94	3.88	5.68
0.65	0.65	0	4.34	5.00	4.00	4.68
		0.5	4.34	5.06	4.00	4.60
0.70	0.70	0	4.70	5.22	4.44	4.78
		0.5	4.70	5.28	4.44	4.74
0.75	0.75	0	4.52	4.64	4.92	5.22
		0.5	4.52	4.62	4.94	5.22
0.80	0.80	0	4.68	4.94	4.56	4.32
		0.5	4.68	4.70	4.56	4.38
0.85	0.85	0	5.60	4.88	4.68	4.76
		0.5	5.62	4.94	4.68	4.76

AUCf = area under the curve for full model with covariates (X, Z);
AUCr = area under the curve for reduced model with covariate X;
ρ = correlation between the covariates (X, Z);
LCCS, linear combination of chi-square random variables;
WALD = Wald statistic.

Table 2. *Power simulations* ($\boldsymbol{\gamma}_0 \neq 0$). *All entries multiplied by 100*

			Logistic		Probit	
δ	AUCr	ρ	LCCS	WALD	LCCS	WALD
0.02	0.55	0	28.80	50.48	30.18	50.12
		0.5	28.00	51.60	29.80	50.24
0.02	0.60	0	63.70	74.68	63.06	74.14
		0.5	65.40	73.40	62.90	74.10
0.01	0.65	0	54.34	61.02	60.56	67.76
		0.5	54.20	60.20	60.38	67.50
0.01	0.70	0	68.46	73.18	75.88	80.82
		0.5	65.00	69.80	75.96	80.72
0.01	0.75	0	81.30	84.38	88.34	91.72
		0.5	80.60	83.40	88.32	91.94
0.005	0.80	0	62.86	65.60	81.68	85.14
		0.5	63.60	65.20	81.66	85.02
0.005	0.85	0	75.14	75.76	96.74	97.82
		0.5	75.60	77.80	96.70	97.70

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.

The coverage properties of the proposed confidence interval are summarized in Table 3. The simulations evaluated the standard asymptotic normal 95% confidence interval for δ (DIFF) and the variance stabilized square root transformed confidence interval for δ (DIFFvst). The variance stabilized confidence interval produced accurate coverage across the simulations explored. In contrast, the untransformed confidence interval was inaccurate. However, at the largest δ (0.05), the difference in coverage between the two methods was small, indicating that as the true difference in AUCs increase, the asymptotic normality of $\hat{\delta}$ improves.

Although the square root transformation produced accurate confidence interval coverage in the simulations, a data-based transformation may prove useful on individual datasets. One approach is to use the Box–Cox transformation

$$\begin{aligned} h(\delta) = & \frac{\delta^{\lambda} - 1}{\lambda} \lambda \neq 0 \\ & \ln(\delta)\ \lambda = 0 \end{aligned}$$

and choose λ to minimize the correlation between $h(\hat{\delta})$ and $\mathrm{var}[h(\hat{\delta})]$ (DiCiccio *and others*, 2006).

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

Table 3. *Coverage estimates for 95% confidence intervals for δ. Average length of intervals in parentheses. All entries multiplied by 100.*

δ	AUCr	ρ	DIFF	DIFFvst	ρ	DIFF	DIFFvst
0.002	0.55	0	94.60 (3.4)	93.62 (6.4)	0.5	94.52 (3.4)	93.68 (6.5)
	0.60	0	91.92 (2.0)	95.70 (4.0)	0.5	92.00 (2.0)	95.72 (4.1)
	0.65	0	89.04 (1.5)	96.54 (3.1)	0.5	88.92 (1.5)	96.48 (3.0)
	0.70	0	87.66 (1.2)	96.28 (2.5)	0.5	87.48 (1.7)	96.18 (2.5)
	0.75	0	86.32 (1.0)	96.20 (2.0)	0.5	86.12 (1.0)	96.16 (2.0)
	0.80	0	86.26 (0.9)	95.88 (1.8)	0.5	86.00 (0.9)	95.76 (1.9)
	0.85	0	85.84 (0.8)	95.80 (1.5)	0.5	85.80 (0.8)	95.68 (1.5)
0.005	0.55	0	89.54 (3.9)	94.06 (6.5)	0.5	89.52 (3.9)	94.12 (6.6)
	0.60	0	86.56 (2.6)	95.34 (4.3)	0.5	86.44 (2.6)	95.46 (4.2)
	0.65	0	86.10 (2.1)	95.92 (3.2)	0.5	86.08 (2.1)	95.90 (3.2)
	0.70	0	85.96 (1.8)	95.54 (2.6)	0.5	85.92 (1.8)	95.42 (2.6)
	0.75	0	87.64 (1.6)	95.48 (2.1)	0.5	87.44 (1.6)	95.52 (2.1)
	0.80	0	89.40 (1.5)	95.10 (1.9)	0.5	89.40 (1.5)	95.22 (1.9)
	0.85	0	89.02 (1.3)	94.18 (1.6)	0.5	89.16 (1.3)	94.20 (1.5)
0.01	0.55	0	86.10 (4.7)	93.74 (7.1)	0.5	86.00 (4.7)	93.72 (7.1)
	0.60	0	86.18 (3.6)	94.52 (4.8)	0.5	86.16 (3.6)	94.46 (4.8)
	0.65	0	88.88 (3.0)	95.14 (3.9)	0.5	88.74 (3.0)	95.16 (3.9)
	0.70	0	88.90 (2.7)	93.90 (3.2)	0.5	89.08 (2.7)	93.88 (3.2)
	0.75	0	90.92 (2.4)	94.16 (2.7)	0.5	90.78 (2.4)	94.20 (2.7)
	0.80	0	92.06 (2.2)	94.40 (2.4)	0.5	91.98 (2.2)	94.38 (2.4)
	0.85	0	91.48 (2.0)	93.90 (2.1)	0.5	91.56 (2.0)	93.96 (2.1)
0.02	0.55	0	86.12 (6.2)	92.04 (8.0)	0.5	86.32 (6.2)	91.94 (8.0)
	0.60	0	88.96 (5.1)	92.86 (6.0)	0.5	88.86 (5.1)	92.96 (6.0)
	0.65	0	90.92 (4.5)	94.32 (4.9)	0.5	91.12 (4.5)	94.38 (4.9)
	0.70	0	91.66 (4.0)	93.70 (4.2)	0.5	91.70 (4.0)	93.78 (4.2)
	0.75	0	92.80 (3.6)	94.42 (3.7)	0.5	92.64 (3.6)	94.44 (3.7)
	0.80	0	93.44 (3.2)	94.38 (3.3)	0.5	93.54 (3.2)	94.42 (3.3)
	0.85	0	92.60 (2.8)	94.06 (2.8)	0.5	92.56 (2.8)	94.08 (2.8)
0.05	0.55	0	90.14 (9.2)	91.40 (9.6)	0.5	90.06 (9.2)	91.34 (9.7)
	0.60	0	92.00 (8.0)	93.20 (8.1)	0.5	92.00 (8.0)	93.32 (8.1)
	0.65	0	93.44 (7.0)	93.90 (7.0)	0.5	93.46 (7.0)	93.86 (7.0)
	0.70	0	93.34 (6.1)	93.82 (6.1)	0.5	93.36 (6.1)	93.86 (6.1)
	0.75	0	93.52 (5.4)	93.90 (5.4)	0.5	93.54 (5.4)	93.84 (5.4)
	0.80	0	94.02 (4.7)	94.02 (4.7)	0.5	94.06 (4.8)	94.04 (4.8)
	0.85	0	93.30 (4.2)	93.54 (4.2)	0.5	93.38 (4.2)	93.58 (4.2)

δ = AUCf − AUCr;
AUCr = area under the curve for reduced model with covariate X;
ρ = correlation between the covariates (X, Z);
DIFF = conventional confidence interval;
DIFFvst = confidence interval using variance stabilizing transformation.

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 *and others*, 2012). Using the data which supported these guidelines one can infer that 40% 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 *and others*, 2013) although they are not yet incorporated into the international guidelines. 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 logistic regression analysis, indicated that recent weight loss is positively associated with high vs. low risk lesions ($p = 0.006$) in the presence of a solid component on imaging, main duct involvement, and the logarithm of lesion size. The MRC AUC estimates from models without and with the weight loss factor were 0.794 and 0.813, 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 confirming the logistic analysis that weight loss is associated with high-risk lesions. The observed difference in model AUCs was $\hat{\delta} = 0.019$ and the test that the added factor increased the population AUC generated a p-value equal to 0.001. The 95% confidence interval for δ, using the variance stabilizing square root transformation, was $(0.0008, 0.052)$. Since the lower bound is close to zero, it is unclear whether adding recent weight loss to the existing clinical factors provides a meaningful benefit to the current surgical risk classification algorithm.

7. DISCUSSION

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 assess the added predictive value of the novel components. This is typically done comparing the AUCs from the full model (containing all variables) and the reduced 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 and if the test is significant compute a confidence interval for the difference in AUCs parameter. 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. In this article we develop the asymptotic theory necessary for the statistical comparison of two AUCs resulting from nested models and provide a method to construct accurate confidence intervals for the difference in AUCs filling another gap in the methodology.

In addition to providing a direct test of equality for the difference in AUCs, the development of the asymptotic distribution theory for the difference in AUCs ($\hat{\delta}$) when its limiting difference (δ) is zero enables the analyst to assess how large $\hat{\delta}$ can be due to sampling variability alone. An upper quantile of this null sampling distribution may be useful when designing future studies to test for the incremental value of new biomarkers. A further usage of this derivation occurs when the objective of the analysis is model selection and the metric used to select variables is AUC. Here the proposed methodology provides a coherent framework for model building and final model selection.

There are other metrics for model performance such as sensitivity and specificity, or more recently introduced metrics such as net benefit (Vickers and Elkin, 2006), net reclassification improvement and integrated discriminant improvement (Pencina *and others*, 2008), and proportion of cases followed and proportion needed to follow-up (Pfeiffer and Gail, 2013; Pfeiffer, 2013). It is noted that the methodological framework, including the smoothing approximation for indicator functions and the distribution theory for nested models, is sufficiently general to be applied to assess the added value of new markers using these metrics. The application of the proposed methodology to these statistics will be explored in future work. These alternative metrics notwithstanding, the AUC remains the most popular measure of medical test performance. It is ubiquitous in clinical, bioinformatic, and radiology journals, and many researchers are familiar with it. The proposed methodology, which provides proper inferential tools to assess the change in AUCs, will prove useful in multiple contexts.

Acknowledgments

Conflict of Interest: None declared.

Funding

National Institutes of Health (grants P30 CA008748 to all and R01 CA136783 to C.S.M.)

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

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

MRC 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 the first component of $\boldsymbol{x}$ conditional on all other covariates is everywhere positive.

Theorem 1 The asymptotic distribution for the difference in AUCs when $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^0)$ are MRC estimates and $\boldsymbol{\gamma}_0 = 0$.

To derive the expansion when $\boldsymbol{\gamma}_0 = 0\,(\boldsymbol{\theta}^0 = \boldsymbol{\theta}_0)$, the difference in AUCs is divided into two components

$$[A_n(\hat{\boldsymbol{\theta}}) - A_n(\boldsymbol{\theta}_0)] - [A_n(\hat{\boldsymbol{\theta}}^0) - A_n(\boldsymbol{\theta}^0)].$$

For the first component, 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 the MRC estimate $\hat{\boldsymbol{\theta}}$ is obtained through maximization of $A_n(\boldsymbol{\theta})$. A similar argument produces a three term expansion of $A_n(\boldsymbol{\theta}^0)$ around $\hat{\boldsymbol{\theta}}^0 = (\hat{\boldsymbol{\eta}}^0, 0)$ for the second component,

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

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

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

Further simplification may be achieved by relating the unrestricted and the restricted MRC estimates $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\eta}}^0$ when $\boldsymbol{\gamma}_0 = 0$ (Cox and Hinkley, 1974, page 308),

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

Thus, the statistic is asymptotically approximated by

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

The quadratic form on the right-hand side asymptotically has a distribution which is a weighted sum of independent chi-square random variables, each with one degree of freedom. Therefore, as $n \to \infty$,

$$\Pr\left(2n[A_n(\hat{\boldsymbol{\theta}}) - A_n(\hat{\boldsymbol{\theta}}^0)] \leq u\right) = \Pr\left(\sum_{j=1}^{q} \lambda_j \chi_j^2 \leq u\right),$$

where the weights $\{\lambda_j\}$ are the eigenvalues of the product matrix $-V_\gamma[D^{\gamma\gamma}]^{-1}$, V_γ is the asymptotic variance of the MRC estimate $\hat{\boldsymbol{\gamma}}$, and D is the second derivative matrix of $A_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$ (Baldessari, 1967).

THEOREM 2 The asymptotic distribution of the difference in AUCs when $\boldsymbol{\gamma}_0 \neq 0$.

Consider the first order asymptotic approximation

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

Because $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\eta}}^0$ maximize their respective smooth AUCs, it follows that

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

Since,

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

is a two-sample U-statistic of degree 2 with no estimated parameters, the asymptotic normality for the difference in AUCs follows from U-statistic theory. Its asymptotic variance is (Wei and Johnson, 1985)

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

which may be estimated with the following components

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

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

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

REFERENCES

BALDESSARI, B. (1967). The distribution of a quadratic form of normal random variables. *Annals of Mathematical Statistics* **38**, 1700–1704.

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. AND 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* **20**, 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.

DICICCIO, T. J., MONTI, A. C. AND YOUNG, G. A. (2006). Variance stabilization for a scalar parameter. *Journal of the Royal Statistical Society, B* **68**, 281–303.

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.

LLOYD, C. J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, **93**, 1356–1364.

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, Jr, R. B. AND RAMACHANDRAN, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**, 157–172.

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.

PFEIFFER, R. M. (2013). Extensions of criteria for evaluating risk prediction models for public health applications. *Biostatistics* **14**, 366–381.

PFEIFFER, R. M. AND GAIL, M. H. (2011). Two criteria for evaluating risk prediction models. *Biometrics* **67**, 1057–1065.

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.

SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer.

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. and YAMAO, K. (2012). International consensus guideline 2012 for the management of IPMN and MCN of the pancreas. *Pancreatology* **12**, 183–197.

VICKERS, A. J. AND ELKIN, E. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* **26**, 565–574.

VICKERS, A. J., CRONIN, A. M. AND BEGG, C.B. (2011). One statistical test is sufficient 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.

WEI, L. J. AND JOHNSON, W. E. (1985). Combining dependent tests with incomplete repeated measurements. *Biometrika* **72**, 359–364.

[*Received February 23, 2016; revised June 27, 2016; accepted for publication July 27, 2016*]