Biometrics 64, 1137–1145
December 2008

DOI: 10.1111/j.1541-0420.2008.01000.x

Nonparametric and Semiparametric Group Sequential Methods for Comparing Accuracy of Diagnostic Tests

Liansheng Tang,[1] Scott S. Emerson,[2] and Xiao-Hua Zhou[3,*]

[1]Department of Statistics, George Mason University, Fairfax, Virginia 22030, U.S.A.
[2]Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.
[3]Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.
HSR&D Center of Excellence, 1100 Olive Way, Seattle, Washington 98101, U.S.A.
[*]*email:* azhou@u.washington.edu

Summary. Comparison of the accuracy of two diagnostic tests using the receiver operating characteristic (ROC) curves from two diagnostic tests has been typically conducted using fixed sample designs. On the other hand, the human experimentation inherent in a comparison of diagnostic modalities argues for periodic monitoring of the accruing data to address many issues related to the ethics and efficiency of the medical study. To date, very little research has been done on the use of sequential sampling plans for comparative ROC studies, even when these studies may use expensive and unsafe diagnostic procedures. In this article we propose a nonparametric group sequential design plan. The nonparametric sequential method adapts a nonparametric family of weighted area under the ROC curve statistics (Wieand et al., 1989, *Biometrika* **76,** 585–592) and a group sequential sampling plan. We illustrate the implementation of this nonparametric approach for sequentially comparing ROC curves in the context of diagnostic screening for nonsmall-cell lung cancer. We also describe a semiparametric sequential method based on proportional hazard models. We compare the statistical properties of the nonparametric approach with alternative semiparametric and parametric analyses in simulation studies. The results show the nonparametric approach is robust to model misspecification and has excellent finite-sample performance.

Key words: Diagnostic accuracy; Proportional hazard model; Weighted AUC.

1. Introduction

Medical tests for diagnosis of disease are often based on the comparison of some biologic measurement to some threshold. In evaluating the utility of a particular test and threshold, it is common to consider its sensitivity (the probability that a truly diseased patient has a “positive” test exceeding the threshold) and the specificity (the probability that a truly healthy patient has a “negative” test in which the measurement does not exceed the threshold). However, the optimal threshold for any such biologic measurement depends on the prevalence of disease in the screened population and costs associated with incorrect diagnoses. It is thus common to compare two diagnostic tests with respect to their receiver operating characteristic (ROC) curves: plots of the relationship between the true-positive rate (the sensitivity) and the false-positive rate (FPR) (one minus the specificity), as the choice of threshold varies. When the ROC curve for one diagnostic test is uniformly greater than the ROC curve for another diagnostic test, the use of the first test will tend to engender lower costs from misdiagnosis no matter what the magnitude of costs assigned to each type of diagnostic error and no matter what the prevalence of disease in the screened population. As a measure of the tendency of the ROC curve for one test to dominate another in this manner, it is common for investigators to consider the area under each ROC curve (AUC). Diagnostic tests with the larger AUC are deemed superior. However, when the ROC curve of a test does not dominate that of another test, the AUC may not be a good measure to use for the comparison of two tests. Other measures, such as the partial area under the ROC curve within some range of acceptable specificity or sensitivity (pAUC), or some weighted average of the ROC curve (wAUC), should be used instead (see Zhou, Li, and Gatsonis [2008] for a more complete discussion).

Scientific studies designed to compare the utility of two diagnostic tests will typically use samples of diseased and healthy subjects as defined by some gold standard. Statistical analysis of the resulting data will focus on whether any difference in AUC (pAUC, or wAUC) is statistically significant. Statistical methods for comparing ROC curves might be parametric (e.g., the binormal model of Dorfman and Alf, 1969), semiparametric (Cai and Pepe, 2002), or nonparametric (Hanley and McNeil, 1982; Wieand et al., 1989). Molodianovitch, Faraggi, and Reiser (2006) provided a comprehensive study on nonparametric approaches for comparing AUCs.

In scientific studies that evaluate diagnostic tests, human experimentation raises issues related to ethics and efficiency. Interim analyses conducted at one or more times during the accrual of data in such studies can greatly improve the ability to address those ethical and efficiency issues. Such group sequential monitoring of clinical trials is commonplace, but the use of sequential sampling when evaluating diagnostic tests has not received much attention to date. Mazumdar and Liu (2003) provided a parametric sequential method for testing the equality of two AUCs when the observations in the healthy and diseased populations follow normal distributions (the "binormal model"). Zhou et al. (2008) proposed a nonparametric method for sequentially comparing AUCs. Mazumdar (2004) provided a general guideline for performing sequential tests for diagnostic accuracy studies based on AUCs.

Sequential tests are particularly valuable for comparative diagnostic trials in medical imaging modalities. Commonly used imaging procedures include computed tomography (CT) and positron emission tomography (PET). However, CT scanners expose the subjects to ionizing radiation from a series of X-rays. In PET, subjects have to undergo the injection of radioactive isotopes in order for their regions of interest to be measured. Unnecessarily exposing the subjects in the scientific study to an inferior diagnostic procedure is clearly undesirable. Furthermore, as both PET and CT are expensive procedures costing thousands of dollars per subject, unnecessarily delaying the identification of a beneficial diagnostic test not only diverts resources from better uses, but also does a disservice to those patients who are not participating in the scientific study, but would benefit from more accurate diagnosis.

Wieand et al. (1989) introduced wAUC estimators and the subsequent Wieand, Gail, James, and James (WGJJ) statistic for comparing two ROC curves is the difference between two wAUC estimators. The contribution made by this article is that we show a sequentially computed modification of this WGJJ statistic has approximately uncorrelated increments covariance structure, which allow the use of all popular group sequential methods introduced in Jennison and Turnbull (2000). One advantage of wAUC estimators over parametric AUC estimators in Mazumdar and Liu (2003) is that the wAUC estimator is distribution-free and includes a large family of statistics in diagnostic tests, such as nonparametric estimators of AUC, partial AUC, and sensitivity at 1- a specified specificity. The wAUC estimator also includes the AUC estimator in Zhou et al. (2008) as a special case.

The article is organized as follows. In Section 2 we show that after a specific transformation the WGJJ statistic is a Brownian motion process as information time grows, therefore, it can be applied in sequential tests. In Section 3 we describe design and monitoring procedures for comparing the wAUCs in group sequential designs. In Section 4 we describe the use in ROC curve analysis of a sequential semiparametric estimator. In Section 5 we conduct simulations to investigate the efficiency of these estimators. The nonparametric method is illustrated in Section 6 in lung cancer diagnosis trials and some discussion is presented in Section 7.

2. Asymptotic Distribution of Sequential WGJJ Statistic

Suppose we have measurements from two diagnostic tests on m diseased subjects and n healthy subjects, where all subjects are independent. Denote the measurements from test ℓ ($\ell =$ 1, 2) on the ith diseased subject as $X_{\ell i}$ and the corresponding measurements on the jth healthy subject as $Y_{\ell j}$. Define joint cumulative survivor distribution functions $(X_{1i}, X_{2i}) \sim \bar{F}(x_1, x_2)$ for the diseased population and $(Y_{1j}, Y_{2j}) \sim \bar{G}(y_1, y_2)$ for the healthy population. Further define marginal survivor distributions $X_{\ell i} \sim \bar{F}_\ell(x)$ and $Y_{\ell j} \sim \bar{G}_\ell(y)$.

Without loss of generality, we assume that measurements tend to be larger for diseased subjects than for healthy subjects. An ROC curve for the ℓth test can be expressed as a plot of sensitivity ($\Pr(X_{\ell i} > c) = \bar{F}_\ell(c)$) versus FPR, or 1 minus the specificity ($\Pr(Y_{\ell j} > c) = \bar{G}_\ell(c)$) as the threshold c varies over the real numbers. Equivalently, we can define the ROC curve for test ℓ as $ROC_\ell(u) = \bar{F}_\ell(\bar{G}_\ell^{-1}(u))$, where $0 \leqslant u \leqslant 1$, noting that in this parameterization u corresponds to the FPR.

Wieand et al. (1989) proposed comparing two ROC curves on the basis of the weighted AUC $\Omega_\ell = \int_0^1 [\bar{F}_\ell\{\bar{G}_\ell^{-1}(u)\}]\, dW(u)$, with a probability measure $W(u)$ defined on the FPR, u, for $u \in (0, 1)$. Included in this class of accuracy measures are the AUC (when $W(u) = u$ for $0 < u < 1$), the pAUC between FPRs u_1 and u_2 (when $W(u) = u$ for $0 < u_1 \leqslant u \leqslant u_2 < 1$), and the sensitivity at a given level of 1-specificity u_0 (when $W(u)$ is a point mass at $1 - u_0$).

In particular, Wieand et al. (1989) considered a nonparametric estimator based on empirical survivor distribution functions $\hat{\bar{F}}_\ell(x)$ and $\hat{\bar{G}}_\ell(y)$. Two diagnostic tests are then compared using the difference Δ in two wAUC estimators as estimated by

$$\hat{\Delta} = \hat{\Omega}_1 - \hat{\Omega}_2 = \int_0^1 \left[\hat{\bar{F}}_1\left\{\hat{\bar{G}}_1^{-1}(u)\right\}\right] dW(u) - \int_0^1 \left[\hat{\bar{F}}_2\left\{\hat{\bar{G}}_2^{-1}(u)\right\}\right] dW(u). \qquad (1)$$

In Wieand et al. (1989)'s paper, they derived the asymptotic property of $\hat{\Delta}$. However, their proof was complicated. The use of empirical theory can greatly simplify the proof. Here we re-derive the asymptotic distribution of $\hat{\Delta}$ in Theorem 1 by applying Lemma 3.9.27 of Van der Vaart and Wellner (1996), because $\hat{\Omega}_\ell$ can be written as the sum of two independent Brownian bridges. The detailed proof is provided in the Web Appendix.

Theorem 1. *Under mild regularity conditions, when* $m/n \rightarrow \lambda < \infty$ *as* $m, n \rightarrow \infty$, *the difference* $\hat{\Delta} = \hat{\Omega}_1 - \hat{\Omega}_2$ *satisfies*

$$\sqrt{m}(\hat{\Delta} - \Delta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, v_X + \lambda v_Y),$$

where v_X *and* v_Y *are given in the Web Appendix.*

The approximate variance σ_Δ^2 of $\hat{\Delta}$ is then $\sigma_\Delta^2 = v_X/m + v_Y/n$ and can be estimated by substituting the corresponding empirical estimates in v_X and v_Y in equation (A.1) in the Web Appendix. Using the above distributional theory, we can make statistical inference about Δ by using the nonparametric estimator $\hat{\Delta}$ with its approximately (large sample) normal distribution, $\mathcal{N}(\Delta, \sigma_\Delta^2)$. In particular, hypothesis tests of $H_0 : \Delta = \Delta_0$ can be based on the normalized statistic $Z = (\hat{\Delta} - \Delta_0)/\hat{\sigma}_\Delta$. In the presence of large sample sizes and the absence of early stopping, the Z statistic has the approximately standard normal distribution under H_0. The use of this statistic in the group sequential setting is described in the next section.

3. Use of the WGJJ Statistic under Group Sequential Sampling

3.1 *Stopping Rules*

We consider a group sequential sampling plan involving up to J analyses of the accruing data. At the time of the jth analysis, we have diagnostic test data available on the first m_j diseased subjects and the first n_j healthy subjects. From these available data, we compute for the ℓth diagnostic test the empirical survivor functions $\hat{\hat{F}}_{\ell j}$ and $\hat{\hat{G}}_{\ell j}$ and wAUC estimators $\hat{\Omega}_{\ell j}$. These interim estimates are then used to compare ROC curves using interim contrast $\hat{\Delta}_j$, its standard error $\sigma_{\Delta j}$, and the interim normalized statistic $Z_j = \hat{\Delta}_j/\sigma_{\Delta j}$.

We consider a group sequential sampling plan defined by up to four boundaries $-\infty \leqslant a_j \leqslant b_j \leqslant c_j \leqslant d_j \leqslant \infty$ at each of the J analyses. In order to uniquely define a stopping rule, we demand for $j < J$ that either $a_j < b_j$ or $c_j < d_j$ (or both) and that at least one of the four boundaries are finite. It is also typical that we obtain termination with a finite sample size by ensuring that $a_J = b_J$ and $c_J = d_J$ for some finite (but possibly random) choices of J, m_J, and n_J.

Sequential sampling proceeds by starting at analysis $j = 1$. At the jth analysis, measurements on the first m_j diseased subjects and n_j healthy subjects are used to compute the interim statistic Z_j. If $Z_j \leqslant a_j$, $b_j < Z_j < c_j$, or $Z_j \geqslant d_j$, the study is stopped without accruing more subjects. Otherwise, the study accrues sufficient subjects to be able to proceed to analysis $j + 1$. We define group sequential statistic $(\tilde{M}, Z)$ by $\tilde{M} = \min\{1 \leqslant j \leqslant J : Z_j \notin (a_j, b_j] \cup [c_j, d_j)\}$ and $Z = Z_{\tilde{M}}$. In the setting of comparing ROC curves, we would most often decide that diagnostic test 1 is superior, approximately equivalent, or inferior to diagnostic test 2 according to whether $Z \geqslant d_{\tilde{M}}, b_{\tilde{M}} < Z < c_{\tilde{M}}$, or $Z \leqslant a_{\tilde{M}}$, respectively.

When frequentist statistical inference is the ultimate goal, it is common to ensure that the experimentwise error is controlled at a desired level (e.g., in a two-sided hypothesis test, choose stopping boundaries to ensure $\Pr(b_{\tilde{M}} < Z < c_{\tilde{M}} \,|\, H_0) = 1 - \alpha$ for a desired type I error α). The dimensionality of the boundary space is reduced through the use of a boundary shape function that defines a relationship between the exact value of the stopping boundaries a_j, b_j, c_j, d_j and the statistical information available at the jth analysis (which is typically $1/\sigma^2_{\Delta j}$ in the case of an approximately normally distributed statistic). For purposes of sample size calculation, the boundary shape function is specified in terms of the proportion τ_j of maximal statistical information available at each analysis (in the case of approximately normally distributed statistics $\tau_j = \sigma^2_{\Delta j}/\sigma^2_{\Delta J}$). Common boundary shape functions include the O'Brien–Fleming, the triangular test, and the Pocock boundaries (see Jennison and Turnbull, 2000).

The statistical literature is replete with alternative strategies for choosing stopping boundaries appropriate for particular scientific and statistical settings (see Jennison and Turnbull, 2000; Emerson, Kittelson, and Gillen, 2007a, 2007b). Almost all of that statistical literature, and all three of the commercially available statistical software capable of implementing the general methods (`S+SeqTrial`, `PEST`, `EaSt`), presume a particular covariance structure for the statistics $Z_1, Z_2, \ldots, Z_J$. In this covariance structure, the statistical information accrued between two successive analyses is independent of all prior information accrued, and it is commonly referred to as an "independent increment" covariance structure (Jennison and Turnbull, 2000, Chapter 11).

The use of the WGJJ statistic in the sequential comparison of ROC curves is greatly facilitated by showing that the statistic has an "independent increment" structure. We let I_j denote the statistical information at the jth analysis, and $\tau_j = I_j/I_J$ denote the proportion of maximal information as before. Define $B(\tau_j) = \sqrt{\tau_j I_j}\hat{\Delta}_j$, which is an asymptotically unbiased estimator for $\sqrt{\tau_j I_j}\Delta = \tau_j\sqrt{I_J}\Delta =$ with asymptotic variance $\text{var}(B(\tau_j)) = \tau_j$.

Theorem 2. *For $j < k$, $\text{cov}\{B(\tau_j), B(\tau_k)\} = \tau_j$.*

The proof is provided in the Web Appendix. Thus $B(\tau_j)$ behaves asymptotically like a Brownian motion process with a drift parameter θ, where $\theta = \Delta\sqrt{I_J}$. The WGJJ estimator can then be readily accommodated by standard group sequential software.

3.2 *Sample Size Determination*

In a wide variety of statistical models, the maximal number $\tilde{N}_J$ of sampling units needed is estimated by $\tilde{N}_J = \delta^2_{\alpha\beta} V/\Delta_1^2$, where $1/V$ is the (average) statistical information contributed by a single sampling unit, Δ_1 is the difference between Ω_ℓ's under the alternative hypothesis to be detected with statistical power $1 - \beta$ in a level α hypothesis test, and $\delta_{\alpha\beta}$ is the design alternative in some standardized version of the test. For instance, in a fixed sample (no interim analyses, $J = 1$) two-sided hypothesis test of the difference in weighted AUCs having equal sample sizes, $\tilde{N}_J$ might be the sample size to be accrued in each group, $\Delta_1 = \Omega_1 - \Omega_2$ might be the difference between group means under the design alternative, $V = \sigma^2_\Delta$ is the variance, and $\delta_{\alpha\beta} = z_{1-\alpha/2} + z_\beta$. This same formula can be used in a group sequential test, providing that the value of $\delta_{\alpha\beta}$ specific to the selected stopping rule is used.

With pilot data, it is trivial to nonparametrically calculate the variance of the WGJJ statistic from equation (A.1) in the Web Appendix. However, if the pilot study is not yet available, we will nevertheless need parametric distribution assumptions to obtain the variance of the WGJJ statistic. This means that we will have to guess explicit distributional models and parameters in the models. In the following context, we will discuss a method to obtain conservative sample sizes based on the conjectured values of AUCs without either having a pilot data or specifying model parameters.

We denote the WGJJ statistic for AUCs as Δ^A. Δ^A is the difference between two Wilcoxon statistics, $\hat{\Omega}^A_1$ and $\hat{\Omega}^A_2$. Hanley and McNeil (1982) showed that under negative exponential models an estimated AUC had a larger variance than under normal or gamma distributions, therefore, a more conservative sample size. By using negative exponential distributions, we can derive the variance of $\hat{\Omega}^A_\ell$ solely from hypothesized AUC values, Ω^A_ℓ, without prespecified parameters. Sample sizes for two arms can then be calculated without knowing parameters in negative exponential distributions as stated in the following theorem.

Theorem 3. *Under mild regularity conditions, as $m/n \to \lambda$ when $m, n \to \infty$, the variance v^2_A of Δ^A is then given by $v^2_A = \text{var}(\hat{\Omega}^A_1) + \text{var}(\hat{\Omega}^A_2) - 2\rho\sqrt{\text{var}(\hat{\Omega}^A_1)\text{var}(\hat{\Omega}^A_2)}$, where $\text{var}(\sqrt{m}\hat{\Omega}^A_\ell) = \lambda Q_{1\ell} + Q_{2\ell} - (\lambda + 1)(\Omega^A_\ell)^2$, with $Q_{1\ell} = \Omega^A_\ell/$*

$(2 - \Omega_\ell^A)$, $Q_{2\ell} = 2(\Omega_\ell^A)^2/(1 + \Omega_\ell^A)$ *and* ρ, *the correlation between two AUCs.*

Proof. With finite sample sizes m and n, we have

$$\begin{aligned}\operatorname{var}\left(\hat{\Omega}_\ell^A\right) = \Big[&\Omega_\ell^A\left(1 - \Omega_\ell^A\right) + (m-1)\left\{Q_{1\ell} - \left(\Omega_\ell^A\right)^2\right\} \\ &+ (n-1)\left\{Q_{2\ell} - \left(\Omega_\ell^A\right)^2\right\}\Big]/mn.\end{aligned}$$

Thus, as $m, n \to \infty$, it is true that $\operatorname{var}(\sqrt{m}\hat{\Omega}_\ell^A) \to \lambda Q_{1\ell} + Q_{2\ell} - (\lambda + 1)(\Omega_\ell^A)^2$.

Denote $\tilde{V}_\ell$ to be the right side of the above equation. Consider a null hypothesis, $\mathrm{H}_0 : \Delta^A = 0$ and the alternative two-sided hypothesis, $H_A : \Delta^A \neq 0$ with Type I error α and power $1 - \beta$ at the conjectured AUC values, Ω_1^A and Ω_2^A. Using a general formula in Zhou, McClish, and Obuchowski (2002, Section 6.2), the required fixed sample sizes for the diseased and healthy subjects, denoted as M_f and N_f, respectively, can be derived by

$$\begin{aligned}M_f &= \lambda N_f \\ &= \frac{\left\{Z_{1-\alpha/2}\sqrt{(2-2\rho)\tilde{V}_1} + Z_{1-\beta}\sqrt{\tilde{V}_1 + \tilde{V}_2 - 2\rho\sqrt{\tilde{V}_1\tilde{V}_2}}\right\}^2}{\left(\Omega_1^A - \Omega_2^A\right)^2}. \qquad (2)\end{aligned}$$

A special case with $\lambda = 1$ and $\rho = 0$ is given in Hanley and McNeil (1982). Subsequently, the maximum sample sizes M_g for the diseased and N_g for the healthy can be obtained for the O'Brien–Fleming test, the triangular test, and the Pocock test by $M_g = \lambda N_g = \delta^2_{\alpha\beta,g}/\delta^2_{\alpha\beta,f}M_f$, where $\delta^2_{\alpha\beta,g}/\delta^2_{\alpha\beta,f}$ is the sample size ratio between a fixed sample design and a sequential design. The maximum sample sizes M_e and N_e can also be obtained for a more flexible error spending function design (Lan and DeMets, 1983) by using $M_e = \lambda N_e = \theta_e^2/\theta_f^2 M_f$, where θ_f and θ_e are drift values for a fixed sample design and a group sequential design, respectively.

4. Semiparametric Partial AUC Estimator

When the measurements of diagnostic tests are from exponential distributions or other power–law distributions, the proportional hazard model assumption is satisfied (for a real example, see Sanchez-Marin and Padilla-Medina, 2006). We can then use a semiparametric estimator for comparing correlated AUCs under proportional hazard models. Let $Z_{\ell i} = 1$ if the ith subject is diseased, 0 otherwise, under the ℓth diagnostic test with $\ell = 1, 2$. Now the ℓth test of the ith subject has a hazard function $\lambda_{\ell i}(t) = \lambda_{\ell 0}(t)\exp\{\gamma_\ell Z_{\ell i}\}$, where $\lambda_{\ell 0}(t)$ is a baseline hazard function and γ_ℓ's are parameters in the Cox regression. The resulting ROC curve for the ℓth test takes on the form of $ROC_\ell(u) = u^{\exp(\gamma_\ell)}$. Its AUC is given by $\Omega^A(\gamma_\ell) = 1/\{\exp(\gamma_\ell) + 1\}$, and its partial AUC between a and b, $0 < a < b \leqslant 1$, is given by $\Omega^{pA}(\hat{\gamma}_\ell) = \{b^{\exp(\hat{\gamma}_\ell)+1} - a^{\exp(\hat{\gamma}_\ell)+1}\}/\{\exp(\hat{\gamma}_\ell) + 1\}$. If $0 = a < b \leqslant 1$, the partial AUC is $\Omega^{pA}(\hat{\gamma}_\ell) = b^{\exp(\hat{\gamma}_\ell)+1}/\{\exp(\hat{\gamma}_\ell) + 1\}$. The covariate-adjusted estimator $\hat{\gamma}_\ell$ is obtained from the marginal Cox regression model for the ℓth test. The asymptotical property of $\hat{\gamma}_\ell$'s is shown in Wei, Lin, and Weissfeld (1989) by

$$n^{1/2}(\hat{\gamma}_1 - \gamma_1, \hat{\gamma}_2 - \gamma_2) \xrightarrow{\mathcal{D}} N\left(0, v_1\Sigma^\gamma v_2\right),$$

where $v_\ell = \operatorname{var}(\hat{\gamma}_\ell)$ and Σ^γ is the correlation matrix of $n^{1/2}(\hat{\gamma}_1 - \gamma_1, \hat{\gamma}_2 - \gamma_2)$.

Theorem 4. *As* $m, n \to \infty$, *the estimated AUC (or partial AUC) difference,* $\Delta(\hat{\gamma}_1, \hat{\gamma}_2)$, *satisfies*

$$n^{1/2}\{\Delta(\hat{\gamma}_1, \hat{\gamma}_2) - \Delta(\gamma_1, \gamma_2)\} \xrightarrow{\mathcal{D}} N\left(0, v_p^2\right),$$

where

$$v_p^2 = \begin{pmatrix} (\Omega)'(\gamma_1) \\ -(\Omega)'(\gamma_2) \end{pmatrix}' \Sigma^\gamma \begin{pmatrix} (\Omega)'(\gamma_1) \\ -(\Omega)'(\gamma_2) \end{pmatrix}.$$

Its proof in a general survival model framework is provided in the Web Appendix. The sequential versions of our semiparametric proportional hazard AUC or pAUC estimators have asymptotically independent increments. Thus, it is straightforward to incorporate the semiparametric estimators in group sequential designs. Our estimator is also easily adapted for sequentially testing survival outcome measurements from two diagnostic tests. The covariance matrix Σ^γ between γ_ℓ's can be consistently estimated using a sandwich estimator (Wei et al., 1989) by $\hat{\Sigma}^\gamma = n\hat{v}_1 W_1 W_2' \hat{v}_2$, where $\hat{v}_\ell$ is the estimated variances for $\hat{\gamma}_\ell$, and W_ℓ is the residual vector calculated from the marginal proportional hazard model for the ℓth test.

5. Finite Sample Property

In this simulation study, we investigated the finite sample performance of the sequential WGJJ statistic and the previously described semiparametric procedure, both in a fixed sample test ($J = 1$), a three-group sequential test ($J = 3$), a four-group sequential test ($J = 4$), and a five-group sequential test ($J = 5$). We also included a common binormal parametric approach for comparisons among these three methods. The null hypothesis of equal AUCs was set to be true and the nominal type I error was set to be 0.05 for two-sided tests. We simulated bivariate normal (Binorm), bivariate lognormal (Bilog), and bivariate exponential (Biexp) data as outcome measurements for two diagnostic tests. The bivariate normal models had the forms of $(X_1, X_2) \sim N\{(11, 1), \Sigma_1\}$ and $(Y_1, Y_2) \sim N\{(10, 0), \Sigma_2\}$, where

$$\Sigma_1 = \begin{pmatrix} 1 & \sqrt{2}\rho \\ \sqrt{2}\rho & 2 \end{pmatrix} \text{ and } \Sigma_2 = \begin{pmatrix} 2 & \sqrt{2}\rho \\ \sqrt{2}\rho & 1 \end{pmatrix}, \text{ with } \rho = 0.5.$$

The AUCs were thus the same from the formula of AUC under binormal models (Zhou et al., 2002): $AUC = \Phi\{(\mu_1 - \mu_0)/(\sqrt{\sigma_1^2 + \sigma_0^2})\}$, where (μ_1, σ_1) and (μ_0, σ_0) are the normal parameters in diseased and healthy groups, respectively. The bivariate lognormal models had the forms of $\exp(X_1, X_2)$ and $\exp(Y_1, Y_2)$ for the diseased and healthy subjects, respectively. The AUCs under simulated lognormal models were also equal, because ROC curves are invariant to monotone transformations. Equal numbers of diseased and healthy subjects were considered in the simulation, that is, m = n = (50, 100, 200).

The bivariate exponential random variables were simulated using an algorithm in Gumbel (1960). The Gumbel's distribution had the form $H(x, y) = H_1(x)H_2(y)[1 + 4\rho\{1 - H_1(x)\}\ \{1 - H_2(y)\}]$, where $H_1(x)$, $H_2(x)$ are exponential functions, and $\rho \in [-0.25, 0.25]$. We set ρ to be 0.25.

Table 1

Type I error with the nominal level $\alpha = 0.05$ in the group sequential designs

	m	WGJJ statistic			Parametric			Semiparametric		
	(n)	Binorm	Bilog	Biexp	Binorm	Bilog	Biexp	Binorm	Bilog	Biexp
					Fixed sample design ($J = 1$)					
	50	6.9%	6.6%	5.6%	4.7%	12.0%	0.8%	31.5%	30.4%	5.5%
	100	4.5%	6.0%	6.2%	4.1%	21.6%	1.4%	81.2%	80.7%	5.6%
	200	5.6%	6.4%	4.0%	4.0%	37.5%	0.8%	88.6%	90.6%	5.5%
					Three-group sequential design ($J = 3$)					
OBF	50	4.3%	5.2%	5.2%	5.0%	30.1%	1.0%	77.0%	75.8%	6.6%
	100	4.2%	3.3%	3.7%	2.3%	43.8%	0.4%	98.3%	97.4%	6.1%
	200	4.7%	5.9%	6.4%	4.5%	62.9%	1.5%	100.0%	100.0%	5.3%
POC	50	5.5%	6.7%	5.7%	4.4%	26.1%	0.8%	69.7%	66.8%	6.8%
	100	5.3%	4.5%	5.0%	3.3%	41.4%	0.3%	96.5%	95.5%	6.5%
	200	4.7%	6.0%	6.5%	4.1%	60.6%	1.3%	100.0%	100.0%	5.0%
					Four-group sequential design ($J = 4$)					
OBF	50	4.7%	3.6%	5.3%	4.2%	33.8%	0.4%	92.5%	92.8%	6.6%
	100	4.3%	4.6%	5.2%	3.9%	52.4%	1.0%	100.0%	99.6%	6.5%
	200	5.8%	4.8%	5.4%	4.1%	72.6%	0.3%	100.0%	100.0%	4.9%
POC	50	4.5%	6.2%	6.0%	3.4%	30.8%	0.7%	88.7%	87.6%	6.2%
	100	5.4%	3.6%	5.8%	4.9%	48.0%	1.2%	99.9%	99.0%	6.1%
	200	4.8%	5.1%	5.7%	3.8%	69.4%	0.5%	100.0%	100.0%	5.3%
					Five-group sequential design ($J = 5$)					
OBF	50	4.9%	4.5%	5.8%	3.9%	40.8%	0.5%	97.4%	97.4%	5.7%
	100	4.4%	5.6%	4.7%	5.1%	60.4%	0.6%	100.0%	99.8%	5.3%
	200	4.5%	4.5%	5.7%	3.5%	78.8%	0.4%	100.0%	100.0%	5.0%
POC	50	5.9%	5.2%	6.0%	4.1%	35.5%	0.7%	95.2%	94.7%	5.3%
	100	4.4%	5.9%	5.1%	3.1%	55.6%	1.2%	100.0%	99.8%	5.0%
	200	5.4%	5.4%	6.0%	3.8%	75.9%	1.4%	100.0%	100.0%	5.0%

The rejection rate with 1000 realizations. The 95% prediction interval is (95.0% ± 1.4%).

Bivariate exponential random variables were simulated with the marginal survival functions $\exp(-\beta_{\ell 1}x)$ and $\exp(-\beta_{\ell 2}y)$ for diseased and healthy subjects respectively, where $\ell = 1$, 2, denotes the types of tests. In the simulation, $(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}) = (1, 2, 2, 4)$. Because the AUCs under these exponential distributions are given by $\exp(\gamma_\ell) = \beta_{\ell 1}/\beta_{\ell 2}$, two resulting AUCs should be the same.

Under each of the above model assumptions, 1000 random variables were simulated and three methods including sequential WGJJ statistic, semiparametric method, and parametric binormal method were fitted to the simulated data. The Z statistics were then calculated based on estimated parameters and their variances. The rejection rates were obtained by comparing the Z statistics with corresponding test boundaries under either the fixed sample design or sequential designs. Table 1 gives the rejection rates of all three methods with a nominal level 0.05 under both Pocock's (POC) and O'Brien and Fleming's (OBF) criterions. In the fixed sample test, the WGJJ statistic gives the rejection rate close to the nominal level under all model specifications for sample sizes in both groups as small as 50, whereas the parametric binormal method greatly inflates rejection rates when the true underlying distribution is bivariate lognormal, and deflates rejection rates when the underlying distribution is in fact bivariate exponential. Under the setting of group sequential designs, the WGJJ statistic also gives correct rejection rates regardless of underlying distribution models. When the underlying distributions are misspecified, the parametric and semiparametric methods inflate the rejection rates further compared with their fixed sample counterparts. In addition, with proportional hazard model correctly specified, the semiparametric method performs well for both the fixed sample design and the sequential designs. In summary, the nonparametric approach based on the WGJJ statistic is robust to model specifications and it performs as well as parametric approach under correct model assumptions. Moreover, the nonparametric approach has excellent small sample performance, which makes it a well-suited method for conducting group sequential diagnostic trials.

6. Examples

6.1 *An Illustration for Sample Size Determination*

We used binormal and biexponential models to illustrate how to determine maximum sample sizes in the fixed sample design and group sequential design using the WGJJ statistic for AUC and partial AUC estimators, which are denoted as Δ^A and Δ^{pA}, respectively. The equally spaced symmetric two-sided error spending (Kim and Demets, 1992) test with the error spending function $f(\tau) = \min(\alpha\ \tau, \alpha)$ was used with power 0.8 and type I error $\alpha = 0.05$.

6.1.1 *Under bivariate normal model assumption.* Suppose the binormal distribution of the test outcomes is given by $(X_1, X_2) \sim N\{(\mu_1, \mu_2), \Sigma\}$, $(Y_1, Y_2) \sim N\{(0, 0), \Sigma\}$, where covariance matrix Σ had common variances 1 and covariances

Table 2

Maximum possible number of subjects in both arms for testing Δ^A or Δ^{pA} in three-group analysis (and fixed sample analysis) with $1 - \beta = 0.80$ and $\alpha = 0.05$ under the bivariate normal distribution

$\Delta^A = \Omega_1^A - \Omega_2^A$						
$\Omega_1^A \backslash \Omega_2^A$	0.750	0.800	0.850	0.900	0.950	0.975
0.700	929 (832)	217 (195)	91 (81)	49 (44)	32 (29)	28 (25)
0.750	NA	793 (710)	182 (163)	76 (68)	43 (38)	36 (32)
0.800		NA	634 (568)	143 (128)	62 (55)	48 (43)
0.850			NA	456 (408)	103 (92)	68 (61)
0.900				NA	267 (239)	116 (103)
0.950					NA	416 (372)
$\Delta^{pA} = \Omega_1^{pA} - \Omega_2^{pA}$						
$\Omega_1^{pA} \backslash \Omega_2^{pA}$	0.350	0.400	0.450	0.500	0.550	0.575
0.300	732 (655)	174 (156)	72 (65)	38 (34)	24 (21)	20 (18)
0.350	NA	675 (604)	156 (140)	64 (57)	35 (31)	28 (25)
0.400		NA	573 (513)	129 (116)	54 (48)	41 (36)
0.450			NA	436 (390)	97 (87)	62 (56)
0.500				NA	273 (244)	115 (103)
0.550					NA	505 (452)

0.5. We let $\lambda = 1$, equivalent to disease prevalence 0.5. Because the distributions were known, we were able to obtain the exact variance of Δ^A or Δ^{pA} from the results in equation (A.1) in the Web Appendix. Under the specified test setting, we obtained sample sizes N_g from the results in Section 3.2. Here, the drift value $\theta_e = 2.96$ for a sequential design and $\theta_f = 2.80$ for a fixed sample design can be calculated using aforementioned software. The sample sizes N_f for the fixed sample design can then be computed. The results for Δ^A are presented in Table 2. We also found that the sample sizes given by equation (2) under the same test setting are slightly more than those in Table 2. This ensures that without the pilot data or prior knowledge of the distributions, the sample size computed by equation (2) can serve as a conservative initial guess. To give an example on how accurate these sample sizes are in maintaining the desired power, we used a sample size 929 in Table 2, which is for comparing AUC 0.75 with AUC 0.70 in a three-group design. We simulated 1000 data sets under the binormal setting, and computed the number of times that the null hypothesis of equal AUCs was rejected. Our result was 803 out of 1000. The resulting power was then 80.3%, very close to the nominal level 80%.

Suppose now we are interested in comparing partial AUCs for the FPR less than 0.6. For our range of FPR, it can be calculated that the partial AUC is between 0.18 and 0.6 (Zhou et al., 2002, Section 4). Table 2 also gives maximum possible sample sizes for testing the difference between partial AUCs.

These two sets of examples gave an insight on deriving maximum sample sizes for three-group sequential designs if the distributions of test outcomes are specified. Given a specified power and a type I error, available software such as `S+SeqTrial`, `PEST`, and `EaSt` can compute sample size ratios for these designs. If the required sample size for a fixed sample design is available, the maximum sample size for any sequential design can be derived by multiplying a specified constant ratio related to the sequential design. In the case that the distributions are unknown but the pilot data are available, v_X and v_Y can be estimated by plugging in the empirical distributions and quantiles and substituting the empirical estimates of $r_1(u)$ and $r_2(u)$ in equation (A.1) in the Web Appendix, respectively.

At the first glance of Table 2, one may notice that the maximum sample sizes of the group sequential test are often larger than the required sample sizes of the fixed sample test. This is because these maximum sample sizes only occur in the worst scenario when group sequential tests are carried all the way to the end. Often sequential tests terminate earlier before the maximum number of subjects are recruited. Therefore, looking at expected sample sizes reveals the advantage of sequential trials. The fact that the expected sizes are about 81% of those of the fixed sample test ensures the early stopping of the group sequential design.

6.1.2 *Under bivariate exponential model assumption.* The aforementioned bivariate exponential distribution in Gumbel (1960) was used to calculate the variance of the WGJJ estimator and that of the semiparametric estimator. We let $\rho = 0.25$. For diseased and healthy groups, bivariate exponential random variables had marginal survival function of $\exp(-\beta_{\ell 1}x)$ and $\exp(-\beta_{\ell 2}y)$ for tests 1 and 2. We let $\beta_{12} = \beta_{22} = 1$. The values of β_{11} and β_{21} corresponded to the AUCs (or pAUCs) in Table 3. We let $\lambda = 1$ in the sample size calculation with the specified test setting. Not surprisingly, the nonparametric and semiparametric methods give the same sample sizes when comparing AUCs or pAUCs. The maximum possible sample sizes are listed in Table 3 for AUCs and partial AUCs.

6.2 *Sequential Lung Cancer Diagnostic Trial*

Lung cancer is one of the most common cancers in the world and is the leading cause of cancer death in the United States. The lung cancer is categorized into two types: small cell and nonsmall cell. The nonsmall-cell lung cancer is the most common type, which is only curable with surgery in its early stages. CT and PET are both the preoperative scanning tests for the staging of nonsmall-cell lung cancer. CT, a traditional diagnostic tool, does not perform well to differentiate between benign and malignant lesions. PET, a new scanning

Table 3

Maximum possible number of subjects in both arms for testing Δ^A or Δ^{pA} in three-group analysis (and fixed sample analysis) with power $1 - \beta = 0.80$ and $\alpha = 0.05$ under the bivariate exponential distribution

$\Delta^A = \Omega_1^A - \Omega_2^A$

$\Omega_1^A \backslash \Omega_2^A$	0.750	0.800	0.850	0.900	0.950	0.975
0.700	1244 (1113)	293 (262)	122 (109)	65 (58)	39 (35)	32 (28)
0.750	NA	1094 (979)	253 (226)	104 (93)	54 (49)	42 (37)
0.800		NA	914 (818)	206 (184)	83 (74)	58 (52)
0.850			NA	705 (631)	152 (136)	90 (80)
0.900				NA	461 (412)	180 (161)
0.950					NA	975 (872)

$\Delta^{pA} = \Omega_1^{pA} - \Omega_2^{pA}$

$\Omega_1^{pA} \backslash \Omega_2^{pA}$	0.350	0.400	0.450	0.500	0.550	0.575
0.300	852 (763)	203 (182)	84 (76)	44 (39)	26 (24)	21 (19)
0.350	NA	785 (703)	183 (164)	75 (67)	39 (35)	30 (27)
0.400		NA	680 (609)	154 (138)	62 (55)	43 (39)
0.450			NA	539 (482)	117 (104)	69 (62)
0.500				NA	362 (324)	142 (127)
0.550					NA	781 (699)

technique, provides higher resolution image with a detailed view of regions of interest. But the results of PET are usually affected by muscle and inflammatory activities, which are considered to be factors of imprecision locations of abnormalities. Comparing the diagnostic performance of these two techniques is of extensive interest of radiologists (Lardinois et al., 2003; Silvestri et al., 2003). The gold standard in these tests is pathology results from biopsy specimens.

The staging accuracy of nonsmall-cell lung cancer is usually between 52% and 85% for CT and between 81% and 96% for PET (Lardinois et al., 2003; and Silvestri et al., 2003). Consider testing the null hypothesis of equal AUCs against the two-sided alternative with power 0.8 and significance level 0.05. A possible initial alternative is likely to be 15%, which is the difference between 70% for CT and 85% for PET. If an investigator is interested in a nonparametric AUC estimator, corresponding to two-sample Wilcoxon statistics (Hanley and McNeil, 1983), as an accuracy measure in a fixed sample lung cancer trial to compare CT and PET, the maximum sample size is 81 under power 0.8, and type I error 0.05 under the binormal assumption according to Table 2. If the outcome measurements are assumed to be from bivariate exponentials, a total of 109 subjects in both arms would be required from Table 3. However, as mentioned in the Introduction, these diagnostic methods are expensive and carry some safety risks. While carrying out clinical trials to compare CT and PET, the results need to be monitored repeatedly to ensure that human subjects are not exposed to inferior scanning techniques. In negative trials that show equivalence of the accuracy of CT and PET, the trials need to be terminated early and the subjects can be switched to compare CT with other scanning techniques that involve combining CT and PET, or recently developed magnetic resonance imaging technique. If PET is found to be more accurate than CT in the early interim analysis, it means that PET scan in staging the nonsmall-cell lung cancer would be preferred and should be performed more frequently in lung cancer diagnosis.

Suppose we want to design a three-group error spending test with the error spending function $f(\tau) = \min(\alpha\ \tau,\ \alpha)$. The equally spaced symmetric two-sided test with significance level 0.05 and power 0.8 would need maximum possible sample sizes of 91 and 62 under binormal and biexponential assumptions, respectively, based on the results from Table 2. In this test the boundaries for the normalized AUC statistic, Z_j, can be calculated as ($c_1 = -b_1 = 2.39$, $c_2 = -b_2 = 2.29$, $c_3 = -b_3 = 2.20$). During the comparative diagnostic trial, Z_j is computed at the jth interim analysis and compared with these critical boundaries. A significant Z_j gives early evidence that CT is different from PET, and the better imaging method should be adapted in detecting lung cancer. If there is no significance and we have not finished the trial on all patients, we will continue recruiting more patients until there is significant evidence in the next analysis. After all patients are scanned, if there is still no significant evidence to support the alternative hypothesis, a conclusion will be made that CT has the same diagnostic ability as PET, and a cheaper CT scanning technique can be recommended for nonsmall-cell lung cancer diagnosis. Because we already noted that the expected sample sizes under sequential tests would be as low as 81% of that under the fixed sample test, the early stopping on the average is ensured by group sequential tests.

AUC is an excellent accuracy measure if two ROC curves do not cross each other. However, when ROC curves cross, they may have similar AUCs but different partial AUCs over a range of specificities. Suppose an investigator is interested in whether there is a difference between partial AUCs of CT and PET over the FPR less than 0.6. In this trial, the WGJJ statistic then becomes the difference between the nonparametric estimators of partial AUCs over high specificities. An initial guess of partial AUCs is 40% for CT and 55% for PET, respectively. With power 0.8, type I error 0.05, and disease prevalence 0.5, a fixed sample trial would require 48 patients under the binormal assumption and 55 patients under the exponential assumption according to Tables 2 and 3, respectively.

If a three-group error spending test is decided for the trial, the maximum possible sample size for both arms would be 54 and 62 under binormal and biexponential assumptions, respectively, based on the results in Tables 2 and 3. The boundaries of three stages for the normalized partial AUC differences are also ($c_1 = -b_1 = 2.39$, $c_2 = -b_2 = 2.29$, $c_3 = -b_3 = 2.20$). Similar to comparative diagnostic trial based on AUC, the normalized test statistic Z_j based on partial AUC is computed based on accruing patients at the jth stage and corresponding decision will be based on whether Z_j crosses the boundaries. Note that the maximum possible sample sizes are smaller for the sequential trials comparing partial AUCs than those for AUCs. This is because partial AUC estimator is less variable than AUC estimator.

To the best of our knowledge, no actual trials in diagnostic medicine have been conducted sequentially. To illustrate the details of calculations of monitoring decisions, we simulated a set of simple hypothetical outcomes from PET and CT scans in aforementioned lung cancer diagnostic trials. Suppose an investigator has decided to conduct a three-group sequential trial to compare the accuracy of CT and PET procedures. Under the binormal assumption the maximum possible sample size was initially 91 in total, based on the 80% power and 0.05 type I error. We simulated binormal data as measurements from 30 subjects at the first look. We then calculated interim contrast $\hat{\Delta}_1 = 0.0259$, its standard error $\sigma_{\Delta 1} = 0.0673$, and the interim normalized statistic $Z_1 = \hat{\Delta}_1/\sigma_{\Delta 1} = 0.3848$. Because Z_1 fell within the boundary ($c_1 = -b_1 = 2.39$), we continued with the second look at 30 more simulated measurements. At the second look, we calculated interim contrast $\hat{\Delta}_2 = 0.1469$, its standard error $\sigma_{\Delta 2} = 0.0534$, and the interim normalized statistic $Z_2 = \hat{\Delta}_2/\sigma_{\Delta 2} = 2.7510$ from all 60 subjects. Now Z_2 was outside the boundary ($c_2 = -b_2 = 2.29$); thus we stopped recruiting more subjects and came to a conclusion that PET has better accuracy to stage lung cancer than CT.

7. Discussion

In this article, we described that after modification the sequential WGJJ statistic behaves like the Brownian motion process, and therefore, can be readily implemented using standard statistical software. The WGJJ statistic includes a large family of nonparametric estimators in comparative diagnostic tests, offering great modeling flexibility. We also proposed a semiparametric method for comparing two diagnostic tests based on multivariate proportional hazard models. With correct model specification, the semiparametric method can allow survival outcome measurements in the presence of censoring.

Calculating sample size is an important issue when performing group sequential trials. We illustrated an example of sample size determination based on binormal and biexponential distribution assumptions. If a pilot study is available, sample sizes can be empirically determined from our results. Otherwise, conservative sample size can be determined with the knowledge of the AUCs and their correlation from equation (2).

Both nonparametric AUC and partial AUC estimators, as special cases of the WGJJ statistics, were illustrated in sequential lung cancer trials for comparing the staging accuracy of nonsmall-cell cancer. Accurate staging diagnosis could guide surgery to help prolong patients' life or even cure the patients at early stages of lung cancer, one of the top killer diseases. Therefore, designing comparative trials for the accuracy of scan-imaging techniques is rather important at this point, and our nonparametric estimator provides a robust and efficient way to sequentially compare techniques in their staging ability. Other examples may be found on `clinicaltrials.gov`, which is a website developed by the National Institutes of Health and the Food and Drug Administration to provide information for federally and privately supported clinical trials. One ongoing trial is titled "Comparison of Cardiac Computed Tomographic Angiography (CTA) to Tc-99m Single Photon Emission Computed Tomography (SPECT)" with the purpose of comparing the accuracy of CT and SPECT. In this trial, diagnostic tests are expensive, and the patients' disease status are obtained before tests. These are good logistics for implementing our sequential methods in this trial.

As a final note, in diagnostic imaging trials the results are usually immediately available. The patients' disease status are obtained before tests or shortly after tests. These provide good logistics for conducting sequential diagnostic trials. However, there are some diagnostic studies in which sequential designs are not applicable. In the studies that obtain marker information from patients' medical files after patients are identified, the logistics of sequential tests appear complicated. In some biomarker studies it may take a long time to verify true disease status, and the sequential designs would be inapplicable. Also, one may want to look at Youden index or find "optimal" thresholds in addition to comparing diagnostic tests. How to solve these issues would be future research topics.

8. Supplementary Materials

The Web Appendix is available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors would like to thank an associate editor and two referees for their constructive comments and suggestions. This work is supported by NIH grant R01EB005829. This report presents the findings and conclusions of the authors. It does not necessarily represent those of VA HSR&D Service.

References

Cai, T. and Pepe, M. S. (2002). Semi-parametric ROC analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association* **97,** 1099–1107.

Dorfman, D. D. and Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals–rating method data. *Journal of Mathematical Psychology* **6,** 487–496.

EaSt. Cytel Software Corporation. Cambridge, Massachusetts.

Emerson, S., Kittelson, J., and Gillen, D. (2007a). Frequentist evaluation of group sequential clinical trial designs. *Statistics in Medicine* **26,** 5047–5080.

Emerson, S., Kittelson, J., and Gillen, D. (2007b). On the use of stochastic curtailment in group sequential clinical trials. *Statistics in Medicine* **26,** 1431–1449.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association* **55,** 698–707.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36.

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

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* New York: Chapman and Hall.

Kim, K. and DeMets, D. L. (1992). Sample size determination for group sequential clinical trials with immediate response. *Statistics in Medicine* **11,** 1391–1399.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70,** 659–663.

Lardinois, D., Walter Weder, W., Hany, T. F., Kamel, E. M., Korom, S., Seifert, B., von Schulthess, G. K., and Steinert, H. C. (2003). Staging of non-small-cell lung cancer with integrated positron-emission tomography and computed tomography. *New England Journal of Medicine* **348,** 2500–2507.

Mazumdar, M. (2004). Group sequential design for comparative diagnostic accuracy studies: Implications and guidelines for practitioners. *Medical Decision Making* **24,** 525–533.

Mazumdar, M. and Liu, A. (2003). Group sequential design for diagnostic accuracy studies. *Statistics in Medicine* **22,** 727–739.

Molodianovitch, K., Faraggi, D., and Reiser, B. (2006). Comparing the areas under two correlated ROC curves: Parametric and non-parametric approaches. *Biometrical Journal* **48,** 745–757.

PEST. Medical and Pharmaceutical Statistics Research Unit. Reading, U.K.: University of Reading.

S+SeqTrial. Insightful Corporation. Seattle, Washington.

Sanchez-Marin, F. J. and Padilla-Medina, J. A. (2006). Alternative performance index to analyze receiver operating characteristic data under the exponential assumption. *Journal of Electronic Imaging* **15,** 1–9.

Silvestri, G. A., Tanoue, L. T., Margolis, M. L., Barker, J., and Detterbeck, F. (2003). Noninvasive staging of non-small cell lung cancer. *Chest* **123,** 137S–146S.

Van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* New York: Springer-Verlag.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84,** 1065–1073.

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

Zhou, X. H., McClish, D. K., and Obuchowski, N. A. (2002). *Statistical Methods in Diagnostic Medicine.* New York: Wiley.

Zhou, X. H., Li, S. M., and Gatsonis, C. A. (2008). Wilcoxon-based group sequential designs for comparison of areas under two correlated ROC curves. *Statistics in Medicine* **27,** 213–223.

Received June 2007. *Revised October* 2007.
Accepted October 2007.