Statistics and Its Interface Volume 9 (2016) 343–354

Multistage nonparametric tests for treatment comparisons in clinical trials with multiple primary endpoints

Peng Huang*,† and Ming T. Tan‡

Many clinical trials, e.g., neurodegenerative disease trials, are conducted to test whether a new treatment could slow or modify disease progression. Multiple primary endpoints are often used since it is difficult to find a single clinical endpoint that summarizes the treatment effect, e.g., the neuroprotective effect. There are three major challenges in the design and analysis of such trials: (1) the presence of nuisance effect regardless whether the desired neuroprotective effect exists; (2) primary endpoints are of mixed type; (3) the need for interim analysis stopping rule for multiple primary endpoints. We propose a simple nonparametric multistage adaptive (group sequential) test to overcome these difficulties. Statistically, this test is another solution to the multivariate nonparametric Behrens-Fisher problem. We provide both large and small sample properties of the proposed test. The methodology is illustrated using data from two randomized clinical trials.

AMS 2000 subject classifications: Primary 62G10, 62H15, 62L05.
Keywords and phrases: Rank-based test, Behrens-Fisher problem, Adaptive group sequential test, Brownian motion.

1. INTRODUCTION

In recent years, multiple primary endpoints are increasingly used in randomized controlled clinical trials to determine whether a new treatment is more efficacious than a control. In addition, multidimensional biomarkers of different data types are also compared between different phenotypes and/or between different treatments. To list a few, the National Institute of Neurological Disorders and Stroke sponsored Neuroprotection Exploratory Trials in Parkinson's Disease (NET-PD) used quality of life, activities of daily living, mobility, and cognition (including modified Rankin, Symbol Digit Modalities, Schwab and England activities of daily living, PDQ-39, and 5 questions on gait from the UPDRS) as the primary endpoints to measure treatment's neuroprotective effect (Olanow, Wunderle, and Kieburtz, 2011). In a randomized controlled trial to determine if a combined pharmacological and behavioral intervention improves both depression and pain in 250 primary care patients with musculoskeletal pain and comorbid depression, the co-primary endpoints were the depression (20-item Hopkins Symptom Checklist), pain severity and interference (Brief Pain Inventory), and global improvement in pain at 12 months (Kroenke et al., 2009). Sankoh, D'Agostino and Huque (2003) have listed several types of disease studies and examples where no consensus on the most important clinical endpoint is available due to disease complexity. In particular, they proposed that multiple primary endpoints should be considered in clinical trials for diseases with unknown etiologies or diseases that manifest in multiple dimensions.

This paper is motivated from the design of a Parkinson's disease (PD) clinical trial to identify the most promising neuroprotective compounds for individuals with PD. In 2003, The National Institute of Neurological Disorders and Stroke (NINDS) launched a major series of cooperative clinical studies designed to evaluate a number of promising compounds for use in slowing the progression of PD. The first two phase II futility studies of the series Neuroprotection Exploratory Trials in PD (NET-PD) were carried out in 5/2003–5/2005 and 3/2004–1/2006 respectively to assess the impact of minocycline and creatine (in the first futility study), or CoQ10 and GPI 1485 (in the second futility study) on the progression of PD in order to assess if it is non-futile to proceed with further study of these agents (The NINDS NET-PD Investigators, 2006, 2007). The NET-PD phase III trial, initiated in March 2007, used a global statistical test (GST) to compare disease progression at 5 years between the creatine and placebo groups (clinicaltrials.gov registration number NCT00449865) based on multiple continuous and ordinal primary endpoints: quality of life, activities of daily living, mobility, and cognition measures of modified Rankin, Symbol Digit Modalities, Schwab and England activities of daily living, PDQ-39, and 5 ordinal measures on gait from the Unified Parkinson's Disease Rating Scale (UPDRS) (Olanow, Wunderle, and Kieburtz, 2011). Two important features must be taken into consideration when designing such type of clinical trials. First, most

*Corresponding author. For preprint, contact: phuang12@jhmi.edu.
†Dr. Huang's research is partially supported by 1R21NS043569-01, P50CA103175, MCRF-FHA05CRF, and P30CA006973.
‡Dr. Tan's research is partially supported by NIH grant R01CA164717.

compounds for Parkinson's disease have various effects other than the desired neuroprotective effect. Ignoring such nuisance effects could result in biased assessment of treatment benefit. A typical example of nuisance effect is the transient symptomatic effect that does not slow the disease progression. Because of this, the null hypothesis of identical distribution between the two treatments being compared is not appropriate to use. Testing treatment effect with the presence of nuisance parameter is called a *Behrens-Fisher Problem*. Second, there is no gold standard single endpoint that is sufficient to summarize treatment's neuroprotective effect. Treatment comparison relies on a global assessment of multiple endpoints. In addition, sequential monitory of treatment effect is often required in large multi-center studies, particularly for phase III trials. A multivariate sequential statistical test for Behrens-Fisher Problem is a useful tool for such clinical trial design and sequential data analysis.

Solutions to parametric multivariate Behrens-Fisher Problem have been well studied for many years since Bennett (1951) and James (1954). Christensen and Renche (1997) have reviewed several commonly used solutions. For multivariate nonparametric Behrens-Fisher Problem, Brunner, Munzel, and Puri (2002) proposed rank based ANOVA-type statistic and Wald-type statistic to test whether there is any difference in at least one of the endpoints. Huang, Tilley, Woolson, and Lipsitz (2005) extended O'Brien's rank-sum global statistical test to Behrens-Fisher Problem in treatment comparison with multiple equally important primary endpoints. Liu et al. (2010) proposed a multivariate test based on the maximum rank sum difference among all endpoints. However, none of these methods is for multistage adaptive or group sequential testing that allows early stopping for pronounced treatment effect or the lack of it thereof. Since a group sequential design with G interim analysis looks is also called an adaptive multistage design with G stages, throughout this paper, we will not distinguish terms between "group sequential" and "multistage"; nor between "t-th interim analysis" and "t-th stage analysis".

For single primary endpoint, group sequential designs and stopping boundaries have been extensively studied in the literature. Whitehead (1997) and Jennison and Turnbull (2000) give a quite comprehensive presentation of different strategies that can be used to develop sequential stopping boundaries such as the repeated significant test (or confidence interval) approach, Lan and DeMets (1983) flexible error-spending function approach, stochastic curtailments, and Bayesian approach. Critical values of the sequential stopping boundaries are often determined numerically. Since such numeric computation could be intensive, many authors have tabulated the critical values for different combinations of design parameters (such as the number of interim analysis looks, type I and type II errors, the shape of alpha spending function or stopping boundaries). Softwares (e.g., EAST, PEST) have been developed to provide sequential stopping boundaries and numerical evaluation of the design operating characteristics.

While it is relatively easy for investigators to choose an appropriate sequential design for testing treatment benefit with single (i.e., uni-dimensional) primary endpoint, and to implement it using existing software, it is less clear what is the most appropriate sequential stopping rule when multiple primary endpoints (i.e., a multi-dimensional endpoint) are analyzed sequentially. This is because it is relatively easy to find a commonly agreed definition of "better" with single endpoint while there are different ways to define improvement with multiple endpoints. Statistical approaches for multiple endpoints can be classified into 3 categories (Kosorok, Shi, and DeMets, 2004): the global, auxiliary, and multiple hypothesis methods. The global method combines all endpoints into a single composite endpoint to access treatment benefit. Examples of global method include Wei and Lachin (1984), O'Brien (1984), Pocock, Geller, and Tsiatis (1987), Tang, Gnecco and Geller (1989a, 1989b), and Lin (1991). The auxiliary method chooses one endpoint as primary, and use information from all other endpoints to increase the power on the primary endpoint. The multiple hypothesis method evaluates treatment benefit on each single endpoint first, then combines these evaluations across all endpoints. Examples include the Bonferroni procedures, Simes test (1986), Benjamini-Hochberg Procedure (1995), Holm step-down (1979), Hochberg step-up (1988), Shaffer procedure (1986), Fallback procedure (Wiens and Dmitrienko, 2005), and various p-value based procedures. These procedures are most suitable when the goal is to find any type of effect (no matter it is beneficial or detrimental) between the two treatments. However, if the goal is to find which treatment should be recommended to use when no single gold standard endpoint is available to summarize treatment benefit, it is more appropriate to use a global statistical method to compare treatment's overall benefit across multiple equally important endpoints (Tilley et al, 1996, 2000, Olanow, Wunderle, and Kieburtz, 2011).

O'Brien (1984) introduced three types of global statistical tests (GSTs) based on ordinal least square (OLS), generalized least square (GLS), and rank-sum respectively. When multiple endpoints have a joint normal distribution, Tang, Gnecco, and Geller (1989b) extended O'Brien's GLS based GST (a fixed sample size test) to a sequential test of the equality of two treatment group means. They have showed that, for any pre-specified directional difference in alternative hypothesis, multivariate test always requires smaller sample size than any univariate test under the same type I and type II error probability requirements. Since O'Brien's GLS multivariate test statistic is a weighted average of univariate test statistics with weights derived from the inverse matrix of their variance-covariances, the resulting test statistic can be difficult to interpret if some of the weights are negative when the objective is to determine whether one treatment is better than the other one. Thus, a test with positive weights for all endpoints is recommended (Pocock, Geller, and Tsiatis, 1987). We focus our discussions on nonparametric tests based on ranks because multiple endpoints

could have quite different continuous and/or ordinal distributions, and assumptions of their marginal and joint distributions can be difficult to make. Based on Wei, Lin, and Weissfeld (1989) proportional hazards regression model, Lin (1991) proposed a sequential test using a weighted sum of linear rank statistics with respect to marginal distributions of all single endpoints. This test is applicable when the stochastic ordering assumption between the two treatments can be made: $F_{1u}(\cdot) \leq F_{2u}(\cdot)$ $(v = 1, \ldots, K)$ with at least one strict inequality, where F_{iu} is the marginal cumulative distribution function of the ith treatment on the vth endpoint. Su and Lachin (1992) generalized Mann-Whitney-Wilcoxon statist with kernel function $\phi(x, y) = I[x < y]$ to test location shift model. Lee and DeMets (1992) used linear rank statistics to compare treatments under location-shift model assumption, and they have established asymptotic normality of sequentially computed linear rank statistics. All these methods require identical null distribution between the two treatments and thus are not solutions to the Behrens-Fisher Problem.

To develop multistage test using sequentially computed global statistical test statistic for Behrens-Fisher Problem, we propose, in Section 2, a simple rank-based test statistic which can be expressed as a linear combination of Wilcoxon test statistics for multiple endpoints. When this rank based test statistic is computed sequentially, it forms a discrete time asymptotic Gaussian process (Theorem 1 in Section 3). If, in addition, sample size ratios between the two groups are about the same in all stages, its limiting process is a Brownian motion measured at a finite time points. We provide the mean and variance-covariance matrix of the joint rank-sum test statistics. The approximation of our stochastic process to a Brownian motion process is controlled through the first two moments: their mean processes are exactly the same, and an upper bound to the relative difference in variance-covariance structures between the two processes is provided. Numerical evaluation of this upper bound is provided through simulation. These quantities are useful to evaluate the finite sample properties of the proposed stochastic process, and how well the Brownian motion process approximation could be. Theorem proof is given in the Appendix. The simulation in Section 4 shows how well the type I error and statistical power are controlled when the proposed stochastic process is applied to sequential designs for univariate endpoint when sample size is moderate. In Section 5, we illustrate its use in DATATOP and QE2 trials, one with large sample size and one with moderate sample size. These two Parkinson disease trials are chosen because information from these two trials were used in the planning of NET-PD trials. We conclude with a discussion.

2. MODEL FORMULATION

To formulate the model, we compare two groups of patients as in a randomized clinical trial to test whether a new treatment is more efficacious than a control, based on K-dimensional outcomes. Let $\mathbf{x_{ij}} = (x_{ij1}, \ldots, x_{ijK})^T$ be the observation vector from the j-th subject in the i-th treatment group, and $\mathbf{x}^T$ be the transpose of $\mathbf{x}$. Vectors $\mathbf{x_{i1}}, \mathbf{x_{i2}}, \ldots$ are iid copies of random vector $\mathbf{X_i} = (X_{i1}, \ldots, X_{iK})^T$. Without loss of generality, all observations are coded so that larger values correspond to better clinical conditions. For the v-th endpoint, Brunner, Munzel, and Puri (2002) proposed to use $p^{(v)} = P(X_{1v} < X_{2v}) + \frac{1}{2}P(X_{1v} = X_{2v})$ as a measure of relative treatment effect. Quantity $\frac{1}{2}P(X_{1v} = X_{2v})$ is used to account for possible discontinuous distribution. This parameter was first introduced by Kruskal (1952) and has been used by many other authors thereafter. We use an equivalent parameter $\theta_v = P(X_{1v} < X_{2v}) - P(X_{1v} > X_{2v}) = 2p^{(v)} - 1$ to measure treatment difference on the v-th endpoint. Lachin (1992) called θ_v a *Mann-Whitney difference* and pointed out that it can be used to summarize the difference on any scale of measurement. Huang, Ou, Piantadosi and Tan (2014) studied relationship between Mann-Whitney type difference and other commonly used treatment differences. Figure 2 from their paper shows that testing Mann-Whitney type difference is more likely to yield an informative conclusion than testing other types of differences in most scenarios. They also provided guidance how to properly define treatment superiority in clinical trials. Suppose the t-th interim analysis will be conducted when data from the first n_{it} subjects in the i-th group are available, $n_{it} < n_{i,t+1}, N_t = n_{1t} + n_{2t}$, $i = 1, 2$, $t = 1, 2, \ldots$. To compare two treatment groups with multiple endpoints nonparametrically, Huang, Tilley, Woolson, and Lipsitz (2005) proposed several single stage nonparametric global statistical tests for the Behrens-Fisher problem whose test statistics have the same asymptotic distribution as the sum of Wilcoxon rank-sum test statistics from all endpoints. The parameter to be tested is $\bar{\theta} = \sum_{v=1}^{K} \theta_v / K$. We consider the same hypothesis of interest:

$$(1) \qquad H_0 : \bar{\theta} = \sum_{v=1}^{K} \theta_v / K \leq 0, \qquad H_1 : \bar{\theta} > 0.$$

A two-sided problem can be formulated in a similar fashion. Parameter $\bar{\theta}$ was called global treatment effect (GTE) in Huang, Woolson, and O'Brien (2008). We choose this parametrization based on the following considerations. First, the scientific question to be addressed is whether one treatment is better than the other treatment when multiple equally important endpoints are evaluated together. In fact, this $\bar{\theta}$ is exact the same parameter used in the landmark study of NET-PD phase III clinical trial. As described in Olanow, Wunderle, and Kieburtz (2011), a rank based global statistical test of this parameter "offers a new comprehensive method for evaluating the progression of movement disorders in areas of functional significance rather than on points on a scale." A treatment may be considered beneficial if it shows improvement

on most endpoints. This is reflected by using $\bar{\theta} > 0$ to denote treatment benefit. Second, all $\theta_1, \ldots, \theta_K$ have the same positive weight. This avoids negative weight that makes the hypothesis test result difficult to interpret (Pocock, Geller, and Tsiatis, 1987). The equal weight reflects the assumption of equal importance of all endpoints. Our test statistic for parameter $\bar{\theta}$ will be constructed through a composite endpoint. As suggested in the FDA's Guidance for Industry: Patient-Reported Outcome Measures, a composite endpoint is suitable to use when all components are "of similar importance to patients, the more important and less important components are equally likely to occur with similar frequency, and the components are likely to have roughly similar treatment effects." (http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf). When some endpoints are considered more important than others, the weights in both $\bar{\theta}$ and our nonparametric test statistics can be adjusted to reflect this difference. The corresponding properties of test statistics discussed in this paper can be readily modified to accommodate unequal weights. Third, $\bar{\theta}$ is invariant to any monotone transformation of the data. This makes it relatively easy to summarize treatment's overall effect across different types of endpoints. More discussions on why $\bar{\theta}$ is suitable to use in clinical trials are given by Huang et al. (2009) and Olanow et al. (2011).

3. SEQUENTIALLY COMPUTED TEST STATISTIC

This section will construct a stochastic process from a sequentially computed rank-sum statistic and show that it is asymptotically a Gaussian process. We will further show that, when sample size ratio is relatively stable in all stages, such a discrete time stochastic process can be approximated by a Brownian motion measured at finite time points. When the Central Limiting Theory is applicable, this Brownian motion approximation provides a foundation to apply many available sequential designs for single primary endpoint, such as O'Brien-Fleming design, for adaptive and nonparametric treatment comparisons with multiple primary endpoints.

Consider the statistical problem of testing the hypothesis (1) in a randomized clinical trial based on K primary endpoints. We continue to use notations in the previous section. The t-th interim analysis will be performed when data from the first n_{it} subjects in the i-th group are available, and $n_{it} < n_{i,t+1}$. Let $F_{iv}(z) = P(X_{iv} < z) + \frac{1}{2}P(X_{iv} = z)$, $\theta = (\theta_1, \ldots, \theta_K)^T$. Throughout the paper, we will impose regularity conditions $Var\{F_{1v}(X_{2v})\} > 0$ and $Var\{F_{2v}(X_{1v})\} > 0$ $(v = 1, \ldots, K)$ to rule out degenerate distributions and redundant parameters. At the t-th stage (interim) analysis, we pool observations from all $N_t = n_{1t} + n_{2t}$ subjects and rank them separately on each of the vth endpoints. Define test statistic at the t-th stage by

$$D(n_{1t}, n_{2t}) = \frac{2n_{1t}}{n_{1t} + n_{2t}}(\bar{R}_{2t} - \bar{R}_{1t}),$$

where $\bar{R}_{it} = \frac{1}{n_i}\sum_{j=1}^{n_{it}}\sum_{v=1}^{K} R_{ijv}(t)$ is the mean rank sum across all K endpoints in the i-th group, $R_{ijv}(t)$ is the rank of x_{ijv} among observations $\{x_{11v}, \ldots, x_{1n_{1t}v}, x_{21v}, \ldots, x_{2n_{2t}v}\}$. Since Wilcoxon rank-sum test statistic on the v-th endpoint is $W_v = \sum_{j=1}^{n_{2t}} R_{2jv}(t)$, test statistic

$$D(n_{1t}, n_{2t}) = \frac{2}{n_{2t}}\sum_{v=1}^{K} W_v - K(N_t + 1)$$

is a linear transformation of the sum of Wilcoxon rank-sum test statistics. The following theorem shows that sequentially computed test statistic $D(n_{1t}, n_{2t})$ forms an asymptotic Gaussian process. A proof is given in Appendix.

Theorem 3.1. *Let J be a K-dimensional vector with all elements equal to one, $\xi(x, y) = I[x < y] - I[x > y]$ where indicator $I[E]$ is defined by $I[E] = 1$ if event E is true, and $I[E] = 0$ otherwise. A, B, and C are three $K \times K$ matrices with (u, v) elements given by $a_{uv} = cov\Big(F_{2u}(X_{1u}), F_{2v}(X_{1v})\Big)$, $b_{uv} = cov\Big(F_{1u}(X_{2u}),\ F_{1v}(X_{2v})\Big)$, and $c_{uv} = cov\Big(\xi(X_{1u}, X_{2u}),\ \xi(X_{1v}, X_{2v})\Big)$ respectively, n_{it} is increasing in t. We assume that $\frac{n_{1t}}{n_{2t}} + \frac{n_{2t}}{n_{1t}} = O_p(1)$ is bounded for $t = 1, \ldots, T$ when $n_{1t} + n_{2t} = N_t \to \infty$. Then*
(i) $E[D(n_{1t}, n_{2t})] = n_{1t}K\bar{\theta}$,

$Var[D(n_{1t}, n_{2t})] = \mathcal{I}(n_{1t}, n_{2t}) = (4n_{1t}/n_{2t})J^T\{(n_{2t} - 1)A + (n_{1t} - 1)B + C/4\}J$.
(ii) For any $s < t$, $Cov\Big(D(n_{1s}, n_{2s}), D(n_{1t}, n_{2t})\Big) = \frac{4n_{1s}(n_{2t}-1)}{n_{2t}}J^TAJ + \frac{4n_{1s}(n_{1t}-1)}{n_{2t}}J^TBJ + \frac{n_{1s}}{n_{2t}}J^TCJ = \mathcal{I}(n_{1s}, n_{2s})(1 + \gamma)$, where

$$|\mathcal{I}(n_{1s}, n_{2s})\ \gamma| \le 8K^2\frac{n_{1s}(n_{2t} - n_{2s})}{n_{2s}n_{2t}} + 4K^2 n_{1s}\left|\frac{n_{1t}}{n_{2t}} - \frac{n_{1s}}{n_{2s}}\right|$$

and

$$|\gamma| \le \frac{1}{\min\{n_{1s}, n_{2s}\} - 1}\left|\left(1 - \frac{n_{2s}}{n_{2t}}\right)\left(1 - \frac{J^TCJ}{4J^T(A + B)J}\right) + \frac{(n_{2s}n_{1t} - n_{1s}n_{2t})J^TBJ}{n_{2t}J^T(A + B)J}\right| = \gamma^*.$$

(iii) Let $N_0 \equiv 0$ and $G \ge 1$ be any finite integer. If $\lim_{N_t\to\infty} n_{2t}/n_{1t} = r_0$ as $N_t = n_{1t} + n_{2t} \to \infty$ for $t = 1, \ldots, G$ and some non-zero constant r_0, then, the random process $\{n_{1t}^{-1/2}D(n_{1t}, n_{2t}), t = 1, ..., G\}$ converges to a Gaussian process with the first two moments given by (i) and (ii).

It is seen that, under the condition of Theorem 3.1(iii), the upper bound γ^* in Theorem 3.1(ii) converges to zero as sample size increases. This implies that the limiting process of $\{n_{1t}^{-1/2}D(n_{1t}, n_{2t}), t = 1, ..., G\}$ has the same variance-covariance structure as a Brownian motion measured at finite time points. We thus obtain the following:

Theorem 3.2. *Under the same assumptions as in Theorem 3.1(iv) and* $r_0 = \lim_{N_t\to\infty} n_{2t}/n_{1t}$ *for all* $t = 1, \ldots, G$, *process* $\{D(n_{1t}, n_{2t}),\ t = 1, \ldots, G\}$ *converges to a Brownian motion with drift* $\delta = \frac{r_0 K\bar{\theta}}{4J^T(r_0A+B)J}$ *at information times* $\{\mathcal{I}(n_{1t}, n_{2t}), t = 1, 2, \ldots, G\}$:

$$\begin{cases} E_A[D(n_{1t}, n_{2t})] = \delta\mathcal{I}(n_{1t}, n_{2t}), \\ Var[D(n_{1t}, n_{2t})] = \mathcal{I}(n_{1t}, n_{2t}) \\ Cov_A\Big(D(n_{1s}, n_{2s}), D(n_{1t}, n_{2t})\Big) = Var[D(n_{1s}, n_{2s})] \end{cases}$$

for any $s < t$. *The subscript A is used to denote "asymptotically true".*

Since $\lim_{N\to\infty} \frac{1}{n_{1t}}\mathcal{I}(n_{1t}, n_{2t}) = 4J^T(A + B/r_0)J$ under the assumption of Theorem 3.2, an immediate result is that sequence $\frac{1}{\sqrt{n_{1t}}}\{D(n_{1t}, n_{2t}) - n_{1t}K\bar{\theta}\}$ converges in distribution to $N(0, 4J^T(A + B/r_0)J)$.

Remark 1. Let $Z_t = D(n_{1t}, n_{2t})/\sqrt{\mathcal{I}(n_{1t}, n_{2t})}$. Under the assumption of Theorem 3.2, asymptotically, process $\{Z_t, t = 1, \ldots, G\}$ has the canonical joint distribution

(2)
$$\begin{cases} Z_t \sim N\Big(\delta\sqrt{\mathcal{I}(n_{1t}, n_{2t})}\ ,\ \ 1\Big), \\ Cov(Z_s, Z_t) = \sqrt{\mathcal{I}(n_{1s}, n_{2s})/\mathcal{I}(n_{1t}, n_{2t})} \quad \text{for any } s < t. \end{cases}$$

described in Jennison and Turnbull (2000).

Remark 2. When $K = 1$ and the endpoint has non-normal distribution, sequential test statistics are sometimes constructed using score function and Fisher information that are derived from the corresponding log-likelihood function. An unexpected property of this approach is that, as sample size increases, the numerical estimate of Fisher information (served as an information time) does not always increase. Sometimes the estimated Fisher information at a latter stage could be even smaller than its estimate at an earlier stage. Such an embarrassing scenario will never happen in our case because $\mathcal{I}(n_{1t}, n_{2t})$ is increasing in t. If we re-scale the process $D(n_{1t}, n_{2t})$ to $D_{t^*} = D(n_{1t}, n_{2t})/\sqrt{\mathcal{I}(n_{1_G}, n_{2_G})}$ using new time scale $t^* = \sqrt{\mathcal{I}(n_{1t}, n_{2t})/\mathcal{I}(n_{1G}, n_{2G})}$ for $t = 1, 2, \ldots, G$, then $0 < t^* \le 1$, and process

$$\{D_{t^*}, t^* = \sqrt{\mathcal{I}(n_{11}, n_{21})/\mathcal{I}(n_{1G}, n_{2G})}\ , \ldots, 1\}$$

is asymptotically a Brownian motion with drift $\delta^* = \delta\sqrt{\mathcal{I}(n_{1G}, n_{2G})}$ in information time interval $t^* \in [0, 1]$: $D_{t^*} \sim N(\delta^* t^*, t^*)$. Here, t^* is sometimes called an information fraction. Consequently, this process can be used in conjunction with many existing univariate multistage sequential procedures to yield multistage test procedures for comparing multidimensional outcomes (see simulation of next section). For example, the conditional likelihood ratio based SCPRT in the stochastic process setting for testing (1), or equivalently, $H_0 : \delta^* \le \delta_0^*$ vs. $H_a : \delta^* > \delta_0^*$, on a set of discrete information time points, $(t_1^*, t_2^*, ..., t_G^*)$, using a G-stage design. Here $\delta^* = \delta_{\bar{\theta}}^* = \frac{r_0 K\bar{\theta}}{4J^T(r_0A+B)J}\sqrt{\mathcal{I}(n_{1G}, n_{2G})}$, t_g^* (≤ 1) is the information fraction $(g = 1, ..., G)$, $t_{g_1}^* \le t_{g_2}^*$ for $g_1 < g_2$, and $t_G^* = 1$. In multistage test, if $D_{t_g^*}$ hits the upper stopping boundary defined by (a_g, b_g), i.e., $D_{t_g^*} \ge b_g$, the null hypothesis is rejected in favor of H_a; If it hits the lower stopping boundary $(D_{t_g^*} \le a_g)$, the null hypothesis is accepted; otherwise, the trial continues. At the final Gth stage, we have $a_G = b_G$ to terminate the trial. Let $L(D_{t^*}|D_1)$ be the likelihood function of D_{t^*} given D_1. The conditional maximum likelihood ratio is given by

$$\lambda(t^*, D_{t^*}|D_1) = \frac{max_{z>z_\alpha} L(D_{t^*}|D_1 = z)}{max_{z\le z_\alpha} L(D_{t^*}|D_1 = z)}, \tag{3}$$

where z_α is the $(1-\alpha)$th percentile of a standard normal distribution. As shown earlier, the sequential process $\{D_{t^*}\}$ is approximately a Brownian motion process. Based on Xiong et al. (2003), the lower and upper stopping boundaries are:

(4)
$$a_g = z_\alpha + \{2at_g(1-t_g)\}^{1/2},\ \text{and}\ b_g = z_\alpha - \{2bt_g(1-t_g)\}^{1/2},$$

where a and b are determined by the probability ρ of discordance between the decisions from interim analysis and final analysis. This probability is chosen as a design parameter. Typically, $\rho = 2\%$ is selected (Tan, Xiong, Kutner, 1998).

Remark 3. In addition to establishing the large sample property, the upper bounds of $|\mathcal{I}(n_{1s}, n_{2s})\ \gamma|$ and γ^* in Theorem 3.1 allow us to assess how well the Brownian motion approximation is when the sample size is only moderate. The regularity conditions $Var\{F_{1v}(X_{2v})\} > 0$ and $Var\{F_{2v}(X_{1v})\} > 0$ $(v = 1, \ldots, K)$ require data from both treatment groups have the same distribution support. This is true for almost all clinical trial applications. Under such regularity conditions, $4J^T(A + B)J$ is bounded away from zero since both A and B are positive definite matrices. The quantity $J^T CJ$ is bounded by K^2 since all elements in matrix C are bounded by 1. For small to moderate sample sizes, e.g., both n_{1s} and n_{2s} are greater than 30, the upper bound γ^* of $|\gamma|$ generally is a small number. Simulation in Table 1 shows the magnitude of γ^* under different distributions and correlation structures among K endpoints. The data were generated using the same method as described in the next section. Equal correlation among K endpoints were generated by setting $r_1 = \cdots = r_K = 0.5$, and unequal correlation among K endpoints were generated by setting $r_1 = \cdots = r_{K-1} = 0.5$ and $r_K = 1$. These upper bounds also provide theoretical bases why the method would work even for small to moderate sample sizes.

Table 1. Simulation evaluation of the upper bound γ^* in Theorem 1(ii), $n_{1t} = 2n_{1s}$, $\Delta_r = \frac{n_{1s}}{n_{2s}} - \frac{n_{1t}}{n_{2t}}$, and 1000 simulations

Distribution	$r_1, \cdots, r_K$	K	$n_{1s} = n_{2s}$	γ^*	
				$\Delta_r = 0$	$\Delta_r = 10\%$
Uniform		2	30	0.0060	0.0529
			100	0.0017	0.0475
	equal	5	30	0.0031	0.0503
			100	0.0009	0.0469
	unequal	5	30	0.0032	0.0499
			100	0.0009	0.0469
Normal		2	30	0.0060	0.0537
			100	0.0018	0.0477
	equal	5	30	0.0034	0.0508
			100	0.0001	0.0471
	unequal	5	30	0.0034	0.0506
			100	0.0010	0.0470
Exponential		2	30	0.0062	0.0535
			100	0.0018	0.0478
	equal	5	30	0.0037	0.0510
			100	0.0011	0.0470
	unequal	5	30	0.0038	0.0514
			100	0.0011	0.0472

4. SIMULATION

The goal of our simulation is to evaluate the performance of multistage test constructed through combining existing univariate sequential stopping boundaries with the proposed rank based stochastic process for multi-dimensional endpoints (see Remarks 1 and 2 of the previous section). In particular, we evaluated how type I and type II errors were controlled when the sample size was small or moderate. We present simulation results for $K = 4$ and 8 endpoints. Results for other choices of K are similar and are not presented here.

The simulation study was modeled after the Parkinson disease trial and designed to evaluate the operating characteristics (e.g., type I and II errors) of the resulting sequential test for testing hypothesis (1) and the finite-sample performance. Let $x_{iju} = r_u y_{ij0} + \sqrt{1 - r_u^2}\, y_{iju}$, where $i = 1, 2$; $j = 1, \cdots, n_i$; and $u = 1, \cdots, K$. The y_{ij0} and y_{iju} were generated independently from exponential and beta distributions, respectively. We considered type I error $\alpha = 0.05$ and type II error $\beta = 0.10$. We set $r_1 = \cdots = r_K = r = 0.5$ or 0.8. Since many Parkinson's disease clinical endpoints are ordinal with 5 different levels ("normal", "mild", "moderate", "severe", and "most serious"), one or two endpoints in each simulated dataset were further discretized into 5 levels: -2,-1, 0, 1 and 2. To introduce unequal covariances between the two groups, the cut-off values used in the two groups were not the same. For exponential distribution, y_{ij0} and y_{iju} under the null hypothesis were generated from the standard exponential distribution Exp(1) for both groups. Under the alternative hypothesis, they were generated from Exp(1) for $i = 1$ and Exp(3/4) for $i = 2$ respectively. For beta distribution, y_{ij0} and y_{iju} under the null hypothesis were generated from Beta(0.43, 1) distribution for both groups. Under the alternative hypothesis, they were generated from Beta(0.43, 1) for $i = 1$ and Beta(0.55, 1) for $i = 2$ respectively. We fixed randomization ratio $r_0 = 1$. When the total number of interim analysis looks (stages) $G = 1$, it reduces to the fixed sample size test. In each setting, the empirical power and type I error rate of the sequential test constructed by combining SCPRT or O'Brien-Fleming stoping boundaries with process $\{D_{t^*}, t^* = \sqrt{\mathcal{I}(n_{11}, n_{21})/\mathcal{I}(n_{1G}, n_{2G})}\ , \ldots, 1\}$ were computed in single stage, 2-stage and 3-stage designs, respectively. All simulations were performed in R, with 10,000 replications.

We used two sequential stopping boundaries to illustrate: the O'Brien-Fleming stopping boundary and the SCPRT stopping boundaries (Xiong, 1995; and Xiong et al., 2003). The latter is derived by requiring the multistage test to have a negligible discordance probability, namely, the probability that the results based on interim data would be reversed should the analysis be performed with the data from all stages. These two stopping boundaries were chosen because they both have the required conservatism for not rejecting the null hypothesis too early, and their required maximum sample sizes are not much larger than that of the conventional single stage test. The O'Brien-Fleming stopping boundary is quite conservative at early stages while the SCPRT is not so conservative at early stages but remains to be conservative throughout all stages much like the Haybittle-Peto procedure. A detailed comparison of the two stopping boundaries can be found in Tan et al. (1998) and Freidlin et al. (1999).

The multistage tests considered in the simulation include both two and three-stage tests. We used $n_{11} = n_{21}$, $n_{12} = n_{22}$, and $n_{13} = n_{23}$ to denote the cumulative sample sizes per group at stages 1, 2, and 3, respectively (Table 2). In addition, the type I and II errors for a single stage test are also included as a reference. The simulation study demonstrates that both the type I and II errors for the multistage tests are reasonably preserved, namely, similar to those of a single stage test (Table 2), although the statistical power for the case of 4 endpoints are slightly (5%) attenuated.

5. APPLICATION TO PARKINSON DISEASE CLINICAL TRIALS

Example 1: The multi-center randomized controlled clinical trial, Deprenyl and Tocopherol Antioxidative Therapy of Parkinsonism (DATATOP), is the NIH sponsored first landmark neuroprotective treatment study. The DATATOP was carried out in 1987-1989 by the Parkinson Study Group to determine whether long-term therapy with deprenyl or tocopherol would postpone the initiation of levodopa therapy for patients with early, untreated PD (The Parkinson Study Group, 1989). Eight hundred patients were randomly

Table 2. Simulation using SCPRT and O'Brien-Fleming stopping boundaries

Distribution	θ	r	K	Stages $G(n_{11},\ldots,n_{1G})$	Type I error $\alpha=0.05$		Power $1-\beta=0.9$	
					SCPRT	OBF	SCPRT	OBF
Exponential	0.22	0.5	4	1 (74)	0.0456	0.0456	0.8539	0.8539
				2 (37, 74)	0.0456	0.0428	0.8539	0.8493
				3 (25, 50, 74)	0.0454	0.0388	0.8529	0.8409
			8	1 (67)	0.0381	0.0381	0.8921	0.8921
				2 (33, 67)	0.0381	0.0355	0.8921	0.8883
				3 (22, 44, 67)	0.0381	0.0339	0.8915	0.8823
		0.8	4	1 (101)	0.0403	0.0403	0.8496	0.8496
				2 (50, 101)	0.0403	0.0379	0.8496	0.8453
				3 (34, 68, 101)	0.0403	0.0356	0.8491	0.8371
			8	1 (98)	0.0371	0.0371	0.8787	0.8787
				2 (49, 98)	0.0371	0.0352	0.8787	0.8741
				3 (33,66,98)	0.037	0.0334	0.8784	0.8659
Beta	0.1563	0.5	4	1(147)	0.0459	0.0459	0.8615	0.8615
				2 (74, 147)	0.0459	0.0438	0.8615	0.8566
				3 (39, 78, 147)	0.0459	0.0398	0.8615	0.8483
			8	1 (132)	0.0393	0.0393	0.9094	0.9094
				2 (66, 132)	0.0393	0.0367	0.9094	0.9054
				3 (44, 88, 132)	0.0391	0.0336	0.9087	0.9004
		0.8	4	1 (199)	0.0373	0.0373	0.8262	0.8262
				2 (100, 199)	0.0373	0.0358	0.8262	0.8212
				3 (66, 132, 199)	0.0373	0.0326	0.826	0.8126
			8	1 (193)	0.0354	0.0354	0.8458	0.8458
				2 (97, 193)	0.0354	0.0336	0.8458	0.8415
				3 (64, 126, 193)	0.0353	0.0313	0.8453	0.8319

assigned to one of four treatment groups: placebo, active tocopherol and deprenyl placebo, active deprenyl and tocopherol placebo, or both active drugs. The primary endpoint in DATATOP was the development of disability necessitating the introduction of levodopa therapy. After a mean follow up of 14 ± 6 months, deprenyl was found to not only significantly delay the need of levodopa, but also significantly slow the progression measured by the total Unified Parkinson's Disease Rating Scale (UPDRS) and its subscales. No tocopherol effect was found in the study. Since the decision to initiating levodopa is not a clinical scale and was later found to be confounded with many factors not related to the disease progression such as patient's social life, loss of job, and family relations (The Parkinson Study Group, 1993), we retrospectively analyzed its 3 key movement dysfunction measures from UPDRS sub-scales: mentation, motor, and activities of daily living (ADL). At the end of two-year study, patients receiving deprenyl were found to have significant better measures on these endpoints as comparing to patients not receiving deprenyl. Our objective was to see whether treatment benefit on these 3 endpoints could be identified earlier by a multistage test with fewer sample size, as compared to the original analysis performed when data from all 800 patients have been collected.

We used a multistage rank-based test with interim analyses performed at six and nine months, respectively, after the first patient was enrolled. The three primary endpoints are mentation, motor and activities of daily living scores from the UPDRS. The single stage test has a significance level of 5% and power of 80% to detect a difference of $\bar{\theta}$ as shown in Table 3. A three multistage sequential tests were derived by combining SCPRT with the proposed rank based stochastic process (Table 3). The calendar time to perform interim analysis was determined by the information fraction t^* described in Remark 2 of Section 3 and patient's randomization date. The SCPRT design has the same type I and II error rates as the single stage test, and it has a discordance probability of less than 1%. With the availability of this multistage test, Table 3 suggests a pronounced early treatment effect since the first stage data has already demonstrated the significance for both the six months and the nine months analyses: the stochastic process hits the early stopping upper bound for efficacy, and the chance of a potential reversion of the conclusion at the end of the study is negligible.

Example 2: To illustrate the proposed method in a randomized trial with moderate sample sizes, we used data from the multicenter controlled clinical trial of Coenzyme Q_{10} in early Parkinson's disease (QE2 trial). Same as the previous example, our goal was to show whether the emerging trend for efficacy was sufficient for an early conclusion for efficacy when using the proposed multistage test. The trial was conducted in 1999-2001 to determine whether Coenzyme Q_{10} could slow the functional decline in Parkinson's dis-

Table 3. Parkinson Disease Trial multistage Test Results

G	Stage	Sample Size	$\bar{\theta}$	Test Statistic	Information Fraction t^*	Lower Boundary	Upper Boundary
		.	Analysis based on 6-month improvement with total sample size 768				
1	1	$(n_{11}, n_{21}) = (387, 381)$	0.2249	8.1150	1.00	1.96	1.96
2	1	$(n_{11}, n_{21}) = (180, 190)$	0.2280	3.8267	0.45	−0.13	1.91
	2	$(n_{12}, n_{22}) = (387, 381)$	0.2249	8.1150	1.00	1.96	1.96
3	1	$(n_{11}, n_{21}) = (120, 130)$	0.2189	2.4498	0.31	−0.46	1.66
	2	$(n_{12}, n_{22}) = (255, 262)$	0.2252	5.3546	0.66	0.19	2.38
	3	$(n_{13}, n_{23}) = (387, 381)$	0.2248	8.1150	1.00	1.96	1.96
		.	Analysis based on 9-month improvement with total sample size 663				
1	1	$(n_{11}, n_{21}) = (355, 308)$	0.2175	7.2230	1.00	1.96	1.96
2	1	$(n_{11}, n_{21}) = (170, 160)$	0.2286	3.6341	0.46	−0.13	1.92
	2	$(n_{12}, n_{22}) = (355, 308)$	0.2175	7.2230	1.00	1.96	1.96
3	1	$(n_{11}, n_{21}) = (110, 105)$	0.1966	2.0225	0.31	−0.46	1.67
	2	$(n_{12}, n_{22}) = (225, 210)$	0.1929	4.0597	0.63	0.12	2.34
	3	$(n_{13}, n_{23}) = (355, 308)$	0.2175	7.2230	1.00	1.96	1.96

ease (Shults, et al., 2002). Sixteen and 64 patients were randomized to receive placebo or Coenzyme Q_{10} respectively, and were followed for up to 16 months or until disability requiring treatment with levodopa had developed. Treatment efficacy was measured by mental (mentation), motor, ADL (average daily living) subscales of the UPDRS, and the Schwarb and England ADL (SEADL) score. The primary outcome was the change in the total UPDRS score (a sum of mental, motor and ADL) from baseline to the last visit in 16 months. At the end of the study, the investigators concluded a significant improvement in the total UPDRS for patients receiving Coenzyme Q_{10} with p-value of 0.09 (Shults et al., 2002).

Instead of testing treatment effect through a sum of mental, motor and ADL scores in UPDRS, we evaluated all 3 components jointly. We derived a two-stage test using a significance level of 5% and power of 80%. The first stage analysis was performed using the first half of the patients with corresponding information fraction $t_1^* = 0.602$. The critical values of the stopping boundaries are 2.001 and 1.6546 for stage 1 and stage 2, respectively. The observed test statistic at the first stage analysis was $D_{t_1^*} = 0.454 < 2.001$, which suggests that the trial should continue to the second stage. At the end of stage 2 analysis, $D_{t_2^*} = 1.66$ was obtained. This is slightly greater than the critical level of 1.6546. Thus, the multistage test gives a stronger evidence of treatment benefit as compared to the original findings from QE2 trial investigators.

6. DISCUSSION

Clinical trials are often conducted to test whether a new treatment could improve clinical outcomes as compared to the standard of care. It is well recognized that many treatments have nuisance effect even if they do not show the desired clinical benefit. In recent years, multiple primary endpoints are increasly used in treatment comparison to obtain a more comprehensive assessment of treatment benefit. However, most multivariate tests require identical distribution under the null hypothesis which is not applicable when treatment only has nuisance effect without the desired effect. For example, a treatment could change the distribution shape or covariance structure among the primary endpoints without changing their mean values. We have addressed these data challenges in the framework of the Behrens-Fisher problem, and proposed a rank-based multistage (sequential) test to accommodate the need for interim analyses in clinical trials. We showed that the stochastic process formed from a sequentially computed global statistical test (GST) statistic converges to a Gaussian process. Furthermore, if the sample size ratios between the two groups are about the same in all stages, we showed that the process can be approximated by a Brownian motion measured at finite time points. We have derived a more accurate inequality for finite sample case and showed that the asymptotic results hold for studies with small to moderate sample sizes. With rich available resources in univariate sequential designs and property in (2), this strategy greatly simplifies the development of stopping rules for multidimensional endpoints with desired operating characteristics. The asymptotic properties in Theorem 3.1 provides a foundation for further developments of sequential tests for multi-dimensional endpoints.

An important extension of current work is to the case when there are confounding covariates and missing observations, or some of the endpoints are survival endpoints. These works are currently in progress and will be reported later. When $K = G = 1$, our test based on $Z_t = D(n_{1t}, n_{2t})/\sqrt{\mathcal{I}(n_{1t}, n_{2t})}$ is a modified Wilcoxon test for the Behrens-Fisher problem that controls type I error asymptotically. This avoids the problem of non-robustness seen in Wilcoxon test (Fagerland, Morten, and Sandvik, 2009). Although our applications have focused on Parkinson disease trials, the method can be used widely in clinical research on stroke, dermatology, multiple sclerosis, asthma, rheumatoid arthritis, and potentially in early stage trials on cancer immunotherapy where multiple immune-response monitoring

parameters are frequentily used as intermediate markers for clinical response.

Zhao, Hu, and Lagakos (2009) argued that, for any J-dimensional parameter (or function) $\mu_0(\cdot)$ of interest and K total number of looks, if we can find consistent estimate $\hat{\mu}_n(\cdot;t)$ such that $\sqrt{n}\{\hat{\mu}_n(s;T_k)-\mu_0(s)\}$ $(k=1,2,\ldots,K)$ converges weakly to a Gaussian process, we can always construct sequential stopping boundaries for $\mu_0(\cdot)$. This is true for the construction of essentially all sequential stopping boundaries when sample size is large enough. However, it is not clear how large the sample size should be in order to have a good numerical approximation to the Gaussian process. In many applications, such an asymptotic Gaussian process may not be readily identifiable. Moreover, numerical computation of critical values of the stopping boundaries is often not trivial when the sample size is relatively small or moderate.

When planning a group sequential trial with multiple primary endpoints, the required sample size depends on the information matrix through $J^{T}(r_0A+B)J$, where $r_0=n_{2t}/n_{1t}$ $(t=1,\ldots,G)$ is a fixed non-zero randomization ratio. Our result can be generalized to the class of multistage tests that allow the test critical levels at each stage and the sample size to depend on the updated estimate of $J^{T}(r_0A+B)J$, which is usually not well estimated at the planning stage of the study. If this value is underestimated, so is the sample size, leading to an underpowered test. Similarly, the updated estimate of $J^{T}(r_0A+B)J$ will affect the critical level of the test at each stage (sequential stopping boundary). Various adaptive designs in sample size re-estimation have been proposed. For example, Wittes and Brittain (1990) and Gould and Shih (1992) proposed an internal pilot data for sample re-estimation where no early stopping is considered, whereas Gould and Shih (1998) allowed early stopping. Denne and Jennison (2000) used Stein's two-stage procedure (Stein, 1945) and its generalization to update sample sizes for group sequential tests (see Chapter 7 of Jennison and Turnbull (2000) for flexible monitoring of group sequential tests). Xiong et al. (2003) proposed a nuisance parameter adaptive design using the updated variance based on data from previous stage by using the power function approximated by a Taylor expansion.

Like the rank-sum test proposed by O'Brien (1984), our hypothesis (1) and multistage procedure are constructed for testing treatment's global benefit across multiple equally important endpoints in clinical trial setting. If the treatment has strong beneficial effects on half of the endpoints and equally strong detrimental effects on the remaining half of the endpoints, our test will lose power. This is a desired property of our proposed test because we do not want to have a high power to claim benefit for such a treatment. If the goal is to test whether the treatment has any type of effect (positive or negative), Liu et al. (2010) proposed an omnidirectional test based on the maximum rank sum difference T_{max} among K endpoints. They have showed, when treatment has strong effects in both positive and negative directions, or it has a strong positive effect on one endpoint but trivial effects on all others, a test based on T_{max} gives a much higher power to reject the null hypothesis than the tests proposed by O'Brien (1984) and Huang et al. (2005).

Our multistage test, suitable for the Behrens-Fisher problem, allows unknown and possibly unequal variances between the two groups. In fact, our test can still be used even when a treatment may demonstrate some unexpected effects or nuisance effects. For example, almost all Parkinson disease agents are known to have certain symptomatic (nuisance) effects. If the goal is to test treatment's neuroprotective effect in a randomized placebo-controlled clinical trial (like DATATOP and NET-PD trials), the null hypothesis of equal joint distribution of multiple endpoints between the two groups is not appropriate to use. In our simulation, the symptomatic effect is demonstrated through the change in distribution shape (or more specifically, the change in variance). Upon the rejection of the null hypothesis, a claim of global treatment benefit can be made. If we are further interested to know which endpoints have stronger effects, we can continue to test treatment effect on each single endpoint. These tests will be considered as secondary analyses and do not affect the type I error of the primary test for hypothesis (1).

APPENDIX. THEOREM PROOF

Proof of Theorem 3.1.

(i) is a direct consequence of $E\xi(x_{1iv},x_{2jv})=\theta_v$, $D(n_{1t},n_{2t})=\frac{2}{n_{2t}}\sum_{v=1}^{K}W_{vt}-K(N_t+1)$, and $Var[(W_{1t},\ldots,W_{Kt})^{T}]=n_{1t}n_{2t}\{(n_{2t}-1)A+(n_{1t}-1)B+C/4\}$, where $W_{vt}=\sum_{j=1}^{n_2}R_{2jv}(t)$ is the Wilcoxon rank-sum test statistic on the v-th endpoint at the t-th interim analysis.

(ii) For any $s<t$,

$$
\begin{aligned}
&cov\Big(D(n_{1t},n_{2t})-D(n_{1s},n_{2s}),\ \ D(n_{1s},n_{2s})\Big)\\
&=cov\left(\frac{2}{1+r_0}(\bar{R}_{22}-\bar{R}_{12})-\frac{2}{1+r_0}(\bar{R}_{21}-\bar{R}_{11}),\right.\\
&\quad\left.\frac{2}{1+r_0}(\bar{R}_{21}-\bar{R}_{11})\right)\\
&=\frac{1}{n_{2s}n_{2t}}J^{T}\Big\{4n_{1s}n_{2s}(n_{2t}-n_{2s})A\\
&\quad+4n_{1s}n_{2s}(n_{1t}-n_{1s})B\Big\}J\\
&\quad+\frac{4n_{1s}^2(n_{2s}-n_{2t})}{n_{2t}N_{1s}}J^{T}\Sigma(n_{1s},n_{2s})J\\
&=n_{1s}(n_{2t}-n_{2s})n_{2s}n_{2t}J^{T}(4A+4B-C)J\\
&\quad+\frac{4n_{1s}(n_{2s}n_{1t}-n_{1s}n_{2t})}{n_{2s}n_{2t}}J^{T}BJ.
\end{aligned}
$$

Now we have

$$
\begin{aligned}
&cov\Big(D(n_{1t},n_{2t}),\ \ D(n_{1s},n_{2s})\Big)\\
&= Var[D(n_{1s},n_{2s})]+\\
&cov\Big(D(n_{1t},n_{2t})-D(n_{1s},n_{2s}),\ \ D(n_{1s},n_{2s})\Big)\\
&= \frac{4n_{1s}}{n_{2s}}J^T\{(n_{2s}-1)A+(n_{1s}-1)B+C/4\}J\\
&\quad+\frac{4n_{1s}(n_{2t}-n_{2s})}{n_{2s}n_{2t}}J^TAJ\\
&\quad+\frac{4n_{1s}(n_{2s}n_{1t}-n_{1s}n_{2t}-n_{2s}+n_{2t})}{n_{2s}n_{2t}}J^TBJ\\
&\quad-\frac{n_{1s}(n_{2t}-n_{2s})}{n_{2s}n_{2t}}J^TCJ\\
&= \frac{4n_{1s}(n_{2t}-1)}{n_{2t}}J^TAJ+\frac{4n_{1s}(n_{1t}-1)}{n_{2t}}J^TBJ\\
&\quad+\frac{n_{1s}}{n_{2t}}J^TCJ\\
&= \mathcal{I}(n_{1s},n_{2s})(1+\gamma),
\end{aligned}
\tag{5}
$$

where

$$
\begin{aligned}
&|\mathcal{I}(n_{1s},n_{2s})\,\gamma|\\
&= \left|\frac{n_{1s}(n_{2t}-n_{2s})}{n_{2s}n_{2t}}J^T(4A+4B-C)J\right.\\
&\quad\left.+\frac{4n_{1s}(n_{2s}n_{1t}-n_{1s}n_{2t})}{n_{2s}n_{2t}}J^TBJ\right|\\
&\le \left|\frac{n_{1s}(n_{2t}-n_{2s})}{n_{2s}n_{2t}}J^T(4A+4B-C)J\right|\\
&\quad+\left|\frac{4n_{1s}(n_{2s}n_{1t}-n_{1s}n_{2t})}{n_{2s}n_{2t}}J^TBJ\right|\\
&\le 8K^2\frac{n_{1s}(n_{2t}-n_{2s})}{n_{2s}n_{2t}}+4K^2n_{1s}\left|\frac{n_{1t}}{n_{2t}}-\frac{n_{1s}}{n_{2s}}\right|
\end{aligned}
$$

and

$$
\begin{aligned}
&|\gamma|\\
&= \left|\frac{n_{1s}(n_{2t}-n_{2s})}{n_{2s}n_{2t}}J^T(4A+4B-C)J\right.\\
&\quad\left.+\frac{4n_{1s}(n_{2s}n_{1t}-n_{1s}n_{2t})}{n_{2s}n_{2t}}J^TBJ\right|\\
&\quad\times\Big((4n_{1s}/n_{2s})J^T\{(n_{2s}-1)A+(n_{1s}-1)B+C/4\}J\Big)^{-1}\\
&\le \frac{\left|\frac{(n_{2t}-n_{2s})}{n_{2t}}J^T(4A+4B-C)J+\frac{4(n_{2s}n_{1t}-n_{1s}n_{2t})}{n_{2t}}J^TBJ\right|}{4J^T\{(n_{2s}-1)A+(n_{1s}-1)B\}J}\\
&\le \frac{1}{\min\{n_{1s},n_{2s}\}-1}\left|\left(1-\frac{n_{2s}}{n_{2t}}\right)\left(1-\frac{J^TCJ}{4J^T(A+B)J}\right)\right.\\
&\quad\left.+\frac{(n_{2s}n_{1t}-n_{1s}n_{2t})J^TBJ}{n_{2t}J^T(A+B)J}\right|\\
&= \gamma^*.
\end{aligned}
$$

(iii) Without the loss of generality, we only need to show that, for any $s<t$, vector $(n_{1s}^{-1/2}D(n_{1s},n_{2s}),n_{1t}^{-1/2}D(n_{1t},n_{2t}))^T$ converges to a bivariate normal distribution as $N_s\to\infty$ and $N_t-N_s\to\infty$. For higher finite dimension case, the proof is similar. The asymptotic normality of $n_{1t}^{-1/2}D(n_{1t},n_{2t})$ was shown by Huang et al. (2008). For joint asymptotic normality, it suffices to show that $Z=\lambda_{10}n_{1s}^{-1/2}D(n_{1s},n_{2s})+\lambda_{20}n_{1t}^{-1/2}D(n_{1t},n_{2t})$ has asymptotic normal distribution for any constants λ_{10} and λ_{20}. Define $\lambda_1=\lambda_{10}n_{1s}^{-1/2}$, $\lambda_2=\lambda_{20}n_{1t}^{-1/2}$, and

$$
\begin{aligned}
U=&-K(\lambda_1+\lambda_2)n_{1s}(\bar\theta-1)\\
&-2(\lambda_1+\lambda_2)\sum_{l=1}^{n_{1s}}\sum_{v=1}^{K}F_{2v}(x_{1lv})\\
&-2\lambda_2\sum_{l=n_{1s}+1}^{n_{1t}}\sum_{v=1}^{K}F_{2v}(x_{1lv})\\
&+2\left(\lambda_1\frac{n_{1s}}{n_{2s}}+\lambda_2\frac{n_{1t}}{n_{2t}}\right)\sum_{l=1}^{n_{2s}}\sum_{v=1}^{K}F_{1v}(x_{2lv})\\
&+2\lambda_2\frac{n_{1t}}{n_{2t}}\sum_{l=n_{2s}+1}^{n_{2t}}\sum_{v=1}^{K}F_{1v}(x_{2lv})
\end{aligned}
$$

From Central Limiting Theorem, U has asymptotic normal distribution with $Var[U]=4(\lambda_1^2n_{1s}+2\lambda_1\lambda_2n_{1s}+\lambda_2^2n_{1t})J^TAJ+4\left(\frac{\lambda_1^2n_{1s}^2}{n_{2s}}+2\lambda_1\lambda_2\frac{n_{1s}n_{1t}}{n_{2t}}+\frac{\lambda_2^2n_{1t}^2}{n_{2t}}\right)J^TBJ$, On the other hand,

$$
\begin{aligned}
&Var[Z]\\
&= Var[\lambda_1\{D(n_{1s},n_{2s})-n_{1s}J^T\theta\}\\
&\quad+\lambda_2\{D(n_{1t},n_{2t})-n_{1t}J^T\theta\}]\\
&= \lambda_1^2\mathcal{I}(n_{1s},n_{2s})+\lambda_2^2\mathcal{I}(n_{1t},n_{2t})+2\lambda_1\lambda_2\{\mathcal{I}(n_{1s},n_{2s})\\
&\quad+O_p(1)\}\\
&= (\lambda_1^2+2\lambda_1\lambda_2)\mathcal{I}(n_{1s},n_{2s})+\lambda_2^2\mathcal{I}(n_{1t},n_{2t})+\lambda_1\lambda_2O_p(1)\\
&= (\lambda_1^2+2\lambda_1\lambda_2)\left\{\frac{4n_{1s}}{n_{2s}}J^T(n_{2s}A+n_{1s}B)J+O_p(1)\right\}\\
&\quad+\lambda_2^2\left\{\frac{4n_{1t}}{n_{2t}}J^T(n_{2t}A+n_{1t}B)J+O_p(1)\right\}+\lambda_1\lambda_2O_p(1)\\
&= 4(\lambda_1^2n_{1s}+2\lambda_1\lambda_2n_{1s}+\lambda_2^2n_{1t})J^TAJ\\
&\quad+4\left\{(\lambda_1^2+2\lambda_1\lambda_2)\frac{n_{1s}^2}{n_{2s}}+\lambda_2^2\frac{n_{1t}^2}{n_{2t}}\right\}J^TBJ\\
&\quad+(\lambda_1+\lambda_2)^2O_p(1)
\end{aligned}
$$

Comparing $Var[U]$ and $Var[Z]$, we have

$$
\begin{aligned}
Var[Z]=&\,Var[U]+8\lambda_1\lambda_2n_{1s}\left(\frac{n_{1s}}{n_{2s}}-\frac{n_{1t}}{n_{2t}}\right)J^TBJ\\
&+(\lambda_1+\lambda_2)^2O_p(1)
\end{aligned}
$$

$$= Var[U] + 8\lambda_{10}\lambda_{20}\sqrt{\frac{n_{1s}}{n_{1t}}}\left(\frac{n_{1s}}{n_{2s}} - \frac{n_{1t}}{n_{2t}}\right) J^T B J$$
$$+ O_p\left(\frac{1}{n_{1s}} + \frac{1}{n_{1t}} + \frac{1}{\sqrt{n_{1s}n_{1t}}}\right)$$
$$= Var[U] + o_p(1) + O_p\left(\frac{1}{n_{1s}} + \frac{1}{n_{1t}} + \frac{1}{\sqrt{n_{1s}n_{1t}}}\right).$$

This implies that $\lim_{N_s,N_t\to\infty}(Var[Z] - Var[U]) = 0$. Since $E[ZU] = E[U^2]$ and $E(Z-U)^2 = E[Z]^2 - E[U]^2 - 2E[(Z-U)U] = Var[Z] - Var[U] \to 0$ as $N_s, N_t \to \infty$, we see that, for any constant $\epsilon > 0$ and as $N_t \to \infty$,

$$P\{|Z - U| > \epsilon\} \le E(Z-U)^2/\epsilon^2 \to 0.$$

That is, $Z - U$ converges in probability to zero. Since $Z = U + (Z - U)$, using Slutsky Theorem, Z and U have the same asymptotic distribution. Thus Z also has asymptotic normal distribution.

ACKNOWLEDGEMENTS

We want to thank the editor and two associate editors for their insightful review and helpful comments that greatly improve the presentation of this paper, the DATATOP trial steering committee, the QE2 trial committee, and Parkinson Study Group for providing the data, and Ms. Yang Yang for statistical computing assistance. This work was supported in part by National Institutes of Health Grants 1R21NS043569-01, P50CA103175, MCRF-FHA05CRF, P30CA006973, and R01CA164717.

Received 1 May 2015

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300. MR1325392

Bennett, B. M. (1951). Note on a solution of the generalized Behrens-Fisher problem, *Ann. Inst. Statist. Math.*, 2:87–90. MR0041399

Brunner, E., Munzel, U., and Puri, M. L. (2002). The multivariate nonparametric Behrens-Fisher problem, *Journal of Statistical Planning and Inference*, 108:37–53. MR1947390

Christensen, W. F. and Rencher, A. C. (1997). A comparison of type i error rates and power levels for seven solutions to the multivariate Behrens-Fisher problem, *Communications in Statistics: Simulation and Computation*, 26:1251–1273.

Denne, J. S. and Jennison, C. (2000). A group sequential t-test with updating of sample size, *Biometrika*, 87:125–134. MR1766833

Fagerland, Morten W. and Sandvik, Leiv (2009). The Wilcoxon-Mann-Whitney test under scrutiny, *Statistics in Medicine*, 28:1487–1497. MR2649707

Freidlin, B., Korn, E. L., and George, S. L. (1999). Data monitoring committees and interim monitoring guidelines, *Controlled Clinical Trials*, 20:395–407.

Gould, A. L. and Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes withunkno wn variance, *Communication in Statistics A*, 21:2833–2853.

Gould, A. L. and Shih, W. J. (1998). Modifying the design of ongoing trials without unblinding, *Statistics in Medicine*, 17:89–100.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika*, 75:800–802. MR0995126

Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics*, 6(2):65–70. MR0538597

Huang, P., Tilley B., Woolson R., and Lipsitz, S. (2005). Adjusting O'Brien's test to control type I error for the generalized nonparametric Behrens-Fisher problem, *Biometrics*, 61:532–539. MR2140925

Huang, P., Woolson, R. F., and O'Brien, P. C. (2008). A rank-based sample size method for multiple outcomes in clinical trials, *Statistics in Medicine*, 27(16):3084–3104. MR2516885

Huang, P., Goetz, C. G., Woolson, R. F., Tilley, B., Kerr, D., Palesch, Y., Elm, J., Ravina, B., Bergmann, K. J., Kieburtz, K., and Parkinson Study Group (2009). Using global statistical tests in long-term Parkinson's disease clinical trials, *Movement Disorders*, 24(12):1732–9.

Huang, P., Ou, A., Piantadosi, S., and Tan, M. (2014). Formulating appropriate statistical hypotheses for treatment comparison in clinical trial design and analysis, *Contemporary Clinical Trials*, 39:294–302.

James, G. S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown, *Biometrika*, 41:19–43. MR0068189

Jennison, Christopher and Turnbull, Bruce W. (2000). *Group sequential methods with applications to clinical trials*, CRC Press Inc. (Boca Raton, FL) MR1710781

Kosorok, M. R., Yuanjun, S., and DeMets, D. L. (2004). Design and analysis of group sequential clinical trials with multiple primary endpoints, *Biometrics*, 60:134–145. MR2043628

Kroenke, K., Bair, M. J., Damush, T. M., Wu, J., Hoke, S., Sutherland, J., and Tu, W. (2009). Optimized antidepressant therapy and pain self-management in primary care patients with depression and musculoskeletal pain: a randomized controlled trial, *The Journal of the American Medical Association*, 301:2099–2110.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem, *Annals of Mathematical Statistics*, 23:525–540. MR0050850

Lachin, J. M. (1992). Some large-sample distribution-free estimators and tests for multivariate partially incomplete data from two populations, *Statistics in Medicine*, 11:1151–1170.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials, *Biomnetrika*, 70:659–663. MR0725380

Lee, J. W. and DeMets, D. L. (1992). Sequential rank tests with repeated measurements in clinical trials, *Journal of the American Statistical Association*, 87:136–142. MR1158631

Lin, D. Y. (1991). Nonparametric sequential testing in clinical trials with incomplete multivariate observations, *Biometrika*, 78:123–131. MR1118237

Liu, A., Li, Q., Liu, C., Yu, K., and Yu, K. F. (2010). A rank-based test for comparison of multidimensional outcomes, *Journal of the American Statistical Association*, 105:578–587. MR2724843

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints, *Biometrics*, 40:1079–1087. MR0786180

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials, *Biometrics*, 35:549–556.

Olanow, C. W., Wunderle, K. B., and Kieburtz, K. (2011). Milestones in movement disorders clinical trials: advances and landmark studies, *Movement Disorders*, May, 26(6):1003–14. doi: 10.1002/mds.23727.

Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials, *Biometrics*, 43:487–498. MR0909756

Sankoh, A. J., D'Agostino, R. B., and Huque, M. F. (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues, *Statistics in Medicine*, 22:3133–3150.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures, *Journal of the American Statistical Association*, 81:826–831.

Shults, C. W., Oakes, D., Kieburtz, K., Beal, M. F., Haas, R., Plumb, S., Juncos, J. L., Nutt, J., Shoulson, I., Carter, J., Kompoliti, K., Perlmutter, J. S., Reich, S., Stern, M., Watts, R. L., Kurlan, R., Molho, E., Harrison, M., Lew, M., and The Parkinson Study Group. (2002). Effects of coenzyme Q_{10} in early Parkinson disease: evidence of slowing of the functional decline, *Archives of Neurology*, 59(10):1541–1550.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance, *Biometrika*, 73(3):751–754. MR0897872

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance, *Annals of Mathematical Statistics*, 16:243–258. MR0013885

Su, J. Q. and Lachin, J. M. (1992). Group sequential distribution-free methods for the analysis of multivariate observations, *Biometrics*, 48(4):1033–1042.

Tan, M., Xiong, X., and Kutner, M. H. (1998). Clinical trial designs based on sequential conditional probability ratio tests and reverse stochastic curtailing, *Biometrics*, 54:682–695.

Tang, D., Gnecco, C., and Geller, N. L. (1989a). An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials, *Biometrika*, 76:577–583. MR1040650

Tang, D., Gnecco, C., and Geller, N. L. (1989b). Design of group sequential clinical trials with multiple endpoints, *Journal of the American Statistical Association*, 84:776–779.

The NINDS NET-PD Investigators (2006). A randomized, double blinded, futility clinical trial of creatine and minocycline in early Parkinson's disease, *Neurology*, 66:664–671.

The NINDS NET-PD Investigators (2007). A randomized clinical trial of coenzyme Q10 and GPI-1485 in early Parkinson's disease, *Neurology*, 68:20–28.

The Parkinson Study Group (1989). Effect of deprenyl on the progression of disability in early Parkinson's disease, *New England Journal of Medicine*, Nov. 16, 321(20):1364–71.

The Parkinson Study Group (1993). Effects of tocopherol and deprenyl on the progression of disability in early Parkinson's disease, *New England Journal of Medicine*, 328:176–183.

Tilley, B. C., Marler, J., Geller, N. L., Lu, M., Legler, J., Brott, T., Lyden, P., and Grotta, J. (1996). Use of a Global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and stroke T-PA stroke trial, *Stroke*, 27(11):2136–42.

Tilley, B. C., Pillemer, S. R., Heyse, S. P., Clegg, D. O., Alarcón, G. S., and The MIRA Trial Group (2000). Global test for comparing multiple outcomes in rheumatoid arthritis trials, *Arthritis and Rheumatism*, 42(9):1879–88.

Wei, L. J. and Lachin, J. M. (1984). Two sample distribution-free tests for in complete multivariate observations, *Journal of the American Statistical Association*, 79:653–661. MR0763584

Whitehead, John (1997). *The design and analysis of sequential clinical trials*, John Wiley & Sons (New York; Chichester). MR0793018

Wiens and Dmitrienko (2005). The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics*, 15:929–942. MR2210804

Xiong, X. (1995). A class of sequential conditional probability ratio tests, *Journal of the American Statistical Association*, 90:1463–1473. MR1379490

Xiong, X., Tan, M., and Boyett, J. (2003). Sequential conditional probability ratio tests for normalized test statistics on information time, *Biometrics*, 59:624–631. MR2004267

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials (with comments), *Statistics in Medicine*, 9:65–72.

Zhao, L., Hu, X. J., and Lagakos, S. (2009). Statistical monitoring of clinical trials with multivariate response and/or multiple arms: a flexible approach, *Biostatistics*, 10(2):310–323.

Peng Huang
Department of Oncology and
Department of Biostatistics
Johns Hopkins University
Baltimore, Maryland 21205
USA
E-mail address: phuang12@jhmi.edu

Ming T. Tan
Department of Biostatistics, Bioinformatics & Biomathematics
Georgetown University
Washington DC 20057
USA
E-mail address: Ming.Tan@georgetown.edu