Published in final edited form as:

J Appl Stat. 2011 August 1; 38(8): 1623–1632. doi:10.1080/02664763.2010.515678.

A sequential conditional probability ratio test procedure for comparing diagnostic tests

Liansheng Tang^{a,*}, Ming Tan^b, and Xiao-Hua Zhou^{c,d}

^aDepartment of Statistics, George Mason University, Fairfax, VA 22030, USA

^bDepartment of Epidemiology and Public Health, University of Maryland Greenebaum Cancer Center, Baltimore, MD 21201, USA

^cDepartment of Biostatistics, University of Washington, Seattle, WA 98195, USA

^dHSR&D Center of Excellence, 1100 Olive Way, Seattle, WA 98101, USA

Abstract

In this paper, we derive sequential conditional probability ratio tests to compare diagnostic tests without distributional assumptions on test results. The test statistics in our method are nonparametric weighted areas under the receiver-operating characteristic curves. By using the new method, the decision of stopping the diagnostic trial early is unlikely to be reversed should the trials continue to the planned end. The conservatism reflected in this approach to have more conservative stopping boundaries during the course of the trial is especially appealing for diagnostic trials since the end point is not death. In addition, the maximum sample size of our method is not greater than a fixed sample test with similar power functions. Simulation studies are performed to evaluate the properties of the proposed sequential procedure. We illustrate the method using data from a thoracic aorta imaging study.

Keywords

diagnostic accuracy; ROC; AUC; weighted AUC; SCPRT

1. Introduction

Magnetic resonance imaging (MRI) is a commonly used routine in disease diagnosis due to its high resolution and relative safety in practice. Unlike traditional computed tomography using radioactive isotopes, MRI machines apply strong magnetic field to align hydrogen atoms in the body and generate 3D images from these aligned atoms. Due to its noninvasive nature, MRI can be a good candidate in diagnosing patients with severe conditions. Diagnostic trials have been conducted to compare MRI with traditional computed tomograph or to compare different MRI modalities. For instance, thoracic aortic dissection is a life-threatening condition. Its mortality rate within 48 h can reach 68% without early diagnosis and prompt treatment [13]. MRI provides the detailed dissecting process and shows higher sensitivity than many other types of imaging modalities [12]. It thus emerges as an excellent diagnostic tool [1,21]. Common MRI techniques include spin-echo MRI (SE-MRI) and cinematic presentation of MRI (CINE-MRI). The former is a conventional technique, and the latter monitors the flow of cerebrospinal fluid in addition to SE-MRI. To

compare these two MRI modalities on their accuracy to detect thoracic aortic dissections, a diagnostic imaging trial was conducted by Van Dyke *et al.* [17].

To summarize test results in these trials, the receiver-operating characteristic (ROC) curve is a commonly used statistical tool [22]. The ROC curve plots the true positive rate (TPR) (the ratio of correctly detected diseased subjects) versus false positive rate (FPR) (the ratio of incorrectly identified nondiseased subjects) in the entire range of threshold values. Summary measures of ROC curves are obtained to evaluate the accuracy of diagnostic tests. The area under the ROC curve (AUC) is one of these summary measures [7]. However, when ROC curves from two tests intersect, resulting AUCs may be equal. This may lead to the false conclusion of the same accuracy of the two tests. Concerned with this limitation, Wieand *et al.* [18] proposed a Δ -statistic to compare the discriminant accuracy of tests at a prespecified range of specificities. A properly defined Δ -statistic also gives the nonparametric sensitivity estimator at a given specificity.

Group sequential designs in these diagnostic imaging trials are logistically feasible since a patient's disease status is usually available by some gold standard when the patient is recruited, and test results in these trials are immediately generated from scans [10]. Parametric and non-parametric group sequential methods have been proposed in diagnostic trials [11,16,23]. Zhou *et al.* [23] proposed a nonparametric sequential AUC difference estimator. They used the fact that an empirical AUC estimator is essentially a Wilcoxon statistic. They then derived its asymptotic property in group sequential designs. A more general sequential ROC summary statistic was introduced by Tang *et al.* [16]. They proposed a sequential Δ -statistic which is asymptotically a Brownian motion process as the information time increases. This desirable property allows the Δ -statistic to be used with standard group sequential monitoring methods such as Pocock, O'Brien-Fleming (OBF), and the error spending function method. The Pocock's test applies the same nominal significance level for each of the sequential stages, while the OBF test tends to make the earlier rejection more difficult and the latter rejection easier. These two tests require a fixed number of groups to compute the stopping boundaries. A more flexible error spending function test does not specify the number of stages. It partitions the overall type I error to derive stopping boundaries for sequentially computed test statistics at every group. The test statistics are compared with the boundaries to determine the acceptance or rejection of the null hypothesis until type I error is totally spent (see, [8] for more detailed discussion). Compared with a fixed sample test, these sequential tests have smaller expected sample sizes as they tend to reject the null hypothesis in earlier stages. However, the maximum sample sizes (MSS) of these tests exceed those in the fixed sample test. This may be considered a distinct disadvantage by medical researchers, and may become a practical obstacle for them to utilize the group sequential test in diagnostic trials. In addition, these methods do not consider likely outcomes if the trial is continued to its planned end.

To address the aforementioned concerns with common sequential tests, a class of stochastic curtailment procedures including the predictive power test (PPT) [2,3,5,14] and the sequential conditional probability ratio test (SCPRT) [19,20] have been proposed. Using the stochastic curtailment tests, we can specify in the trial protocol a pre-trial discordance probability that the significance of the trial based on the interim data will be reversed should the trial continue to the end for a given interim analysis plan. So the confidence of the early stopping decision will be enhanced should such a discordance probability is negligible (say, 1%). The discordance probability reflects the chance we are willing to take in making a decision based on interim data. If we want to be more conservative, we set a small discordance probability (say) 1%, and we can be less conservative in an individual trial, we can set a discordance probability of 3–5% to derive the monitoring boundary. The PPT requires a prior distribution for the parameter of interest, while the SCPRT is virtually

parameter-free and can obtain almost identical stopping boundaries as the predictive power approach. Furthermore, the SCPRT has MSS no larger than a fixed sample test with the same type I and II error rates. Since the test is conditional on the total information at the planned end, the decision reached by the test during interim analysis is unlikely to be reversed should the trial continue to the end. Due to the conditioning, SCPRT tends to have larger expected sample sizes than common group sequential design methods as shown in [15].

The purpose of this article is to derive a SCPRT test in comparing diagnostic imaging tests, to evaluate their properties and to compare their performance with the classic Pocock and OBF tests in diagnostic trials. Utilizing the SCPRT framework for a class of stochastic processes based on information time [20], we can derive the boundaries for the test statistic of interest if the sequentially computed test statistic can be approximated by a Brownian motion. Therefore, based on the Brownian motion property of Δ -statistics, we can derive a new nonparametric SCPRT test to sequentially compare accuracy of diagnostic tests. Our method will not only enjoy appealing properties of SCPRTs, but also include a general family of ROC summary measures. The rest of the paper is organized as follows. In Section 2, we briefly introduce the SCPRT procedure based on information time and the Δ -statistics. We then combine a sequential version of Δ -statistics with SCPRT tests to form a new class of sequential procedures. In Section 3, we evaluate their properties in simulation studies. In Section 4, we illustrate our method with the thoracic aorta imaging trial. Finally, discussion is given in Section 5.

2. Methods

We first briefly introduce SCPRT tests and the Δ -statistic. We will then combine the sequential Δ -statistic and SCPRT tests, and propose a new sequential procedure to have common MSS as the fixed sample test under the same specification of power, type I error, and alternatives.

2.1 SCPRT tests on information time

Without loss of generality, we define the null $H_0: \theta = \theta_0$ versus the alternative $H_a: \theta > \theta_0$. In this section, we will describe SCPRT tests under H_0 versus H_a in a set of discrete information time points, $(t_1, t_2, \dots, t_K)$, in a group sequential design with maximum K analyses, where t_k is the information time, $k = 1, \dots, K$ and $t_{k_1} \leq t_{k_2}$ for $k_1 < k_2$. Let w_k be a sufficient test statistic for the parameter θ at the k th stage. In a group sequential test, if w_k is outside of stopping boundaries defined by (a_k, b_k) , then the trial is stopped. H_0 is rejected in favor of H_a if $w_k \geq b_k$; and H_0 is accepted if $w_k \leq a_k$. Otherwise, the trial is continued. At the final stage, we let $a_K = b_K$ to ensure termination of the trial. $w_K \geq b_K$ allows us to reject H_0 , and $w_K < a_K$ allows the acceptance of H_0 . Let $L(w_k|w_K)$ be the likelihood function of w_k given w_K . The conditional maximum-likelihood ratio is given by

$$\lambda(t_k, w_k|t_K, z_\alpha) = \frac{\max_{w > z_\alpha} L(w_k|w_K = w)}{\max_{w \leq z_\alpha} L(w_k|w_K = w)}, \quad (1)$$

where z_α is the $(1 - \alpha)$ th percentile of a standard normal distribution, which is also the stopping boundary at the last time point t_K . Assume the sequential w_k statistic is approximately a Brownian motion process, the lower and upper stopping boundaries based on Equation (1) are derived by Xiong [19] and Xiong *et al.* [20] as follows:

$$a_k = z_\alpha - \{2at_k(1 - t_k)\}^{1/2}, \quad \text{and} \quad b_k = z_\alpha + \{2bt_k(1 - t_k)\}^{1/2}, \quad (2)$$

where a and b are determined by the probability of discordance between the conclusion by the sequential test and the conclusion should the trial be carried out to the planned end.

2.2 Δ -Statistic

Suppose two diagnostic tests are conducted on m diseased subjects and n nondiseased subjects. We denote the measurements from test ℓ ($\ell = 1, 2$) on the i th diseased subject as $X_{\ell i}$, where $i = 1, \dots, m$, and the measurements on the j th nondiseased subject as $Y_{\ell j}$, where $j = 1, \dots, n$. Define the joint cumulative survival functions $(X_{1i}, X_{2i}) \sim F(x_1, x_2)$ for the diseased population with marginal survival functions $X_{\ell i} \sim F_{\ell}(x)$. Similarly, define $(Y_{1j}, Y_{2j}) \sim G(y_1, y_2)$ for the nondiseased population with marginal survival functions $Y_{\ell j} \sim G_{\ell}(y)$.

An ROC curve for the ℓ th test can be expressed as a plot of the TPR versus the FPR, as the threshold varies over the real numbers. Equivalently, we can define the ROC curve for test ℓ as $\text{ROC}_{\ell}(u) = F_{\ell}(G_{\ell}^{-1}(u))$, where $0 \leq u \leq 1$, and $G_{\ell}^{-1}(u) = \inf\{y: G_{\ell}(y) < u\}$. Here, u corresponds to the FPR. Wieand *et al.* [18] proposed comparing two ROC curves on the basis of the

weighted area under the ROC curve (wAUC) $\Omega_{\ell} = \int_0^1 [F_{\ell}\{G_{\ell}^{-1}(u)\}] dW(u)$, 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 partial area under the curve (pAUC) between FPRs u_1 and u_2 (when $W(u) = (u - u_1)/(u_2 - u_1)$ for $0 < u_1 \leq u \leq u_2 \leq 1$), and the sensitivity at a given level of specificity u_0 (when $W(u)$ is a point mass at u_0). A natural nonparametric estimator for Ω_{ℓ} is given by

$$\widehat{\Omega}_{\ell} = \int_0^1 [\widehat{F}_{\ell}\{\widehat{G}_{\ell}^{-1}(u)\}] dW(u), \quad (3)$$

based on empirical survivor distribution functions $\widehat{F}_{\ell}(x)$ and $\widehat{G}_{\ell}(y)$, where $\widehat{G}_{\ell}^{-1}(u) = \inf\{y: \widehat{G}_{\ell}(y) < u\}$. Wieand *et al.* [18] used the difference between two wAUCs, $\Delta = \Omega_1 - \Omega_2$, as a statistical measure to compare diagnostic tests. Its nonparametric estimator, $\widehat{\Delta}$, is given by

$$\widehat{\Delta} = \widehat{\Omega}_1 - \widehat{\Omega}_2 \quad (4)$$

From results in [16,18], the asymptotic variance of $\widehat{\Delta}$ takes the form

$$\sigma_{\Delta}^2 = \frac{v_X}{m} + \frac{v_Y}{n},$$

where v_X and v_Y are given by

$$\begin{aligned} v_X &= \sum_{\ell=1}^2 \left(\int_0^1 \int_0^1 F_{\ell}\{G_{\ell}^{-1}(u_1 \wedge u_2)\} dW(u_1) dW(u_2) - \left[\int_0^1 F_{\ell}\{G_{\ell}^{-1}(u_1)\} dW(u_1) \right]^2 \right) \\ &\quad - 2 \int_0^1 \int_0^1 [F\{G_1^{-1}(u_1), G_2^{-1}(u_2)\} - F_1\{G_1^{-1}(u_1)\} F_2\{G_2^{-1}(u_2)\}] dW(u_1) dW(u_2), \\ v_Y &= \sum_{\ell=1}^2 \left[\int_0^1 \int_0^1 r_{\ell}(u_1) r_{\ell}(u_2) (u_1 \wedge u_2) dW(u_1) dW(u_2) - \left\{ \int_0^1 r_{\ell}(u) u dW(u) \right\}^2 \right] \\ &\quad - 2 \int_0^1 \int_0^1 r_1(u_1) r_2(u_2) [G\{G_1^{-1}(u_1), G_2^{-1}(u_2)\} - u_1 u_2] dW(u_1) dW(u_2), \end{aligned} \quad (5)$$

with

$$r_{\ell}(u) = \frac{F'_{\ell}\{G_{\ell}^{-1}(u)\}}{G'_{\ell}\{G_{\ell}^{-1}(u)\}}.$$

2.3 Our method

We define the following symbols for the k th stage of a SCPRT test with a maximum K analyses, $k = 1, \dots, K$:

- m_k, n_k are the numbers of available observations for diseased and nondiseased groups, respectively
- $\hat{F}_{\ell k}, \hat{G}_{\ell k}$ are respective empirical survival functions
- $\hat{\Delta}_k = \hat{\Omega}_{k1} - \hat{\Omega}_{k2}$ where $\hat{\Omega}_{k\ell}$ is the ℓ th empirical wAUC
- $\sigma_{\Delta k}$ is the variance of $\hat{\Delta}_k$ at the k th look, its estimate is $\hat{\sigma}_{\Delta k}$
- $Z_k = \hat{\Delta}_k / \hat{\sigma}_{\Delta k}$
- $I_k = 1/\sigma_{\Delta k}$, statistical information, consequently, $I_k \leq I_{k+1}$, $k = 1, \dots, K$
- $\tau_k = I_k / I_K$.

Define a modified sequential Δ -statistic $B_k = \sqrt{\tau_k I_k} \hat{\Delta}_k$, which is an asymptotically unbiased estimator for $\sqrt{\tau_k I_k} \Delta_k$, with asymptotic variance $\text{var}(B_k) = \tau_k$. B_k behaves asymptotically like a Brownian motion process with the drift parameter $\Delta \sqrt{I_k}$ [16]. We propose to use the modified sequential Δ -statistic, B_k , in SCPRT tests in the following paragraph.

At the planning stage of a trial, MSS of a SCPRT test are the same as those of a fixed sample size test with level α and power $1 - \beta$ under the same H_a . For instance, suppose we are interested in a one-sided test of $H_0: \Delta = 0$ versus $H_a: \Delta > 0$. The rejection of H_0 will show significant evidence of the superiority of test 1 over test 2 in diagnostic accuracy to discriminate the diseased from the non-diseased. Let $\lambda = m_K/n_K$. Given an alternative value δ_a , the MSS m_K and n_K are calculated by

$$m_K = \lambda n_K = \frac{(z_{1-\alpha} + z_{\beta})^2 V^2}{\delta_a^2},$$

where V^2 is an initial guess for $v_X + v_Y$ by investigators or calculated from preliminary results. As the trial is carried out to the k th stage ($1 \leq k < K$), supposedly m_k diseased and n_k nondiseased subjects are recruited and two diagnostic tests of interest are conducted on these subjects. From available test outcomes, we estimate the sequential Δ -statistic $\hat{\Delta}_k$, its standard error $\sigma_{\Delta k}$, and the modified interim statistic B_k . It can be shown from Equation (5) that the statistical information time at the k th stage is the ratio of current sample sizes and MSS, i.e. $\tau_k = m_k/m_K$, or $\tau_k = n_k/n_K$, as we maintain the sample size ratio $\lambda = m_k/n_k$ of the diseased and the nondiseased as a constant. The estimated modified Δ -statistic then becomes

$$B_k = \frac{m_k \hat{\Delta}_k}{\sqrt{m_k (v_{Xk} + \lambda v_{Yk})}},$$

where $\hat{v}_{Xk}$ and $\hat{v}_{Yk}$ are respective estimates of v_{Xk} and v_{Yk} . In our simulation studies and the examples, the density functions in $r_t(u)$ are estimated using the Epanechnikov kernel function $E(x) = 3/4(1 - x^2)I(|x| \leq 1)$ with the bandwidth of $4/\max(\min(m_k, n_k)^{4/5}, 50)$. The stopping boundaries a_k and b_k are obtained via Equation (2) with a certain probability of discordance. [20, Table 1] provides the exact values of design parameters a and b under various discordance probabilities. According to Tan *et al.* [15], if B_k is outside of (a_k, b_k) , then the trial is stopped without accruing more subjects. The null is rejected if $B_k \geq b_k$, and the null is accepted if $B_k \leq a_k$. Otherwise, more subjects are recruited to continue to the $k + 1$ th interim analysis. If at the final interim analysis $\hat{B}_K \geq b_K$, we would conclude that test 1 has superior diagnostic accuracy than test 2. Note that our method is not limited in the aforementioned one-sided test. We can apply it to two-sided test of equal diagnostic accuracy or other hypothesis tests. The test procedure is essentially the same.

Our method is summarized in the following steps:

1. At the k th stage, determine boundaries a_k and b_k , and calculate B_k
2. Reject H_0 if $B_k \geq b_k$, and Accept H_0 if $B_k \leq a_k$
3. Otherwise, continue to the $(k + 1)$ th stage
4. At the final stage, conclude that test 1 has superior diagnostic accuracy than test 2 if $\hat{B}_K \geq b_K$

3. Simulation studies

3.1 Type I error rate

We applied our SCPRT method to simulated data sets to evaluate its finite-sample performance. To simulate type I error rate, the hypothesis was set to be $H_0: \Delta = 0$ versus $H_0: \Delta > 0$. Under the true H_0 , we simulated test results using the following three parametric models:

- Bivariate normal: $(X_1, X_2) \sim N((11, 1), \Sigma_1)$ and $(Y_1, Y_2) \sim N((10, 0), \Sigma_2)$, where

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

We chose these mean vectors so that true AUCs are the same for two tests.

- Bivariate lognormal: $\exp(X_1, X_2)$ and $\exp(Y_1, Y_2)$, where $(X_1, X_2), (Y_1, Y_2)$ are from the above bivariate normal.
- Bivariate exponential [6]: $(X_1, X_2) \sim H_x(x_1, x_2)$, and $(Y_1, Y_2) \sim H_y(y_1, y_2)$. Here, $H_x(x_1, x_2) = \exp(-\beta_{11}x_1) \exp(-\beta_{21}x_2)[1 + 4\rho\{1 - \exp(-\beta_{11}x_1)\}\{1 - \exp(-\beta_{21}x_2)\}]$, and $H_y(y_1, y_2) = \exp(-\beta_{12}y_1) \exp(-\beta_{22}y_2)[1 + 4\rho\{1 - \exp(-\beta_{12}y_1)\}\{1 - \exp(-\beta_{22}y_2)\}]$. In the simulation, $\rho = 0.25$ and $\beta = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}) = (1, 2, 2, 4)$. Again, the β vector was chosen so that true AUCs for two tests are the same.

In the simulation, we considered three-group and four-group equal-space SCPRT tests. Different discordance probabilities, p_d , ranging from 0.001 to 0.04 were specified. The nominal significance level was set to be $\alpha = 0.025$. The simulated significance levels were calculated from 1000 replicates under each setting and are shown in Table 1 ($m = n$) and in Table 2 ($m \neq n$). They show that most of the simulated levels are within the 95% prediction interval. This illustrates the good finite-sample performance of our procedure.

3.2 Maximum and average sample sizes

We compared MSS for fixed-sample design (FSD), SCPRT, Pocock and OBF designs with the Δ -statistic for binormal data. We used power 0.8 and type I error rate $\alpha = 0.05$ for the sample size calculation. 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 0.5. We let $\lambda = 1$. Since the distributions were known, we were able to obtain the exact variance of $\hat{\Delta}$ from the results in Equation (5). We obtained MSS for three-group SCPRT, Pocock and OBF designs. Let Δ be the difference between AUCs, or partial AUCs (pAUCs). The MSS for comparing AUCs and pAUCs are presented in Table 3. These tables indicate that our SCPRT design has smaller MSS than Pocock and OBF designs.

We also conducted a simulation study to compare the average sample number (ASN) and simulated power of the proposed method with those of Pocock and OBF methods. The same model configuration for calculating the MSS in the preceding paragraph was used to simulate data. For the comparison of AUCs, we selected $\mu_1 = 0.7500$ and $\mu_2 = 0.9655$ to have $\Omega_1 = 0.75$, $\Omega_2 = 0.7$. For the comparison of pAUCs, we selected $\mu_1 = 0.6142$ and $\mu_2 = 0.8653$ to have $\Omega_1 = 0.35$, $\Omega_2 = 0.3$. We simulated 1000 data sets for each setting, and counted the number of rejections. Dividing the number by 1000 gives the simulation power. The simulated powers and ASNs are presented in Table 4. The results indicate that the proposed SCPRT method has the smallest MSS and maintains the nominal power well. The SCPRT method tends to have larger ASNs than Pocock and OBF methods when the discordance probability p_d takes smaller values such as 0.005 or 0.01, but it has the lowest ASNs among all three methods when $\rho = 0.04$.

4. A real example

The aforementioned aortic dissection diagnostic trial recruited 45 patients with and 69 without a dissection identified by surgery [17]. Their MRI images were presented to radiologists to be rated in a five-point scale according to radiologists' confidence for the presence of a dissection. 1 indicated "definitely no aortic dissection", 2 indicated "probably no aortic dissection", 3 indicated "unsure about aortic dissection", 4 indicates "probably aortic dissection", and 5 indicates "definitely aortic dissection".

Since both the Δ measure and its variance are estimated nonparametrically using rank statistics, the proposed method can be used for categorical scores in the example. We applied our method to results by radiologist 1 to sequentially compare the diagnostic accuracy of SE-MRI and CINE-MRI, and tried to decide whether the diagnostic trial could be stopped earlier by our method. Let Ω_1 and Ω_2 be AUCs for SE-MRI and CINE-MRI, respectively. Because of relatively small sample sizes, we applied our method with two looks. Let Δ be the AUC difference. The hypothesis in the study is $H_0: \Delta = 0$ versus $H_a: \Delta > 0$. We assume the original sample sizes were determined by the authors to maintain type I error rate, $\alpha = 0.05$ and get their desired power, $1 - \beta$. Our sequential SCPRT test need the same MSS as the original fixed sample test. At the first stage of our analysis, we randomly chose half observations of the first 25 patients with and 35 without dissection to estimated the modified Δ -statistic, B_1 . The information time t_1 is approximately 0.5. Using results from Table 1 in [20], the upper and lower stopping boundaries $a_1 = -0.5611$ and $b_1 = 2.5211$ are calculated from the maximum discordance probability 0.001. We then calculated the modified Δ -statistic, B_1 and compared it with the boundaries. From the data at the first stage, we got AUC estimates $\hat{\Omega}_{11} = 0.6051$ for CINE-MRI and $\hat{\Omega}_{12} = 0.7125$ for SE-MRI. The estimated variance of $\hat{\Delta}_1 = \hat{\Omega}_{11} - \hat{\Omega}_{12}$ is $\hat{v}_{\Delta} = 0.00735$. Thus, it follows that $B_1 = -0.8852$. Since $B_1 < a_1$, the trial could have been stopped earlier to accept H_0 , that is, *SE-MRI is not more accurate than CINE-MRI to detect thoracic aortic dissections*. Scanning the

rest of the 25 patients with and 34 patients without the condition could have been unnecessary.

5. Discussion

In this paper, we have proposed a sequential SCPRT test based on a Δ -statistic for comparing diagnostic accuracy of two tests. Our method is applicable to most diagnostic trials in which patients' disease status and test results are immediately available at interim analysis. We applied the newly proposed method to a real MRI example for diagnosing aortic dissection and showed that our method ensures possible early stopping of the trial while ensuring a negligible discordance probability (of 0.001). Our sequential test addresses both ethical and cost concerns, which frequently arise in diagnostic imaging trials. In the example we analyzed, the MRI machine generates a huge noise during its operation. It is ethical to stop a trial earlier should significant evidence against H_0 be found. The MRI typically costs thousands of dollars per subject. Stopping a trial earlier will divert the resources for better use, especially when the stopping rule is such that the significant efficacy would sustain (with 0.999 probability) should the trial continue to the planned end.

When we applied the proposed SCPRT design to the aortic dissection example, we only considered ratings from one radiologist, although several radiologists provided image ratings in the example. The effect of radiologists complicates the design of diagnostic trials. Further topics are to develop new test statistics, which allow more than one radiologists and have approximately independent incremental variance structures, for designing sequential diagnostic trials. In addition, the proposed method may be applied for the comparison among more than two tests when the statistic of interest is a linear contrast of several estimated wAUCs. Such a contrast of AUCs has been considered in [4]. Furthermore, an adaptive procedure similar to the Jennison and Turnbull [9] t -test boundaries is available for the SCPRT [20].

Acknowledgments

The authors thank three anonymous referees for their constructive comments and useful suggestions.

References

1. Bitar R, Moody AR, Leung G, Kiss A, Gladstone D, Sahlas DJ, Maggisano R. In vivo identification of complicated upper thoracic aorta and arch vessel plaque by MR direct thrombus imaging in patients investigated for cerebrovascular disease. *Am J Roentgenol*. 2006; 187:228–234. [PubMed: 16794181]
2. Choi SC, Pepple PA. Monitoring clinical trials based on predictive probability of significance. *Biometrics*. 1989; 45:317–323. [PubMed: 2720057]
3. Choi SC, Smith PJ, Becker DP. Early decision in clinical-trials when the treatment differences are small –experience of a controlled trial in head trauma. *Control Clin Trials*. 1985; 6:280–288. [PubMed: 4075806]
4. DeLong ER, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*. 1988; 44:837–845. [PubMed: 3203132]
5. Freidlin B, Korn EL, George SL. Data monitoring committees and interim monitoring guidelines. *Control Clin Trials*. 1999; 20:395–407. [PubMed: 10503800]
6. Gumbel EJ. Bivariate exponential distributions. *J Am Stat Assoc*. 1960; 55:698–707.
7. Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143:29–36. [PubMed: 7063747]
8. Jennison, C.; Turnbull, B. *Group Sequential Methods*. Chapman and Hall; New York: 2000.

9. Jennison C, Turnbull B. On group sequential tests for data in unequally sized groups and with unknown variance. *J Statist Plan Inference*. 2001; 96:263–288.
10. Kidwell CS, Chalela JA, Saver JL. Comparison of mri and ct for detection of acute intracerebral hemorrhage. *J Am Med Assoc*. 2004; 292:1823–1830.
11. Liu A, Wu C, Schisterman EF. Nonparametric sequential evaluation of diagnostic biomarkers. *Stat Med*. 2008; 27:1667–1678. [PubMed: 18241080]
12. Nienaber C, Kodolitsch Y, Kodolitsch von Y, Siglow V, Piepho A, Jaup T, Nicolas V, Weber P, Triebel H, Bleifeld W. The diagnosis of thoracic aortic dissection by noninvasive imaging procedures. *N Engl J Med*. 1993; 328:1–9. [PubMed: 8416265]
13. Shiga T, Wajima Z, Apfel CC, Inoue T, Ohe Y. Diagnostic accuracy of transesophageal echocardiography, helical computed tomography, and magnetic resonance imaging for suspected thoracic aortic dissection: Systematic review and meta-analysis. *Arch Intern Med*. 2006; 166:1350–1356. [PubMed: 16831999]
14. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical-trials – conditional or predictive power. *Control Clin Trials*. 1986; 7:8–17. [PubMed: 3956212]
15. Tan M, Xiong X, Kutner MH. Clinical trial designs based on sequential conditional probability ratio tests and reverse stochastic curtailing. *Biometrics*. 1998; 54:682–695. [PubMed: 9629648]
16. Tang L, Emerson SS, Zhou XH. Nonparametric and semiparametric group sequential methods for comparing accuracy of diagnostic tests. *Biometrics*. 2008; 64:1137–1145. [PubMed: 18371124]
17. Van Dyke, C.; White, R.; Obuchowski, N.; Geisinger, MA.; Lorig, RJ.; Mezziane, MA. Cine MRI in the Diagnosis of Thoracic Aortic Dissection. 79th RSNA Meetings; Chicago, IL. 1993.
18. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*. 1989; 76:585–592.
19. Xiong X. A class of sequential conditional probability ratio tests. *J Am Stat Assoc*. 1995; 90:1463–1473.
20. Xiong X, Tan M, Boyett J. Sequential conditional probability ratio tests for normalized test statistic on information time. *Biometrics*. 2003; 59:624–631. [PubMed: 14601763]
21. Yoshida S, Akiba H, Tamakawa M, Yama N, Hareyama M, Morishita K, Abe T. Thoracic involvement of type A aortic dissection and intramural hematoma: Diagnostic accuracy–comparison of emergency helical CT and surgical findings. *Radiology*. 2003; 228:430–435. [PubMed: 12819341]
22. Zhou, XH.; McClish, DK.; Obuchowski, N. *Statistical Methods in Diagnostic Medicine*. Wiley; New York: 2002.
23. Zhou XH, Li SM, Gatsonis CA. Wilcoxon-based group sequential designs for comparison of areas under two correlated ROC curves. *Stat Med*. 2008; 27:213–223. [PubMed: 17357988]

Table 1

Simulated significance level ($m = n$).

p_d	$m(n)$	Three-group SCPT test			Four-group SCPT test		
		Norm	Lognorm	Exp	Norm	Lognorm	Exp
0.001	50	0.022	0.032	0.032	0.027	0.032	0.028
	100	0.030	0.029	0.022	0.025	0.031	0.027
	200	0.023	0.020	0.024	0.023	0.023	0.025
0.010	50	0.022	0.033	0.032	0.028	0.032	0.029
	100	0.032	0.029	0.024	0.029	0.033	0.029
	200	0.024	0.022	0.025	0.025	0.025	0.025
0.040	50	0.034	0.041	0.036	0.040	0.032	0.030
	100	0.037	0.035	0.030	0.037	0.035	0.035
	200	0.026	0.025	0.025	0.027	0.025	0.031

Note: The 95% prediction interval is (0.025 ± 0.010) .

Table 2

Simulated significance level ($m \neq n$).

p_d	m	n	Three-group SCPRT test			Four-group SCPRT test		
			Norm	Lognorm	Exp	Norm	Lognorm	Exp
0.001	50	60	0.021	0.028	0.030	0.024	0.016	0.028
	50	80	0.018	0.017	0.027	0.015	0.021	0.020
0.010	50	60	0.022	0.028	0.030	0.025	0.016	0.028
	50	80	0.019	0.017	0.028	0.016	0.021	0.021
0.040	50	60	0.027	0.033	0.036	0.025	0.019	0.030
	50	80	0.021	0.021	0.034	0.022	0.023	0.025

Table 3

Maximum possible number of subjects in both arms for comparing AUCs or pAUCs.

	Comparing AUCs											
	$\Omega_2 = 0.7$				$\Omega_2 = 0.75$				$\Omega_2 = 0.8$			
	$\Delta = 0.05$	0.1	0.15	0.2	0.05	0.1	0.15	0.2	0.05	0.1	0.15	0.2
FSD	832	195	81	44	710	163	68	38	568	128	55	43
SCPRT	832	195	81	44	710	163	68	38	568	128	55	43
Pocock	970	227	94	51	828	190	79	44	662	149	64	50
OBF	847	199	83	45	723	166	70	39	578	131	56	45
	Comparing pAUCs											
	$\Omega_2 = 0.3$				$\Omega_2 = 0.35$				$\Omega_2 = 0.4$			
	$\Delta = 0.05$	0.1	0.15	0.2	0.05	0.1	0.15	0.2	0.05	0.1	0.15	0.2
FSD	655	156	65	34	604	140	57	31	513	116	48	36
SCPRT	655	156	65	34	604	140	57	31	513	116	48	36
Pocock	764	182	76	40	704	163	67	36	598	135	56	42
OBF	667	159	66	35	614	143	59	32	522	118	50	37

Table 4

Simulated powers and ASNs.

	Comparing AUCs			Comparing pAUCs		
	$\Omega_1 = 0.75, \Omega_2 = 0.70$			$\Omega_1 = 0.35, \Omega_2 = 0.30$		
	MSS	Power	ASN	MSS	Power	ASN
Pocock	970	79.2%	680.08	764	80.3%	517.89
OBF	847	81.1%	710.28	667	81.2%	557.09
SCPRT ($p_d = 0.001$)	832	80.6%	793.13	655	80.5%	630.30
SCPRT ($p_d = 0.010$)	832	80.7%	750.04	655	80.5%	592.90
SCPRT ($p_d = 0.040$)	832	81.1%	670.54	655	81.6%	524.70

Note: MSS, maximum sample size.