Published in final edited form as:
J Am Stat Assoc. 2008 ; 103(482): 705–713.

A Unified Approach to Nonparametric Comparison of Receiver Operating Characteristic Curves for Longitudinal and Clustered Data

Gang Li and
Professor, Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095-1772 (E-mail: vli@ucla.edu)

Kefei Zhou
Biostatistics Manager, Amgen Inc., Thousand Oaks, CA 91320 (E-mail: kzhou@amgen.com)

Abstract

We present a unified approach to nonparametric comparisons of receiver operating characteristic (ROC) curves for a paired design with clustered data. Treating empirical ROC curves as stochastic processes, their asymptotic joint distribution is derived in the presence of both between-marker and within-subject correlations. A Monte Carlo method is developed to approximate their joint distribution without involving nonparametric density estimation. The developed theory is applied to derive new inferential procedures for comparing weighted areas under the ROC curves, confidence bands for the difference function of ROC curves, confidence intervals for the set of specificities at which one diagnostic test is more sensitive than the other, and multiple comparison procedures for comparing more than two diagnostic markers. Our methods demonstrate satisfactory small-sample performance in simulations. We illustrate our methods using clustered data from a glaucoma study and repeated-measurement data from a startle response study.

Keywords

Area under the receiver operating characteristic curve; Clustered data; Confidence band; Intersection-union tests; Longitudinal data; Multiple comparison; Paired design; Partial area under the receiver operating characteristic curve; Quantile process; Repeated measurement

1. INTRODUCTION

Receiver operating characteristic (ROC) curves are commonly used to evaluate and compare diagnostic markers in various fields, such as signal detection and medicine (Green and Swets 1966; Zhou, McClish, and Obuchowski 2002; Pepe 2003). The ROC curve of a diagnostic marker is a plot of the true positive rate (sensitivity) against the false-positive rate (1 – specificity) at different threshold values. It has the appealing property of describing the discrimination capacity of a diagnostic marker without linking it to any specific threshold. This enables direct comparisons of diagnostic markers even if they are on different measurement scales.

The most popular nonparametric methods for comparing ROC curves are based on some summary indexes, such as the area under the ROC curve (AUC) and the partial area under the ROC curve (pAUC) (see Hanley and McNeil 1982, 1983; McClish 1987; DeLong, DeLong, and Clarke-Pearson 1988; and Wieand, Gail, Barray, and James 1989 for independent data, and Obuchowski 1997; Emir, Wieand, Su, and Cha 1998; and Emir, Wieand, Jung, and Ying

2000 for clustered data). Although simple to use, these summary indexes are not without limitations; for example, a large value of the AUC or pAUC does not necessarily imply high sensitivity at a prespecified specificity. Many medical applications require comparisons of sensitivities of diagnostic markers at a prespecified specificity that is of clinical relevance. Confidence bands for the difference of two ROC curves also are desirable in practice for comparing sensitivities of two diagnostic markers over a prespecified range of specificities. In addition, interest may lie in estimating the set of specificities with differential sensitivity. Unfortunately, very few nonparametric comparison procedures beyond the use of AUC and pAUC have been rigorously developed in the literature. The problems can become more complicated when a paired design is used in which both markers are performed on the same study subjects and when there are multiple measurements for each marker from repeated-measurement or clustered data studies. In such situations, the between-marker and within-subject correlations must be accounted for. Some examples of a paired design with clustered data are given in Section 5.

The purpose of this article is to develop a unified approach to the analysis of ROC curves that allows a variety of non-parametric comparisons in the presence of between-marker and within-cluster correlations. We first derive the asymptotic distribution theory for correlated empirical ROC processes, taking into account of both between-marker and within-cluster correlations. This extends the work of Hsieh and Turnbull (1996) and Li, Tiwari, and Wells (1996b), who derived the asymptotic distribution of a single empirical ROC curve. The variance–covariance functions of the limiting ROC processes are shown to involve unknown densities that are difficult to estimate nonparametrically. By extending an idea of Keaney and Wei (1994), we propose to approximate the joint distribution of the empirical ROC processes by some Gaussian random processes whose distribution can be calculated by computer Monte Carlo simulations without involving density estimation. We provide a theoretical justification for this approach by proving that the Monte Carlo Gaussian processes have the same limiting distribution as the original empirical ROC processes. These results allow us to approximate the distribution of any continuous functional of the empirical ROC curves. This provides a general framework that enables various ROC curve comparisons. As examples, we apply the developed theory to obtain confidence bands for the difference of two ROC curves over a given range of specificities. We also derive nonparametric tests and confidence intervals for comparing weighted areas under the ROC curves that include AUCs, pAUCs, and sensitivities at a fixed specificity as special cases. In addition, we apply the intersection-union test concept (cf. Berger and Boos 1999; Berger and Hsu 1996) to obtain confidence intervals for the set of specificities at which one diagnostic test is more sensitive than the other. We also extend our theory to deal with multiple comparison problems involving more than two diagnostic markers.

We note that Uno, Cai, Tian, and Wei (2007) described a similar perturbation-resampling method to evaluate an ROC curve for censored data that include uncensored binary outcomes as a special case. As pointed by a referee, it may be possible to extend their method to correlated ROC curves and to clustered data. But our work is the first to clearly adapt these ideas to analyses of correlated ROC curves with clustered data, and we describe additional measures, such as confidence intervals, for the set of specificities with differential sensitivity.

The rest of the article is organized as follows. Section 2 gives the general asymptotic distribution theory for empirical ROC curves. Section 3 applies the developed theory to derive a number of inferential procedures for comparing ROC curves. Section 4 presents a simulation study to evaluate our methods' performance. Section 5 illustrates our methods using a cluster data set from an ophthalmology study and a repeated-measurement data set from a neuroscience study. Section 6 gives some closing remarks.

2. LARGE–SAMPLE DISTRIBUTION THEORY FOR RECEIVER OPERATING CHARACTERISTIC CURVES

For simplicity, we consider only two diagnostic tests in this section.

2.1 The Receiver Operating Characteristic Curve

Let $X^{(v)}$ denote the continuous outcome of the *v*th diagnostic marker for which a value greater than a cutoff value *t* indicates a positive test result, $v = 1, 2$. The sensitivity and specificity are then given by $1 - G^{(v)}(t)$ and $F^{(v)}(t)$, where $F^{(v)}$ and $G^{(v)}$ are the cumulative distribution functions of $X^{(v)}$ for the healthy and diseased populations, $v = 1, 2$. The ROC curve for the *v*th diagnostic test is defined by

$$ROC^{(v)}(p)=1-G^{(v)}\{F^{(v)^{-1}}(1-p)\}, \qquad p \in [0,1], \tag{1}$$

where $F^{-1}(p) = \inf(t\colon F(t) \geq p)$ for any function *F*. Equivalently, this is the plot of sensitivity against 1 − specificity, as the cutoff value *t* varies. Clearly, the closer to the upper left corner of the unit box, the greater the discriminating power.

2.2 The Data

Assume that there are a total of *n* subjects in the study. Suppose that we observe $X_{ij}^{(v)} \sim F^{(v)}, j=1,\ldots,m_i^{(v)}$, representing the measurements of the *v*th marker from $m_i^{(v)}$ healthy units within subject *i*, and $Y_{ij}^{(v)} \sim G^{(v)}, j=1,\ldots,n_i^{(v)}$, the measurements of the *v*th marker from $n_i^{(v)}$ diseased units within subject *i*, $i = 1, \ldots, n$ and $v = 1, 2$. Assume further that measurements from different subjects are independent and measurements within a subject are possibly correlated. These types of data commonly arise from clustered data or repeated-measurement studies. Our data setting allows for both between-marker and within-subject correlations. We also allow different markers to have different numbers of measurements per subject.

2.3 The Estimators

Let $M^{(v)}=\sum_{i=1}^{n} m_i^{(v)}$ and $N^{(v)}=\sum_{i=1}^{n} n_i^{(v)}$. Define the empirical ROC curves

$$\widehat{ROC}^{(v)}(p)=1-\widehat{G}^{(v)}\{\widehat{F^{(v)}}^{-1}(1-p)\}, \qquad v=1,2, \tag{2}$$

where $\widehat{F}^{(v)}(t)=\sum_{i=1}^{n}\sum_{j=1}^{m_i^{(v)}} I(X_{ij}^{(v)} \leq t)/M^{(v)}$ and $\widehat{G}^{(v)}(t)=\sum_{i=1}^{n}\sum_{j=1}^{m} I(X_{ij}^{(v)} \leq t)/N^{(v)}$.

2.4 The Joint Limiting Distribution of ($\widehat{ROC}^{(1)}(p), \widehat{ROC}^{(2)}(p)$)

For any interval *I*, let *D*(*I*) denote the cadlag space of all right-continuous functions on *I* that have left-side limits, equipped with supremum norm $\| \cdot \|_\infty$. The following lemma is needed to establish the limiting distribution of $(ROC^{(1)}(p), ROC^{(2)}(p))$.

Lemma 1—Assume that as $n \to \infty$, $n^{-1}\sum_{i=1}^{n} m_i^{(v)^k} \to \lambda_k^{(v)}$, and $n^{-1}\sum_{i=1}^{n} n_i^{(v)^k} \to \gamma_k^{(v)}$ for some positive constants $\lambda_k^{(v)}$ and $\gamma_k^{(v)}$, $v = 1, 2$ and $k = 1, 2, 3$. Then

$$\sqrt{n}\begin{pmatrix}\widehat{F}^{(1)}(t)-F^{(1)}(t)\\ \widehat{F}^{(2)}(t)-F^{(2)}(t)\\ \widehat{G}^{(1)}(t)-G^{(1)}(t)\\ \widehat{G}^{(2)}(t)-G^{(2)}(t)\end{pmatrix}\xrightarrow{d}\begin{pmatrix}W_{F^{(1)}}(t)\\ W_{F^{(2)}}(t)\\ W_{G^{(1)}}(t)\\ W_{G^{(2)}}(t)\end{pmatrix}\quad \text{as } n\to\infty, \tag{3}$$

in $D(\mathbb{R})^4 = D(\mathbb{R}) \times \cdots \times D(\mathbb{R})$, where $\mathbb{R} = (-\infty, \infty)$, $(W_{F^{(1)}}(t), W_{F^{(2)}}(t), W_{G^{(1)}}(t), W_{G^{(2)}}(t))'$ is a vector of mean-0 Gaussian processes defined in (A.1) of the Appendix.

The joint limiting distribution of ($\widehat{ROC}^{(1)}(p), \widehat{ROC}^{(2)}(p)$) is given next.

Theorem 1—Assume that the assumptions of Lemma 1 hold. Assume further that for $v = 1, 2$, $F^{(v)}$ and $G^{(v)}$ have derivatives $F^{(v)'}$ and $G^{(v)'}$ that are positive and continuous on $[F^{(v)^{-1}}(a) - \varepsilon, F^{(v)^{-1}}(b) + \varepsilon]$, for some $0 < a < b < 1$ and $\varepsilon > 0$. Then, as $n \to \infty$,

$$\sqrt{n}\begin{pmatrix}\widehat{ROC}^{(1)}(p)-ROC^{(1)}(p)\\ \widehat{ROC}^{(2)}(p)-ROC^{(2)}(p)\end{pmatrix}\xrightarrow{d}\begin{pmatrix}Z^{(1)}(1-p)\\ Z^{(2)}(1-p)\end{pmatrix}, \tag{4}$$

in $D[1 - b, 1 - a] \times D[1 - b, 1 - a]$, where for $v = 1, 2$, $Z^{(v)}(p) = - s^{(v)}(p) \cdot W_{F^{(v)}}(F^{(v)^{-1}}(p)) + W_{G^{(v)}}(F^{(v)^{-1}}(p))$, $s^{(v)}(p) = G^{(v)'}(F^{(v)^{-1}}(p))/F^{(v)'}(F^{(v)^{-1}}(p))$. Recall that $W_{F^{(1)}}(t)$, $W_{F^{(2)}}(t)$, $W_{G^{(1)}}(t)$, and $W_{G^{(2)}}(t)$ are the limiting Gaussian processes in (3).

The foregoing theorem allows us to obtain the limiting distribution of any continuous functional of ($\widehat{ROC}^{(1)}(p), \widehat{ROC}^{(2)}(p)$). For simplicity, hereinafter we focus on the difference function of the two ROC curves.

Corollary 1—Let $D(p) = ROC^{(1)}(p) - ROC^{(2)}(p)$ and $\widehat{D}(p)=\widehat{ROC}^{(1)}(p)-\widehat{ROC}^{(2)}(p)$. Then, as $n \to \infty$,

$$\sqrt{n}(\widehat{D}(p)-D(p))\xrightarrow{d}U(p)=Z^{(2)}(1-p)-Z^{(1)}(1-p).$$

2.5 Approximate the Limiting Process *U*(*p*)

Note that the result of Corollary 1 cannot be readily used to make inference for $D(p)$. For instance, to construct simultaneous confidence bands for $D(p)$ over $[p_1, p_2]$, we need to know the distribution of $\sup_{p\in[p_1,p_2]} |U(p)|$, which depends on some unknown quantities and is intractable. Variance estimation also is problematic, because it involves density estimation. Next, we study how to approximate the distribution of $U(\cdot)$ without using density estimation.

As a building block, we first describe a standard Monte Carlo method for approximating the empirical distributions. For $v = 1, 2$, define

$$W^{*}_{F^{(v)}}(t)=\frac{\sqrt{n}}{M^{(v)}}\sum_{i=1}^{n}\eta_i\sum_{j=1}^{m_i^{(v)}}\{I(X_{ij}^{(v)}\le t)-\widehat{F}^{(v)}(t)\}$$

and

$$W^*_{G^{(v)}}(t)=\frac{\sqrt{n}}{N^{(v)}}\sum_{i=1}^{n}\eta_i\sum_{j=1}^{n_i^{(v)}}\{I(X_{ij}^{(v)}\le t)-\widehat{G}^{(v)}(t)\},$$

where the η_i's are iid standard normal random variables. Then, conditional on the observed data,

$$\begin{aligned}&(W^*_{F^{(1)}}(t),W^*_{F^{(2)}}(t),W^*_{G^{(1)}}(t),W^*_{G^{(2)}}(t))\\ &\xrightarrow{d}(W_{F^{(1)}}(t),W_{F^{(2)}}(t),W_{G^{(1)}}(t),W_{G^{(2)}}(t))\end{aligned}\tag{5}$$

in $D(\mathbb{R})^4$ for almost all data realizations, where the limiting processes in (5) are defined in (3). This can be proven by observing that, conditional on the data, the left side is a Gaussian random field whose covariance function converges to that of the right side.

A naive method for approximating the limiting process U is to replace $(W_{F^{(v)}}(t), W_{F^{(v)}}(t))$ with $(W^*_{F^{(v)}}(t), W^*_{G^{(v)}}(t))$ and then replace $s^{(v)}(p)$ and $F^{(v)}(t)$, $v = 1, 2$, by their corresponding sample estimates. But estimating $s^{(v)}(p)$ nonparametrically is difficult. Keaney and Wei (1994) presented a novel method to approximate the distribution of a median survival time without requiring density estimation. We extend their idea to approximate the process $U(p)$.

Note that the distribution of $\sqrt{n}\{\widehat{F}^{(v)}(F^{(v)^{-1}}(p))-p\}$ can be estimated by that of $W^*_{F^{(v)}}(F^{(v)^{-1}}(p))$. Define $\xi^{(v)*}(p)$ by $\sqrt{n}\{\widehat{F}^{(v)}(\xi^{(v)*}(p))-p\}=W^*_{F^{(v)}}(\widehat{F}^{(v)-1}(p))$, or

$$\xi^{(v)*}(p)=\widehat{F}^{(v)-1}(p+n^{-1/2}W^*_F(\widehat{F}^{(v)-1}(p))).\tag{6}$$

Then it can be shown that the conditional distribution of the process $\sqrt{n}\{\xi^{(v)*}(p)-\widehat{F}^{(v)-1}(p)\}$ given the data is asymptotically equivalent to $\sqrt{n}\{\widehat{F}^{(v)-1}(p)-F^{(v)-1}(p)\}$.

Similarly, define

$$\zeta^{(v)*}(t)=\widehat{G}^{(v)}(t)-n^{-1/2}W_G^{(v)*}(t).\tag{7}$$

Then it is easy to see that the distribution of $\sqrt{n}\{\widehat{G}^{(v)}(t)-G^{(v)}(t)\}$ can be estimated by the conditional distribution of $\sqrt{n}\{\zeta^{(v)*}(t)-\widehat{G}^{(v)}(t)\}$ given the data.

Finally, let $Q^{(v)*}(p) = \zeta^{(v)*}(\xi^{(v)*}(p))$, $ROC^{(v)*}(p) = 1 - Q^{(v)*}(1 - p)$, and $D^*(p) = ROC^{(1)^*}(p) - ROC^{(2)^*}(p)$. We then have the following theorem.

Theorem 2—Assume that the conditions of Theorem 1 hold. Then, for every subsequence $\{n_j\}$ of $\{n\}$, there is a further subsequence $\{m_k\} \subset \{n_j\}$ such that, conditional on the data,

$$\sqrt{n}(D^*(p) - \widehat{D}(p)) \xrightarrow{d} U(p)$$

in $D[p_1, p_2]$ along the subsequence $\{m_k\}$, for almost all data realizations, where $[p_1, p_2] \subset (1 - b, 1 - a)$ and $U(p)$ is the limiting process defined in Theorem 1.

Although the foregoing result is established along subsequences, it leads to the following corollary for any continuous functional of $\sqrt{n}(D^* - \widehat{D})$ along the entire sequence $\{n\}$.

Corollary 2—Assume that the conditions of Theorem 1 hold. Let T: $D[p_1, p_2] \to \mathbb{R}$ be a continuous mapping from $D[p_1, p_2]$ to the real line $\mathbb{R}$. Let $H_n^*(t) = P(T(\sqrt{n}(D^* - \widehat{D})) \le t | data)$. Then, as $n \to \infty$,

$$H_n^* t \to H(t) \equiv P(T(U) \le t)$$

uniformly in t on any compact interval for almost all data realizations.

Corollary 2 allows us to approximate the distribution of $T(\sqrt{n}(\widehat{D} - D))$ by that of $T(\sqrt{n}(D^* - \widehat{D}))$ given the data. Some useful examples of T include $T_1(f) = f(p_0)$ for a fixed p_0, $T_2(f) = \int_{p_1}^{p_2} f(p)dp$, and $T_3(f) = \sup_{p \in [p_1, p_2]} |f(p)|$, which can be used to compare sensitivities at a prespecified specificity p_0, compare partial areas under the curve (pAUC) and construct simultaneous confidence bands for $D(p)$.

3. RECEIVER OPERATING CHARACTERISTIC CURVE ANALYSIS

In this section we apply the theory developed in the previous section to derive a number of inferential procedures for comparing ROC curves. Specifically, we develop statistical tests for comparing weighted areas under ROC curves, construct confidence bands for the difference of two ROC curves, estimate the set of specificities at which one test is more sensitive than the other, and discuss multiple comparison procedures for more than two diagnostic markers.

3.1 Comparing the Weighted Areas Under Receiving Operating Characteristic Curves

Let $\Delta = \int_0^1 D(p)dw(p)$ be the difference between the weighted areas under the two ROC curves, where $w(p)$ is a prespecified weight function. Note that Δ is the difference between two sensitivities for a fixed specificity p_0 if $w(p)$ is a degenerate distribution function at p_0, the difference between the total AUCs if $w(p)$ is the uniform distribution function on [0, 1], and the difference between the pAUCs if $w(p)$ is the uniform distribution function on $[p_1, p_2] \subset [0, 1]$ multiplied by $p_2 - p_1$.

Let $\widehat{\Delta} = \int_0^1 \widehat{D}(p)dw(p)$. From Section 2, we see that the distribution of $\sqrt{n}(\widehat{\Delta} - \Delta)$ can be approximated by the conditional distribution of $\sqrt{n}(\Delta^* - \widehat{\Delta})$ given data. Therefore, a confidence interval for Δ can be computed as follows:

Step 1. Generate K(say K = 1,000) independent standard normal samples $\eta_1^{(k)}, \ldots, \eta_n^{(k)}$, $k = 1, \ldots, K$. For each k, compute $\Delta_k^* = \int_0^1 D_k^*(p)dw(p)$, based on $\eta_1^{(k)}, \ldots, \eta_n^{(k)}$. Let S_Δ be the sample standard deviation of $\{\Delta_k^* - \widehat{\Delta}, k = 1, \ldots, K\}$.

Step 2. A 100(1 − α)% confidence interval for Δ is given by

$$\widehat{\Delta} \pm z_{1-\alpha/2} S_{\Delta},$$

where $z_{1-\alpha/2}$ is the 1 − α/2 percentile of the standard normal distribution.

One-sided confidence intervals and hypothesis tests for Δ can be obtained similarly.

3.2 Confidence Bands for *D*(*p*) = *ROC*$^{(1)}$(*p*) − *ROC*$^{(2)}$(*p*)

The following theorem is a direct consequence of Theorem 2 and Corollary 2.

Theorem 3—For 0 < α < 1 and [p_1, p_2] ⊂ [0, 1], let C_α be determined by

$$\lim_{n\to\infty} P\left\{ \sup_{p\in[p_1,p_2]} \left| \frac{D^*(p) - \widehat{D}(p)}{S_D(p)} \right| \le C_\alpha | data \right\} = 1 - \alpha,$$

where $S_D^2(p)$ is a consistent variance estimate of $\hat{D}(p)$. Then, as $n \to \infty$,

$$P\{\widehat{D}(p) - C_\alpha S_D(p) \le D(p) \le \widehat{D}(p) + C_\alpha S_D(p), \\ \text{for all } p \in [p_1, p_2]\} \to 1 - \alpha.$$

The confidence band can be computed as follows:

Step 1. Generate K_1(say K_1 = 500) independent standard normal samples $\eta_1^{(k)}, \ldots, \eta_n^{(k)}$, k = 1, …, K_1. For each k, compute $D_k^*(p)$, based on $\eta_1^{(k)}, \ldots, \eta_n^{(k)}$. Let $S_D(p)$ be the sample standard deviation of { $D_k^*(p) - \widehat{D}(p)$, k = 1, …, K_1}

Step 2. Generate another K_2(say K_2 = 500) independent realizations $D_1^*(p), \ldots, D_{K_2}^*(p)$. Compute C_α as the 100(1 − α)th percentile of $\sup_{p_1\le p\le p_2} |(D_1^*(p) - \widehat{D}(p))/S_D(p)|, \ldots, \sup_{p_1\le p\le p_2} |(D_{K_2}^*(p) - \widehat{D}(p))/S_D(p)|$.

Step 3. A 100(1 − α)% simultaneous confidence band for $D(p)$ over [p_1, p_2] is

$$(\widehat{D}(p) - n^{-1/2} C_\alpha S_D(p), \widehat{D}(p) + n^{-1/2} C_\alpha S_D(p)), \; p \in [p_1, p_2].$$

3.3 Determine the Set of Specificities With Differential Sensitivity

All of the preceding results are focused on comparisons of ROC curves for a prespecified set of specificities. In practice, it often is of interest to find the set of specificities at which one test is more sensitive than the other. Specifically, let $R = \{p \in [0, 1]: ROC_1(p) > ROC_2(p)\}$. We wish to find an estimated set $\hat{R}$ such that $P(\hat{R} \subset R) = 1 - \alpha$.

Note that R is an unknown set of specificities, not a parameter. Next we outline a procedure for estimating R based on an idea of Berger and Boos (1999) and Berger and Hsu (1996) who studied the problem of estimating the onset and duration of a treatment effect. Our procedure is as follows:

Step 1. For each p, conduct a one-sided level $\alpha/2$ test of H_{0p}: $D(p) = 0$ versus H_{ap}: $D(p) > 0$, as discussed in Section 3.1.

Step 2. Let p_0 be an a priori fixed starting value. If H_{0p0} is accepted, then no confidence statement is made. If H_{0p0} is rejected, then test sequentially downward from p_0, and let P_1 be the first p for which the hypothesis H_{0p} is accepted. Also test sequentially upward from p_0, and let P_2 be the first p for which the hypothesis H_{0p} is accepted. Consequently, $\hat{R}$ =[P_1, P_2] is the largest interval containing p$_0$ for which H_{0p} is rejected for all $p \in$ [P_1, P_2].

Theorem 4—Let $\hat{R}$ = [P_1, P_2] be defined by the foregoing algorithm. Then, as $n \to \infty$,

$$P(\widehat{R} \subset R) = P(ROC^{(1)}(p) > ROC^{(2)}(p) \text{ for all } P_1 \le p \le P_2) \\ \ge 1 - \alpha.$$

The proof of this theorem is essentially the same as that of Berger and Boos (1999) and thus is omitted. Note that in Step 1, a one-sided hypothesis is tested at level $\alpha/2$ instead of level α; this is necessary to achieve the overall confidence level of $1 - \alpha$ for $\hat{R}$. As argued by Berger and Boos (1999), the starting value p_0 should be chosen beforehand and in a region likely to produce significant results. A statistical approach that requires less knowledge is to repeat the method at k different starting points using level α/k for each one.

3.4 Multiple Comparisons of More Than Two Diagnostic Tests

In this section we illustrate how to extend our method to multiple comparisons of more than two ROC curves. Without loss of generality, we consider three diagnostic tests. Similar to Theorem 1, it can be shown that

$$\sqrt{n}\begin{pmatrix} \widehat{D}^{(12)}(p) - D^{(12)}(p) \\ \widehat{D}^{(13)}(p) - D^{(13)}(p) \\ \widehat{D}^{(23)}(p) - D^{(23)}(p) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} U^{(12)}(p) \\ U^{(13)}(p) \\ U^{(23)}(p) \end{pmatrix},$$

where $D^{(ij)}(p) = ROC^{(j)}(p) - ROC^{(i)}(p)$, $\hat{D}^{(ij)}(p)$ is the sample estimate of $D^{(ij)}(p)$, $U^{(ij)}(p) = Z^{(i)}(1 - p) - Z^{(j)}(1 - p)$, and $Z^{(v)}(p)$ is as defined in Theorem 1 for $v = 1, 2, 3$.

As in Section 2.5, we construct $D^{(12)*}(p)$, $D^{(13)*}(p)$, and $D^{(23)*}(p)$ so that the joint distribution of ($U^{(12)}$, $U^{(13)}$, $U^{(23)}$) is approximated by that of ($U^{(12)*}$, $U^{(13)*}$, $U^{(23)*}$) given the data, where $U^{(ij2)*}(p) = D^{(ij)*}(p) - \hat{D}^{(ij)}(p)$. This allows us to perform various multiple comparisons. The following algorithm can be used to construct joint confidence intervals for three pairwise comparisons of weighted areas under the curve: $\Delta^{(12)}$, $\Delta^{(13)}$, and $\Delta^{(23)}$, where $\Delta^{(ij)} = \int_0^1 D^{(ij)}(p) dw(p)$:

Step 1. Generate K_1(say K_1 = 500) realizations of $Z^{(12)*} \equiv \Delta^{(12)*} - \hat{\Delta}^{(12)}$, $Z^{(13)*} \equiv \Delta^{(13)*} - \hat{\Delta}^{(13)}$, and $Z^{(23)*} \equiv \Delta^{(23)*} - \hat{\Delta}^{(23)}$ by generating K_1 independent standard normal samples. Compute their sample standard deviations $S_{\Delta^{(12)}}$, $S_{\Delta^{(13)}}$, and $S_{\Delta^{(23)}}$

Step 2. Generate another K_2(say K_2 = 500) independent realizations of $Z^{(12)*}$, $Z^{(13)*}$, and $Z^{(23)*}$ and let A_k be the maximum of $|\frac{Z^{(12)*}}{S_{\Delta(12)}}|$, $|\frac{Z^{(13)*}}{S_{\Delta(13)}}|$, and $|\frac{Z^{(23)*}}{S_{\Delta(23)}}|$ for the kth realization, k = 1, …, K_2.

Step 3. Let C_α be the 100(1 − α)th percentile of A_1, …, AK_2. The 100(1 − α)% joint confidence intervals for $\Delta^{(12)}$, $\Delta^{(13)}$, and $\Delta^{(23)}$ are given by

$$\widehat{\Delta}^{(12)} \pm C_\alpha S_{\Delta(12)}, \quad \widehat{\Delta}^{(13)} \pm C_\alpha S_{\Delta(13)}, \quad \text{and} \quad \widehat{\Delta}^{(23)} \pm C_\alpha S_{\Delta(23)}.$$

Other multiple inferences, such as multiple confidence bands, can be obtained similarly.

4. SIMULATIONS

In this section we report a simulation study conducted to evaluate the finite-sample performance of our methods. We also study the consequence of ignoring the within-subject correlations.

We generated repeated measurements using a setting similar to Emir et al. (2000). A failure time was generated for each subject from an exponential distribution with an expected value of 4. A subject was classified as “healthy” before the failure and as “diseased” after the failure. We then generated $\mathbf{X}^{(1)}=\boldsymbol{Y}_1\sqrt{\lambda}+\boldsymbol{Y}_2\sqrt{(1-\lambda)}$ and $\mathbf{X}^{(2)}=\boldsymbol{Y}_1\sqrt{\lambda}+\boldsymbol{Y}_3\sqrt{(1-\lambda)}$, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i4})'$, i = 1, 2, 3, are iid multivariate normal random vectors with mean $\mathbf{0}$ and cov(Y_{ij}, Y_{ik}) = $\rho^{|j-k|}$, for j, k = 1, 2, 3, 4. The value for subject i for marker v at visit j is $X_{ij}^{(v)}$ if the subject is “healthy” at the jth visit and $X_{ij}^{(v)}+1$ if he or she is “diseased.” Note that λ and ρ measure the between-marker and within-marker correlations.

Table 1 reports the achieved significance level of our test for the equality of pAUCs over three intervals (.1,.3), (.1,.5), and (.05,.95) for various sample sizes and between-marker and within-marker correlations. Each entry is based on 500 simulations. It is observed that the achieved significance level agrees with the nominal level fairly well even for small samples ($n_D = n_H$ = 30).

Next, we illustrate the consequence of ignoring the within-subject correlations. Table 2 reports the achieved significance levels of our method (designated method 1) compared with the method that treats the observations within each cluster as independent (designated method 2) for testing the equality of the AUCs. We set the between-marker correlation λ =.5 and vary the within-marker correlation ρ from.05,.75 to.99. It can be seen that the achieved significance level of our method is always close to the nominal level.05, whereas ignoring the correlations within the cluster can lead to an unreasonably high probability of type I error (e.g.,.39) as the within-subject correlation ρ and cluster size grow. Therefore, ignoring the correlations within the cluster may lead to seriously biased conclusions.

5. EXAMPLES

5.1 Detection of Glaucomatous Deterioration

A visual field test is a technique for measuring an individual’s entire scope of vision. It maps the visual fields of each eye individually. Longitudinal visual field image data can be used for early diagnosis of glaucomatous progression. But reliable detection remains one of the most difficult challenges for clinicians, due to the complex structure and high level of noise in the longitudinal visual field image data. Together with researchers at UCLA's Jules Stein Eye Institute, Jiang (2005) developed some outlier statistics to use as a diagnostic marker for detecting visual field deterioration in glaucoma patients based on Bayesian hierarchical modeling. Here we illustrate how to use our methods to compare the diagnostic power of different models in discriminating between stable and progressive eyes. Our data set comprises a visual field series of 188 eyes of 171 patients over 8 years of follow-up study; it is independent of the training sample used by Jiang (2005) for model building. Apparently these are clustered data, because some data are from both eyes of the same patient. A paired design is appropriate for ROC analysis, because both models are applied to the same set of eyes.

Figure 1 depicts the empirical ROC curves of the diagnostic markers based on two candidate models, referred to as models 1 and 2 here, which correspond to models 1 and 11 of Jiang (2005, p. 45). Figure 1 shows that the diagnostic marker based on model 2 has better power than model 1 for detecting glaucomatous progression. We performed a test for the equality of the total AUCs using the method in Section 3.1; the two-sided *p* value was.024, confirming a significant difference between the two diagnostic markers.

Figure 2 shows the confidence band for the difference function of the ROC curves between models 1 and 2 over the specificities.1–.9. Because the span of the confidence band over the specificities is too broad, it is too conservative to detect any significant difference.

Finally, we applied the method in Section 3.3 to estimate the range of specificities at which the diagnostic marker based on model 2 is more sensitive than that based on model 1. We had $\hat{R}$ = [.34,.92]; therefore, with 95% confidence, model 2 had a higher sensitivity than model 1 at specificities between.34 and.92.

5.2 Acoustic Startle Response Data

The acoustic startle response in human studies is quantified by the startle blink response to a startling stimulus. It is typically measured and summarized by the orbicularis oculi electromyogram (ooEMG), a time series plot. But high variabilities make it difficult to evaluate the ooEMG and distinguish a normal response (healthy) from no response (diseased).

In this example we consider four diagnostic markers developed by researchers at UCLA's Semel Institute for Neuroscience: the *peak* magnitude after the stimulus, the *duration* of the maximum magnitude above a certain threshold, the *area* under the maximum magnitude, and the *ratio* representing the amount exceeding a threshold. The data set comprises 37 participants from a total of 229 experiments. The maximum number of repeated experiments per person is 16. A detailed description of the study has been given by Waters and Ornitz (2008). Results of the same subject from different experiments are considered to be highly correlated; thus we adopt the clustered data setting. This is also a paired design or, more accurately, a block design, because each ooEMG yields values for all four markers.

Figure 3 displays the empirical ROC curves for the four markers. We can see that *ratio* appears to be the best and *peak* the worst. We used the multiple comparison method outlined in Section 3.5 to construct 95% joint confidence intervals for six pairwise differences of the pAUC, $pAUC_{.05,.95}$, for these four markers. The results, summarized in Table 3, show that *area*, *ratio*, and *duration* are all significantly better than *peak*. No significant difference is found among *area*, *ratio*, and *duration*.

Figure 4 gives the 95% confidence band for the difference of ROC curves between the *peak* and *area* over specificities of.1–.9. The confidence bands again seem to be conservative.

We also applied the method in Section 3.3 to estimate the set of specificities at which *area* is more sensitive than *peak*. We found $\hat{R}$ = [.38,.95]; therefore, with 95% confidence, *area* has a higher sensitivity than *peak* at specificities between.38 and.95.

6. DISCUSSION

The Monte Carlo resampling method is commonly used to approximate an empirical process, such as the empirical distribution process, that can be represented as the sum of iid random variables. But this article is the first to rigorously demonstrate that this method also can be extended to approximate quantile processes and ROC curves without the need to estimate density functions for a paired design. Clustered observations also are allowed. The result allows us to approximate the distribution of any continuous functional of correlated empirical ROC curves and thus provides a unified approach to the nonparametric comparison of ROC curves. Our method demonstrates satisfactory performance for both large and relatively small samples in simulations. We have developed software to implement the proposed procedures using R; this software can be downloaded at http://roc.cluster.googlepages.com/.

Note that much effort has been directed at approximating a quantile process without involving density estimation. For example, Li, Hollander, McKeague, and Yang (1996a) developed a nonparametric likelihood ratio method; Doss and Gill (1992) studied the bootstrap method, which was later adapted for ROC analysis by Li et al. (1996b); and Keaney and Wei (1994) introduced a Monte Carlo method for estimating a median survival time. But all of the earlier methods consider only iid observations and thus do not account for within-subject correlation for clustered data.

An alternative approach is to use the bootstrap method to approximate ROC processes by resampling the independent subjects. Although the bootstrap method has been commonly suggested in practice and is simple to use in principle, its consistency for the paired design with clustered data has not been established in the ROC analysis literature. For the upaired case where different markers are performed on different sets of subjects, Li et al. (1996b, 1999) showed that the bootstrap method is consistent in a sense that the bootstrapped ROC process has the same limiting process as the empirical ROC process. Extending their result to the paired design with clustered data does not appear to be trivial. Future work is needed to develop a theoretical justification for the bootstrap method for clustered data under the paired design.

Acknowledgments

The authors thank the joint editors and two referees for their insightful comments and suggestions, Luohua Jiang and Ed M. Ornitz for providing the data in the examples, and Tianxi Cai for helpful discussions. Li's research was supported in part by a National Institutes of Health grant.

References

Andersen et al. (1993), ???.

Berger R, Boos D. Confidence Limits for the Onset and Duration of Treatment Effect. Biometrical Journal 1999;41:517–531.

Berger RL, Hsu JC. Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets. Statistical Science 1996;11:283–319.

Billingsley, P. Convergence of Probability Measures. 2. New York: Wiley; 1999.

DeLong ER, DeLong DM, Clarke-Person DL. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 1988;44:837–845. [PubMed: 3203132]

Doss H, Gill RD. An Elementary Approach to Weak Convergence for Quantile Processes With Applications to Censored Survival Data. Journal of American Statistical Association 1992;87:869–877.

Emir B, Wieand S, Jung SH, Ying Z. Comparison of Diagnostic Markers With Repeated Measurements: A Nonparametric ROC Curve Approach. Statistics in Medicine 2000;19:11–23.

Emir B, Wieand S, Su JQ, Cha S. Analysis of Repeated Markers Used to Predict Progression of Cancer. Statistics in Medicine 1998;17:2563–2578. [PubMed: 9839348]

Green, DM.; Swets, JA. Signal Detection: Theory and Psychophysics. New York: Wiley; 1966.

Hanley JA, McNeil BJ. The Meaning and Use of the Area Under an ROC Curve. Radiology 1982;143:29–36. [PubMed: 7063747]

Hanley JA, McNeil BJ. A Method of Comparing the Area Under Two ROC Curves Derived From the Same Cases. Radiology 1983;148:839–843. [PubMed: 6878708]

Hsieh F, Turnbull BW. Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve. The Annals of Statistics 1996;24:25–40.

Jiang, LH. unpublished doctoral dissertation. University of California; Los Angeles: 2005. Bayesian Hierarchical Modelling of Glaucomatous Visual Field Data. ???

Keaney KM, Wei LJ. Interim Analysis Based on Median Survival Times. Biometrica 1994;81:279–286.

Li G, Hollander M, McKeague I, Yang J. Nonparametric Likelihood Ratio Confidence Bands for Quantile Functions From Incomplete Survival Data. The Annals of Statistics 1996a;24:628–640.

Li G, Tiwari RC, Wells MT. Quantile Comparison Functions in Two-Sample Problems, With Application to Comparisons of Diagnostic Markers. Journal of American Statistical Association 1996b;91:689–698.

Li G, Tiwari RC, Wells MT. Semiparametric Inference for Shift Functions: With Applications to Receiver Operating Characteristic Curves. Biometrika 1999;86:487–502.

McClish DK. Comparing the Areas Under More Than Two Independent ROC Curves. Medical Decision Making 1987;7:149–155. [PubMed: 3613915]

Obuchowski NA. Nonparametric Analysis of Clustered ROC Curve Data. Biometrics 1997;53:567–578. [PubMed: 9192452]

Pepe, MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford, U.K: Oxford University Press; 2003.

Shorack, GR.; Wellner, JA. Empirical Processes With Applications to Statistics. New York: Wiley; 1984.

Uno H, Cai T, Tian L, Wei LJ. Evaluating Prediction Rules for t -Year Survivors With Censored Regression Models. Journal of American Statistical Association 2007;103:527–537.

Waters AM, Ornitz EM. When the Orbicularis Oculi Response to a Startling Stimulus Is Zero, the Vertical EOG May Reveal That a Blink Has Occurred. Clinical Neurophysiology 2008;??:??–??.

Wieand HS, Gail MH, Barray RJ, James KL. A Family of Nonparametric Statistics for Comparing Diagnostic Markers With Paired or Unpaired Data. Biometrica 1989;76:585–592.

Zhou, XH.; McClish, DK.; Obuchowski, NA. Statistical Methods in Diagnostic Medicine. New York: Wiley; 2002.

APPENDIX: PROOFS

Proof of Lemma 1

Note that

$$\sqrt{n}\begin{pmatrix} \widehat{F}^{(1)}(t)-F^{(1)}(t) \\ \widehat{F}^{(2)}(t)-F^{(2)}(t) \\ \widehat{G}^{(1)}(t)-G^{(1)}(t) \\ \widehat{G}^{(2)}(t)-G^{(2)}(t) \end{pmatrix}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{V}_i(t),$$

where

$$\mathbf{V}_i(t) = \begin{pmatrix} \frac{n}{M^{(1)}} \sum_{j=1}^{m_i^{(1)}} \{I(X_{ij}^{(1)} \le t) - F^{(1)}(t)\} \\ \frac{n}{M^{(2)}} \sum_{j=1}^{m_i^{(2)}} \{I(X_{ij}^{(2)} \le t) - F^{(2)}(t)\} \\ \frac{n}{N^{(1)}} \sum_{j=1}^{n_i^{(1)}} \{I(X_{ij}^{(1)} \le t) - G^{(1)}(t)\} \\ \frac{n}{N^{(2)}} \sum_{j=1}^{n_i^{(2)}} \{I(X_{ij}^{(2)} \le t) - G^{(2)}(t)\} \end{pmatrix}, \quad i = 1, \dots, n,$$

are independent random vectors. Applying the Cramer–Rao device and the Lyapunov central limit theorem, and following along the lines of the approach of Billingsley (1999) for empirical distribution processes, it can be shown that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{V}_i(t) \xrightarrow{d} \mathbf{W}(t) \ \text{in}\, D(\mathbb{R})^4, \tag{A.1}$$

where $\mathbf{W}(t) = (W_{F^{(1)}}(t), W_{F^{(2)}}(t), W_{G^{(1)}}(t), W_{G^{(2)}}(t))'$ is a Gaussian process in $D(\mathbb{R})^4$ whose variance–covariance function is the limit of $\frac{1}{n}\sum_{i=1}^{n} \text{cov}(\mathbf{V}_i(t), \mathbf{V}_i(t))$ as $n \to \infty$.

Proof of Theorem 1

By (A.1), the compact differentiability of the inverse function and the functional delta method (see, e.g., Andersen et al. 1993), we have

$$\sqrt{n} \begin{pmatrix} \widehat{F^{(1)}}^{-1} - F^{(1)^{-1}} \\ \widehat{F^{(2)}}^{-1} - F^{(2)^{-1}} \\ \widehat{G^{(1)}} - G^{(1)} \\ \widehat{G^{(2)}} - G^{(2)} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \frac{\tilde{W}_{F^{(1)}}(F^{(1)^{-1}})}{F^{(1)\prime}(F^{(1)^{-1}})} \\ \frac{\tilde{W}_{F^{(2)}}(F^{(2)^{-1}})}{F^{(2)\prime}(F^{(2)^{-1}})} \\ \tilde{W}_{G^{(1)}} \\ \tilde{W}_{G^{(2)}} \end{pmatrix} \tag{A.2}$$

in $D[a, b] \times D[a, b] \times D[F^{(1)-1}(a), F^{(1)-1}(b)] \times D[F^{(2)-1}(a), F^{(2)-1}(b)]$, as $n \to \infty$. This, along with lemma A.1 of Li et al. (1996a,b) and the functional delta method, implies that

$$\sqrt{n} \begin{pmatrix} \widehat{G}^{(1)}(\widehat{F^{(1)}}^{-1}(p)) - G^{(1)}(F^{(1)^{-1}}(p)) \\ \widehat{G}^{(2)}(\widehat{F^{(2)}}^{-1}(p)) - G^{(2)}(F^{(2)^{-1}}(p)) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z^{(1)}(p) \\ Z^{(2)}(p) \end{pmatrix} \tag{A.3}$$

in $D[a, b]$ where $Z^{(v)}$, $v = 1, 2$ are as defined in Theorem 1.

Finally, combining (A.3) and the continuous mapping theorem proves the theorem.

Proof of Theorem 2

First, note that

$$\sqrt{n}\{D^*(p)-\widehat{D}(p)\}=\sqrt{n}\{Q^{(2)*}(1-p)-\widehat{Q}^{(2)}(1-p)\}\\ -\sqrt{n}\{Q^{(1)*}(1-p)-\widehat{Q}^{(1)}(1-p)\},$$

where for $v = 1, 2$, $Q^{(v)}(p) = G^{(v)}(F^{(v)-1}(p))$, $\hat{Q}^{(v)}(p) = \hat{G}^{(v)}(\hat{F}^{(v)-1}(p))$, and $Q^{(v)*}(p) = \zeta^{(v)*}(\hat{\zeta}^{(v)}(p))$ are as defined before Theorem 2. Moreover, by (A.3),

$$\sqrt{n}\begin{pmatrix}\widehat{Q}^{(1)}(p)-Q^{(1)}(p)\\ \widehat{Q}^{(2)}(p)-Q^{(2)}(p)\end{pmatrix}\xrightarrow{d}\begin{pmatrix}Z^{(1)}(p)\\ Z^{(2)}(p)\end{pmatrix}\quad \text{in } D[a,b]. \tag{A.4}$$

Thus, to prove Theorem 2, it suffices to show that for every subsequence $\{n_j\}$ of $\{n\}$, there is a further subsequence $\{m_k\} \subset \{n_j\}$ such that, conditional on the data,

$$\sqrt{n}\begin{pmatrix}Q^{*(1)}(p)-\widehat{Q}^{(1)}(p)\\ Q^{*(2)}(p)-\widehat{Q}^{(2)}(p)\end{pmatrix}\xrightarrow{d}\begin{pmatrix}Z^{(1)}(p)\\ Z^{(2)}(p)\end{pmatrix}\quad \text{in } D[q_1,q_2], \tag{A.5}$$

along the subsequence $\{m_k\}$ for almost all data realizations and $a < q_1 < q_2 < b$.

Here we prove only the convergence of a marginal process in (A.5), because the joint convergence is derived along the same lines. For simplicity, we also omit the superscript.

We first show that for any sequence $\delta_n(p)$ satisfying $\sup_{p\in\,[q_1,q_2]} |\delta_n(p)| = O(n^{-1/2})$,

$$\sup_{p\in[q_1,q_2]}|\sqrt{n}\{\widehat{Q}(p+\delta_n(p))-\widehat{Q}(p)\}+Q'(p)\sqrt{n}\delta_n(p)|\xrightarrow{P}0. \tag{A.6}$$

Let $Z_n(p)=\sqrt{n}(\widehat{Q}(p)-Q(p))$. Then, by (A.4), $Z_n\xrightarrow{d}Z$ in $D[a, b]$ as $n \to \infty$. By the Skorohod–Dudley–Wichura representation theorem (cf. thm. 4 of Shorack and Wellner 1984), there exists a sequence $Z_n^\dagger, Z^\dagger$ of random elements in $D[a, b]$, defined in a common probability space, such that $Z_n^\dagger\overset{d}{=}Z_n, Z^\dagger\overset{d}{=}Z$, and $Z_n^\dagger\xrightarrow{\|\cdot\|_\infty}Z^\dagger$ almost surely in $D[a, b]$ as $n \to \infty$, where $Z^\dagger$ has continuous sample paths on $[a, b]$ and $\|\cdot\|_\infty$ represents the supremum norm. Because the sample paths of $Z^\dagger$ are continuous on a compact interval $[a, b]$, they also are uniformly continuous on $[a, b]$. Thus

$$\begin{aligned}&\sup_{p\in[q_1,q_2]}|Z_n^\dagger(p+\delta_n(p))-Z_n^\dagger(p)|\\ &\le\sup_{p\in[q_1,q_2]}|Z_n^\dagger(p)-Z^\dagger(p)|\\ &+\sup_{p\in[q_1,q_2]}|Z^\dagger(p+\delta_n(p))-Z^\dagger(p)|\\ &+\sup_{p\in[q_1,q_2]}|Z_n^\dagger(p+\delta_n(p))-Z^\dagger(p+\delta_n(p))|\\ &\xrightarrow{a.s.}0,\end{aligned} \tag{A.7}$$

where the first and third terms converge to 0 because $Z_n^\dagger \xrightarrow{\|\cdot\|_\infty} Z^\dagger$ almost surely in $D[a, b]$ and the second term goes to 0 because $Z^\dagger$ has uniformly continuous sample paths on $[a, b]$. Because $Z_n^\dagger \overset{d}{=} Z_n$, (A.7) implies that as $n \to \infty$,

$$\sup_{p\in[q_1,q_2]} |Z_n(p+\delta_n(p)) - Z_n(p)| \overset{P}{\to} 0. \tag{A.8}$$

Moreover, by the mean value theorem, we have that $\sqrt{n}\{Q(p+\delta_n(p)) - Q(p)\}=Q'(\delta_n^*)\sqrt{n}\delta_n(p)$, where δ_n^* lies between p and $p + \delta_n(p)$. This, together with the fact that $Q'(p)$ is uniformly continuous on a compact interval, implies that as $n \to \infty$,

$$\sup_{p\in[q_1,q_2]} |\sqrt{n}\{Q(p+\delta_n(p)) - Q(p)\} - Q'(p)\sqrt{n}\delta_n(p)| \to 0. \tag{A.9}$$

Combining (A.8) and (A.9) proves (A.6).

For the foregoing sequence $\delta_n(p)$, we also have

$$\begin{aligned} &\sup_{p\in[q_1,q_2]} |\widehat{F}^{-1}(p+\delta_n(p)) - F^{-1}(p)| \\ \leq &\sup_{p\in[q_1,q_2]} |\widehat{F}^{-1}(p+\delta_n(p)) - F^{-1}(p+\delta_n(p))| \\ &+ \sup_{p\in[q_1,q_2]} |F^{-1}(p+\delta_n(p)) - F^{-1}(p)| \\ &\overset{P}{\to} 0, \end{aligned} \tag{A.10}$$

where the convergence of the first term follows from the uniform consistency of $\hat{F}^{-1}$ on $[a, b]$ and the second term goes to 0 because F^{-1} is uniformly continuous on $[q_1, q_2]$.

By (A.6) and (A.10), for any subsequence $\{n_j\}$ of $\{n\}$, there is a further subsequence $\{m_k\} \subset \{n_j\}$ and a subsample space $\Omega_0 \subset \Omega$ such that $P(\Omega_0) = 1$ and, for every $\omega \in \Omega_0$,

$$\sup_{p\in[q_1,q_2]} |\sqrt{n}\{\widehat{Q}(p+\delta_n(p)) - \widehat{Q}(p)\}+Q'(p)\sqrt{n}\delta_n(p)|(\omega) \to 0 \tag{A.11}$$

and

$$\sup_{p\in[q_1,q_2]} |\widehat{F}^{-1}(p+\delta_n(p)) - F^{-1}(p)|(\omega) \to 0 \tag{A.12}$$

along the subsequence $\{m_k\}$.

Next, we note from direct calculations that

$$\sqrt{n}\{Q^*(p) - \widehat{Q}(p)\} = \sqrt{n}\{\widehat{Q}(p+n^{-1/2}W_F^*(\widehat{F}^{-1}(p))) - \widehat{Q}(p)\}$$
$$-W_G^*(\widehat{F}^{-1}(p+n^{-1/2}W_F^*(\widehat{F}^{-1}(p)))).$$

By (5) and the Korohod–Dudley–Wichura representation theorem (cf. thm. 4 of Shorack and Wellner 1984), there exists a sequence $(W_F^{*\dagger}, W_G^{*\dagger}), (W_F^{\dagger}, W_G^{\dagger})$ of random elements in $D(\mathbb{R})$ defined in a common probability space such that, conditional on the data,

$$(W_F^{*\dagger}, W_G^{*\dagger}) \overset{d}{=} (W_F^*, W_G^*), \qquad (W_F^{\dagger}, W_G^{\dagger}) \overset{d}{=} (W_F, W_G), \tag{A.13}$$

and

$$(W_F^{*\dagger}, W_G^{*\dagger}) \xrightarrow{\|\cdot\|_\infty} (W_F^{\dagger}, W_G^{\dagger}) \text{ almost surely in } D(\mathbb{R}) \tag{A.14}$$

as $n \to \infty$ for almost all data realizations, where $W_F^{\dagger}$ and $W_G^{\dagger}$ have continuous sample paths and $\|\cdot\|_\infty$ is the supremum norm. Thus, conditional on the observed data,

$$\begin{aligned} &\sup_{q_1 \le p \le q_2} |W_F^{*\dagger}(\widehat{F}^{-1}(p)) - W_F^{\dagger}(F^{-1}(p))| \\ \le &\sup_{q_1 \le p \le q_2} |W_F^{*\dagger}(\widehat{F}^{-1}(p)) - W_F^{\dagger}(\widehat{F}^{-1}(p))| \\ + &\sup_{q_1 \le p \le q_2} |W_F^{\dagger}(\widehat{F}^{-1}(p)) - W_F^{\dagger}(F^{-1}(p))| \\ &\xrightarrow{a.s.} 0, \end{aligned} \tag{A.15}$$

as $n \to \infty$ for almost all data realizations, where the first term on the right side of the inequality converges to 0 because $W_F^{*\dagger} \xrightarrow{\|\cdot\|_\infty} W_F^{\dagger}$ almost surely in $D(\mathbb{R})$, and the second term goes to 0 because of the uniform consistency of $\hat{F}^{-1}$ and the uniform continuity of $W_F^{\dagger}(t)$ on any compact interval.

Finally, by combining (A.11)–(A.15), we can conclude that for every subsequence $\{n_j\}$ of $\{n\}$, there is a further subsequence $\{m_k\} \subset \{n_j\}$ such that, conditional on the data, $\sqrt{n}(Q^* - \widehat{Q}) \xrightarrow{d} Z$ in $D[q_1, q_2]$, along the subsequence $\{m_k\}$ for almost all data realizations and $a < q_1 < q_2 < b$. This concludes Theorem 2, as argued at the beginning of the proof.

Proof of Corollary 1

By Theorem 2 and the continuity of $T(\cdot)$, for every subsequence $\{n_j\}$ of $\{n\}$ there is another subsequence $\{m_k\} \subset \{n_j\}$ such that, conditional on the data, $T(\sqrt{n}(D^* - \widehat{D})) \xrightarrow{d} T(U)$ along the subsequence $\{m_k\}$, for almost all data realizations.

We first show that as $n \to \infty$,

$$H_n^*(t) \to H(t) \ \text{for every } t, \tag{A.16}$$

for almost all data realizations. This can be proved by contradiction. If for some t, $H_n^*(t)$ does not converge to $H(t)$, then there exist an $\varepsilon > 0$ and a subsequence $\{n_j\}$ of $\{n\}$ such that $|H_{n_j}^*(t) - H(t)| \geq \varepsilon > 0$ for all n_j. This implies that no further subsequence of $H_{nj}(t)$ converges to $H(t)$, which contradicts the result in the first paragraph; thus (A.16) holds. The uniform convergence $H_n^*(t)$ to $H(t)$ in t follows from (A.16) and the fact the $H_n^*(t)$ and $H(t)$ are continuous nondecreasing distribution functions.

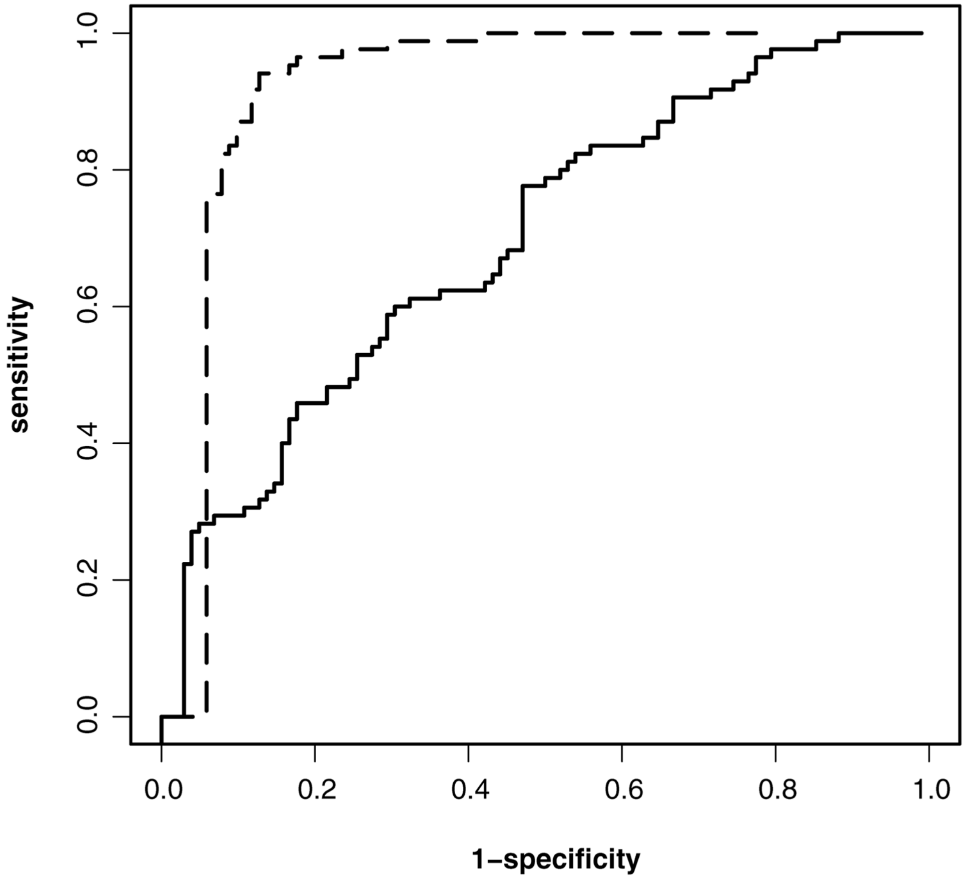

Figure 1.

Empirical ROC curves of two diagnostic markers for detecting glaucoma deterioration (——, model 2;— —, model 1).

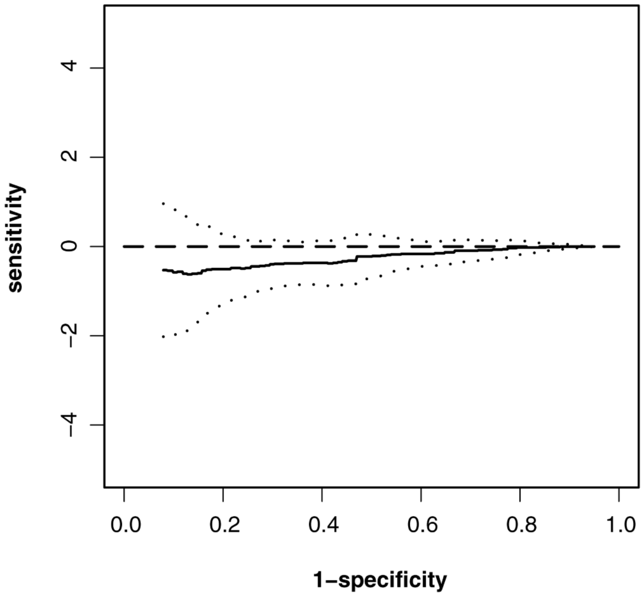

Figure 2.

95% confidence band for the difference of the ROC curves between models 1 and 2 from.1 to. 9 for the glaucoma data (——, difference; · · · ·, lower bound; · · · ·, upper bound).

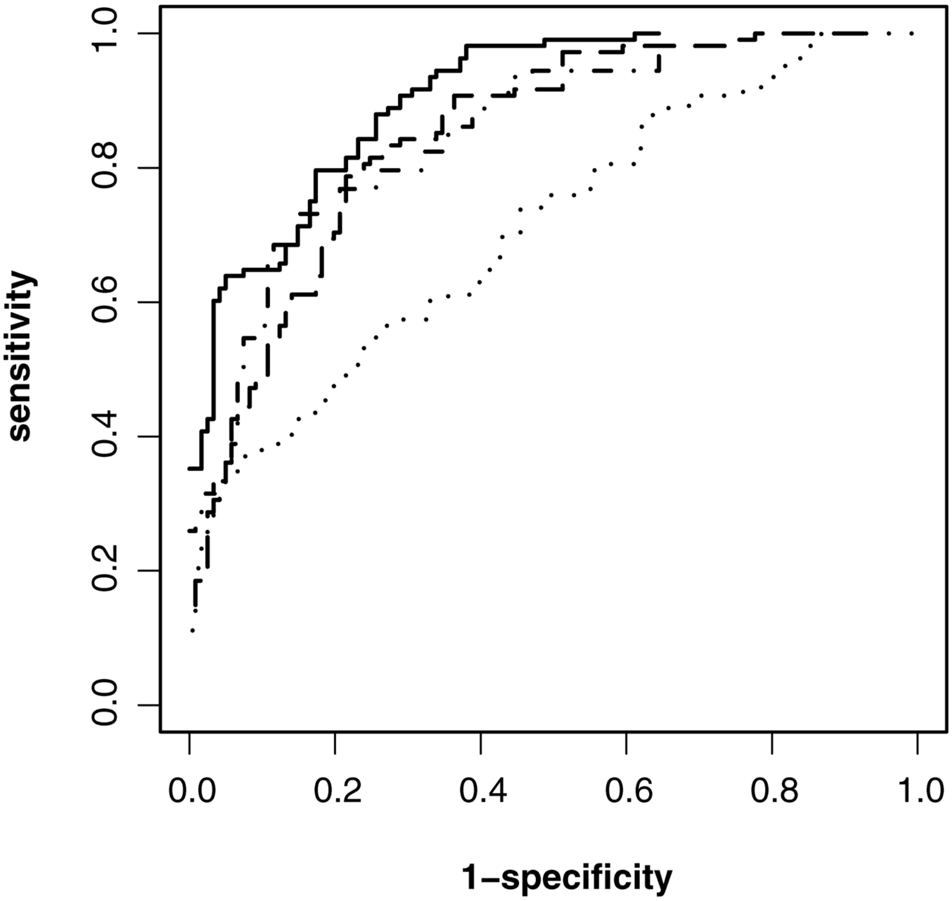

Figure 3.

Empirical ROC curves of four diagnostic markers for the acoustic startle response data (——, peak; · – ·, area; · · · ·, ratio; · – ·, duration).

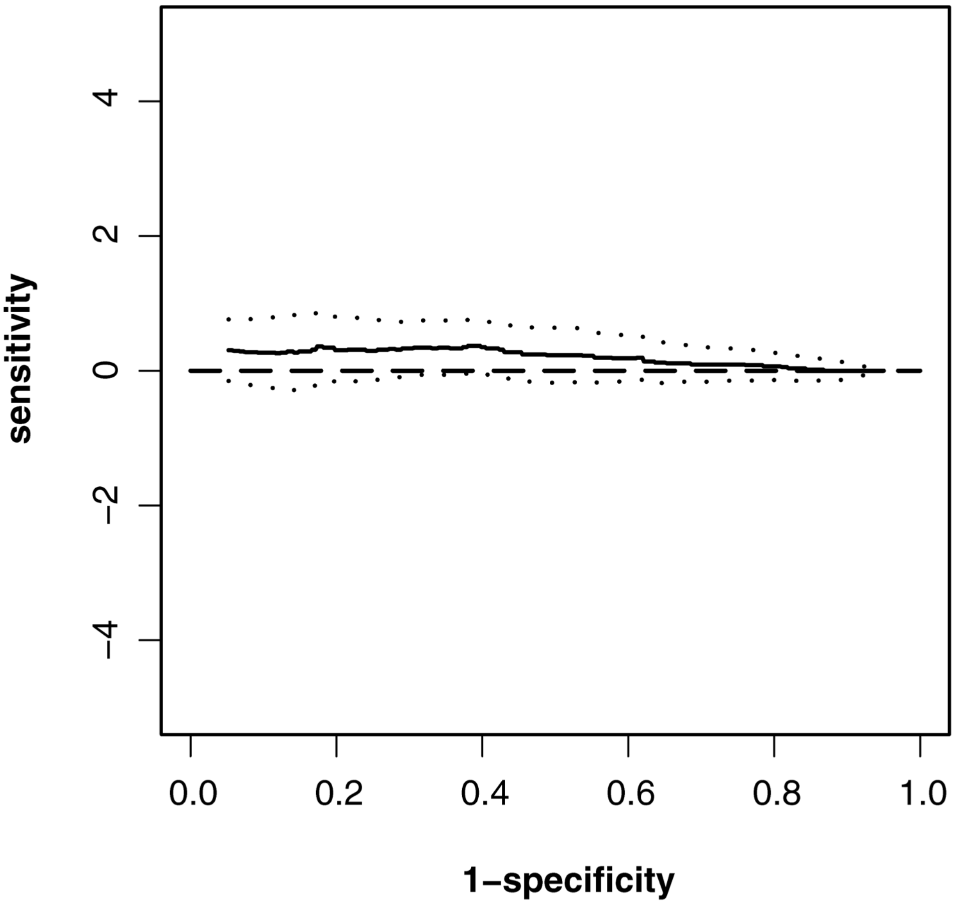

Figure 4.

95% confidence band of the difference of the ROC curves between *peak* and *area* from.1 to. 9 for the acoustic startle response data (——, difference; · · · ·, lower bound; · · · ·, upper bound).

Table 1

Achieved significance level of our method for two-sided test of equality of partial areas under the curves ($pAUC_{(p_1,\, p_2)}$) (nominal level =.05)

		Achieved significance level		
(ρ, λ)	**(n_D, n_H)**	**$pAUC_{(.05,.35)}$**	**$pAUC_{(.05,.55)}$**	**$pAUC_{(.05,.95)}$**
(0, 0)	(30, 30)	.07	.07	.06
	(50, 50)	.06	.05	.05
	(100, 100)	.05	.05	.06
(.25,.5)	(30, 30)	.05	.05	.04
	(50, 50)	.04	.05	.05
	(100, 100)	.04	.06	.04
(.50,.75)	(30, 30)	.05	.05	.06
	(50, 50)	.05	.04	.04
	(100, 100)	.05	.04	.04
(.75,.75)	(30, 30)	.05	.05	.06
	(50, 50)	.04	.04	.04
	(100, 100)	.04	.05	.05

NOTE: ρ and λ are measures of the within-subject and between-marker correlations, and (n_D, n_H) are the number of clusters for diseased and healthy subjects.

Table 2

Achieved significance level of the proposed method (method 1) versus the method ignoring clustering (method 2) for testing the equality of the total areas under the two ROC curves [nominal level =.05, (n_D, n_H) = (100, 100), λ =.5]

		Achieved significance level	
ρ	Cluster size	Method 1	Method 2
.05	4	.05	.04
	8	.06	.05
.75	4	.05	.13
	8	.06	.26
.99	4	.05	.22
	8	.04	.39

NOTE: ρ is a measure of within-subject correlation, and (n_D, n_H) are the number of diseased and healthy subjects.

Table 3

Simultaneous 95% confidence intervals for the pairwise differences of $pAUC_{(.05,.95)}$ of four tests for the acoustic response data

	Area	Ratio	Duration
Peak	(.03,.33)	(.05,.35)	(.01,.30)
Area		(−.03,.08)	(−.10,.06)
Ratio			(−.12,.03)

NOTE: $pAUC_{peak}$ =.66, $pAUC_{area}$ =.84, $pAUC_{ratio}$ =.86, $pAUC_{duration}$ =.81, and C_{α} = 2.62.