COMPARISON OF PAIRED ROC CURVES THROUGH A TWO-STAGE TEST

Wenbao Yu¹, Eunsik Park¹, and Yuan-Chin Ivan Chang²

¹Department of Statistics, Chonnam National University, Gwangju, South Korea
²Institute of Statistics Science, Academia Sinica, Taipei, Taiwan

The area under the receiver operating characteristic (ROC) curve (AUC) is a popularly used index when comparing two ROC curves. Statistical tests based on it for analyzing the difference have been well developed. However, this index is less informative when two ROC curves cross and have similar AUCs. In order to detect differences between ROC curves in such situations, a two-stage nonparametric test that uses a shifted area under the ROC curve (sAUC), along with AUCs, is proposed for paired designs. The new procedure is shown, numerically, to be effective in terms of power under a wide range of scenarios; additionally, it outperforms two conventional ROC-type tests, especially when two ROC curves cross each other and have similar AUCs. Larger sAUC implies larger partial AUC at the range of low false-positive rates in this case. Because high specificity is important in many classification tasks, such as medical diagnosis, this is an appealing characteristic. The test also implicitly analyzes the equality of two commonly used binormal ROC curves at every operating point. We also apply the proposed method to synthesized data and two real examples to illustrate its usefulness in practice.

Key Words: AUC; Nonparametric test; sAUC; Two-stage test.

1. INTRODUCTION

Evaluating the diagnostic performance of markers or classifiers becomes increasingly important when new diagnostic tools are developed or new markers are proposed to enhance existing diagnostic methods. The receiver operating characteristic (ROC) curve is a popular statistical tool for this purpose. For a given diagnostic test, the ROC curve expresses the true positive rate (TPR) as a function of its corresponding false-positive rate (FPR). It demonstrates the tradeoff between TPR (sensitivity) and FPR(1-specificity).

Since ROC curves involve all possible threshold values for a decision, calibration of these values is embedded in ROC analysis. Beam and Wieand (1991) approached this problem by comparing the TPRs for fixed FPRs. However, it is more common in practice to use methods that consider the ROC curve in its entirety. Under the assumption that the test scores (or after a monotonic transformation) of both diseased and non-diseased subjects follow normal distributions, the ROC curve is characterized by two parameters: a slope and an intercept. A comparison of two ROC curves can be accomplished by
developing a test of the equality of two slopes and two intercepts. This type of ROC curve is called a binormal ROC curve and has been studied by Metz et al. (1984, 1998); Cai and Moskowitz (2004), and Pepe (2004). A likelihood ratio test for the comparison of two ROC curves was developed by Metz et al. (1984) for paired data (i.e., in cases where each patient received both tests).

An alternative approach, which is frequently used in practice, is to consider the area under the curve (AUC) as a measure of accuracy. The AUC summarizes the overall validity of a diagnostic test and represents the probability that the test assigns a higher score to a diseased individual than to a healthy individual (Bamber, 1975). For paired experiments, Hanley and McNeil (1983) developed a test that accounts for the correlations induced by the paired design. DeLong et al. (1988) developed a fully nonparametric approach with nonparametric estimates of all covariances, which leads to a test statistic with an asymptotically standard normal distribution. More recent studies on the AUC include three permutation tests by Bandos et al. (2005), Bandos et al. (2006), and Braun and Alonzo (2008). Several software programs have been released that compare the differences between AUCs, including the pROC package in R and S+ by Robin et al. (2011).

However, statistical tests based only on the AUC are not appropriate when the two ROC curves cross each other and their AUCs are similar. Venkatraman and Begg (1996) suggested a distribution-free procedure to test the equality of two ROC curves at every operating point in the paired design for continuous markers. Their test is more powerful than the AUC test of DeLong et al. (1988) when two curves have similar AUCs. However, as pointed out by Bandos et al. (2005), this test is not as powerful as the conventional AUC test when one ROC curve is uniformly superior to another (i.e., one ROC curve is always located in a higher position than another), and it is computationally burdensome when the sample size is large. Although a general hypothesis test in which the null hypothesis represented equality of the entirety of the ROC curves was constructed, the test was unable to indicate which part of the difference between the ROC curves might occur so that the null hypothesis would be rejected.

The partial area under the ROC curve (pAUC) is a more practical index than AUC in situations when some specific range of FPR is of interest. For example, the area over a limited region of very low FPRs is of great interest to those investigating cancer screenings. However, the inference on pAUC is more difficult than on AUC. McClish (1989) and Thompson and Zucchini (1989) have proposed a parametric estimator of pAUC and its variance using a bi-normal model. Unfortunately, Walsh (1997) demonstrated that inferences derived from the bi-normal assumption are sensitive to mode mis-specification and to the location of decision thresholds. Wieand et al. (1989) have proposed a non-parametric method for pAUC and its asymptotic variance, but their method has not been effectively applied because of its mathematical complexity (He and Escobar, 2008). Zhang et al. (2002) suggested a much simpler nonparametric method than that of Wieand et al. (1989) for deriving variance of pAUC. However, their method was developed for ordinal data. Dodd and Pepe (2003) proposed estimating the variance of pAUC by bootstrap resampling for continuous data and Robin et al. (2011) built such a test of pAUC in their package “pROC”, but it suffers from the same computational intensity as the permutation test.

By definition, the pAUC completely ignores the performance of a marker outside the specified interval. In some cases, it is useful to retain some information from the entire FPR range, with emphases on clinically interesting ranges. In this sense, we propose using a shifted area under the ROC curve (sAUC), which focuses on the differences of two
groups that are bigger than a given level. sAUC is a special case of a modified area under the ROC curve defined by Yu et al. (2013). The sAUC has an interesting property: it associates with the pAUC within a range of low FPRs. This property is important when differentiating two ROC curves, especially when two diagnostic tests have similar AUCs while their ROC curves are quite different.

Considering this, we propose a two-stage test to determine the difference between two ROC curves under different situations. First, a conventional, nonparametric AUC test (as is found in DeLong et al., 1988) will be performed, and then a sAUC test will be conducted if the AUC test does not significantly reject the null hypothesis. The sAUC test is also fully nonparametric and easily implemented. If the conventional AUC test does not reject the null hypothesis and the sAUC test does, then it can be concluded that the diagnostic test/marker with a higher sAUC also has a larger partial AUC in the lower FPR range. The global nominal test size can be allocated differently in two stages based on the asymptotic joint distribution of the two test statistics.

The rest of this article is organized as follows: we define sAUC, demonstrate its properties, and establish a two-stage test in Section 2. Simulation studies are given in Section 3.1 and examples with real datasets are reported in Section 3.2. The discussions are provided in Section 4.

2. METHOD

Throughout the remainder of this article, we refer to the diagnostic tests, biomarkers, and classifiers as “markers” for simplicity. Suppose that $Y$ and $X$ are random scores for diseased and non-diseased subjects, respectively. Without loss of generality, a sample will be classified as a positive one if its score is greater than a threshold $c$. Then, TPR and FPR are defined as $TPR(c) = Pr(Y > c)$ and $FPR(c) = Pr(X > c)$, respectively, and AUC is defined as $\int_0^1 TPR(FPR^{-1}(t))dt$. The partial AUC over a specific false positive rate range $(a, b)$ is defined as $\int_a^b TPR(FPR^{-1}(t))dt$.

2.1. AUC and Shifted AUC

Following the above notations, AUC can be written as $P(Y > X)$ (Bamber, 1975). Since AUC cannot differentiate two crossed ROC curves, we defined an auxiliary index—the shifted area under an ROC curve (sAUC)—as

$$sAUC = Pr(Y > X + \delta),$$

where $\delta = z_{1-\alpha}/\sigma_X$, with $z_{\alpha}$ is the $\alpha$th percentile of a standard normal distribution, and $\sigma_X$ is the standard deviation of the non-diseased group. If we treat location-shifted scores for non-diseased subjects $X + \delta$ as scores from another non-diseased group, we can define a new “ROC curve” by $Y$ and $X + \delta$. We name it a shifted ROC curve.

Suppose two markers, $M_1$ and $M_2$, follow bi-normal distributions. Their ROC curves are shown in Fig. 1(a) and their shifted ROC curves in Fig. 1(b). The two AUCs are set to be equal, but $M2$ has a larger area under its sAUC. This demonstrates that sAUC can be related to pAUC with FPR in the low range.
As seen in Fig. 1, for both ROC curves and shifted ROC curves, $M_2$ has higher curves at the left side of the cross point. Furthermore, the cross point of the shifted ROC curves is shifted to the right from the cross point of the original ROC curves. Thus, sROC curves emphasize their differences more than ROC curves do when FPR is relatively low. More precisely, in Fig. 1, by simple arithmetic, the FPR at the cross point of the ROC curves and the FPR at the cross point of the shifted ROC curves are $\Phi(A)$ and $\Phi(A + z_{1-\alpha/2}/\sigma_X)$ (see Lemma 2.3 in Yu et al. (2013)), respectively. This implies that the cross point of the shifted ROC curves is shifted to the right from the cross point of the ROC curves and the shift size is determined by the parameter $\delta$ or $z_{1-\alpha/2}/\sigma_X$.

Suppose that two ROC curves (with equal AUCs) crossed at FPR = 0.5. In this extreme case, $A = 0$ and the shifted ROC curves crossed at FPR = $\Phi(z_{1-\alpha/2})$. Then, the area of the right segment of the shifted ROC curve is less than $1 - \Phi(z_{1-\alpha/2}) = \alpha/2$. Therefore, if $\alpha$ is small, the sAUC will be close to the area of the left segment of the shifted ROC, as the right segment area becomes negligible. However, we should not specify $\alpha$ to be arbitrarily small. To ensure the sAUC is close to the left segment area, we expect the sAUC be at least as large as $\alpha$, that is, $\Phi(\frac{\mu - z_{1-\alpha/2}\sigma_X}{\sqrt{1+\sigma^2}}) > \alpha$. (Here we assume that the scores are standardized such that the non-diseased scores have mean zero and variance 1, where $\mu$ and $\sigma$ are the mean and variance of the diseased scores). Simple calculations show that $\alpha$ should satisfy

$$z_{1-\alpha/2} < \mu - \sqrt{1 + \sigma^2 z_{\alpha}}.$$  

Therefore, we should select the $\alpha$ to be as small as possible while satisfying the inequality 2.2. $\alpha = 0.05$ is a good choice when $\mu > 0$ and $\sigma^2 > 0.5$, and the corresponding $\delta = 1.96\sigma_X$. $\mu > 0$ is a reasonable assumption since a practical AUC value should be greater than 0.5. Note that if $\sigma^2 < 0.5$, a smaller $\alpha$ should be selected. Since the variance of the diseased in our numerical study is larger than 0.5, we choose $\delta = 1.96\sigma_X$. 

![Figure 1](image_url)
The weighted average of AUC and its shifted AUC together defined the modified area under ROC curve (mAUC)

$$\text{mAUC}(\lambda) = (1 - \lambda) \text{AUC} + \lambda s\text{AUC},$$

by Yu et al. (2013), where $0 \leq \lambda \leq 1$, and a similar graphic illustration can be found therein. Note that $s\text{AUC} = \text{mAUC}(1)$.

**Relationship with pAUC.** So far, we have illustrated that the $s\text{AUC}$ can be related to pAUC and the shift size could be specified to ascertain this relationship. Now, we demonstrate this relationship under bi-normal distributions.

Suppose that the scores of diseased and non-diseased subjects, or after monotonic transformations, follow $N(\mu_Y, \sigma_Y^2)$ and $N(\mu_X, \sigma_X^2)$, respectively, for marker $M_1$, and $N(\mu_{\tilde{Y}}, \sigma_{\tilde{Y}}^2)$ and $N(\mu_{\tilde{X}}, \sigma_{\tilde{X}}^2)$ for marker $M_2$, respectively. Let $\text{AUC}_i$ and $s\text{AUC}_i$ denote AUC and $s\text{AUC}$ for marker $M_i$, respectively, $i = 1, 2$. We can give the analytical formula for AUC and mAUC for each marker: for example, for $M_1$,

$$\text{AUC}_1 = \Phi\left( \frac{\mu_Y - \mu_X}{\sqrt{\sigma_Y^2 + \sigma_X^2}} \right) = \Phi(\Delta)$$

and

$$s\text{AUC}_1 = \Phi(\Delta - \frac{\delta}{\sqrt{\sigma_Y^2 + \sigma_X^2}}),$$

where $\Delta = (\mu_Y - \mu_X)/\sqrt{\sigma_Y^2 + \sigma_X^2}$.

If $\text{AUC}_1 = \text{AUC}_2$, then $s\text{AUC}_1 > s\text{AUC}_2 \iff \text{ROC}_1(\Phi(\Delta)) > \text{ROC}_2(\Phi(\Delta))$, where $\Phi(\Delta)$ is the FPR of the cross point of two ROC curves and $A = (\frac{\sigma_X}{\sigma_Y} - \frac{\sigma_Y}{\sigma_X})^{-1}(\frac{\mu_Y - \mu_X}{\sigma_Y} - \frac{\mu_{\tilde{Y}} - \mu_{\tilde{X}}}{\sigma_{\tilde{Y}}})$. The notation $\text{ROC}_1(t) > \text{ROC}_2(t)$ denotes that the ROC curve of $M_1$ is above the curve of $M_2$ when $0 \leq \text{FPR} \leq t$. The proof is a special case of the result described in Yu et al. (2013).

Additionally, if $\text{AUC}_1 = \text{AUC}_2 = a$, then $s\text{AUC}_1 > s\text{AUC}_2$ implies $\text{ROC}_1(t) > \text{ROC}_2(t)$ for $t = 1 - a$. (Corollary 2.5 in Yu et al. (2013)). This corollary asserts that when AUCs fail to detect the differences between two ROC curves, the $s\text{AUC}$ can still tell the difference between them. The one with a larger $s\text{AUC}$ has a larger pAUC when $\text{FPR} \leq 1 - a (= 1 - \text{AUC})$.

### 2.2. Nonparametric Test Using AUC and sAUC

Suppose that there are $n$ diseased subjects and $m$ non-diseased subjects. Let $Y_i$ and $X_j$ denote the risk scores of the $i$-th and $j$-th subjects in the diseased and the non-diseased groups, respectively. As a special case of mAUC, a nonparametric estimator of sAUC would be defined as

$$s\text{AUC} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \psi(Y_i, X_j + \delta)}{mn}$$

where $\psi(Y_i, X_j + \delta)$ represents the kernel function.
where

$$\psi(y, x) = \begin{cases} 
1 & y > x \\
0.5 & y = x \\
0 & y < x 
\end{cases} \quad (7)$$

In order to construct the covariance of \(\tilde{\mu}_{\text{AUC}}\)s for paired designs, we use the asymptotic property of the linear combination of empirical AUCs under the paired designs of DeLong et al. (1988). First, we establish the asymptotic distribution for the linear combination of \(\mu_{\text{AUC}}\). Then, the distribution for \(\tilde{\mu}_{\text{AUC}}\) follows as a special case.

We adopt the notations of DeLong et al. (1988) and let \(\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p)'\) be a vector of the empirical mAUCs derived from the scores \(\{Y^r_i\}, \{X^r_j\}(i = 1, \ldots, n; j = 1, \ldots, m; 1 \leq r \leq p)\) of \(p\) markers. Let \(\theta\) be the true mAUC vector. \(L = (L^1, L^2, \ldots, L^p)\) is the coefficient vector. \(\tilde{L} = ((1 - \lambda)L', \lambda L')\), which is a \(2p\) dimension vector. Let \(A = (A_1, \ldots, A_{2p})\), where \(A_k\) is the empirical mAUC derived from \(\{Y^k_i\}, \{X^k_j\}\), and \(\hat{A}_{k+p}\) be the empirical AUC derived from \(\{Y^k_i\}, \{X^k_j + \delta^k\}(i = 1, \ldots, n; j = 1, \ldots, m; 1 \leq k \leq p)\). Then, \(L'\hat{\theta} = \tilde{L}'A\). Define

$$S_{10}^r = \frac{1}{n-1} \sum_{i=1}^{n} [V'_{10}^r(Y_i) - \hat{A}_r][V'_{10}(Y_i) - \hat{A}_r]$$

$$S_{01}^r = \frac{1}{m-1} \sum_{j=1}^{m} [V'_{01}^r(X_j) - \hat{A}_r][V'_{01}(X_j) - \hat{A}_r],$$

where \(r, t = 1, \ldots, 2p\), \(V^k_{10}(Y_i) = \sum_{j=1}^{m} \psi(Y^k_i > X^k_j)/m, \quad V^k_{01}(X_j) = \sum_{i=1}^{n} \psi(Y^k_i > X^k_j)/n, \quad V^{p+k}_{10}(Y_i) = \sum_{j=1}^{m} \psi(Y^k_i > X^k_j + \delta^k)/m, \quad V^{p+k}_{01}(X_j) = \sum_{i=1}^{n} \psi(Y^k_i > X^k_j + \delta^k)/n, \quad i = 1, \ldots, n, j = 1, \ldots, m, k = 1, \ldots, p.\)

**Theorem 2.1** Let \(m\) and \(n\) be the sample sizes for the diseased and the non-diseased groups, respectively, and \(\theta, \hat{\theta}, L, \tilde{L}\) be defined as above. Let \(N = m + n\). Then, if \(\lim_{N \to \infty} m/n\) is bounded and non-zero, the asymptotic distribution of linear combination of \(\mu_{\text{AUC}}(\lambda)\)s is as follows

$$\frac{\tilde{L}'\hat{\theta} - L'\theta}{\sqrt{\tilde{L}'(S_{10}/n + S_{01}/m)\tilde{L}}} \tilde{N}(0, 1), \quad \text{as} \quad N \to \infty, \quad (8)$$

where \(S_{10}\) and \(S_{01}\) are \(2p \times 2p\) matrices, with the \((r, t)\)th elements \(S_{10}^r\) and \(S_{01}^r\), respectively \((r, t = 1, 2, \ldots, 2p)\).

Because \(L'\hat{\theta} = \tilde{L}'A\), which is a linear combination of empirical AUCs, Theorem 2.1 directly follows DeLong et al. (1988). In particular, if \(L = (1, -1)'\) and \(\lambda = 1(0)\), then \(L'\hat{\theta}\) becomes the difference of the two \(s\mu_{\text{AUC}}\)s. According to Theorem 2.1, we can perform a large-sample z-test using \(L'\hat{\theta}\) and construct a tolerance interval of \(L'\theta\).
2.3. A Two-Stage Test

Before defining our two-stage test, we argue that neither a single AUC test nor a single sAUC test is appropriate for detecting the difference between the two ROC curves. As mentioned earlier, AUCs can be similar when two ROC curves are quite different; the same is true for sAUCs. Generally speaking, a ROC curve cannot be characterized by a single parameter. Even for “binormal” ROC curves, the curve is decided by two parameters, a slope and an intercept (Cai and Moskowitz, 2004; Metz et al., 1984, 1998; Pepe, 2004). Therefore, to declare the equivalence of two ROC curves, we require at least two equations. For this reason, single AUC and sAUC tests are both invalid in certain situations (see Section 3.1 for a detailed explanation).

However, a sAUC could be used as an auxiliary index of an AUC when the AUC test is not informative. The following theorem implies that identical sAUCs and AUCs together guarantee the identical ROC curves when normal distributions are satisfied.

**Theorem 2.2** Suppose there are two markers, M₁ and M₂. Diseased and non-diseased scores follow $N(\mu_Y, \sigma^2_Y)$ and $N(\mu_X, \sigma^2_X)$, respectively, for marker M₁, and $N(\mu_{\tilde{Y}}, \sigma^2_{\tilde{Y}})$ and $N(\mu_{\tilde{X}}, \sigma^2_{\tilde{X}})$, respectively, for marker M₂. Let $\delta = \sigma_X$ and $\tilde{\delta} = \sigma_{\tilde{X}}$, where c is a positive constant such as 1.96. The following two declarations are equivalent.

a. $AUC_1 = AUC_2$ and $sAUC_1 = sAUC_2$.

b. $ROC_1(t) = ROC_2(t)$ for all $t \in [0,1]$.

The proof is given in the supplemental material. Theorem 2.2 explains that, if both scores of diseased and non-diseased subjects follow (or after a monotonic transformation follow) normal distributions, identical sAUCs and AUCs simultaneously guarantee the identical binormal ROC curves, and vice versa.

The two-stage, two-sided testing procedure that we propose is composed of two nonparametric tests, as follows:

1. Delong’s AUC test: $H_{10}$: $AUC_1 = AUC_2$ vs. $H_{11}$: $AUC_1 \neq AUC_2$. The test statistic is $Z_1 = (\hat{AUC}_1 - \hat{AUC}_2)/se(\hat{AUC}_1 - \hat{AUC}_2)$.

2. sAUC test: $H_{20}$: $sAUC_1 = sAUC_2$ vs. $H_{21}$: $sAUC_1 \neq sAUC_2$. The test statistic is $Z_2 = (s\hat{AUC}_1 - s\hat{AUC}_2)/se(s\hat{AUC}_1 - s\hat{AUC}_2)$, where the standard errors (se) of $\hat{AUC}_1 - \hat{AUC}_2$ and $s\hat{AUC}_1 - s\hat{AUC}_2$ are obtained from the denominator of the left side of equation 2.8. The marginal distributions of $Z_1$ and $Z_2$ (determined by Theorem 2.1) are both standard normal under $H_{10}$ and $H_{20}$, respectively. We illustrated the proposed two-stage test in Fig. 2. If $H_{10}$ is rejected, we can conclude that the two markers have different AUCs and therefore different ROC curves. If we fail to reject $H_{10}$ but succeed in rejecting $H_{20}$, then we can conclude that the two markers have similar AUCs, but the one with the larger sAUC has a larger partial AUC in a range of low FPRs. Or, to be more precise, if both markers are normally distributed, then by Corollary 2.5 in Yu et al. (2013), the one with the larger sAUC will also have a larger partial AUC at $FPR \leq 1 - \hat{AUC}$. 
2.3.1. Asymptotic Joint Distribution of Two Test Statistics. Our procedure is a two-stage hierarchical test. The global test size \( \alpha \) is defined as the probability of incorrectly rejecting the null hypothesis \( H_{10} \) or \( H_{20} \). The size \( \alpha \) can be split into two parts in various ways; different splitting methods lead to different testing powers. To approximate the global test size \( \alpha \), we require the following theorem, which indicates that the asymptotic joint distribution of \( Z_1 \) and \( Z_2 \) is a bivariate normal distribution.

We consider two markers: let \( \{Y_i\}, i = 1, \ldots, n \) and \( \{\bar{X}_j\}, j = 1, \ldots, m \) denote the first marker scores of the diseased and non-diseased subjects, respectively. Similarly, let \( \{\bar{Y}_k\}, k = 1, \ldots, n \) and \( \{\bar{X}_l\}, l = 1, \ldots, m \) for the corresponding scores of the second marker. Define \( \phi^{(1)}(Y_i; X_j; \bar{Y}_k; \bar{X}_l) = \psi(Y_i, X_j) - \psi(\bar{Y}_k, \bar{X}_l) \) and \( \phi^{(2)}(Y_i; X_j; \bar{Y}_k; \bar{X}_l) = \psi(Y_i, X_j + \delta) - \psi(\bar{Y}_k, \bar{X}_l + \delta) \), where \( \psi(y, x) \) was defined in equation 7. Denote \( \delta = cS_X \) and \( \bar{\delta} = cS_{\bar{X}} \) with \( S_X \) and \( S_{\bar{X}} \) for the sample standard deviations of \( Xs \) and \( \bar{X}s \), and \( c \) is any given positive constant. Then, we have the following theorem whose proof is given in the supplemental material.

**Theorem 2.3** Let \( \{Y_i\}, \{X_j\}, \{\bar{Y}_k\}, \) and \( \{\bar{X}_l\} \) be independent samples from distributions \( F, G, \bar{F}, \bar{G} \), respectively \( (i = 1, \ldots, n_1; j = 1, \ldots, n_2; k = 1, \ldots, n_3; \) and \( l = 1, \ldots, n_4) \). Let \( N = \sum_{i=1}^{4} n_i \). If \( n_i/N \) \( (i = 1, \ldots, 4) \) is bounded and non-zero, as \( N \to \infty \), then the joint distribution of \( \sqrt{N}(\bar{AUC}_1 - \bar{AUC}_2 - (AUC_1 - AUC_2)) \) and \( \sqrt{N}(sAUC_1 - sAUC_2 - (sAUC_1 - sAUC_2)) \) is asymptotically bivariate normal with a mean of zero and covariance matrix \( \Sigma = (\sigma_{rs}) \) where

\[
\sigma_{rs} = (N/n_1)\xi_{1000}^{rs} + (N/n_2)\xi_{0100}^{rs} + (N/n_3)\xi_{9001}^{rs} + (N/n_4)\xi_{0001}^{rs}.
\]

Here, \( \xi_{ab}^{rs} \) is the covariance between \( \phi^{(r)}(Y_i; X_j; \bar{Y}_k; \bar{X}_l) \) and \( \phi^{(s)}(Y_i; X_j; \bar{Y}_k; \bar{X}_l) \), \( r, s = 1, 2 \), \( a = I(i = i') \), \( b = I(j = j') \), \( c = I(k = k') \), and \( d = I(l = l') \) where \( I(\cdot) \) is the indicator function.

Moreover, we can estimate the covariance components by extending the estimation method proposed by DeLong et al. (1988), where \( \xi_{1000}^{rs}, \xi_{0100}^{rs}, \xi_{9001}^{rs}, \) and \( \xi_{0001}^{rs} \) are

---

**Figure 2** A two-stage test procedure.
consistently approximated by $S_{1000}^{rs}, S_{0100}^{rs}, S_{0010}^{rs},$ and $S_{0001}^{rs}$, respectively. Let $\Delta_1 = \widehat{\mathrm{AUC}}_1 - \widehat{\mathrm{AUC}}_2$ and $\Delta_2 = \widehat{s\mathrm{AUC}}_1 - \widehat{s\mathrm{AUC}}_2$. Define

$$S_{1000}^{rs} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} [V_{1000}^{rs}(Y_i) - \Delta_t][V_{1000}^{rs}(Y_i) - \Delta_s]$$

where

$$V_{1000}^{rs}(Y_i) = \frac{1}{n_2n_3n_4} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \sum_{l=1}^{n_4} \phi^{(r)}(Y_i; X_j; \hat{Y}_k; \hat{X}_l).$$

$S_{0100}^{rs}, S_{0010}^{rs},$ and $S_{0001}^{rs}$ are defined in a similar way.

Suppose that the two markers $W_i$ and $V_j$ are independent but obtained in a paired design where $W_i = (Y_i, \hat{Y}_i)$ and $V_j = (X_j, \hat{X}_j)$ ($i = 1, \ldots, n$ and $j = 1, \ldots, m$). Define $\phi^{(1)}(W_i; V_j) = \psi(Y_i, X_j) - \psi(\hat{Y}_i, \hat{X}_j)$ and $\phi^{(2)}(W_i; V_j) = \psi(Y_i, X_j + \delta) - \psi(\hat{Y}_i, \hat{X}_j + \hat{\delta})$, where $\psi(y,x)$ was defined in equation 7. Denote $\hat{\delta} = cS_X$ and $\hat{\delta} = cS_X$ with $S_X$ and $S_X$ for the sample standard deviations of $X$s and $\hat{X}$s, and $c$ is any given positive constant. If this is the case, then we have the following theorem:

**Theorem 2.4** Let $\{W_i\}$ and $\{V_j\}$ be independent samples from distributions $F$ and $G$, respectively, with $W_i = (Y_i, \hat{Y}_i)$ and $V_j = (X_j, \hat{X}_j)$ ($i = 1, \ldots, n$ and $j = 1, \ldots, m$). Let $N = n + m$. If $n/N$ is bounded and non-zero, as $N \to \infty$, then the joint distribution of $\sqrt{N}(\widehat{\mathrm{AUC}}_1 - \widehat{\mathrm{AUC}}_2 - (\widehat{\mathrm{AUC}}_1 - \widehat{\mathrm{AUC}}_2))$ and $\sqrt{N}(\widehat{s\mathrm{AUC}}_1 - \widehat{s\mathrm{AUC}}_2 - (\widehat{s\mathrm{AUC}}_1 - \widehat{s\mathrm{AUC}}_2))$ is asymptotically bivariate normal with a mean of zero and covariance matrix $\Sigma = (\sigma_{rs})$ where

$$\sigma_{rs} = \left(\frac{N}{n}\right)\xi_{11}^{rs} + \left(\frac{N}{m}\right)\xi_{01}^{rs}.$$ 

Here $\xi_{ab}^{rs}$ is the covariance between $\phi^{(r)}(W_i; V_j)$ and $\phi^{(s)}(W_i; V_j)$, $r, s = 1, 2$, $a = l(i = r)$ and $b = l$ ($j = f$), where $l(f)$ is the indicator function.

Theorems 2.3 and 2.4 are valid whether $s\widehat{\mathrm{AUC}}$ is defined assuming known ($\delta$) or estimated ($\hat{\delta}$) shift sizes. Proofs for Theorems 2.3 and 2.4 are given in the supplemental material. Similarly, we can estimate the covariance components by extending the estimation method proposed by DeLong et al. (1988), where $\xi_{10}^{rs}$ and $\xi_{01}^{rs}$ could be consistently approximated by $S_{10}^{rs}$ and $S_{01}^{rs}$, respectively. Let

$$\Delta_1 = \widehat{\mathrm{AUC}}_1 - \widehat{\mathrm{AUC}}_2 \quad \Delta_2 = \widehat{s\mathrm{AUC}}_1 - \widehat{s\mathrm{AUC}}_2.$$ 

Define

$$S_{10}^{rs} = \frac{1}{n - 1} \sum_{i=1}^{n} [V_{10}^{rs}(W_i) - \Delta_t][V_{10}^{rs}(W_i) - \Delta_s] \quad \text{and}$$

$$S_{01}^{rs} = \frac{1}{m - 1} \sum_{j=1}^{m} [V_{01}^{rs}(V_j) - \Delta_t][V_{01}^{rs}(V_j) - \Delta_s],$$ 

where
According to Theorem 2.4, the joint distribution of \( Z_1 \) and \( Z_2 \) follows a standard bivariate normal distribution with a correlation \( \gamma \) obtained using the above covariances. Then, the overall test size \( \alpha \) is decided by

\[
\alpha = 1 - pr(\left| Z_1 \right| < c_1, \left| Z_2 \right| < c_2 | H_{10}, H_{20}) \]

\[
= 1 - \int_{-c_1}^{c_1} \int_{-c_2}^{c_2} f(x, y | \gamma) dx dy
\]

(9)

where \( f(x, y | \gamma) \) is the pdf of the standard bivariate normal distribution with the correlation \( \gamma \). The test sizes of the AUC and sAUC tests are denoted by \( \alpha_1 \) and \( \alpha_2 \), respectively. We can approximate \( \gamma \) using the estimated asymptotic covariance components \( S_{10}^{rs} \) and \( S_{01}^{rs} \) by

\[
\hat{\gamma} = \frac{S_{10}^{12} + S_{01}^{12}}{\sqrt{(S_{10}^{11} + S_{01}^{11})(S_{10}^{21} + S_{01}^{21})}}.
\]

(10)

Simulation studies indicate that this asymptotic estimator, \( \hat{\gamma} \), leads to a test size similar to the nominal size (see Table 1).

### 3. NUMERICAL STUDY

To investigate the proposed two-stage test, we verified that the empirical test size was close to the nominal size and then compared the powers of the two-stage test (TwoT) and four single stage tests, including Delong’s AUC test (AUCT), sAUC test (sAUCT),
mAUC(0.5) Test (mAUC_{0.5,T}) and Venkatraman’s permutation test (PermT). Here, mAUC(t) means mAUC with \( \lambda = t \).

### 3.1. Simulation Study

Consider two markers \( M_1 \) and \( M_2 \) whose correlation coefficients are \( \rho = 0 \) or 0.5. Suppose that the scores for non-diseased individuals are generated from \( N(0,1) \), and the scores for diseased ones follow \( N(u_i, \sigma^2_i) \) for the marker \( M_i \), where \( i = 1, 2 \).

To examine the test size, allow the two markers to be identically distributed. For the power comparison, we first studied four scenarios: in scenario \( S_1 \), the two corresponding ROC curves crossed each other while having an identical AUC; in scenario \( S_2 \), one marker was uniformly superior to the other; in scenario \( S_3 \), two ROC curves crossed and mAUCs were similar; and in scenario \( S_4 \), two ROC curves crossed and the sAUCs were similar. Examples of the ROC curves in the four scenarios are drawn in Fig. 3.

The empirical test sizes are shown in Table 1; in all cases, the global nominal size \( \alpha \) is set to 0.05 or 0.1 and \( \alpha_1 = \alpha_2 \). The empirical test size was the rejection rate of 5000 simulations. The empirical test sizes were similar to the nominal levels. The critical values for the two-stage test procedure are listed in Table 2 for the different global test sizes and different correlations between \( Z_1 \) and \( Z_2 \).

The power comparison of the two-stage test procedure with the AUC, sAUC, mAUC(0.5), and permutation tests are listed in Tables 3–6, for scenarios \( S_1-S_4 \), respectively. Generally speaking, no single test outperformed the others in all of the scenarios. The single-stage tests using AUC, sAUC, and mAUC(0.5) were invalid in scenarios \( S_1 \), \( S_3 \), and \( S_4 \), respectively, and the two-stage test was very powerful in all scenarios even though it did not obtain the highest power in some scenarios. The permutation test was far less powerful than the two-stage test in scenarios \( S_1 \) and \( S_2 \). In all scenarios, correlated markers led to higher powers than independent markers for all tests. This was consistent with the finding of Venkatraman and Begg (1996) and Bandos et al. (2005). The difference between the correlated markers of the AUCs (sAUCs) was less variant than that of the independent markers; therefore, the Z-statistics of the AUC (sAUC) became larger than those of the independent markers and the power (i.e., the rejection rate) correspondingly increased.

In scenario \( S_1 \) (Table 3), the power of the AUC test was close to the nominal test size when the compared AUCs were equal. This means that the AUC test was not appropriate for detecting different ROC curves in this situation. In this case, the permutation test was more powerful than the AUC test but less powerful than the three mAUC-based tests. The sAUC test displayed the highest power in this scenario and the two-stage test gave slightly less power than the sAUC.

In scenario \( S_2 \), all three tests based on the AUC or sAUC detected the difference between two ROC curves effectively. The mAUC(0.5) test had the highest power; the two-stage test had a similar power. As seen in Table 4, the permutation test was not as efficient as the conventional AUC test, which had been pointed out by Bandos et al. (2005). Finally, the permutation test was less powerful than all of the other tests.

In scenario \( S_3 \) (Table 5), the mAUC(0.5) test was less efficient than the others. The two-stage test and the permutation test performed similarly and had higher powers than the others. In scenario \( S_4 \) (Table 6), the sAUC test was extremely inferior to the others.
We next considered the ROC curves of three scenarios when the underline scores are asymmetrically distributed: mild skewed, moderately skewed, and bimodal generated from mixture normal distributions. Two compared markers have scores in the form of $0.4X_1 + 0.6X_2$ and $0.6X_1 + 0.4X_2$ for the diseased, respectively, where $X_1$ and $X_2$ are from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. The non-diseased follow standard normal distributions. The distribution parameters and the power for each test are given in Table 7. All sAUC-based tests outperform AUC test and the permutation test, and the two-stage test still achieves the second-most optimum power.

In summary, the two-stage test attained a high power in all scenarios, while each of the four single-stage tests was extremely uninformative in certain specific scenarios. Additionally, our two-stage test had a more meaningful characteristic than the other tests. In scenario $S_1$, we noticed that marker $M_2$ with a larger sAUC had a higher partial
ROC curve when the FPR was less than $1 - AUC$. For instance, in Figure 3(a), the sAUCs were 0.2 and 0.36 for $M_1$ and $M_2$, respectively, and both of the AUCs were equal to 0.70. The $M_2$ had a higher partial ROC curve when the FPR was less than 0.3, as expected based on Corollary 2.5 in Yu et al. (2013). This characteristic is meaningful, especially when the feature selection method is based on the AUC and shares many markers with similar AUCs. In these cases, our test procedure could be applied to differentiate the markers.

Our two-stage test incorporates the single-stage AUC test and the sAUC test; in each scenario explained above, either the AUC or sAUC tests were very powerful. This might be the reason that the two-stage test had uniform, powerful performances. The two-stage test is a broad testing framework that consists of the AUC and sAUC tests, examining the “path” between them. That is, if the $\alpha_2$ increases from 0 to $\alpha$, the two-stage test would shrink from the AUC test to the sAUC test. Assigning different sizes in each stage seems to be better, depending on the scenario. The following are guidelines to split the test size based on these findings:

- If there is prior knowledge that the two ROC curves cross with similar AUCs, or if lower ranges of FPRs are of interest, give the larger size to the sAUC test or only conduct one sAUC test.
- If there is prior knowledge that the two ROC curves cross with different AUCs, give the larger test size to the AUC test or only conduct one AUC test.
- If you have no prior knowledge of the ROC curves, assign equal sizes to each stage for simplicity.

In most practical situations, before implementing the test, people are unsure about which scenario the two compared ROC curves belong to; if they were sure, the two-stage test would not be necessary. Therefore, we highly recommend the proposed two-stage test with identical sizes for each stage.
Table 3  Empirical test power comparison for scenario $S_1$, where two ROC curves cross and their AUCs are identical when $\alpha_1 = \alpha_2$ and $\alpha = 0.05$

| AUC$_1$ | AUC$_2$ | sAUC$_1$ | sAUC$_2$ | mAUC$_1(0.5)$ | mAUC$_2(0.5)$ | $\mu_1$ | $\sigma_1$ | $\mu_2$ | $\sigma_2$ | $n$ | $\rho$ | TwoT | AUCT | sAUCT | mAUC$_{0.5T}$ | PermT |
|---------|---------|---------|---------|---------------|---------------|--------|---------|--------|---------|-----|------|------|------|------|-------------|-------|
| 0.6     | 0.6     | 0.13    | 0.27    | 0.37          | 0.44          | 0.36   | 1       | 0.56   | 2       | 40  | 0.0  | 0.422 | 0.053 | 0.481 | 0.140 | 0.146 |
|         |         |         |         |               |               |        |         |        |         |     | 0.5  | 0.623 | 0.049 | 0.700 | 0.226 | 0.291 |
|         |         |         |         |               |               |        |         |        |         |     | 80   | 0.0  | 0.747 | 0.050 | 0.793 | 0.239 | 0.359 |
|         |         |         |         |               |               |        |         |        |         |     | 0.5  | 0.924 | 0.052 | 0.947 | 0.393 | 0.686 |
| 0.7     | 0.7     | 0.19    | 0.37    | 0.45          | 0.54          | 0.74   | 1       | 1.19   | 2       | 40  | 0.0  | 0.493 | 0.049 | 0.557 | 0.201 | 0.131 |
|         |         |         |         |               |               |        |         |        |         |     | 0.5  | 0.724 | 0.052 | 0.789 | 0.334 | 0.280 |
|         |         |         |         |               |               |        |         |        |         |     | 80   | 0.0  | 0.816 | 0.050 | 0.848 | 0.359 | 0.321 |
|         |         |         |         |               |               |        |         |        |         |     | 0.5  | 0.964 | 0.050 | 0.975 | 0.578 | 0.631 |
| 0.8     | 0.8     | 0.29    | 0.49    | 0.55          | 0.65          | 1.19   | 1       | 1.88   | 2       | 40  | 0.0  | 0.508 | 0.047 | 0.575 | 0.253 | 0.114 |
|         |         |         |         |               |               |        |         |        |         |     | 0.5  | 0.756 | 0.054 | 0.809 | 0.408 | 0.229 |
|         |         |         |         |               |               |        |         |        |         |     | 80   | 0.0  | 0.833 | 0.051 | 0.856 | 0.437 | 0.287 |
|         |         |         |         |               |               |        |         |        |         |     | 0.5  | 0.977 | 0.045 | 0.982 | 0.678 | 0.541 |
| 0.9     | 0.9     | 0.46    | 0.66    | 0.68          | 0.78          | 1.81   | 1       | 2.87   | 2       | 40  | 0.0  | 0.521 | 0.044 | 0.588 | 0.338 | 0.095 |
|         |         |         |         |               |               |        |         |        |         |     | 0.5  | 0.768 | 0.049 | 0.818 | 0.518 | 0.165 |
|         |         |         |         |               |               |        |         |        |         |     | 80   | 0.0  | 0.842 | 0.054 | 0.863 | 0.562 | 0.227 |
|         |         |         |         |               |               |        |         |        |         |     | 0.5  | 0.978 | 0.046 | 0.983 | 0.799 | 0.370 |

Note. Two markers have follow distribution $N(u_i, \sigma^2_i)$ for diseased subjects, respectively ($i = 1, 2$) with their correlation $\rho$. The power is the rejection rate over 5000 simulations. TwoT, AUCT, sAUCT, mAUC$_{0.5T}$, and PermT stand for the two-stage test, AUC test, sAUC test, mAUC$(0.5)$ test, and permutation test, respectively.
Table 4  Empirical test power comparison for scenario $S_2$, where two ROC curves parallel when $\alpha_1 = \alpha_2$ and $\alpha = 0.05$

| $AUC_1$ | $AUC_2$ | $sAUC_1$ | $sAUC_2$ | $mAUC_1(0.5)$ | $mAUC_2(0.5)$ | $\mu_1$ | $\sigma_1$ | $\mu_2$ | $\sigma_2$ | $n$ | $\rho$ | $\text{TwoT}$ | $\text{AUCT}$ | $sAUCT$ | $mAUC_{0.5T}$ | $\text{PermT}$ |
|--------|--------|--------|--------|-------------|-------------|--------|--------|--------|--------|----|-----|----------|----------|--------|-------------|----------|
| 0.6    | 0.7    | 0.13   | 0.19   | 0.37        | 0.45        | 0.36   | 1      | 0.74   | 1      | 40 | 0.0 | 0.200   | 0.210   | 0.171 | 0.216      | 0.186    |
|        |        |        |        |             |             |        |        |        |        |    | 0.5 | 0.343   | 0.344   | 0.264 | 0.380      | 0.291    |
|        |        |        |        |             |             |        |        |        |        |    | 80  | 0.0    | 0.371   | 0.372 | 0.314      | 0.392    | 0.335    |
|        |        |        |        |             |             |        |        |        |        |    | 0.5 | 0.598   | 0.610   | 0.508 | 0.653      | 0.536    |
| 0.7    | 0.8    | 0.19   | 0.29   | 0.45        | 0.55        | 0.74   | 1      | 1.19   | 1      | 40 | 0.0 | 0.269   | 0.260   | 0.255 | 0.291      | 0.248    |
|        |        |        |        |             |             |        |        |        |        |    | 0.5 | 0.465   | 0.418   | 0.419 | 0.509      | 0.375    |
|        |        |        |        |             |             |        |        |        |        |    | 80  | 0.0    | 0.497   | 0.459 | 0.460      | 0.510    | 0.439    |
|        |        |        |        |             |             |        |        |        |        |    | 0.5 | 0.748   | 0.713   | 0.717 | 0.803      | 0.653    |
| 0.8    | 0.9    | 0.29   | 0.46   | 0.55        | 0.68        | 1.19   | 1      | 1.81   | 1      | 40 | 0.0 | 0.461   | 0.389   | 0.460 | 0.490      | 0.369    |
|        |        |        |        |             |             |        |        |        |        |    | 0.5 | 0.720   | 0.579   | 0.739 | 0.774      | 0.548    |
|        |        |        |        |             |             |        |        |        |        |    | 80  | 0.0    | 0.761   | 0.661 | 0.754      | 0.775    | 0.643    |
|        |        |        |        |             |             |        |        |        |        |    | 0.5 | 0.956   | 0.879   | 0.959 | 0.971      | 0.848    |
| 0.6    | 0.8    | 0.13   | 0.29   | 0.37        | 0.55        | 0.36   | 1      | 1.19   | 1      | 40 | 0.0 | 0.708   | 0.695   | 0.646 | 0.749      | 0.671    |
|        |        |        |        |             |             |        |        |        |        |    | 0.5 | 0.936   | 0.921   | 0.886 | 0.956      | 0.897    |
|        |        |        |        |             |             |        |        |        |        |    | 80  | 0.0    | 0.948   | 0.936 | 0.916      | 0.958    | 0.918    |
|        |        |        |        |             |             |        |        |        |        |    | 0.5 | 0.999   | 0.998   | 0.994 | 1.000      | 0.996    |
| 0.7    | 0.9    | 0.19   | 0.46   | 0.45        | 0.68        | 0.74   | 1      | 1.81   | 1      | 40 | 0.0 | 0.902   | 0.854   | 0.902 | 0.921      | 0.842    |
|        |        |        |        |             |             |        |        |        |        |    | 0.5 | 0.995   | 0.979   | 0.994 | 0.997      | 0.970    |
|        |        |        |        |             |             |        |        |        |        |    | 80  | 0.0    | 0.996   | 0.990 | 0.995      | 0.998    | 0.984    |
|        |        |        |        |             |             |        |        |        |        |    | 0.5 | 1.000   | 1.000   | 1.000 | 1.000      | 0.999    |

Note: Two markers follow distribution $N(u_i, \sigma_i^2)$ for diseased $N$ subjects with their correlation $\rho$ and AUC value $AUC_i$, respectively ($i = 1, 2$). The power is the rejection rate over 5000 simulations. TwoT, AUCT, sAUCT, mAUC_{0.5T}, and PermT stand for the two-stage test, AUC test, sAUC test, mAUC(0.5) test, and permutation test, respectively.
Table 5: Empirical test power comparison for scenario S3 that two mAUC(0.5)s are close, with $\alpha_1 = \alpha_2$ and $\alpha = 0.05$

| AUC1 | AUC2 | sAUC1 | sAUC2 | mAUC1(0.5) | mAUC2(0.5) | $\mu_1$ | $\sigma_1$ | $\mu_2$ | $\sigma_2$ | n | TwoT | AUCL | sAUCL | mAUC0.5L | PermT |
|------|------|-------|-------|------------|------------|--------|-----------|--------|-----------|----|-------|------|-------|-----------|-------|
| 0.7  | 0.63 | 0.19  | 0.29  | 0.45       | 0.46       | 1      | 0.74      | 2      | 40        | 0  | 0.270 | 0.125 | 0.242 | 0.050     | 0.240 |
|      |      |       |       |           |           |        |           |        |           | 0.5| 0.414 | 0.191 | 0.375 | 0.059     | 0.399 |
|      |      |       |       |           |           |        |           |        |           | 80 | 0.518 | 0.203 | 0.447 | 0.062     | 0.475 |
|      |      |       |       |           |           |        |           |        |           | 0.5| 0.763 | 0.325 | 0.647 | 0.067     | 0.765 |
|      |      |       |       |           |           |        |           |        |           | 80 | 0.406 | 0.363 | 0.188 | 0.060     | 0.507 |
|      |      |       |       |           |           |        |           |        |           | 0.5| 0.768 | 0.633 | 0.339 | 0.064     | 0.875 |
| 0.5  | 0.5  | 0.29  | 0.37  | 0.55       | 0.54       | 1.19   | 1         | 1.19   | 2         | 40 | 0.267 | 0.239 | 0.133 | 0.055     | 0.336 |
|      |      |       |       |           |           |        |           |        |           | 0.5| 0.406 | 0.363 | 0.188 | 0.060     | 0.507 |
|      |      |       |       |           |           |        |           |        |           | 80 | 0.512 | 0.422 | 0.234 | 0.057     | 0.610 |
|      |      |       |       |           |           |        |           |        |           | 0.5| 0.768 | 0.633 | 0.339 | 0.064     | 0.875 |
| 0.7  | 0.59 | 0.19  | 0.35  | 0.45       | 0.47       | 0.74   | 1         | 0.74   | 3         | 40 | 0.538 | 0.224 | 0.481 | 0.056     | 0.525 |
|      |      |       |       |           |           |        |           |        |           | 0.5| 0.722 | 0.328 | 0.660 | 0.067     | 0.786 |
|      |      |       |       |           |           |        |           |        |           | 80 | 0.905 | 0.393 | 0.766 | 0.075     | 0.903 |
|      |      |       |       |           |           |        |           |        |           | 0.5| 0.994 | 0.579 | 0.922 | 0.080     | 0.990 |
| 0.8  | 0.65 | 0.29  | 0.40  | 0.55       | 0.58       | 1.19   | 1         | 1.19   | 3         | 40 | 0.548 | 0.450 | 0.232 | 0.117     | 0.656 |
|      |      |       |       |           |           |        |           |        |           | 0.5| 0.770 | 0.642 | 0.336 | 0.113     | 0.862 |
|      |      |       |       |           |           |        |           |        |           | 80 | 0.913 | 0.751 | 0.411 | 0.131     | 0.951 |
|      |      |       |       |           |           |        |           |        |           | 0.5| 0.994 | 0.911 | 0.604 | 0.142     | 0.997 |

Note: Two markers follow distribution $N(u_i, \sigma_i)$ for diseased subjects with their correlation $\rho$ and AUC value $AUC_i$ respectively ($i = 1, 2$). The power is the rejection rate over 5000 simulations. TwoT, AUCL, sAUCL, mAUC0.5L, and PermT stand for the two-stage test, AUCL test, sAUCL test, mAUC(0.5) test and permutation test, respectively.
Table 6 Empirical test power comparison for scenario $S_4$ that two ROC curves are different and sAUCs are similar with $\alpha = \alpha_2$ and $\alpha = 0.05$

| AUC_1 | AUC_2 | sAUC_1 | sAUC_2 | mAUC_1(0.5) | mAUC_2(0.5) | $\mu_1$ | $\sigma_1$ | $\mu_2$ | $\sigma_2$ | $n$ | $\rho$ | TwoT | AUCT | sAUCT | mAUC(0.5)_T | PermT |
|-------|-------|--------|--------|-------------|-------------|--------|----------|--------|----------|-----|-------|-------|-------|-------|-----------|-------|
| 0.88  | 0.74  | 0.19   | 0.19   | 0.54        | 0.47        | 0.74   | 1        | 0.03   | 2        | 40  | 0.0   | 0.516 | 0.585 | 0.045 | 0.251 | 0.661 |
|       |       |        |        |             |             | 0.5    | 0.761    | 0.817  | 0.048    | 0.425| 0.873 | 0.0   | 0.832 | 0.863 | 0.053 | 0.461 | 0.917 |
|       |       |        |        |             |             | 80     | 0.0      | 0.974  | 0.980    | 0.048| 0.693 | 0.0   | 0.974 | 0.980 | 0.053 | 0.693 | 0.992 |
| 0.8   | 0.63  | 0.29   | 0.29   | 0.55        | 0.46        | 1.19   | 1        | 0.74   | 2        | 40  | 0.0   | 0.488 | 0.554 | 0.045 | 0.199 | 0.626 |
|       |       |        |        |             |             | 0.5    | 0.710    | 0.779  | 0.052    | 0.323| 0.828 | 0.0   | 0.809 | 0.831 | 0.049 | 0.354 | 0.907 |
|       |       |        |        |             |             | 80     | 0.0      | 0.963  | 0.972    | 0.051| 0.563 | 0.0   | 0.963 | 0.972 | 0.051 | 0.563 | 0.987 |
| 0.88  | 0.74  | 0.46   | 0.46   | 0.67        | 0.60        | 1.81   | 1        | 1.72   | 2        | 40  | 0.0   | 0.398 | 0.465 | 0.051 | 0.130 | 0.528 |
|       |       |        |        |             |             | 0.5    | 0.568    | 0.652  | 0.055    | 0.194| 0.717 | 0.0   | 0.724 | 0.769 | 0.055 | 0.216 | 0.841 |
|       |       |        |        |             |             | 80     | 0.0      | 0.904  | 0.927    | 0.050| 0.343 | 0.0   | 0.904 | 0.927 | 0.050 | 0.343 | 0.962 |
| 0.89  | 0.77  | 0.74   | 0.74   | 0.82        | 0.76        | 2.87   | 1        | 3.40   | 2        | 40  | 0.0   | 0.140 | 0.202 | 0.052 | 0.064 | 0.259 |
|       |       |        |        |             |             | 0.5    | 0.159    | 0.255  | 0.048    | 0.071| 0.316 | 0.0   | 0.401 | 0.500 | 0.057 | 0.086 | 0.546 |
|       |       |        |        |             |             | 80     | 0.0      | 0.510  | 0.618    | 0.059| 0.103 | 0.0   | 0.510 | 0.618 | 0.059 | 0.103 | 0.715 |

Note. Two markers follow distribution $N(u_i, \sigma^2_i)$ for diseased subjects with their correlation $\rho$ and AUC value $\text{AUC}_\rho$ respectively ($i = 1, 2$). The power is the rejection rate over 5000 simulations. TwoT, AUCT, sAUCT, mAUC(0.5)_T, and PermT stand for the two-stage test, AUC test, sAUC test, mAUC(0.5) test and permutation test, respectively.
Table 7  Empirical test power comparison for ROC curves generated from asymmetric distributions, with $\alpha_1 = \alpha_2$ and $\alpha = 0.05$

| skew    | $AUC_1$ | $AUC_2$ | $sAUC_1$ | $sAUC_2$ | $mAUC_1(0.5)$ | $mAUC_2(0.5)$ | $\mu_1$ | $\sigma_1$ | $\mu_2$ | $\sigma_2$ | $n$ | TwoT | AUCT | sAUCT | mAUC0.5T | PermT |
|---------|---------|---------|----------|----------|---------------|---------------|---------|------------|---------|------------|-----|-------|------|--------|-----------|-------|
| Mild    | 0.75    | 0.73    | 0.36     | 0.30     | 0.55          |               | 0.51    | 0.7        | 1       | 1.8        | 2   | 40    | 0.100| 0.059 | 0.121    | 0.090 | 0.058 |
|         | 0.90    | 0.90    | 0.58     | 0.54     | 0.74          |               | 0.72    | 1.8        | 1       | 2.9        | 2   | 40    | 0.069| 0.044 | 0.083    | 0.068 | 0.058 |
|         | 0.90    | 0.90    | 0.58     | 0.54     | 0.74          |               | 0.72    | 1.8        | 1       | 2.9        | 2   | 40    | 0.069| 0.044 | 0.083    | 0.068 | 0.058 |
| Moderate| 0.75    | 0.71    | 0.39     | 0.31     | 0.57          |               | 0.51    | 0.4        | 1       | 2.3        | 2   | 40    | 0.168| 0.105 | 0.196    | 0.164 | 0.105 |
|         | 0.91    | 0.89    | 0.65     | 0.52     | 0.78          |               | 0.69    | 1.5        | 1       | 3.5        | 2   | 40    | 0.128| 0.036 | 0.160    | 0.135 | 0.077 |
|         | 0.91    | 0.89    | 0.65     | 0.52     | 0.78          |               | 0.69    | 1.5        | 1       | 3.5        | 2   | 40    | 0.128| 0.036 | 0.160    | 0.135 | 0.077 |
| Bimodal | 0.75    | 0.64    | 0.50     | 0.35     | 0.63          | -0.3          | 0.50    | -0.3       | 1       | 3.5        | 1.5 | 40    | 0.359| 0.306 | 0.371    | 0.362 | 0.275 |
|         | 0.90    | 0.85    | 0.65     | 0.52     | 0.78          |               | 0.69    | 1         | 1       | 4.5        | 1.5 | 40    | 0.287| 0.155 | 0.337    | 0.295 | 0.150 |
|         | 0.90    | 0.85    | 0.65     | 0.52     | 0.78          |               | 0.69    | 1         | 1       | 4.5        | 1.5 | 40    | 0.287| 0.155 | 0.337    | 0.295 | 0.150 |

Note. Two markers independently follow $0.4N(\mu_1, \sigma_1^2) + 0.6N(\mu_2, \sigma_2^2)$ and $0.6N(\mu_1, \sigma_1^2) + 0.4N(\mu_2, \sigma_2^2)$ for diseased subjects with their AUC value $AUC_i$, respectively. The power is the rejection rate over 5000 simulations. TwoT, AUCT, sAUCT, mAUC(0.5) T, and PermT stand for the two-stage test, AUC test, sAUC test, mAUC(0.5) test, and permutation test, respectively.
3.2. Real Examples

Two examples from published studies are provided with our two-stage test in this section. The first one was conducted by Venkatraman and Begg (1996). In the study, two techniques for diagnosing melanoma were compared. Seventy-two lesions were examined: 21 from a diseased group and 51 from a non-diseased group. A clinical scoring scheme without a dermoscope and a dermoscopic scoring scheme were used. The purpose of the analysis was to determine whether the dermoscope contributed to the diagnostic information gathered. The null hypothesis was that the dermoscope contributed no useful information. Further details and data can be found in Venkatraman and Begg (1996).

The test used by Venkatraman and Begg depended on a permutation randomization step. As was mentioned, the two diagnostic scores for each individual were derived from different systems and were not exchangeable, meaning that they needed a second randomization step to construct their test. As such, the test was very complex and computationally burdensome.

We applied our two-stage test and gathered the following data. We plotted the empirical ROC curves and the smoothed binormal ROC curves for both tests using the “pROC” package (Robin et al., 2011) of the R program, changing the coordinate labels to FPR vs. TPR (the original plot is coordinated with specificity vs. sensitivity) in Fig. 4. The empirical AUCs were 0.906 and 0.900 for Test X and Test Y, respectively, and the corresponding sAUCs were 0.437 and 0.374, respectively. The sample correlation between Test X and Test Y was 0.84. The two Z-statistics ($Z_1$ and $Z_2$) of our two-stage test were 0.148 and 1.041, respectively. The asymptotically estimated correlation coefficient $\hat{\gamma}$ between $Z_1$ and $Z_2$ was 0.54. Referring to Table 2, both the AUC and sAUC tests were nonsignificant at $\alpha = 0.1$; therefore, the null hypothesis was not rejected, indicating that the dermoscope contributed no useful information to the diagnosis of melanoma, which was consistent with the previous results.

In Example 2, yeast data from the University of California, Irvine (UCI) machine learning repository (Asuncion and Newman, 2007) was used. The dataset contained 1484 samples, 8 features, and 10 classes. In order to best illustrate a two-class classification problem, we used the cytosolic or cytoskeletal (CYT) and mitochondrial (MIT) classes as one group and the other eight classes as the other group. Here, CYT and MIT
were treated as the “healthy” group, and the other classes were treated as the diseased group. All eight attributes were real numbers. We studied the discrimination characteristics of two linear models. The first one was the linear discriminant analysis (LDA) classifier. The second was the single feature “alm,” the score of the ALOM membrane spanning region prediction program (for the details of this dataset, please refer to the UCI machine learning repository Web site). We referred to this single feature model as SFM. We used the first two-thirds of the samples as a training set; the rest were the testing set. For the testing set, the empirical AUCs were 0.739 and 0.775 for SFM and LDA, respectively. Their empirical sAUCs were 0.294 and 0.174, respectively. The sample correlation between the two models was 0.64. When testing these two markers with our two-stage test procedure, the two Z-statistics were −1.703 and 5.013, respectively, and their correlation coefficient was \( \hat{\gamma} = 0.5 \). This implies that their AUCs were not significantly different according to critical values shown in Table 2; however, their sAUCs were significantly different at \( a = 0.01 \). This suggests that the single feature “alm” had higher ROC curves than the full LDA classifier when the FPR was 0.26 or lower, as seen in Fig. 5.

4. DISCUSSION

The proposed two-stage test provides a powerful comparison of two correlated ROC curves for a broad range of scenarios, especially when comparing two crossed ROC curves with similar AUCs. This procedure consists of an AUC and sAUC test. We derived the asymptotic joint distribution for these two test statistics. Constructed using two nonparametric tests based on Z-statistics, our test outperforms a conventional nonparametric AUC test (DeLong et al., 1988) in terms of power; it outperforms a conventional permutation-based ROC test (Venkatraman and Begg, 1996) in terms of power and its computational efficiency. Furthermore, we have demonstrated that a single-stage test based on either the AUC or the mAUC was not powerful in certain situations. On the other hand, the proposed two-stage test displayed a high
level of power in broad scenarios. If two binormal ROC curves were compared, our
two-stage test would be equivalent to testing the identity of the entire range of
ROC curves.

In this article, we use identical test sizes at each stage for simplicity. However, we
conclude that different allocations of the test size at each stage may lead to different
overall power, depending on how the two ROC curves are shaped. If there is prior
information regarding the curves, it is possible to allocate the test size differently.
Otherwise, equal allocation is the best option. Our definition of sAUC relies on a positive
shift, $\delta$, which is set to be $\delta = 1.96\sigma_X$. For the ordinal scale data, this setting might be
inappropriate. We will pursue this problem in the future.

We argue that the sAUC is not invariant to a monotonic transformation, such as
logarithmic transformation, although it is invariant to location and/or scale transformation.
Suppose that we compare two markers, where one is the logarithmic transformation of the
other. Then, the two sAUCs could be different. In such cases, our two-stage test might
increase Type I errors. This criticism could be alleviated if we note that such comparisons
are rare in practice. In addition, since these two markers have exactly the same ROC
curves, we could also avoid such comparisons by plotting the curves before performing
the test.

It is well-known that specifying a data-driven upper limit of the false positive rate in
pAUC is not trivial. However, if a pre-specified region of the ROC curves is of interest,
the statistical test about the partial area under the ROC curve (pAUC) is also available for
comparing two correlated ROC curves. In such cases, our test may not as powerful as a
pAUC-based test such as the method proposed in Wieand et al. (1989). However, in the
sense of detecting differences between entire ROC curves in varying situations, the two-
stage test could be more useful.

FUNDING

This research was supported by the Basic Science Research Program through the
National Research Foundation of Korea (NRF) funded by the Ministry of Education,
Science and Technology (2010-0006158).

SUPPLEMENTARY MATERIAL

Supplemental data for this article can be accessed on the publisher’s website.

REFERENCES

Asuncion, A. and Newman, D. J. (2007). UCI Machine Learning Repository Irvine. University of
California, School of Information and Computer Science. http://www.ics.uci.edu/mlearn/
milepository.html.
Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver
operating characteristic graph. Journal of Mathematical Psychology 12(4):387–415.
Bandos, A. I., Rockette, H. E., Gur, D. (2005). A permutation test sensitive to differences in
areas for comparing roc curves from a paired design. Statistics in Medicine 24
(18):2873–2893.
Bandos, A. I., Rockette, H. E., Gur, D. (2006). A permutation test for comparing roc curves in
multireader studies a multi-reader roc, permutation test. Academic Radiology 13(4):414.
Beam, C. A., Wieand, H. S. (1991). A statistical method for the comparison of a discrete diagnostic test with several continuous diagnostic tests. *Biometrics* 47(3):907–919.

Braun, T. M., Alonzo, T. A. (2008). A modified sign test for comparing paired ROC curves. *Biostatistics* 9(2):364–372.

Cai, T., Moskowitz, C. S. (2004). Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. *Biostatistics* 5(4):573–586.

DeLong, E. R., DeLong, D. M., Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44(3):837–845.

Dodd, L. E., Pepe, M. S. (2003). Partial AUC estimation and regression. *Biometrics* 59(3):614–623.

Hanley, J. A., McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148(3):839–843.

He, Y., Escobar, M. (2008). Nonparametric statistical inference method for partial areas under receiver operating characteristic curves, with application to genomic studies. *Statistics in Medicine* 27(25):5291–5308.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* 9(3):190–195.

Metz, C. E., Herman, B. A., Shen, J.-H. (1998). Maximum likelihood estimation of receiver operating characteristic(roc) curves from continuously-distributed data. *Statistics in Medicine* 17(9):1033–1053.

Metz, C. E., Wang, P., Kronman, H. B. (1984). A new approach for testing the significance of differences between ROC curves measured from correlated data. *Information Processing in Medical Imaging* 8:286–295.

Pepe, M. S. (2004). *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Oxford University Press.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M. (2011). proc: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12(1):77.

Thompson, M. L., Zucchini, W. (1989). On the statistical analysis of roc curves. *Statistics in Medicine* 8(10):1277–1290.

Venkatraman, E., Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* 83(4):835–848.

Walsh, S. J. (1997). Limitations to the robustness of binormal roc curves: effects of model misspecification and location of decision thresholds on bias, precision, size and power. *Statistics in Medicine* 16(6):669–679.

Wieand, S., Gail, M. H., James, B. R., James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 76(3):585–592.

Yu, W., Chang, Y.-C. I., Park, E. (2013). A modified area under the ROC curve and its application to marker selection and classification. *Journal of the Korean Statistical Society* 43(2):161–175.

Zhang, D. D., Zhou, X.-H., Freeman, D. H., Freeman, J. L. (2002). A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Statistics in Medicine* 21(5):701–715.