Published in final edited form as:

Stat Med. 2008 May 10; 27(10): 1667–1678. doi:10.1002/sim.3203.

Nonparametric sequential evaluation of diagnostic biomarkers

Aiyi Liu^{*,†}, Chengqing Wu, and Enrique F. Schisterman

Biometry and Mathematical Statistics Branch, National Institute of Child Health and Human Development, Department of Health and Human Services, 6100 Executive Blvd., Rockville, MD, U.S.A.

SUMMARY

We consider evaluation and comparison of the diagnostic accuracy of biomarkers with continuous test outcomes, possibly correlated due to repeated measurements. We develop nonparametric group sequential testing procedures to evaluate and compare the area of biomarkers under their receiver operating characteristic curves, with either independent or paired test outcomes. These procedures rely on the construction of a two-dimensional statistic of Whitehead so that design methods based on Brownian motion can be applied.

Keywords

biomarkers; correlated measurements; diagnostic tests; interim looks and early stopping; correlated U-statistics

1. INTRODUCTION

In the area of diagnostic medicine, research frequently focuses on assessing how well a diagnostic biomarker performs in distinguishing subjects with certain conditions (diseased subjects or cases) from those without such conditions (nondiseased subjects or controls). When the outcomes of the biomarker are measured on a continuous scale, evaluation of the biomarker's diagnostic accuracy is often based on analysis of its receiver operating characteristic curve, which is the plot of the marker's true positive fraction (TPF), i.e. the probability that it correctly identifies a diseased subject, *versus* its false positive fraction (FPF), i.e. the probability that it incorrectly identifies a nondiseased subject as being diseased. Each point on the curve corresponds to a specific cutpoint value based on which the decision is made; if the biomarker's outcome exceeds the cutpoint value, then the subject is classified as a case, assuming that the biomarker's values from cases tend to be larger than from controls.

Often the performance of a biomarker is measured by the area under its receiver operating characteristic (ROC) curve, especially when evaluation of TPF and FPF at only certain cutpoint values is inadequate. If a biomarker has larger area under its ROC curve than the other biomarker, then we say that the former has higher diagnostic accuracy than the latter. Both parametric and nonparametric methods have been proposed for estimation and hypotheses testing regarding the area under the ROC curves. See, e.g. [1-7].

In practice, assessing the usefulness of a diagnostic biomarker can be very costly. For example, measuring the amount of F2 isoprostanes, an important biomarker for oxidative stress, costs \$130 each assay. Certain diagnostic procedures such as taking tumor biopsy for cancer diagnosis can be very painful to the patients and hence should be stopped for the evaluation if early evidence shows the procedure to be inefficient. Taking into consideration of these cost and ethical concerns, the need for interim analysis and early stopping is apparent in the design and analysis of studies to evaluate newly developed diagnostic biomarkers. Mazumdar and Liu [8] presented group sequential methods for comparison of two ROC areas based on likelihood theory, assuming that the values of the biomarkers are normally distributed. Their methods can be extended straightforwardly, using Whitehead's [9] unified theory, to other parametric settings where likelihood function is workable.

In this paper we present nonparametric sequential procedures that are robust against distributional assumptions for the evaluation and comparison of the area under ROC curves. Extending the unified theory of Whitehead [9], which is likelihood based and parametric in essence, to a nonparametric setting suitable for ROC curve analysis, we construct the two statistics, Z and V , that allow the use of Brownian motion theory to derive the stopping boundaries or critical values of a specific sequential testing procedure, such as those in [10-13] that are developed in the context of clinical trials. In Section 2 we briefly describe Whitehead's [9] methods in sequential testing. Allowing each subject to have multiple measurements, we obtain in Section 3 Whitehead's (Z, V) statistics based on the Mann-Whitney statistic for the area under a single ROC curve. In Section 4 we derive the (Z, V) statistics for testing equal area of two or more ROC curves, either correlated (paired) or independent. Section 5 presents some simulation results to demonstrate the efficiency of the proposed distribution-free approach, followed by illustration of the proposed sequential method using data from a study to investigate the association of several biomarkers with cardiovascular disease (CVD).

2. WHITEHEAD'S UNIFIED APPROACH FOR THE SEQUENTIAL TESTING OF A HYPOTHESIS

Suppose, without loss of generality, we wish to test the null hypothesis concerning a parameter θ , $H_0: \theta=0$, with a two-sided alternative $\theta \neq 0$. Whitehead's [9] unified approach involves constructing of two statistics, expressed throughout as Z and V in accordance with Whitehead's notation. It requires that at least asymptotically $(Z-V\theta)/\sqrt{V}$ follow the standard normal distribution, i.e. $Z \sim N(\theta V, V)$. The V -statistic, or the estimated information, increases as the sample size becomes larger. The increments in the Z -statistic, corresponding to the increase of the V -statistic, are independent and normally distributed, so that Z resembles a Brownian motion with V being viewed as the time and θ the drift parameter; see also [14]. With fixed sample size, the hypothesis H_0 is rejected if $|Z/\sqrt{V}| > \Phi^{-1}(1-\alpha/2)$, where Φ is the standard normal distribution function and α is the nominal type I error rate. A K -stage group sequential test rejects H_0 at the first stage if $|Z_1/\sqrt{V_1}| > c_1$ and at stage $1 < k \leq K$ if

$$\{|Z_i/\sqrt{V_i}| \leq c_i, i=1, \dots, k-1, |Z_k/\sqrt{V_k}| > c_k\}$$

Here $Z_1, \dots, Z_K$ and $V_1 < \dots < V_K$ are the values of Z and V calculated at the k th stage (based on all available data), and $\{c_k\}$ are the critical values chosen so to satisfy the error requirements. Popular methods to choose c_k include, among others, Pocock's [10], O'Brien and Fleming's [11], Lan and DeMets' [12], and Whitehead and Stratton's [13] methods. Both Pocock's and O'Brien and Fleming's designs are equally spaced, but i.e. $V_1 = V_k - V_{k-1}$, $2 \leq k \leq K$, but the former sets $c_1 = \dots = c_K$ and the latter sets $c_1 \sqrt{V_1} = \dots = c_K \sqrt{V_K}$. Tables of these critical values for various α and K are available; see e.g. Jennison and Turnbull [15]. The method of Lan and DeMets is

much more flexible, allowing the design to be unequally spaced and the critical values to be determined by an error spending function, which specifies the Type I error portion 'to be spent' at each stage of the interim analysis. Whitehead and Stratton's method uses a continuous triangular stopping boundary to determine the critical values.

Whitehead [9] proposed a unified likelihood-based theory to construct the test statistic Z and V . Let $l(\theta)$ be the log-likelihood of θ based on the data available. Then

$$Z = l'(\theta)|_{\theta=0}, \quad V = l''(\theta)|_{\theta=0}$$

assuming that the likelihood function can be differentiated at least twice with respect to θ . Whitehead [9] also presented an alternative approach: Let $\hat{\theta}$ be an efficient estimate (e.g. maximum likelihood estimate) of θ , then V is the reciprocal of its estimated variance, and $Z = \hat{\theta}V$. In the calculation of Z and V , nuisance parameters, if any, are replaced by their maximum likelihood estimates obtained from maximizing the likelihood function with $\theta=0$.

The fact that Z and V resembles a Brownian motion is ensured by the large sample theory of the likelihood-based inference. However, as Whitehead [9] pointed out, if a nonparametric estimate of θ (e.g. the Kaplan-Meier estimate of survival rate) is used, then special treatment is needed to warrant such resemblance. The nonparametric estimation and testing of the ROC curve areas is another example of such a case, as demonstrated in the following two sections.

3. NONPARAMETRIC ESTIMATION AND HYPOTHESIS TESTING OF A SINGLE ROC AREA

Let X be the biomarker value from a diseased subject with distribution function F and Y be the biomarker value from a nondiseased subject with distribution function G . Let us express $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$. At a specific cutpoint c , the biomarker's true positive fraction is $\text{TPF}(c) = P(X > c) = \bar{F}(c)$ and its false positive fraction $\text{FPF}(c) = P(Y > c) = \bar{G}(c)$. Plotting $\bar{F}(c)$ versus $\bar{G}(c)$ over all possible values of c shows the ROC curve of the biomarker. The area A under the receiver operating characteristic curve is thus

$$A = \left| \int_{-\infty}^{\infty} \bar{F}(c) d\bar{G}(c) \right| = \int_{-\infty}^{\infty} \bar{F}(c) dG(c) = \int_0^1 \bar{F}(G^{-1}(p)) dp$$

where G^{-1} denotes the quartile function of G , i.e. $G(G^{-1}(p)) = p$ for all $0 \leq p \leq 1$. The fact that $A = \text{Pr}(X > Y)$ was first discovered by Bamber [1] and forms the basis for nonparametric inference on the area under a ROC curve.

Now suppose that m diseased and n nondiseased subjects are randomly selected from the population, and the biomarker is measured r times for each of the diseased subject and s times for each of the nondiseased subject, and thus yielding data $\mathbf{X}_i = (X_{i1}, \dots, X_{ir})$, $i = 1, \dots, m$, and $\mathbf{Y}_j = (Y_{j1}, \dots, Y_{js})$, $j = 1, \dots, n$. (Here, to simplify lengthy notations, we assume equal number of repeated measurements per subject. The Brownian motion property derived below holds if the number of repeated measurements differs.) Such data sets are frequently encountered when the concentration of a biomarker (e.g. F2 isoprostanes) in blood samples is measured repeatedly a few times to investigate the disease-exposure association. Let I be the 0-1 indicator function, so that $I(x > y) = 1$ if $x > y$ and 0 if otherwise, and define, for a vector of order r , $\mathbf{a} = (a_1, \dots, a_r)$, and, a vector of order s , $\mathbf{b} = (b_1, \dots, b_s)$,

$$\psi(\mathbf{a}, \mathbf{b}) = \frac{1}{rs} \sum_{h=1}^r \sum_{l=1}^s I(a_h > b_l)$$

Then with these repeated measurements the area under the ROC curve can be estimated by

$$\hat{A} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \left\{ \frac{1}{rs} \sum_{t=1}^r \sum_{k=1}^s I(X_{it} > Y_{jk}) \right\} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \psi(\mathbf{X}_i, \mathbf{Y}_j) \quad (1)$$

Following [16, pp. 362-371], the asymptotic distribution of $\hat{A}$ can be obtained using Hajek's method of projection. Let us express

$$\psi_{10}(\mathbf{X}_i) = \frac{1}{r} \sum_{h=1}^r [G(X_{ih}) - A] \quad \text{and} \quad \psi_{01}(\mathbf{Y}_j) = \frac{1}{s} \sum_{l=1}^s [\bar{F}(Y_{jl}) - A] \quad (2)$$

then, as $m, n \rightarrow \infty$,

$$\hat{A} = A + \frac{1}{m} \sum_{i=1}^m \psi_{10}(\mathbf{X}_i) + \frac{1}{n} \sum_{j=1}^n \psi_{01}(\mathbf{Y}_j) + \frac{1}{mn} O_p(1) \quad (3)$$

where $O_p(1)$ denotes a sequence that is bounded in probability [17]. Therefore, if $m/n \rightarrow \lambda$, a nonzero constant, as m and $n \rightarrow \infty$, then

$$\sqrt{m}(\hat{A} - A) \xrightarrow{d} N(0, \sigma^2) \quad (4)$$

where $\sigma^2 = \sigma_{10}^2 + \lambda \sigma_{01}^2$ with

$$\sigma_{10}^2 = \text{Var}(\psi_{10}(\mathbf{X})) = \frac{1}{r} E G^2(X_{11}) + \frac{2}{r^2} \sum_{1 \leq h_1 < h_2 \leq r} E G(X_{1h_1}) G(X_{1h_2}) - A^2 \quad (5)$$

$$\sigma_{01}^2 = \text{Var}(\psi_{01}(\mathbf{Y})) = \frac{1}{s} E \bar{F}^2(Y_{11}) + \frac{2}{s^2} \sum_{1 \leq l_1 < l_2 \leq s} E \bar{F}(Y_{1l_1}) \bar{F}(Y_{1l_2}) - A^2 \quad (6)$$

Note that, with $1 \leq l_1 \neq l_2 \leq s$,

$$E G^2(X_{11}) = P\{Y_{1l_1} < X_{11}, Y_{2l_2} < X_{11}\}$$

the probability that two nondiseased subjects both have lower biomarker values than a diseased subject. This quantity can be estimated empirically as

$$\frac{2}{mn(n-1)s^2} \sum_{i=1}^m \sum_{h=1}^r \sum_{j_1 < j_2}^s \sum_{l_1=1}^s \sum_{l_2=1}^s I(Y_{j_1 l_1} < X_{ik}) I(Y_{j_2 l_2} < X_{ik})$$

Similarly, for $1 \leq h_1 \neq h_2 \leq r$,

$$EG(X_{1h_1})G(X_{1h_2}) = P\{Y_{j_1 l_1} < X_{1h_1}, Y_{j_2 l_2} < X_{1h_2}\}$$

where $1 \leq j_1 \neq j_2 \leq n$ and $1 \leq l_1, l_2 \leq s$. The empirical estimation of this quantity is

$$\frac{2}{mn(n-1)s^2} \sum_{i=1}^m \sum_{j_1 < j_2}^s \sum_{l_1=1}^s \sum_{l_2=1}^s I(Y_{j_1 l_1} < X_{ik_1}) I(Y_{j_2 l_2} < X_{ik_2})$$

Thus, the variance σ_{10}^2 can be empirically estimated as

$$\widehat{\sigma_{10}^2} = \frac{2}{mn(n-1)} \sum_{i=1}^m \sum_{1 \leq j_1 < j_2 \leq n} \psi(\mathbf{X}_i, \mathbf{Y}_{j_1}) \psi(\mathbf{X}_i, \mathbf{Y}_{j_2}) - \widehat{A}^2 \quad (7)$$

and similarly the variance σ_{01}^2 can be estimated by

$$\widehat{\sigma_{01}^2} = \frac{2}{mn(m-1)} \sum_{j=1}^n \sum_{1 \leq i_1 < i_2 \leq m} \psi(\mathbf{X}_{i_1}, \mathbf{Y}_j) \psi(\mathbf{X}_{i_2}, \mathbf{Y}_j) - \widehat{A}^2 \quad (8)$$

yielding an estimate of the asymptotic variance of $\hat{A}$ as $\widehat{\sigma_{10}^2}/m + \widehat{\sigma_{01}^2}/n$.

Therefore, to test the hypothesis $H_0 : A = A_0$ versus $H_1 : A > A_0$, we apply a specific sequential procedure to the test statistics

$$V = \frac{mn}{n\widehat{\sigma_{10}^2} + m\widehat{\sigma_{01}^2}}, \quad Z = V(\widehat{A} - A_0) \quad (9)$$

That Z behaves like a Brownian motion with time V and drift $A - A_0$ and has independent normally distributed increment is readily seen from (3).

4. COMPARISON OF THE AREA UNDER CORRELATED ROC CURVES

Studies in diagnostic medicine often involving testing the null hypothesis that several markers have the same diagnostic accuracy, i.e. the areas under the ROC curves of these markers are equal. In most cases these markers are applied on the same subject and thus generate correlated test outcomes ('paired' if there are only two markers). This introduces another layer of correlation, the inter-marker correlation, in addition to the intra-marker correlation among repeated test outcomes from a single marker applied to the same subject.

Suppose there are T diagnostic markers under consideration. For a specific marker t , $t=1, \dots, T$, the r repeated test outcomes from the i th diseased subject are $\mathbf{X}_i^t = (X_{i1}^t, \dots, X_{ir}^t)$, and the s repeated test outcomes from the j th nondiseased subject are $\mathbf{Y}_j^t = (Y_{j1}^t, \dots, Y_{js}^t)$. Let A^t be the area under its ROC curve and $\hat{A}^t$ be its estimate given by (1), that is,

$mn\hat{A}^t = \sum_{i=1}^m \sum_{j=1}^n \psi(\mathbf{X}_i^t, \mathbf{Y}_j^t)$, and put $\mathbf{A} = (A^1, \dots, A^T)$ and $\hat{\mathbf{A}} = (\hat{A}^1, \dots, \hat{A}^T)$. Then as in the case of a single ROC curve, $\hat{\mathbf{A}}$ can be expressed as

$$\hat{\mathbf{A}} = \mathbf{A} + \frac{1}{m} \sum_{i=1}^m \Psi_{10}(\mathbf{X}_i) + \frac{1}{n} \sum_{j=1}^n \Psi_{01}(\mathbf{Y}_j) + \frac{1}{mn} O_p(1) \quad (10)$$

Where $\Psi_{10}(\mathbf{X}_i) = (\psi_{10}(\mathbf{X}_i^1), \dots, \psi_{10}(\mathbf{X}_i^T))'$ and $\Psi_{01}(\mathbf{Y}_j) = (\psi_{01}(\mathbf{Y}_j^1), \dots, \psi_{01}(\mathbf{Y}_j^T))'$.

Denote, respectively, by Σ_{10} and Σ_{01} the covariance matrices of $\Psi_{10}(\mathbf{X}_i)$ and $\Psi_{01}(\mathbf{Y}_j)$. Then the (t_1, t_2) th element of Σ_{10} is

$$\sigma_{10}^{t_1, t_2} = \frac{1}{r^2} \sum_{h_1=1}^r \sum_{h_2=1}^r P(Y_{j_1 h_1}^{t_1} < X_{i k_1}^{t_1}, Y_{j_2 h_2}^{t_2} < X_{i k_2}^{t_2}) - A^{t_1} A^{t_2}, \quad j_1 \neq j_2$$

and the (t_1, t_2) th element of Σ_{01} is

$$\sigma_{01}^{t_1, t_2} = \frac{1}{s^2} \sum_{h_1=1}^s \sum_{h_2=1}^s P(Y_{j_1 h_1}^{t_1} < X_{i_1 h_1}^{t_1}, Y_{j_2 h_2}^{t_2} < X_{i_2 h_2}^{t_2}) - A^{t_1} A^{t_2}, \quad i_1 \neq i_2$$

The two matrices can be estimated by $\widehat{\Sigma}_{10}$ and $\widehat{\Sigma}_{01}$ with their (t_1, t_2) th elements given, respectively, by

$$\widehat{\sigma}_{10}^{t_1, t_2} = \frac{2}{mn(n-1)} \sum_{i=1}^m \sum_{1 \leq j_1 < j_2 \leq n} \psi(\mathbf{X}_i^{t_1}, \mathbf{Y}_{j_1}^{t_1}) \psi(\mathbf{X}_i^{t_2}, \mathbf{Y}_{j_2}^{t_2}) - \widehat{A}^{t_1} \widehat{A}^{t_2}$$

$$\widehat{\sigma}_{01}^{t_1, t_2} = \frac{2}{mn(m-1)} \sum_{j=1}^n \sum_{1 \leq i_1 < i_2 \leq m} \psi(\mathbf{X}_{i_1}^{t_1}, \mathbf{Y}_j^{t_1}) \psi(\mathbf{X}_{i_2}^{t_2}, \mathbf{Y}_j^{t_2}) - \widehat{A}^{t_1} \widehat{A}^{t_2}$$

Thus, an estimate of the asymptotic covariance matrix of $\hat{\mathbf{A}}$ is

$$S = \frac{1}{m} \widehat{\Sigma}_{10} + \frac{1}{n} \widehat{\Sigma}_{01} \quad (11)$$

4.1. Comparison of two ROC curves

Suppose that investigators wish to know whether one marker has higher accuracy than the other in identifying diseased and nondiseased subjects. In this case, the hypotheses of interest are $H_0: A^1 - A^2 = 0$ versus $H_1: A^1 - A^2 \neq 0$. The covariance matrix of $\mathbf{A} (A^1, A^2)'$ can be estimated as in (11). Let us express $L = (1, -1)'$ the vector of contrast, then the test statistics (Z, V) are $V = 1/$

$\{L'SL\}$ and $Z=VL'\hat{\mathbf{A}}$. With the decomposition in (10), Z asymptotically has independent increment and behaves like a Brownian motion with time V and drift A^1-A^2 .

4.2. Comparison of more than two ROC curves

When comparison involves more than two biomarkers, the Brownian motion paradigm for testing of a single parameter does not apply. Instead, we can use sequential χ^2 -test to test the null hypothesis that the markers have equal area under the ROC curves. This subsection compares the discriminating ability of more than two biomarkers. The problem of interest is whether these markers have the same areas under their ROC curves. Let L be a $T \times (T-1)$ matrix of contrast with the upper $(T-1)$ rows form an identity matrix, and the elements in the last row are all equal to -1. From the decomposition in (10), the statistic $(L'SL)^{-1}L'\mathbf{A}^\wedge$ asymptotically follows a multivariate normal distribution with independent increment. Thus, asymptotically the statistic

$$\Omega = \widehat{\mathbf{A}}'L(L'SL)^{-1}L'\widehat{\mathbf{A}}$$

has a χ^2 distribution with $T-1$ degrees of freedom, and noncentrality parameter $\mathbf{A}'L(L'SL)^{-1}L'\mathbf{A}$, which reduces to a central χ^2 distribution under the null hypothesis $H_0: \mathbf{L}'\mathbf{A}=0$.

The group sequential χ^2 test is then carried out as follows. At the k th stage, the χ^2 statistic

$$\Omega_k = \widehat{\mathbf{A}}'_k L(S_k L)^{-1} L' \widehat{\mathbf{A}}_k$$

is computed, where $\mathbf{A}_k^\wedge$ and S_k are the corresponding estimates based on all available data collected up to stage k . The null hypothesis is rejected at the first stage if $\Omega_1 > c_1$ and at stage $1 < k \leq K$ if

$$\{\Omega_i \leq c_i, i=1, \dots, k-1, \Omega_k > c_k\}$$

The critical values c_k are chosen to satisfy error requirements. To this end, we assume that the allocation ratio $\lambda = n_k/m_k$ remains constant. Put $\Lambda = \Sigma_{10} + \lambda \Sigma_{01}$. Since $m_k S_k$ is a consistent estimate of Λ , $\Omega_k, k=1, \dots, K$ are asymptotically equivalent in distribution to $\|W_k\|^2$, where

$$W_k = \sqrt{m_k} (L' \Lambda L)^{-1/2} L' \widehat{\mathbf{A}}_k$$

which asymptotically has a multivariate normal distribution with mean vector

$\sqrt{m_k} (L' \Lambda^{-1/2} L) L' \widehat{\mathbf{A}}$ and covariance matrix I_{T-1} , where, for a positive-definite matrix S of order p , $S^{-1/2}$ is a symmetric matrix such that $S^{-1/2} S S^{-1/2} = I_p$.

From (10), the statistics W_k are asymptotically jointly normal. Moreover, the sequence $W_k, k=1, \dots, K$, is Markov. Under $H_0: \mathbf{L}'\mathbf{A}=0$, the conditional distribution of W_{k+1} given W_k is

$$W_{k+1}|W_k \sim N\left(\frac{m_k}{m_{k+1}} W_k, \frac{g_{k+1}+1}{m_{k+1}} I_{T-1}\right), \quad k=1, \dots, K-1 \quad (12)$$

Here $g_1=m_1$ and for $k=2, \dots, K$, $g_k=m_k-m_{k-1}$, the group size of diseased subjects at stage k . Then, under $H_0: L'A=0$, the sequence of random variables Ω_k , $k=1, \dots, K$ is asymptotically Markov with $\Omega_1|\chi_{T-1}^2$ and for $k=1, \dots, K-1$,

$$\Omega_{k+1}|\Omega_k \left(\frac{g_{k+1}}{m_{k+1}} \right) \chi_{T-1, m_k \Omega_k / g_{k+1}}^2 \quad (13)$$

where $\chi_{p,\lambda}^2(\cdot)$ is the noncentral χ^2 distribution with p degrees of freedom and noncentrality parameter λ .

Let M be the (random) number of analyses performed at stopping, and let us express $q_{p,\lambda}(\cdot)$ as the density function of the noncentral χ^2 distribution with p degrees of freedom and noncentrality parameter λ . Then the joint density of (M, Ω_M) at $M=k$ and $\Omega_M=\omega$ is

$$f_k(\omega) = q_T - 1, 0(\omega) h_k(\omega)$$

with

$$h_1(\omega) = 1 \quad \text{and for } k > 1$$

$$h_k(\omega) = \int_0^{c_{k-1}} h_{k-1}(u) \frac{m_k}{g_k} \chi_{T-1, m_{k-1} \omega / g_k}^2 \left(\frac{m_k}{g_k} u \right) du$$

In particular, for $k=2$,

$$f_2(\omega) = q_T - 1, 0(\omega) \chi_{T-1, m_1 \omega / g_2}^2 \left(\frac{c_1 g_2}{m_2} \right)$$

These equations can then be used to numerically compute the error rate or critical values c_k , similar to the exact calculation approach by Jennison and Turnbull [18] for the case of identically distributed observations. As an example, the type I error for the simple case of two-stage test ($K=2$) is

$$\begin{aligned} P_{H_0} \{ \Omega_1 > c_1 \} + P_{H_0} \{ \Omega_1 \leq c_1, \Omega_2 > c_2 \} \\ &= \int_{c_1}^{\infty} f_1(\omega) d\omega + \int_{c_2}^{\infty} f_2(\omega) d\omega \\ &= 1 - \chi_{T-1,0}^2(c_1) + \int_{c_2}^{\infty} q_T - 1, 0(\omega) \chi_{T-1, m_1 \omega / g_2}^2 \left(\frac{c_1 g_2}{m_2} \right) d\omega \end{aligned}$$

5. SIMULATION AND EXAMPLE

In this section we first present some simulation results and then apply the sequential procedures developed in the previous section to compare the performance of two biomarkers in association with CVD.

5.1. A simulation study

To investigate the performance in controlling the type I error rate of the proposed nonparametric group sequential procedure, simulation studies were conducted to compare the area under two curves. For simplicity, we set $r=s=2$, that is, two repeated measurements are collected from each subject. We used a O'Brien-Fleming-type boundary with the number of analyses $K=2,3,4,5$, respectively. For each design the type I error rate α is set to be 0.05, and the critical values for the standardized test statistics are given in [15]. For example, the critical values for $K=2$ are 2.796 for the first stage and 1.977 for the second stage. At each interim analysis, there are 20 diseased subjects and 20 nondiseased subjects. The type I error rate was estimated from 10 000 simulations.

In one simulation study, we assume that the underlying measurements from the two tests are continuous and follow a bivariate normal distribution with $N(\mu_x, \Sigma_x)$ for diseased subjects and $N(\mu_y, \Sigma_y)$ for nondiseased subjects, where $\mu_x=(\mu_{x1}, \mu_{x2})$, $\mu_y=(0,0)$, and

$$\Sigma_x = \begin{pmatrix} 1 & \rho_x \\ \rho_x & 1 \end{pmatrix}, \quad \Sigma_y = \begin{pmatrix} 1 & \rho_y \\ \rho_y & 1 \end{pmatrix}$$

Simulation was conducted for each selected value of the triplet $(A_1 = A_2, \rho_x, \rho_y)$. The results are shown in Table I.

In another simulation study, the underlying measurements from the two tests are assumed to follow bivariate exponential distributions of Marshall and Olkin [19,20]. According to Marshall and Olkin's definition, two random variables W_1 and W_2 follow a bivariate exponential distribution if their joint survival function is $\exp(-\lambda_1 w_1 - \lambda_2 w_2 - \lambda_{12} \max(w_1, w_2))$, which will be denoted hereafter by $\text{BiExp}(\lambda_1, \lambda_2, \lambda_{12})$. The correlation coefficient between W_1 and W_2 is thus $\lambda_{12}/(\lambda_1 + \lambda_2 + \lambda_{12})$. Random sample of this bivariate exponential distribution can be drawn using the fact that if U_1, U_2 and U_{12} are mutually independent random variables following univariate exponential distribution with parameter λ_1, λ_2 and λ_{12} , respectively, then $W_1 = \min(U_1, U_{12})$ and $W_2 = \min(U_2, U_{12})$ follow a bivariate exponential distribution $\text{BiExp}(\lambda_1, \lambda_2, \lambda_{12})$.

In our simulation, we assume that outcomes from the two tests follow a bivariate exponential distribution with $\text{BiExp}(\lambda_{x1}, \lambda_{x2}, \lambda_{x12})$ for diseased subjects and $\text{BiExp}(\lambda_{y1}, \lambda_{y2}, \lambda_{y12})$ for nondiseased subjects. To ease the simulation, we assume $\lambda_{x1}=\lambda_{x2}$ and $\lambda_{y1}=\lambda_{y2}$. The correlation coefficients are set to be equal to ρ for the two distributions. For simplicity we fix $\lambda_{y1}+\lambda_{y12}=\lambda_{y2}+\lambda_{y12}=10$. Table II summarizes the simulation results.

Under bivariate normal assumption, the type I error rates in most cases are quite close to the nominal level of 0.05. Although within certain level of tolerance, the results in Table II for bivariate exponential distributions are slightly worse than those in Table I, reflecting the effect on approximation of relatively small sample size and the skewness of the exponential distributions. The results are expected to be improved if groups sizes are larger.

5.2. An example

CVD is the leading cause of death in men and women in the United States. The accurate diagnosis of an individuals' cardiac death risk is crucial for the appropriate risk management of patients. To find and evaluate the risk factors, 5053 consecutive outpatients free of known cardiomyopathy or valvular disease and who underwent rest Tl-201/stress Tc-99m sestamibi dual-isotope myocardial perfusion SPECT with exercise (60 per cent) or pharmacological stress (40 per cent) between January 1991 and December 1996 were enrolled into the study.

Cases were defined as individuals experiencing a coronary event including myocardial infarction and death. Controls were free of disease at the end of the study.

To exemplify the use of a group sequential procedure, we consider comparing the diagnostic accuracy of two risk factors, 'PRETEST', which is a linear combination of cardiovascular risk factors based on the Framingham equations and includes cholesterol levels, race, gender, smoking status and blood pressure, and 'POSTECG', the electro cardiogram that measures the sympathetic cardiac function.

For illustration, we sequentially randomly selected 100 cases and 100 controls according to their entry of the study. We assume that a five-stage group sequential test with O'Brien and Fleming type of stopping boundaries is used. The critical values for the standardized test statistic are, respectively, 4.5616, 3.2255, 2.6336, 2.2808, and 2.0400. Suppose the null hypothesis that 'PRETEST' and 'POSTECG' have equal ROC area is tested at each interim analysis after an additional number of 20 cases and 20 controls are enrolled between the previous analysis and the current analysis. For the first three interim analyses, the ROC area is found to be 0.6419, 0.6315, 0.5439 for 'POSTECG' and 0.6394, 0.6558, 0.6292 for 'PRETEST'. The standardized test statistic is found to be 0.0322, 0.5256, and 2.8705, respectively. Hence, the null hypothesis is rejected at the third interim analysis, concluding that 'POSTECG' has higher diagnostic accuracy than 'PRETEST'.

6. DISCUSSION

In this paper we have proposed nonparametric group sequential testing procedures for comparing the area under several ROC curves. The key development is that the nonparametric test statistics resemble a Brownian motion when sample sizes are relatively large and hence standard methods can be applied to carry out the tests.

When only the area under a particular part of the ROC curve is of interest, these testing procedures can be readily extended to test the equality of several partial areas. This is because the empirical estimate of the partial area also behaves asymptotically like a Brownian motion; see [21].

The proposed group sequential procedures have been shown to control the type I error rate at a reasonable level. In planning a comparative diagnostic study, sample sizes need to be estimated. To do so, we need to first specify the type of stopping boundary to be used (e.g. via an error spending function) and the number of analyses to be conducted. Then using a standard software for group sequential test designs, the maximum information $V_{\max}$ can be calculated, provided that the overall significance level, the difference between two ROC areas, and the power to detect such difference are specified. With a known allocation ratio λ , we can then find the sample sizes m and n using the following equation:

$$V = \frac{mn}{n\sigma_{10}^2 + m\sigma_{01}^2}$$

This of course requires the nuisance parameters (e.g. $EG^2(X_{11})$) in the variance function of σ^2 to be pre-specified.

From (4) it follows (e.g. Bickel and Doksum [22]) that for a well-smoothed function $H(x)$ with first derivative $H_x = dH(x)/dx$:

$$\sqrt{m}(H(\widehat{A}) - H(A)) \xrightarrow{d} N(0, \sigma^2 H_A^2)$$

A properly chosen function H can improve the normal approximation and stabilize the variance (in that $\sigma H_A = \text{constant}$). The complication, however, is that the variance σ^2 depends not only on A but also on some other parameters (e.g. $EG^2(X_{11})$); see Equations (5) and (6). It remains to be seen whether such a variance stabilizing function can be found, or in a broad sense, whether the popular variance stabilization method can be generalized to cases in which nuisance parameters are present.

ACKNOWLEDGEMENTS

This research was supported by the Intramural Research Program of the National Institute of Child Health and Human Development, National Institutes of Health. The opinions expressed in the article are not necessarily of the National Institutes of Health. The authors thank a referee and an associate editor for their helpful comments and suggestions.

Contract/grant sponsor: National Institute of Child Health and Human Development, National Institutes of Health

REFERENCES

1. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristics graph. *Journal of Mathematical Psychology* 1975;12:387–415.
2. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–845. [PubMed: 3203132]
3. Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Statistics in Medicine* 2002;21/20:3093–3106. [PubMed: 12369084]
4. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;148:839–843. [PubMed: 6878708]
5. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics* 1997;53:567–578. [PubMed: 9192452]
6. Zou KH, Hall WJ. Semiparametric and parametric transformation models for comparing diagnostic markers with paired design. *Journal of Applied Statistics* 2002;29:803–816.
7. Zou KH, Hall WJ, Shapiro DE. Smooth nonparametric receiver operating characteristic curves for continuous diagnostic tests. *Statistics in Medicine* 1997;16:2143–2156. [PubMed: 9330425]
8. Mazumdar M, Liu A. Group sequential design for comparative diagnostic accuracy studies. *Statistics in Medicine* 2003;22:727–739. [PubMed: 12587102]
9. Whitehead J. A unified theory for sequential clinical trials. *Statistics in Medicine* 1999;18:2271–2286. [PubMed: 10474138]
10. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977;64:191–199.
11. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1977;35:549–556.
12. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983;70:659–663.
13. Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions. *Biometrics* 1983;39:227–236. [PubMed: 6871351]
14. Lan KKG, Zucker D. Sequential monitoring of clinical trials: the role of information and Brownian motion. *Statistics in Medicine* 1993;12:753–765. [PubMed: 8516592]
15. Jennison, C.; Turnbull, BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC Press; New York: 2000.
16. Lehmann, EL.; D'Abrera, HJM. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day; San Francisco: 1975.

17. van der Vaart, AW. Asymptotic Statistics. Cambridge University Press; New York: 1998.
18. Jennison C, Turnbull BW. Exact calculations for sequential t , χ^2 and F tests. *Biometrika* 1991;78:133–141.
19. Marshall AW, Olkin I. A multivariate exponential distribution. *Journal of the American Statistical Association* 1967;62:30–44.
20. Marshall AW, Olkin I. A generalized bivariate exponential distribution. *Journal of Applied Probability* 1967;4:291–302.
21. Liu A, Schisterman EF, Wu CQ. Nonparametric estimation and hypothesis testing on the partial area under receiver operating characteristic curves. *Communications in Statistics—Theory and Methods* 2005;34:2077–2088.
22. Bickel, PJ.; Doksum, KA. *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day; San Francisco: 1977.

Table I

Type I error rate: bivariate normal case

ρ_x	ρ_y	$A^1=A^2$	Reject rate	$A^1=A^2$	Reject rate	$A^1=A^2$	Reject rate
$K = 2$							
0.5	0.5	0.7	0.0484	0.8	0.0461	0.9	0.0544
-0.5	0.5		0.0535		0.0504		0.0519
0.9	0.9		0.0478		0.0497		0.0521
-0.9	0.9		0.0514		0.0507		0.0484
$K = 3$							
0.5	0.5	0.7	0.058	0.8	0.061	0.9	0.062
-0.5	0.5		0.054		0.059		0.066
0.9	0.9		0.042		0.052		0.039
-0.9	0.9		0.062		0.047		0.047
$K = 4$							
0.5	0.5	0.7	0.043	0.8	0.041	0.9	0.059
-0.5	0.5		0.051		0.045		0.045
0.9	0.9		0.061		0.058		0.061
-0.9	0.9		0.058		0.056		0.064
$K = 5$							
0.5	0.5	0.7	0.061	0.8	0.052	0.9	0.06
-0.5	0.5		0.055		0.043		0.042
0.9	0.9		0.049		0.057		0.058
-0.9	0.9		0.043		0.06		0.055

Table II

Type I error rate: bivariate exponential case

ρ	$A^1=A^2$	Reject rate	$A^1=A^2$	Reject rate	$A^1=A^2$	Reject rate
$K=2$						
0.1	0.7	0.069	0.8	0.063	0.9	0.076
0.5		0.067		0.072		0.062
0.9		0.058		0.057		0.057
$K=3$						
0.1	0.7	0.071	0.8	0.076	0.9	0.064
0.5		0.069		0.068		0.073
0.9		0.072		0.081		0.062
$K=4$						
0.1	0.7	0.069	0.8	0.078	0.9	0.079
0.5		0.082		0.071		0.085
0.9		0.078		0.075		0.076
$K=5$						
0.1	0.7	0.070	0.8	0.081	0.9	0.068
0.5		0.076		0.063		0.074
0.9		0.065		0.078		0.077