Published in final edited form as:

Stat Med. 2008 July 20; 27(16): 3084–3104. doi:10.1002/sim.3182.

A rank-based sample size method for multiple outcomes in clinical trials

Peng Huang^{1,*†}, Robert F. Woolson¹, and Peter C. O'Brien²

¹Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, SC, U.S.A.

²Section of Biostatistics, Mayo Clinic, Rochester, MN, U.S.A.

SUMMARY

O'Brien introduced a rank-sum-type global statistical test to summarize treatment's effect on multiple outcomes and to determine whether a treatment is better than others. This paper presents a sample size computation method for clinical trial design with multiple primary outcomes, and O'Brien's test or its modified test is used for the primary analysis. A new measure, the *global treatment effect* (GTE), is introduced to summarize treatment's efficacy from multiple primary outcomes. Computation of the GTE under various settings is provided. Sample size methods are presented based on prespecified GTE both when pilot data are available and when no pilot data are available. The optimal randomization ratio is given for both cases. We compare our sample size method with the Bonferroni adjustment for multiple tests. Since ranks are used in our derivation, sample size formulas derived here are invariant to any monotone transformation of the data and are robust to outliers and skewed distributions. When all outcomes are binary, we show how sample size is affected by the success probabilities of outcomes. Simulation shows that these sample size formulas provide good control of type I error and statistical power. An application to a Parkinson's disease clinical trial design is demonstrated. Splus codes to compute sample size and the test statistic are provided.

Keywords

effect size; global statistical test; global treatment effect; multivariate nonparametric Behrens–Fisher problem; rank sum; sample size determination

1. INTRODUCTION

Sample size computation is an important component in clinical trial design. When a single outcome is used as a primary outcome, the computation of sample size has been well studied in the literature [1]. For clinical studies where it is difficult to obtain a single summary score for the treatment comparison, multiple outcomes are usually studied jointly. This is true, for example, in Parkinson's disease (PD) clinical trials, when the objective is to identify a neuroprotective treatment that can slow or stop PD progression, and no single outcome is sufficient to capture PD disability in all its aspects, e.g. motor and nonmotor functions, cognition, and drug complications. Furthermore, a treatment may be considered beneficial if it shows improvement in most outcome measures. This work was motivated by the following challenges in clinical trial designs when the goal is to identify a better (or nonfutile) treatment.

First, it is often unrealistic to characterize a joint parametric distribution of multiple primary outcomes, even when all marginal parametric distributions of single outcomes are known. Second, since outcomes are often measured in different scales and may be of different types (continuous, binary, or ordered categorical outcomes), it can be difficult to summarize a treatment's global effect across different outcomes parametrically. Finally, if outcomes have quite skewed distributions, models based on outcome means are not robust to outliers. The aim of this paper is to present sample size computation formulas for the design of clinical trials using nonparametric techniques to cope with these challenges.

Suppose that we wish to compute sample size for a future clinical trial that compares two groups using K primary outcomes. Throughout the paper, we assume that all outcomes are coded in the way that larger observations correspond to better medical conditions. We use capital letter $X_{iv}^{(o)}$ (or X_{iv}) to denote the v th random variable in the i th group in a pilot (or the future) study and small letter $X_{i,jv}^{(o)}$ (or $x_{i,jv}$) to denote the corresponding realization of $X_{iv}^{(o)}$ (or X_{iv}) from the j th subject in the i th group, $v=1, \dots, K$; $i=1,2$. Brunner *et al.* [2] have formulated hypotheses

for this setting as $H_0: p_v = P(X_{1v} \leq X_{2v}) + \frac{1}{2}P(X_{1v} = X_{2v}) = \frac{1}{2}$, $v=1, \dots, K$, and characterize this as the multivariate Behrens–Fisher problem. The advantage of using p_v as a measure of treatment's effect on the v th outcome was discussed in [3–5]. In clinical trials, a weighted average of p_v with weights representing the relative importance of the outcomes could be of interest to investigators in assessing a treatment's global effect on multiple outcomes. In this paper, we confine our attention to cases when the goal is to test whether a treatment is preferred to another. We thus formulate the one-sided hypotheses of the nonparametric problem as

$$H_0: \bar{\theta} = \sum_{v=1}^K \theta_v / K \leq 0, \quad H_1: \bar{\theta} > 0 \quad \text{where} \quad \theta_v = 2p_v - 1 \quad (1)$$

Since exchanging the labeling of 'group 1' and 'group 2' will result in a change of sign in $\bar{\theta}$, we can confine our attention to problem (1) with statistical power controlled at a given $\bar{\theta}_a = \sum_{v=1}^K \theta_{va} / K > 0$. Sample size for a two-sided problem can be derived using parallel methods. Note that $\theta_v = P(X_{1v} \leq X_{2v}) - P(X_{1v} > X_{2v})$. We define $\bar{\theta}$ as the *global treatment effect* (GTE) on all K outcomes and θ_v as the treatment effect on the v th outcome. This GTE is interpretable even if different outcomes are measured in different scales.

O'Brien [6] introduced a simple rank-sum-type global procedure for testing problem (1) to summarize a treatment's global effect across different outcomes and to answer questions like 'Should we give this treatment to patients?' O'Brien's test has been used extensively in clinical research in neurologic diseases, HIV, cancer, health services research, psychiatric diseases, and autoimmune diseases. In spite of its wide applications, sample size computation method using O'Brien's test in clinical trial design has not been developed. This paper provides simple analytic sample size computation method when O'Brien's rank-sum test or its modified test [7] is used for clinical trial primary analysis. Although equal weights for θ_v s are used in (1), results in this paper can be extended to cases when other prespecified weights are chosen. When all outcomes are normally distributed, Tang *et al.* [8] discussed how to compute sample sizes in group-sequential designs when outcome means $\mu_1, \dots, \mu_K$ are related through $\mu_v = \lambda \delta_v$ for known positive numbers $\delta_1, \dots, \delta_K$ when testing $H_0: \lambda = 0$ vs $H_1: \lambda > 0$. Whitehead [9] derived a sample size computation formula for univariate ordered categorical outcome under the assumption of proportional odds. If all categories are equiprobable, the required sample size decreases as the number of categories increases. The limiting case is the sample size from the Wilcoxon–Mann–Whitney test where a full ranking of outcome values was achieved. For

multivariate outcomes, Rochon [10] presented a table of minimum sample size for binary outcomes when a GEE procedure is used for data analysis. We will show that, when all outcomes are binary, the sample sizes based on Rochon's GEE method and ours are driven by outcomes with success probabilities close to 0.5.

In this paper, using O'Brien's rank-sum-type test or its modified test [7], we present sample size computation for problem (1) when the goal is to identify a beneficial (or nonfutile) treatment using multiple primary outcomes. Outcomes can be continuous, discrete, and a mixture of continuous and discrete random variables. We will control type I error rate, α , at the null hypothesis specified by (1) and power, $(1-\beta)$, for a given alternative GTE value $\bar{\theta}_a$ [mvg] 0. Methods to choose $\bar{\theta}_a$ are discussed in Section 2. We derive sample size expressions for rank-sum-type test and its modified test both when no historical data are available (Section 3) and when some historical data are available (Section 4). We compare our sample size formula with Noether's result [11] for a univariate outcome in Section 5. In many nonparametric tests, since the null hypothesis often assumes that the groups have a common distribution, we show a reduction in sample size under this more restricted null assumption in Section 6. For multiple outcomes, Section 7 compares rank-sum-type test with the Bonferroni adjustment both in sample size and in scientific questions to be addressed. When all outcomes are binary, Section 8 examines how sample sizes in both Rochon's formula [10] and our formula are affected by outcomes whose success probabilities are close to 0.5. Simulation in Section 9 shows that sample size derived from this paper gives good control of type I error and statistical power even for moderate sample sizes. Finally, an application to a clinical trial design is demonstrated in Section 10.

2. DETERMINATION OF $\bar{\theta}_a$ FROM PRIOR DATA OR LITERATURE

The definition of GTE, $\bar{\theta}$, is similar to Kendall's τ . Both of them take values between -1 and 1 . Kendall's τ measures the degree of similarity between two groups, whereas GTE measures the degree of dissimilarity between two groups. When $\bar{\theta} = 0$, there is no global preference between groups 1 and 2. When $\bar{\theta} = -1$, group 1 is preferred the most. When $\bar{\theta} = 1$, group 2 is preferred the most. Larger positive $\bar{\theta}$ corresponds to a higher degree of group 2 preference. The selection of a parameter value $\bar{\theta}_a$ as a 'clinically meaningful treatment difference' for the control of statistical power depends in practice on many factors including considerations of scientific rationale, budget, pilot data, or some other source of information. When no pilot data are available, investigators sometimes choose this clinically meaningful treatment difference value using a value suggested in the literature. In this section, we discuss how to incorporate available treatment information when choosing a reasonable $\bar{\theta}_a$ value for the sample size computation. Since $\bar{\theta}_a$ is an average of the GTEs from single outcomes $\theta_{1a}, \dots, \theta_{Ka}$, we need to discuss only the case when $K=1$ without a loss of generality. We thus use θ_a to denote $\bar{\theta}_a$ and drop the subscript v in $X_{iv}^{(o)}$. When some pilot data are available with sample size m_i in the i th group ($i=1,2$), an unbiased estimate of $\theta^{(o)} = P(X_1^{(o)} [tdb] X_2^{(o)}) - P(X_1^{(o)} [mvg] X_2^{(o)})$ is given by $\hat{\theta} = (1/m_1 m_2) \{2 \sum_{j=1}^{m_2} R_{2j}^{(o)} - m_2 (m_1 + m_2 + 1)\}$ where $R_{2j}^{(o)}$ is the rank of the observation from the j th subject in group 2. Based on an investigator's goal of the study, a value of θ_a might be generated from $\hat{\theta}$, but the cautious of Kraemer *et al.* [12] regarding pilot study data generally should be noted.

Treatment difference is often expressed in terms of means and standard deviations using normal approximation or through summary statistics using other parametric approximation. The remainder of this section provides computational formulas or numerical values of $\theta^{(o)}$ when

some parametric assumptions to the distributions of $X_1^{(o)}$ and $X_2^{(o)}$ are made. Investigators can choose θ_a using $\theta^{(o)}$ as reference. Note that such a parametric approximation is used only to help investigators select a reasonable θ_a value. This is *fundamentally different* from making a parametric assumption in the joint distribution of the multiple outcomes for the derivation of the test statistic and sample size computation.

In the literature, a monotone increasing transformation (e.g. log-transformation or square-root transformation) is sometimes applied to the data to obtain a better parametric approximation, and summary statistics are reported using the transformed data. Since GTE is invariant to any monotone increasing transformation g of the data, we do not need to distinguish between

distributional assumptions of $(X_1^{(o)}, X_2^{(o)})$ and $(g(X_1^{(o)}), g(X_2^{(o)}))$ for the computation of $\theta^{(o)}$. The advantage of using GTE to measure treatment difference is that, even if different monotone transformations have been applied to different outcomes in the literature, we can still calculate and interpret GTE.

If both $X_1^{(o)}$ and $X_2^{(o)}$ have symmetric distributions with identical means, then GTE $\tau^{(o)}=0$. For other cases, we discuss continuous outcomes and discrete outcomes separately:

1. If $X_i^{(o)}$ has a normal distribution with density function $f_i(x) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-(x-\mu_i)^2/(2\sigma_i^2)}$ then $\theta^{(o)} = 2\Phi\left((\mu_2 - \mu_1) / \sqrt{\sigma_1^2 + \sigma_2^2}\right) - 1$.
2. If $X_i^{(o)}$ has an exponential distribution with density function $f_i(x) = \alpha_i e^{-\alpha_i x} I(x > 0)$, then $\theta^{(o)} = (\alpha_1 - \alpha_2) / (\alpha_1 + \alpha_2)$.
3. When $X_i^{(o)}$ has a continuous but skewed distribution, its distribution sometimes can be approximated by a Gamma distribution. The density of a Gamma(α, β) distribution is $f(x) = \beta^{-\alpha} x^{\alpha-1} e^{-x/\beta} I(x > 0) / \Gamma(\alpha)$. Suppose that $X_i \sim \text{Gamma}(\alpha_i, \beta)$ with a common scale parameter but different shape parameters. We consider the cases when $\beta=0.5$ or 2, α_1 ranging from 0.5 to 10 and α_2/α_1 ranging from 0.3 to 0.8. These parameters are chosen to capture different shapes of the density function. The corresponding $\theta^{(o)}$ values are given in Table I.
4. If $X_i^{(o)}$ is continuous with finite domain, one can apply a monotone transformation to confine its support to interval $[0, 1]$. A Beta(α, β) distribution with density function $f(x) = x^{\alpha-1} (1-x)^{\beta-1} \Gamma(\alpha+\beta) / \{\Gamma(\alpha)\Gamma(\beta)\}$ can be used to approximate a continuous distribution defined in interval $[0, 1]$. When $\alpha=\beta$, $f(x)$ has a symmetric shape. Function $f(x)$ evolves from U-shape when $\alpha=\beta=0.5$ to bell shape when $\alpha=\beta=2.0$. It is skewed to the right when $\alpha < \beta$ and skewed to the left when $\alpha > \beta$. Since $X \sim \text{Beta}(\alpha, \beta)$ is equivalent to $1-X \sim \text{Beta}(\beta, \alpha)$, w.l.o.g., we consider only the case when $X_i^{(o)} \sim \text{Beta}(\alpha_i, \beta_i)$ with $\alpha_1 < \beta_1$. Table I also lists values of $\theta^{(o)}$ for different combinations of $\alpha_1, \beta_1, \alpha_2$, and β_2 .
5. If $X_i^{(o)}$ is binary such that $X_i^{(o)} \sim \text{Bernoulli}(p_i)$ with $P(X_i^{(o)}=1) = p_i$, then $\theta^{(o)} = p_2 - p_1$.
6. If $X_i^{(o)}$ is discrete and is approximated by a geometric distribution: $X_i^{(o)} \sim \text{Geometric}(p_i)$ with $P(X_i^{(o)}=x) = p_i(1-p_i)^{x-1}$ ($x=1, 2, \dots$), then $\theta^{(o)} = (p_1 - p_2) / (p_1 + p_2 - p_1 p_2)$.
7. Suppose that $X_i^{(o)}$ has a Poisson distribution: $P(X_i^{(o)}=x) = \lambda_i^x e^{-\lambda_i} / x!$. When λ increases, Poisson(λ) approaches to a normal distribution. When $\lambda \geq 10$, it resembles the bell-

shaped distribution. Table I lists $\theta^{(o)}$ values when $\lambda_1=0.25, 0.5, 1, 3, 6, 10$ and $\lambda_2/\lambda_1=0.3, 0.5, 0.7, 0.9, 0.95$. If $\lambda_2 > \lambda_1$, we can switch the roles of X_1 and X_2 and then apply Table I. To switch back the roles of $X_1^{(o)}$ and $X_2^{(o)}$, we can simply change the sign of $\theta^{(o)}$.

The sign of $\theta^{(o)}$ is also consistent with the sign of mean change, the sign median difference, and the stochastic ordering. More specifically, when both $X_1^{(o)}$ and $X_2^{(o)}$ are from the same exponential family with probability distribution functions $f(x)=\exp[\omega_i x - b(\omega_i) - a(x)]$ ($i=1, 2$), respectively, for some convex function $b(\cdot)$ and continuous function $a(\cdot)$, then $\theta^{(o)} > 0$ if and only if $E[X_2^{(o)}] - E[X_1^{(o)}] [mvg] 0$. As pointed out by one of our reviewers, prior data may guide what to expect for a difference, but it does not tell what is ‘clinically’ meaningful. In application, one might incorporate GTE computed from pilot data with information from clinical resources when determining a clinically meaningful $\bar{\theta}_a$ value, but overall this is a clinical assessment.

3. SAMPLE SIZE ESTIMATE WHEN NO HISTORICAL DATA ARE AVAILABLE

Suppose that $(x_{ij1}, \dots, x_{ijK})$ will be the future observation from the j th subject in the i th group, $i=1, 2; j=1, \dots, n_i$. For each i , it is assumed that $\{(x_{ij1}, \dots, x_{ijK}), j=1, \dots, n_i\}$ are independent identically distributed realizations of a random vector $\mathbf{X}_i=(X_{i1}, \dots, X_{iK})$. Denote $N=n_1+n_2$,

$F_{iu}(t) = P(X_{iu} \leq t) + \frac{1}{2} P(X_{iu} = t)$, $\theta_v = P(X_{1v} < X_{2v}) - P(X_{1v} > X_{2v})$, and $\theta = (\theta_1, \dots, \theta_K)^T$. Throughout this paper, we impose the regularity conditions $\text{var}\{F_{1v}(X_{2v})\} > 0$ and $\text{var}\{F_{2v}(X_{1v})\} > 0$ ($v=1, \dots, K$) to rule out degenerate distributions and redundant parameters. Let $r_0 = n_2/n_1$ be a pre-specified (known) randomization ratio ($0 < r_0 < \infty$) in the future study, R_{ijv} be the rank of x_{ijv} among $\{x_{11v}, \dots, x_{1n_1v}, \dots, x_{21v}, \dots, x_{2n_2v}\}$, $\bar{R}_i = \sum_{j=1}^{n_i} \sum_{v=1}^K R_{ijv} / n_i$ be the mean rank sum in group i , $\hat{\sigma}_i^2$ be the sample variance of rank sums $\{\sum_{v=1}^K R_{ijv}, j=1, \dots, n_i\}$, and $\hat{\sigma}_p^2$ be the pooled sample variance of rank sums in the two groups. Then, O’Brien’s rank-sum-type test statistic can be expressed as

$$T_1 = \frac{\bar{R}_2 - \bar{R}_1}{\hat{\sigma}_p \sqrt{1/n_1 + 1/n_2}} \quad \text{or} \quad T_2 = \frac{\bar{R}_2 - \bar{R}_1}{\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}}$$

respectively, when treating two samples with equal or unequal variance. If F_1 and F_2 are not equal under H_0 , Huang *et al.* [7] proposed to modify T_1 and T_2 by

$$T_{1a} = \frac{\bar{R}_2 - \bar{R}_1}{\hat{\sigma}_p \sqrt{\hat{h}_1 (1/n_1 + 1/n_2)}} \quad \text{or} \quad T_{2a} = \frac{\bar{R}_2 - \bar{R}_1}{\sqrt{\hat{h}_2 (\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2)}}$$

where $\hat{h}_1$ and $\hat{h}_2$ are consistent estimators of

$$h_1 = \frac{\sum_{u=1}^k \sum_{v=1}^k (1+r_0)^2 (r_0 a_{uv} + b_{uv})}{\sum_{u=1}^k \sum_{v=1}^k [e_{uv} + (b_{uv} + 2f_{uv}) r_0 + (a_{uv} + 2q_{uv}) r_0^2 + p_{uv} r_0^3]}$$

and

$$h_2 = \frac{\sum_{u=1}^k \sum_{v=1}^k (1+r_0)^2 (r_0 a_{uv} + b_{uv})}{\sum_{u=1}^k \sum_{v=1}^k [b_{uv} + (e_{uv} + 2q_{uv}) r_0 + (p_{uv} + 2f_{uv}) r_0^2 + a_{uv} r_0^3]}$$

respectively, where $a_{uv} = \text{cov}\{F_{2u}(X_{1u}), F_{2v}(X_{1v})\}$, $b_{uv} = \text{cov}\{F_{1u}(X_{2u}), F_{1v}(X_{2v})\}$, $e_{uv} = \text{cov}\{F_{1u}(X_{1u}), F_{1v}(X_{1v})\}$, $f_{uv} = \text{cov}\{F_{1u}(X_{1u}), F_{2v}(X_{1v})\}$, $p_{uv} = \text{cov}\{F_{2u}(X_{2u}), F_{2v}(X_{2v})\}$ and $q_{uv} = \text{cov}\{F_{2u}(X_{2u}), F_{1v}(X_{2v})\}$. Let $\zeta(x, y) = I[(x < y)] - I[(x > y)]$, where indicator $I[E]$ is defined by $I[E] = 1$ if event E is true, and $I[E] = 0$ otherwise. Using similar method in Theorem 1 of [7], we can show that distributions of $T_i / \sqrt{h_i}$, T_{1a} and T_{2a} converge, as N increases, to a normal distribution with mean $K \bar{\theta} \sqrt{N} / (2 \sqrt{J^T \Sigma J})$ and variance 1, where J is a K -dimensional vector with all its elements equal to one

$$\sum = \frac{N(n_2 - 1)}{n_1 n_2} A + \frac{N(n_1 - 1)}{n_1 n_2} B + \frac{N}{4n_1 n_2} C \quad (2)$$

Here A , B , and C are $K \times K$ covariance matrices with (u, v) elements given by a_{uv} , b_{uv} , and $c_{uv} = \text{cov}\{\zeta(X_{1u}, X_{2u}), \zeta(X_{1v}, X_{2v})\}$, respectively. A proof of (2) is available from the authors. Another asymptotically equivalent test that avoids the computation of h_1 or h_2 is to use the following test statistic:

$$Z = \frac{\sqrt{N}}{n_1 n_2 \sqrt{J^T \widehat{\Sigma} J}} \left\{ \frac{1}{n_1} \bar{R}_1 - \frac{1}{n_2} \bar{R}_1 - \frac{(n_2 - n_1)(N+1)}{2n_1 n_2} \right\} \quad (3)$$

where $\widehat{\Sigma}$ is an unbiased estimate of Σ . Computation of matrix $\widehat{\Sigma}$ is discussed in Remark 4. Statistic Z has the same asymptotic distribution as $T_i / \sqrt{h_i}$, T_{1a} , and T_{2a} . One advantage of using Z is that it does not need to check whether two groups have the same variance. To facilitate implementation, Splus code to compute Z , $\widehat{\Sigma}$, $\widehat{h}_1$, and $\widehat{h}_2$ is provided in Appendix C.

Sample size generally depends on A , B , and C evaluated at $\bar{\theta} = \sum_{v=1}^K \theta_{v0} / K = 0$ and $\bar{\theta} = \bar{\theta}_a = \sum_{v=1}^K \theta_{va} / K$. We consider alternatives where $\theta_a = (\theta_{1a}, \dots, \theta_{Ka})$ does not differ too much from $\theta_0 = (\theta_{10}, \dots, \theta_{K0})$. In this case, it will often be appropriate to assume that A , B , and C are held constant for the derivation of sample size. We thus make this assumption throughout this paper. This method of dealing with covariances was first suggested by Noether [11] and is commonly employed in nonparametric applications. For moderate-to-large sample size N , Σ in (2) can be approximated by $((1+r_0)/r_0)(r_0 A + B)$. When A and B are known, the required total sample size N to control type I error rate, α , at the null hypothesis of (1) and power, $(1-\beta)$, at a given alternative GTE $\bar{\theta}_a = \sum_{v=1}^K \theta_{va} / K$ can be approximated by

$$N = \frac{4(1+r_0)}{K^2 \theta_a r_0} (z_\alpha + z_\beta)^2 J^T (r_0 A + B) J \quad (4)$$

where z_α and z_β are the $(1-\alpha)$ th and $(1-\beta)$ th percentiles, respectively, of the standard normal distribution. When A and B are unknown and no pilot data are available to estimate them, one can use an upper bound of $J^T (r_0 A + B) J$. If we can find a relatively small upper bound σ^2 for the diagonal elements of A and B , the off-diagonal elements of A and B will be upper bounded by $\rho \sigma^2$ for some $0 < \rho \leq 1$. When no such σ^2 value is available, one can use

$\sigma^2 = (1 - \min_{(1 \leq v \leq K)} \theta_{va}^2) / 4$ (a proof is given in Appendix A). On the basis of Taylor's expansion, one can approximate ρ by an upper bound of $\{\text{cor}(X_{iu}, X_{iv}), i=1,2; u \neq v; u,v=1, \dots, K\}$ found in the literature, such as papers that report similar measures. Simulation in Section 9 shows that this approximation results in a slightly conservative estimate of sample size. In cases when no such of information on ρ is available, one can always use the most conservative upper bound $\rho=1$. Since $\lim_{N \rightarrow \infty} J^T \Sigma J / [(N^2/n_1 n_2) \sigma^2 K \{(K-1)\rho+1\}] \leq 1$, we can use a more conservative estimate of the total sample size:

$$N = 4 \left\{ \frac{\sigma}{\bar{\theta}_a} \right\}^2 \left\{ \rho + \frac{1}{K} (1 - \rho) \right\} \frac{(1+r_0)^2}{r_0} (Z_\alpha + Z_\beta)^2 \quad (5)$$

The quantity $(\bar{\theta}_a / \sigma) / \sqrt{2\{\rho + (1 - \rho)/K\}}$ is an analogue the 'effect size' in parametric models. To see this, let $D_v = F_{1v}(X_{2v}) - F_{2v}(X_{1v})$ and $D = \sum_{v=1}^K D_v / K$. Then, D_v is a random measure of treatment difference in the v th outcome. Under the alternative $\theta_v = \theta_{va}$ ($v=1, \dots, K$), we have $E(D_v) = \theta_{va}$ (a difference between two areas under the curve), $E(D) = \bar{\theta}_a$, and $\text{var}(D) \leq 2\sigma^2\{\rho + (1 - \rho)/K\}$. Quantity $(\bar{\theta}_a / \sigma) / \sqrt{2\{\rho + (1 - \rho)/K\}}$ is a lower bound of $E(D) / \sqrt{\text{var}(D)}$ and thus a measure of treatment difference we can detect with the given sample size. Since limited information of A and B is available in this case, the total sample size N in (5) depends on r_0 through $(1+r_0)^2/r_0$. The optimal randomization ratio to achieve the smallest total sample size N is $r_0=1$.

4. SAMPLE SIZE ESTIMATE WHEN SOME HISTORICAL DATA ARE AVAILABLE

Let $\{(x_{ij1}^{(o)}, \dots, x_{ijK}^{(o)}) : j=1, \dots, m_i\}$ be random observations from the i th group in a pilot study,

$i=1,2$. Define $F_{iu}^{(o)}(t) = P(X_{iu}^{(o)} \leq t) + \frac{1}{2}P(X_{iu}^{(o)} = t)$; $R_i(x_{3-i,jv}^{(o)})$ is the rank of $x_{3-i,jv}^{(o)}$ among $\{x_{3-i,jv}^{(o)}, x_{i1v}^{(o)}, \dots, x_{imiv}^{(o)}\}$; $\bar{R}_i(x_{3-i,v}^{(o)})$

$$\begin{aligned} &= (1/m_{3-i}) \sum_{j=1}^{m_{3-i}} R_i(x_{3-i,jv}^{(o)}) ; \bar{\xi}_v^{(o)} \\ &= (1/m_1 m_2) \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \xi(x_{1iu}^{(o)}, x_{2ju}^{(o)}) ; S_{3-i,uv} \\ &= (1/m_{3-i}) \sum_{j=1}^{m_{3-i}} \left\{ R_i(x_{3-i,jv}^{(o)}) - \bar{R}_i(x_{3-i,v}^{(o)}) \right\} \left\{ R_i(x_{3-i,jv}^{(o)}) - \bar{R}_i(x_{3-i,v}^{(o)}) \right\} S_{3-iv} \\ &= (1/m_1 m_2) \sum_{i=1}^{m_1} \left\{ \xi(x_{1iu}^{(o)}, x_{2iu}^{(o)}) - \bar{\xi}_u^{(o)} \right\} \left\{ \xi(x_{1iv}^{(o)}, x_{2jv}^{(o)}) - \bar{\xi}_v^{(o)} \right\} ; d_1 \\ &= m_1 m_2 r_0 (m_1 - 1) (m_2 - 1) ; d_2 \\ &= (z_\alpha + z_\beta)^2 / \left(K^2 \theta_a^2 \right) \end{aligned} \quad ; S_1, S_2, \text{ and } S_3$$

are $K \times K$ matrices with (u,v) elements s_{1uv} , s_{2uv} , and s_{3uv} , respectively. If the pilot data are used to determine the alternative $\theta_a = (\theta_{1a}, \dots, \theta_{Ka})$ values, it will be reasonable to assume that correlations between the pilot data and the future data at θ_a are associated. The following theorem gives an unbiased estimate of Σ when $\theta = \theta_a$ using the pilot data.

Theorem 1

Let $(X_{i1}^{(o)}, \dots, X_{ik}^{(o)})$ and $(X_{i1}, \dots, X_{ik})$ represent random vectors of the i th group in a pilot study and the future study, respectively. Observations $\{(x_{ij1}^{(o)}, \dots, x_{ijk}^{(o)}): j=1, \dots, m_i\}$ are independent identically distributed realizations of $(X_{i1}^{(o)}, \dots, X_{ik}^{(o)})$, $i=1, 2$. Suppose that covariance matrices between the pilot data and the future data when $\theta=\theta_a=(\theta_{1a}, \dots, \theta_{Ka})$ are associated through the following:

$$\begin{aligned} \text{cov} \left\{ F_{iu}^{(o)}(X_{ju}^{(o)}), F_{iv}^{(o)}(X_{jv}^{(o)}) \right\} &= \text{cov} \left\{ F_{iu}(X_{jv}), F_{iv}(X_{jv}) \right\} | \theta_a, & i \neq j, \quad i, j=1, 2 \\ \text{cov} \left\{ \xi(X_{1u}^{(o)}, X_{2u}^{(o)}), \xi(X_{1v}^{(o)}, X_{2v}^{(o)}) \right\} &= \text{cov} \left\{ \xi(X_{1u}, X_{2u}), \xi(X_{1v}, X_{2v}) \right\} | \theta_a \end{aligned} \quad (6)$$

Let $\Gamma_1 = (4m_1 + 4m_1^2 r_0) S_1 + (4m_2^2 + 4m_2 r_0) S_2 - (m_1 m_2^2 + m_1^2 m_2 r_0) S_3$ and $\Gamma_2 = 4m_1^2 S_1 + 4m_2^2 S_2 - m_1^2 m_2^2 S_3$. Then an unbiased estimate of Σ under $\theta=\theta_a$ is given by

$$\widehat{\Sigma} = \frac{1+r_0}{4d_1} \left(\Gamma_1 - \Gamma_2 \frac{1}{n_1} \right) \quad (7)$$

A proof of this theorem is given in Appendix B. Requirement (6) applies to many situations. For example, it applies to the commonly used assumption that future data relate to pilot data through a location-scale transformation, that is, when there exist two (possibly unknown)

constants η_1 and η_2 such that $\{\eta_1(X_{iu}^{(o)} - \eta_2), \eta_1(X_{iv}^{(o)} - \eta_2), \eta_1(X_{ik}^{(o)} - \eta_2)\}$ has the same distribution as $\{X_{iu}, X_{iv}, X_{ik}\}$. Furthermore, it applies to all cases when there exists some (possibly unknown) monotonic increasing function g such that $\{g(X_{iu}^{(o)}), g(X_{iv}^{(o)}), g(X_{ik}^{(o)})\}$ has the same distribution as $\{X_{iu}, X_{iv}, X_{ik}\}$ ($i=1, 2; u, v, k=1, \dots, K$). The function g can be introduced by the improvement of medical care techniques (that could result in a reduction in variation) or the change in clinical practice over time (that could result in a shift in outcome means). Define

$$\tilde{E} = \frac{1+r_0}{d_1} \left\{ m_1 \left(1+m_1 r_0 - \frac{m_1}{n_1} \right) S_1 + m_2 \left(m_2 + r_0 - \frac{m_2}{n_1} \right) S_2 + \frac{m_1^2 m_2^2}{4n_1} S_3 \right\}$$

Then $\tilde{\Sigma} = \widehat{\Sigma} + ((1+r_0)/4d_1) (m_1 m_2^2 + m_1^2 m_2 r_0) S_3$. When m_1 and m_2 increase such that m_1/m_2 goes to a positive constant, the difference $\tilde{\Sigma} - \widehat{\Sigma}$ is of order $O(1/m_1)$. Let γ_i be the sum of all K^2 elements in Γ_i , $i=1, 2$. Replacing $((1+r_0)/r_0)(A+r_0B)$ in (4) by $\widehat{\Sigma}$ in (7) gives a total sample size estimation:

$$N = (1+r_0) \frac{d_2 \gamma_1}{2d_1} \left\{ 1 + \sqrt{1 - \frac{4d_1 \gamma_2}{d_2 \gamma_1^2}} \right\} \quad (8)$$

If, instead, when Σ in (4) is replaced by $\tilde{\Sigma}$, we obtain a more conservative estimate of the sample size:

$$N = (1+r_0) \frac{d_2 d_3}{2d_1} \left\{ 1 + \sqrt{1 - \frac{4d_1 \gamma_2}{d_2 d_3^2}} \right\} \quad (9)$$

where $d_3 = \gamma_1 + m_1 m_2 (m_2 + m_1 r_0) J^T S_3 J > \gamma_1$.

Remark 1

Let $M = m_1 + m_2$. Formulas (8) and (9) are derived under the assumption that $M/N = o(1)$ or N is at least the same magnitude as M , which is reasonable in real applications since data from a smaller pilot study are often used for the design of a future larger clinical trial. On the other hand, if M is much larger than N , conducting a new clinical trial for the sample patient population will not be considered necessary.

We can also express $\widehat{\Sigma}$ in Theorem 1 as

$$\widehat{\Sigma} = (1+r_0) \left\{ \frac{m_1(n_2-1)+n_1}{(m_1-1)(m_2-1)m_2n_2} S_1 + \frac{m_2(n_1-1)+n_2}{(m_1-1)m_1(m_2-1)n_2} S_2 + \frac{m_1(m_2-n_2)-m_2n_1}{4(m_1-1)(m_2-1)n_2} S_3 \right\}$$

Large M ensures matrix $\widehat{\Sigma}$ to be positive definite since S_1 and S_2 are of order $O(M^2)$ but $S_3 = O(1)$. However, when M is small (such as $M < 20$), $\widehat{\Sigma}$ may not be positive definite. In this case, one can either set $S_3 = 0$ in $\widehat{\Sigma}$ to obtain a consistent (positive definite) estimate of Σ using

$$(1+r_0) \left\{ \frac{m_1(n_2-1)+n_1}{(m_1-1)(m_2-1)m_2n_2} S_1 + \frac{m_2(n_1-1)+n_2}{(m_1-1)m_1(m_2-1)n_2} S_2 \right\}$$

or use $\tilde{\Sigma}$ to estimate Σ since $\tilde{\Sigma}$ is always positive definite.

Remark 2

In case when $1 - 4d_1\gamma_2 / (d_2\gamma_1^2) [tdb] 0$, one can approximate $((1+r_0)/r_0)(A+r_0B)$ in (4) by $(1+r_0)\Gamma_1/(4d_1)$ and use a more conservative sample size estimation $N = (1+r_0)d_2\gamma_1/d_1$.

Remark 3

The use of matrix notation in Theorem 1 makes it convenient to program in Splus, R, Matlab, or others. It is seen that S_{3-i} ($i=1,2$) and S_3 are sample variance-covariance matrices of the column vectors in matrices

$$\begin{pmatrix} R_i(x_{3-i,11}^{(0)}) & \cdots & R_i(x_{3-i,1K}^{(0)}) \\ R_i(x_{3-i,21}^{(0)}) & \cdots & R_i(x_{3-i,2K}^{(0)}) \\ \vdots & \cdots & \vdots \\ R_i(x_{3-i,m_{3-i}^{(0)}}^{(0)}) & \cdots & R_i(x_{3-i,m_{3-i}^{(0)}K}^{(0)}) \end{pmatrix} \text{ and } \begin{pmatrix} \xi(x_{111}^{(0)}, x_{211}^{(0)}) & \cdots & \xi(x_{11K}^{(0)}, x_{21K}^{(0)}) \\ \vdots & \cdots & \vdots \\ \xi(x_{1m_11}^{(0)}, x_{2m_11}^{(0)}) & \cdots & \xi(x_{1m_1K}^{(0)}, x_{2m_1K}^{(0)}) \\ \vdots & \cdots & \vdots \\ \xi(x_{111}^{(0)}, x_{2m_21}^{(0)}) & \cdots & \xi(x_{11K}^{(0)}, x_{2m_2K}^{(0)}) \\ \vdots & \cdots & \vdots \\ \xi(x_{1m_11}^{(0)}, x_{2m_21}^{(0)}) & \cdots & \xi(x_{1m_1K}^{(0)}, x_{2m_2K}^{(0)}) \end{pmatrix}$$

respectively. Note that the denominators in these covariances are m_{3-i} and $m_1 m_2$, respectively, not the (m_1-1) and $(m_1 m_2-1)$ as the default sample covariances given by most software.

We can improve the sample size estimate by bootstrapping using the pilot data. First, take bootstrap samples with sample sizes m_1 and m_2 for the two groups, respectively. Second, obtain bootstrap estimates of $J^T S_1 J$, $J^T S_2 J$, and $J^T S_3 J$. Finally, substitute these bootstrap estimates into (8) or (9) to compute N . To search for the optimal randomization ratio $r_0 = n_2/n_1$, one can

plot N versus r_0 and find the value of r_0 that corresponds to the smallest N value. Our simulation shows that the optimal r_0 can slightly deviate from the balanced design ($r_0=1$) in either direction, but the loss of efficiency of a balanced design compared with the optimal one is minor—often within 1.5 per cent.

Remark 4

An immediate application of Theorem 1 is to obtain an unbiased estimate of Σ for the computation of test statistic Z in (3) after all data have been collected. Replacing m_1 and m_2 in (7) by n_1 and n_2 , respectively, and computing S_1 , S_2 , and S_3 using current data instead of the pilot data, expression (7) is simplified to

$$\widehat{\Sigma} = \frac{N}{(n_1 - 1)(n_2 - 1)} \left(\frac{1}{n_2} S_1 + \frac{1}{n_1} S_2 - \frac{1}{4} S_3 \right)$$

This $\widehat{\Sigma}$ can be used to compute test statistic Z in (3).

5. UNIVARIATE CASES: $K=1$

When F_1 and F_2 are absolutely continuous and are identical under the null hypothesis,

substituting $\sigma^2 = \frac{1}{12}$ and $K=1$ into (5), we obtain exactly the same sample size formula as

Noether's [11] formula: $N = (z_\alpha + z_\beta)^2 (1 + r_0)^2 / (3r_0\theta_a^2)$. Simulations of Sundukchi and Guenther [13] also showed that the accuracy of the asymptotic approximation to the distribution of the Wilcoxon–Mann–Whitney test statistic was surprisingly good for data from various distributions. For univariate ordered categorical data, Whitehead [9] assumed proportional odds among the ordered categories. His sample size computation requires the proportions in all categories to be known in advance for both groups under both the null hypothesis and the alternative. He further showed that, if the proportional odds assumption is violated (e.g. due to the measurement errors in the data), the sample size could increase considerably. We note that under the proportional odds assumption $P(X_1 \leq i) = \exp(\alpha_i) / \{1 + \exp(\alpha_i)\}$ and $P(X_2 \leq i) = \exp(\alpha_i - \theta_R) / \{1 + \exp(\alpha_i - \theta_R)\}$ condition $\theta_R > 0$ implies $P(X_1 < X_2) - P(X_1 > X_2) > 0$. Thus, our sample size computation formula can also be applied to cases when the proportional odds assumption is made.

6. MULTIVARIATE CASE: IDENTICAL DISTRIBUTION UNDER THE NULL HYPOTHESIS

Let $F_i(t_1, \dots, t_K)$ be the joint distribution of $X_{i1}, \dots, X_{iK}$. A special case of H_0 in (1) is to assume identical distribution between the two groups $H_0: F_1 \equiv F_2$ and assume that the variance–covariance matrix Σ is the same under both H_0 and the local alternative $\bar{\theta}_a$. These assumptions have been made in many traditional nonparametric hypothesis tests, for example, in [14–16]. For this special case, we provide a smaller upper bound σ^2 for (5) when no prior data are available. When some prior data are available, we will provide a sample size expression that uses an estimate of Σ .

When no preliminary data are available, we use the fact that $\text{var}[F_{iu}(X_{iu})] \leq \frac{1}{12}$ under the null hypothesis $F_1 \equiv F_2$ [17]. This gives a tighter upper bound σ^2 in (5):

$\sigma^2 = \min \left\{ 1/12, \left(1 - \min_{1 \leq v \leq K} \theta_{va}^2 \right) / 4 \right\}$. When some preliminary data are available, we can

obtain a better estimate of Σ under these more restricted assumption. We continue to use notations in Section 3 and assume that (6) holds. To simplify our notation, we further define $M=m_1+m_2$,

$$y_{iv}^{(0)} = \begin{cases} x_{1iv}^{(0)} & \text{if } i=1, \dots, m_1 \\ x_{2,(i-m_1)v}^{(0)} & \text{if } i=m_1+1, \dots, M \end{cases}$$

R_{iv} = rank of $y_{iv}^{(o)}$ among $\{y_{1v}^{(o)}, \dots, y_{Mv}^{(o)}\}$ and $\bar{R}_v = (1/M) \sum_{i=1}^M R_{iv}$, $s_{1uv} = (1/M) \sum_{i=1}^M (R_{iu} - \bar{R}_u) (R_{iv} - \bar{R}_v)$. Then, an unbiased estimate of Σ under the alternative $\theta = \theta_a$ is given by

$$\widehat{\Sigma} = \frac{1+r_0}{4m_1m_2(m_1+m_2-1)(m_1+m_2-2)} \left(\Gamma_1 - \Gamma_2 \frac{1}{n_1} \right) \quad (10)$$

Here, Γ_1 and Γ_2 are defined by $\Gamma_1 = 4(m_1m_2-1)(1+r_0)S_1 - (1+r_0)m_1m_2(m_1+m_2-1)S_3$ and $\Gamma_2 = 4(2m_1m_2-m_1-m_2)S_1 - m_1m_2(m_1+m_2)(m_1+m_2-1)S_3$. We continue to use γ_i to denote the sum of all K^2 elements in Γ_i , $i=1,2$. Replacing Σ in (4) by $\widehat{\Sigma}$ in (10), solve for N :

$$N = (1+r_0) \frac{d_2\gamma_1}{2d_1'} \left\{ 1 + \sqrt{1 - \frac{4d_1'\gamma_2}{d_2\gamma_1^2}} \right\}$$

which has exactly the same form as (8) but with d_1 replaced by

$d_1' = m_1m_2r_0(m_1+m_2-1)(m_1+m_2-2)$. A conservative estimate is to replace γ_1 by $d_3 = \gamma_1 + (1+r_0)m_1m_2(m_1+m_2-1)J^T S_3 J$.

7. COMPARISON WITH BONFERRONI ADJUSTMENT

Bonferroni adjustment is a simple method to control the overall type I error when multiple tests are performed for testing the hypothesis $H_0: \theta_1 = \dots = \theta_K = 0$ and controlling power at the alternative H_1 : at least one $\theta_v = \theta_{va} > 0$ ($v=1, \dots, K$). Since conclusions from Bonferroni adjustment depend on the smallest observed p -value, sample size using Bonferroni adjustment can be approximated by $N_B = \min_{(1 \leq v \leq K)} 4 \left(\sigma_v^2 / \theta_{va}^2 \right) \left\{ (1+r_0)^2 / r_0 \right\} \left(Z_{\alpha/K} + z_\beta \right)^2$. Substituting $\sigma_v^2 = (1 - \theta_{va}^2) / 4$ into N_B and $\sigma_2^2 = (1 - \min_{(1 \leq v \leq K)} \theta_{va}^2) / 4$ into N in (5), one obtains

$$\frac{N}{N_B} = \frac{(1 - \min_{(1 \leq v \leq K)} \theta_{va}^2) (Z_\alpha + Z_\beta)^2 \{ \rho + (1 - \rho) / K \}}{\min_{(1 \leq v \leq K)} (1 - \theta_{va}^2) \left(\bar{\theta} / \theta_{va} \right)^2 (Z_{\alpha/K} + Z_\beta)^2}$$

When values of $\theta_{1a}^2, \dots, \theta_{Ka}^2$ are close to each other, $N/N_B \approx (z_\alpha + z_\beta)^2 \{ \rho(1-\rho)/K / (z_{\alpha/K} + z_\beta)^2 \} \leq (z_\alpha + z_\beta)^2 / (z_{\alpha/K} + z_\beta)^2$. For fixed α and β , it decreases to zero quickly as K increases. Thus, if a treatment shows similar magnitude of benefit across different outcomes, Bonferroni adjustment becomes more conservative as the number of outcomes K increases. On the other hand, when values of $\theta_{1a}^2, \dots, \theta_{Ka}^2$ are quite different from each other, especially when one of them is approaching to one, N_B is decreasing quickly (no matter what K value is), and N/N_B can be much greater than one. This is intuitive since Bonferroni adjustment is designed to test whether

a treatment shows any effect on at least one of the K outcomes while rank-sum-type test is designed to evaluate its average effect measured by GTE. It is worth noting that many extended Bonferroni procedures (e.g. [18-20]) are proposed to increase the test power when all θ_{va}^2 are close to each other, since they also compare p -values from univariate tests with values greater than α/K . If the goal is to test whether a treatment has any effect on K outcomes, the use of Bonferroni adjustment or its extensions will be more efficient. However, when the goal is to test whether a treatment has benefit on most of the outcomes, the use of rank-sum test (or Z) and (5) will be more efficient.

8. MULTIVARIATE CASES: MULTIPLE BINARY OUTCOMES

Using GEE, Rochon [10] presented a sample size computation algorithm for a binary outcome measured longitudinally when the correlation between the measure at time τ_t and the measure at time τ_s has the form $\phi^{|\tau_t - \tau_s|}$ ($s \neq t$). Since GEE models treatment means and correlations independently and ignores the restrictions on the correlation by the means, Rochon [10] acknowledged the difficulties in applying his approach for repeated binary data. In general, when the form of correlation structure among multiple binary outcomes is unknown, it is difficult to extend Rochon's sample size procedure for the multiple binary outcome case. Under the special case when the correlation matrix among outcomes is a known function of some unknown parameters, Rochon's procedure can be applied for sample size computation with multiple binary outcomes measured at a single time point. A typical feature of a binary outcome is that the outcome variance is a function of outcome mean (the proportion of success). If the parameter to be tested is a function of the outcome means, then both test statistic and the sample size will be driven by the magnitude of the means. This is not a desired property for many study designs. In this section, we show how sample size from Rochon's procedure and our formulas (8) and (9) in Section 4 are affected by the outcome means.

Denote $\mu_{iv} = P(X_{iv}=1)$, $\sigma_{iv}^2 = \mu_{iv}(1 - \mu_{iv})$ and $\mu = (\mu_{11}, \dots, \mu_{1K}, \mu_{21}, \dots, \mu_{2K})^T$. Then $\theta_v = \mu_{2v} - \mu_{1v}$. Consider the case when $\text{corr}(X_{iv}, X_{iv'}) = \rho$ ($u \neq v, i=1,2$) is known. The design matrix X in [10] is a $(2K) \times 2$ matrix of the form

$$X^T = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix}$$

Using the logit link $h(\mu) = X\beta$ where $\beta = (\beta_1, \beta_2)^T$, the model-based covariance matrix of $\hat{\beta}$ from GEE procedure is

$$\text{var}(\hat{\beta}) = \frac{1}{1 + 2(K-1)\rho - (2K-1)\rho^2} \times \begin{bmatrix} \rho \left(\sum_{j=1}^K \sigma_{1j} \right)^2 + \{1 + (2K-3)\rho\} \sum_{j=1}^K \sigma_{1j}^2 & 0 \\ 0 & \rho \left(\sum_{j=1}^K \sigma_{2j} \right)^2 + \{1 + (2K-3)\rho\} \sum_{j=1}^K \sigma_{2j}^2 \end{bmatrix}$$

Testing $\bar{\theta} = 0$ is equivalent to testing $\beta_1 - \beta_2 = 0$. Now we have $\{1 + 2(K-1)\rho - (2K-1)\rho^2\}$
 $\text{var}(\hat{\beta}_1 - \hat{\beta}_2) = \rho \left[\left(\sum_{j=1}^K \sigma_{1j} \right)^2 + \left(\sum_{j=1}^K \sigma_{2j} \right)^2 \right] + \{1 + (2K-3)\rho\} \sum_{j=1}^K (\sigma_{1j}^2 + \sigma_{2j}^2)$. It is seen that if all outcomes form two clusters, one with success probabilities very close to 0.5, the other with success probabilities close to 0 or 1, then, no matter whether $\tau_v = \mu_{2v} - \mu_{1v}$ remains the same across the K outcomes, the variance of $\hat{\beta}_1 - \hat{\beta}_2$, and thus the sample size or the test statistic, will mainly be determined by the variances from the outcomes in the first cluster. In other words, both test statistic and sample size computation will be driven by the outcomes whose

success rates are close to 0.5. In sample size computation, whether outcomes form two such clusters may be discovered by looking at whether prespecified values $\theta_{1a}, \dots, \tau_{Ka}$ form two such clusters.

The power of the rank-sum-type test and the sample size formulas (8) and (9) when some pilot data are available also have the same feature. However, because of the presence of matrix C in (2), (8) and (9) could be less seriously affected than the GEE approach. To see this, consider

the simple case when all binary outcomes are independent. Let $\bar{x}_{iu}^{(o)} = (1/m_i) \sum_{j=1}^{m_i} x_{iju}^{(o)}/m_i$. Then $s_{1uu} = m_2^2 \bar{x}_{1u}^{(o)} (1 - \bar{x}_{1u}^{(o)})/4$, $s_{2uu} = m_1^2 \bar{x}_{2u}^{(o)} (1 - \bar{x}_{2u}^{(o)})/4$ and $s_{3uu} = \left(1 + \bar{x}_{2u}^{(o)} - \bar{x}_{1u}^{(o)}\right)^2 \bar{x}_{1u}^{(o)} (1 - \bar{x}_{2u}^{(o)}) + \left(1 - \bar{x}_{2u}^{(o)} + \bar{x}_{1u}^{(o)}\right)^2 \left(1 - \bar{x}_{1u}^{(o)}\right) \bar{x}_{2u}^{(o)}$. Sample size, N , in (8) or (9) depends on data through

$$J^T S_1 J = \sum_{u=1}^K s_{1uu} = \sum_{u=1}^K \left(m_2^2/4\right) \bar{x}_{1u}^{(o)} (1 - \bar{x}_{1u}^{(o)}), J^T S_2 J = \sum_{u=1}^K s_{2uu} = \left(m_1^2/4\right) \bar{x}_{2u}^{(o)} (1 - \bar{x}_{2u}^{(o)}) \text{ and}$$

$J^T S_3 J = \sum_{u=1}^K s_{3uu}$. Thus, values of both $J^T S_1 J$ and $J^T S_2 J$ will be driven by outcomes in the first cluster. This situation is encountered because, with binary outcomes, the variance is a function of the mean. This may be viewed as an undesirable feature of both our approach and Rochon's approach.

9. SIMULATION

To evaluate the performance of sample size formulas, we simulated data from distributions

with different shapes. Let $x_{iju} = w_u y_{ij0} + \sqrt{1 - w_u^2} y_{iju}$, $i=1,2; j=1, \dots, n_i; u=1, \dots, K$. The y variables are generated independently from normal, exponential, and beta distributions, respectively. We consider $K=2,5,10$, $\alpha=0.05$, and $\beta=0.10, 0.20$. Different correlation structures are considered: $w_1=\dots=w_K=0.5$ for equal correlation and $w_1=\dots=w_{K-1}=0.5$ or 0.8 , $w_K=0$ for unequal correlation. For beta distribution, data are further discretized into five ordered values depending on which of the five intervals they are belonging: $[0,0.2]$, $(0.2,0.4]$, $(0.4,0.6]$, $(0.6,0.8]$, and $(0.8,1]$. We fix $r_0=1$. For normal distribution, data from both groups are generated from the standard normal distribution under null hypothesis H_0 , and 0.5 is added to data in group 2 under the alternative hypothesis H_1 . For exponential distribution, data from both groups are generated from the standard exponential distribution with density $f(x)=e^{-x}$ under H_0 , and data from group 2 are generated from the distribution with density $f(x)=2e^{-2x/3}/3$ under H_1 . For beta distribution, data from both groups are generated from Beta(0.5, 1) distribution under H_0 , and data from group 2 are generated from Beta(0.7, 1) under H_1 .

Sample size $n_1=N/(1+r_0)=N/2$ in Table II is computed using (5) with $\rho=0.5$ or 0.8 , $\sigma^2=\frac{1}{12}$ and $\bar{\theta}_a$ being computed using methods in Section 2. Using these sample sizes, 5000 simulations were performed to evaluate the type I error and statistical power at $\bar{\theta}_a$ of rank-sum-type test and the test based on Z in (3). Since data from both groups are generated from the same distribution under H_0 , we have $h_1=h_2=1$ [7]; thus, both rank-sum-type test and its modified test are identical.

Table II shows that sample size computed from (5), with ρ replaced by the maximum correlation among $X_{i1}, \dots, X_{iK}$, gives good control of type I error and statistical power under various distribution settings. In all cases, sample size computed from formula (5) is slightly conservative. For unequal correlation case, the K th outcome is simulated independent of the other $K-1$ outcomes. It is not surprising to see that the resulting statistical power is higher than the corresponding equal correlation case, since we still substitute ρ by the same maximum

correlation among $X_{i1}, \dots, X_{iK}$ in sample size computation. This is reasonable because adding independent outcome with the same magnitude of treatment effect as other outcomes will always increase the statistical power.

10. APPLICATION IN PD TRIAL DESIGN

The neuroprotection exploratory trials in Parkinson's disease (NET-PD) are a series of PD clinical trials sponsored by NIH/NINDS to identify the most promising neuroprotective compounds for individuals with PD [21]. A phase III trial in NET-PD was designed to compare creatine with placebo in the following five primary outcomes: the changes from baseline visit to the end of 5-year follow-up in the measures of modified Rankin, symbol digit modalities (verbal), Schwab and England ADL scale (SEADL), PDQ-39, and ambulatory capacity. The study sets power at 85 per cent when the creatine group would achieve an overall mean reduction computed from the mean reductions over placebo of 3, 0.2, 2, 0.33, and 1.5, respectively, among these five primary outcomes.

Based on the literature and discussions among trial investigators, the common standard deviations for these five outcomes were set at 9, 1, 11, 2.11, and 8.0, respectively. Using normal approximation formula in Section 2, we computed the alternative $(\theta_{1a}, \dots, \theta_{5a}) = (0.1863, 0.1125, 0.1023, 0.0881, \text{ and } 0.1055)$. This gives $\bar{\theta}_a = 0.1189$. Bonferroni adjustment to the five one-sided tests required a sample size of 925 patients per group. Since no pilot data were available to estimate covariance matrices A and B in (4) in 5-year outcome changes, formula (5) was used to estimate sample size. It was assumed that, under the null hypothesis, creatine

group and placebo group would have the same distribution. The upper bound $\sigma^2 = \frac{1}{4}$ was substituted into (5). The trial uses randomization ratio $r_0 = 1$. The following table gives a range of sample size per group, using formula (5), when ρ ranges from 0.1 to 1.0 and $(\alpha, \beta) = (0.05, 0.15)$:

ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Sample size per group	238	305	373	441	509	576	644	712	780	847

In a pilot 2-year phase II study, a 1-year change of modified Rankin, SEADL, PDQ-39, and ambulatory capacity was measured. The absolute values of correlation coefficients among these four outcomes ranged from 0.002 to 0.35. If it was assumed that correlations among all five outcomes in their 5-year changes were not greater than 0.5, we could choose a sample size of 509 patients per treatment group.

11. DISCUSSION

The clinical trials' literature is replete with discussions of sample size determination. While there are publications describing the computations for multivariate settings, appropriately the vast body of literature is focused on the univariate case in which a single outcome variable has been identified for the primary efficacy assessment. Lachin [1] has catalogued a number of useful expressions for clinical trial sample size calculations in this single variable situation. Little work has been done in sample size computation even for single ordered categorical outcome [22]. In PD and in many other diseases, it is difficult, if not impossible, to select a single measured variable that will capture the multi-dimensional features associated with disease improvement or deterioration. Multiple continuous and ordered categorical outcomes are often required to be analyzed jointly to determine whether a new treatment should be recommended. In these cases, the approach described in this paper may have merit. In addition,

many clinical trials are designed on the premise that the two treatment groups have identical distributions of response variables under the null hypothesis. Again, the strategy used in this paper does not require such strict equality, but permits inequality of variance–covariance structures between the two treatment groups.

When faced with several important variables for a trial, there are a number of approaches that might be followed to plan the sample size for the study. One approach is to select a single variable as primary and consider the others as important secondary variables. The sample size planning is then driven by the one identified primary variable, and once a sample size is calculated, it is used to calculate what power the study will have with that N for the secondary variables. Another approach is to identify two primary variables and then calculate sample size for each while being certain to incorporate the formal testing strategy. That is, will the trial require differences for both variables or for at least one of the two? A Bonferroni multiple test adjustment might be employed. We have seen that the Bonferroni sample size calculations will require greater numbers than the rank-sum-type test when treatment effects are close to each other across variables.

A final benefit of this approach is its applicability to a variety of settings involving continuous or discrete random variables. This renders substantial flexibility for use in planning a trial with a complex constellation of outcomes. Admittedly, our approach is based on a parameterization that focuses on the GTE and tacitly assumes that it makes clinical sense to combine effects in the manner proposed through the GTE. If nothing else, this approach can be used to calculate sample size and then can be contrasted to the result from calculations performed using the more traditional approaches assuming normality, equality of variance, and other restrictions.

There are limitations to the development outlined herein. First, like many sample size papers we have not focused on the many issues that can complicate the design of a trial. These include: issues of noncompliance, attrition, the registration of ineligible of patients, and similar pragmatic matters. Furthermore, the development focuses on the GTE and is essentially aggregating effects across variables. If the variables interact with treatment producing qualitative interactions, then the interpretation of this planning exercise will be unclear. This, however, is true for any similar multivariate analysis that summarizes an effect into a main effect averaged over the levels of an interacting factor. Hence, knowledge of the clinical application and understanding the variables are critical aspects to the use of these results.

We believe the approach outlined here represents a new way of thinking about this problem, and it has the potential to be applied to a number of clinical trial settings, particularly for chronic diseases.

Acknowledgments

The authors acknowledge the research support of National Institutes of Health Grants #AG023630-01A29002 from NIA, #NS43127 from NINDS, #RR01070 from NCRR, and #HL072377 from NHLBI.

Contract/grant sponsor: NIA; contract/grant number: AG023630-01A29002

Contract/grant sponsor: NINDS; contract/grant number: NS43127

Contract/grant sponsor: NCRR; contract/grant number: RR01070

Contract/grant sponsor: NHLBI; contract/grant number: HL072377

APPENDIX A: PROOF OF $\sigma^2 = (1 - \min_{(1 \leq v \leq K)} \theta_{va}^2)/4$ IN SECTION 3

Consider the case when $\theta_v = \theta_{va}$ ($v=1, \dots, K$). Applying the fact that $E[F_{1v}(X_{2v})] = (1 + \theta_{va})/2$ and $E[F_{2v}(X_{1v})] = (1 - \theta_{va})/2$ ($v=1, \dots, K$), we obtain

$$\begin{aligned} \text{var}\{F_{2v}(X_{1v})\} &= E\{F_{2v}(X_{1v})\}^2 - [E\{F_{2v}(X_{1v})\}]^2 \\ &= \frac{1 - \theta_{va}^2}{4} - \left(\frac{1 - \theta_{va}}{2}\right)^2 \\ &= \frac{1 - \theta_{va}^2}{4} \end{aligned}$$

Similarly, we have $\text{var}\{F_{1v}(X_{2v})\} \leq (1 - \theta_{va}^2)/4$. Now,

$$\sigma^2 = \max_{(1 \leq v \leq K)} (1 - \theta_{va}^2)/4 = (1 - \min_{(1 \leq v \leq K)} \theta_{va}^2)/4.$$

APPENDIX B: PROOF OF THEOREM 1 IN SECTION 4

First, compute

$$\begin{aligned} E\left[\frac{1}{m_2^2} S_1\right] &= \frac{(m_1-1)(m_2-1)}{m_1 m_2} A - \frac{m_1-1}{m_1 m_2} B + \frac{m_1-1}{4 m_1 m_2} C \\ E\left[\frac{1}{m_1^2} S_2\right] &= \frac{(m_1-1)(m_2-1)}{m_1 m_2} B - \frac{m_2-1}{m_1 m_2} A + \frac{m_2-1}{4 m_1 m_2} C \\ E[S_3] &= -\frac{4(m_2-1)}{m_1 m_2} A - \frac{4(m_1-1)}{m_1 m_2} B + \frac{m_1 m_2 - 1}{m_1 m_2} C \end{aligned}$$

Second, substitute the above into the expressions of $E[\Gamma_1]$ and $E[\Gamma_2]$. Third, compute the expectation $E[\widehat{\Sigma}]$ using (7). Finally, apply (2) to complete the proof.

APPENDIX C: SPLUS CODE

The following function **GST.parameter** is used to compute test statistic Z and sample size N :

```
GST.parameter<- function(data, n1)
{
  # N=n1+n2, n1=sample size from group 1, n2=sample size from group 2.
  # data is a N-by-k matrix whose first n1 rows are observations
  # from group 1, and the rest n2 rows are observations from group 2.
  # No missing observation is allowed!

  # when data contain observations from current study,
  # Sigma.hat is an unbiased estimate of Sigma that can be used in
  # test statistic Z, theta.hat is an unbiased estimate of theta.

  # when data contains observations from a pilot study and n1=m1 is
  # the sample size in group 1 in the pilot study, then
  # S1=S1.n1n2*n1*n2, S2=S2.n1n2*n1*n2, S3=S3.n1n2*n1*n2
  # are matrices S1, S2, S3 defined in Theorem 1.

  data<- as.matrix(data)
  if (is.na(sum(data))==T){stop("NAs are not allowed in data")}
```

```

N<- dim(data)[1]
n2<- N-n1
k<- dim(data)[2]
Sigma.hat<- matrix(NA,ncol=k,nrow=k)
theta<- rep(NA,k)

ind.2<- data*0
ind.2[(n1+1):N,]<- 1 # indicate observations in group 2.

# compute the total ranks for group 2
R<- apply(data, 2, rank)
R<- apply(R * ind.2, 2, sum) # group 2 rank sums

# compute an unbiased estimate of theta
theta.hat<- (2*R-n2*(N+1))/n1/n2
A1<- array(NA, dim=c(n1,k))
B1<- array(NA, dim=c(n2,k))

# To avoid numerical overflow, we do not compute S1, S2, and S3
# but compute their ratios with n1*n2

# compute matrix S2.n1n2=S2/n1/n2:
x2<- array(NA, dim=c(n2,k))
for(i in (n1+1):N)
{
  x1<- data[c(i,1:n1),]
  Rx<- apply(x1, 2, rank)
  x2[i-n1,]<- Rx[1,]
  B1[i-n1,]<- 2*Rx[1,]-2-n1 -n1*theta.hat
}
S2.n1n2<- var(x2, unbiased=FALSE)/n1/n2

# compute matrix S1.n1n2=S1/n1/n2:
x2<- array(NA, dim=c(n1,k))
for( i in 1:n1)
{
  x1<- data[c(i,(n1+1):N),]
  Ry<- apply(x1, 2, rank)
  x2[i,]<- Ry[1,]
  A1[i,]<- 2*Ry[1,]-2-n2 +n2*theta.hat
}
S1.n1n2<- var(x2, unbiased=FALSE)/n1/n2

# compute matrix S3.n1n2=S3/n1/n2:
x1<- data[rep(1:n1, times=n2),]
x2<- data[rep(((n1+1):N), each=n1),]
x3<- 1*(x1<x2)-1*(x1>x2)
S3.n1n2<- var(x3, unbiased=FALSE)/n1/n2

# compute an unbiased estimate of Sigma
# Sigma.hat= N/(n1-1)/(n2-1)*(S1/n2+S2/n1-S3/4)
Sigma.hat<- N/(n1-1)/(n2-1)*(S1.n1n2*n1+S2.n1n2*n2-S3.n1n2/4*n1*n2)

```

```

if(sum(Sigma.hat)<=0)
{
  # the following Sigma.hat is only a consistent estimate of Sigma
  # Since matrix S3 is ignored:
  Sigma.hat<- N/(n1-1)/(n2-1)*(S1.nln2*n1+S2.nln2*n2)
}
# compute h1, h2 for adjusted rank-sum-type test
x1<- data[1:n1,]
Rx<- apply(x1,2,rank)
A2<- 2*Rx-1-n1
y1<- data[(n1+1):N,]
Ry<- apply(y1,2,rank)
B2<- 2*Ry-1-n2
h1<- N^2/n1/n2*sum(t(A1)%%% A1+ t(B1)%%% B1)/
  sum(t(A1+A2)%%% (A1+A2)+ t(B1+B2)%%% (B1+B2))
h2<- N^2*sum(t(A1)%%% A1+ t(B1)%%% B1)/
  sum(n2^2*t(A1+A2)%%% (A1+A2)+ n1^2*t(B1+B2)%%% (B1+B2))

list(theta=theta.hat, Sigma=Sigma.hat,
      S1.nln2=S1.nln2, S2.nln2=S2.nln2, S3.nln2=S3.nln2,
      h1=h1,h2=h2)
}

```

C.1. Computation of Z , $\hat{\Sigma}$, $\hat{h}_1$, and $\hat{h}_2$

```

x<- GST.parameter(data, n1)
Z<- sum(x$theta)*sqrt(N)/2/sum(x$Sigma) Sigma.hat<- x$Sigma
h1.hat<- x$h1
h2.hat<- x$h2

```

C.2. Computation of sample size N using formula (8)

Suppose that observations from a pilot study are saved in *Splus* $(m_1+m_2) \times k$ matrix **pilot.data** whose first m_1 rows are observations from group 1 and the rest m_2 rows are observations from group 2.

```

z.alpha<- qnorm(1-alpha)
z.beta<- qnorm(1-beta)
x<- GST.parameter(pilot.data, m1)
d1.m1m2<- r0*(m1-1)*(m2-1) #=d1/m1/m2
d2<- (z.alpha+z.beta)^2/k^2/theta.bar^2
# compute gamma1.m1m2=gamma1/m1/m2 and gamma2.m1m2=gamma2/m1/m2
gamma1.m1m2<- (4*m1 + 4*m1^2*r0)*sum(x$S1.nln2) +
  (4*m2^2 + 4*m2*r0)*sum(x$S2.nln2) -
  (m1*m2^2 + m1^2*m2*r0)*sum(x$S3.nln2)
gamma2.m1m2<- 4*m1^2*sum(x$S1.nln2) +
  4*m2^2*sum(x$S2.nln2) - m1^2*m2^2*sum(x$S3.nln2)
# The required total sample size N is:
N<- ceiling(((1+r0)*d2*gamma1.m1m2/2/d1.m1m2*(1+
  sqrt(1-4*d1.m1m2*gamma2.m1m2/d2/gamma1.m1m2^2))))

```

REFERENCES

1. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*. 1981; 2:93–113. [PubMed: 7273794]
2. Brunner E, Munzel U, Puri ML. The multivariate nonparametric Behrens–Fisher problem. *Journal of Statistical Planning and Inference*. 2002; 108:37–53.
3. Wolfe DA, Hogg RV. On constructing statistics and reporting data. *The American Statistician*. 1971; 25:27–30.
4. O'Brien P, Geller N. Interpreting tests for efficacy in clinical trials with multiple endpoints. *Controlled Clinical Trials*. 1987; 18:222–227.
5. Acion L, Peterson J, Temple S, Arndt S. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*. 2007; 25:591–602. [PubMed: 16143965]
6. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics*. 1984; 40:1079–1087. [PubMed: 6534410]
7. Huang P, Tilley B, Woolson RF, Lipsitz S. Adjusting O'Brien's test to control type I error for the generalized nonparametric Behrens–Fisher problem. *Biometrics*. 2005; 61:532–539. [PubMed: 16011701]
8. Tang DI, Gnecco L, Geller NL. Design of group sequential clinical trials with multiple endpoints. *Journal of the American Statistical Association*. 1989; 84:776–779.
9. Whitehead J. Sample size calculations for ordered categorical data (Corr: 94V13 p871). *Statistics in Medicine*. 1993; 12:2257–2271. [PubMed: 8134732]
10. Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. *Statistics in Medicine*. 1998; 17:1643–1658. [PubMed: 9699236]
11. Noether GE. Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association*. 1987; 82:645–647.
12. Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archive of General Psychiatry*. 2006; 63:484–489.
13. Sundukchi, MS.; Guenther, WC. Some results on sample size problems for inferences based on the rank-sum Wilcoxon statistics; ASA Proceedings of the Sections on Statistical Computing and Statistical Graphics; American Statistical Association: Alexandria, VA. 2000; p. 92-97.
14. Lehmann, EL.; D'Abrera, HJM. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day Inc.; San Francisco: 1975.
15. Hettmansperger, TP. *Statistical Inference Based on Ranks*. Wiley; New York: 1984.
16. Puri, ML.; Sen, PK. *Nonparametric Methods in Multivariate Analysis*. Krieger Publishing Company; Malabar, FL: 1993.
17. Agresti, A. *Categorical Data Analysis*. Wiley; New York: 2002.
18. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 1979; 6:65–70.
19. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*. 1986; 73(3):751–754.
20. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988; 75:800–802.
21. Current ongoing NIH sponsored Neuroprotection Exploratory Trials in Parkinson's Disease. Available at: NIH websites <http://www.parkinsontrial.ninds.nih.gov/> and <http://www.nih.gov/news/pr/mar2007/ninds-22.htm>
22. Rabbee N, Coull BA, Mehta C, Patel N, Senchaudhuri P. Power and sample size for ordered categorical data. *Statistical Methods in Medical Research*. 2003; 12:73–84. [PubMed: 12617509]

Table I

Values of $\theta^{(o)} = P(X_1^{(o)} [ldb] X_2^{(o)}) - P(X_1^{(o)} [ldb] X_2^{(o)})$.

$X_I^{(o)}$ Poisson(λ_I)				$X_I^{(o)}$ Beta(a_r, β_l)				$X_I^{(o)}$ Gamma(a_r, β)			
λ_1	λ_2/λ_1	$\theta^{(o)}$	α_1	β_1	α_2	β_2	$\theta^{(o)}$	α_1	a_2/α_1	β	$\theta^{(o)}$
0.25	0.30	-0.1503	0.5	0.5	0.7	0.5	0.1391	0.5	0.5	0.5	-0.3662
0.25	0.50	-0.1052	0.5	0.5	0.9	0.5	0.2353	0.5	0.8	0.5	-0.1269
0.25	0.70	-0.0619	0.5	1.0	0.7	1.0	0.1667	1.0	0.5	0.5	-0.4142
0.25	0.90	-0.0202	0.5	1.0	0.7	1.5	0.0091	1.0	0.8	0.5	-0.1487
0.25	0.95	-0.0100	0.5	1.5	0.5	1.0	0.1512	3.0	0.5	0.5	-0.5689
0.50	0.30	-0.2628	0.5	1.5	0.5	1.7	-0.0456	3.0	0.8	0.5	-0.2202
0.50	0.50	-0.1820	0.5	1.5	0.7	1.7	0.1300	5.0	0.5	0.5	-0.6716
0.50	0.70	-0.1059	0.5	2.0	0.5	1.7	0.0582	5.0	0.8	0.5	-0.2734
0.50	0.90	-0.0342	0.5	2.0	0.5	1.5	0.1035	10.0	0.5	0.5	-0.8204
0.50	0.95	-0.0170	0.5	2.0	0.7	1.7	0.2435	10.0	0.8	0.5	-0.3709
1	0.3	-0.4201	0.5	2.0	1.0	1.5	-0.0183	0.5	0.3	1	-0.5727
1	0.5	-0.2892	1.0	1.5	1.0	1.7	-0.0625	0.5	0.5	1	-0.3662
1	0.7	-0.1669	1.0	1.5	1.5	1.7	0.1619	0.5	0.8	1	-0.1269
1	0.9	-0.0535	1.0	2.0	1.0	1.7	-0.3134	1.0	0.3	1	-0.6245
1	0.95	-0.0265	1.0	2.0	1.5	1.7	0.3170	1.0	0.5	1	-0.4142
3	0.3	-0.7129	1.0	2.0	1.5	2.0	0.5344	1.0	0.8	1	-0.1487
3	0.5	-0.5133	1.0	2.0	0.5	1.0	-0.0667	2.0	0.3	1	-0.7154
3	0.7	-0.3033	1.5	1.5	1.0	2.0	-0.3750	2.0	0.5	1	-0.5000
3	0.9	-0.0978	1.5	1.5	1.5	1.7	-0.07150	2.0	0.8	1	-0.1876
3	0.95	-0.0484	1.5	1.5	1.5	2.0	-0.1641	3.0	0.3	1	-0.7832
6	0.3	-0.8758	1.5	1.5	1.7	2.0	-0.0936	3.0	0.5	1	-0.5689
6	0.5	-0.6825	1.5	2.0	1.7	2.0	0.0771	3.0	0.8	1	-0.2202
6	0.7	-0.4236	1.5	2.0	2.0	1.7	0.2682	0.5	0.5	1.5	-0.3662
6	0.9	-0.1395	1.5	2.0	1.0	1.7	-0.4732	0.5	0.8	1.5	-0.1269
6	0.95	-0.0691	1.5	2.0	1.0	1.5	-0.4248	1.0	0.5	1.5	-0.4142
10	0.3	-0.9545	1.5	2.0	0.5	2.0	-0.5625	1.0	0.8	1.5	-0.1487
10	0.5	-0.8058	2.0	2.0	1.0	1.5	-0.5344	1.5	0.5	1.5	-0.4594

$X_i^{(\phi)}$ -Poisson(λ_i)			$X_i^{(\phi)}$ -Beta(a_r, β_l)			$X_i^{(\phi)}$ -Gamma(a_r, β)					
10	0.7	-0.5315	2.0	2.0	1.0	2.0	-0.6400	1.5	0.8	1.5	-0.1690
10	0.9	-0.1803	2.0	2.0	1.7	2.0	-0.2407	2.0	0.5	1.5	-0.5000
10	0.95	-0.0896	2.0	2.0	1.7	1.5	0.0771	2.0	0.8	1.5	-0.1876

Note: Parameters are chosen to capture different shapes of the density function.

Table II

Sample size n_1 is computed from (5) with $r_0=1$, $\alpha = 0.05$, $\sigma^2 = \frac{1}{12}$ and $\bar{\theta}_a$ being computed using methods in Section 2.

Distribution	Correlation	$\bar{\theta}_a$	K	β	$n_1=n_2$	Simulated type I error		Simulated power	
						$T_1=T_{1a}$	Z	$T_1=T_{1a}$	Z
Normal	Equal $\rho=0.5$	0.2763	2	0.10	58	0.0470	0.0520	0.9130	0.9174
				0.20	42	0.0492	0.0574	0.8270	0.8330
			5	0.10	46	0.0480	0.0569	0.9181	0.9240
				0.20	34	0.0462	0.0576	0.8192	0.8318
	Unequal $\rho=0.5$	0.2763	10	0.10	42	0.0486	0.0560	0.9160	0.9243
				0.20	31	0.0518	0.0578	0.8270	0.8462
			5	0.10	46	0.0506	0.0596	0.9698	0.9726
				0.20	34	0.0478	0.0652	0.9184	0.9242
	Unequal $\rho=0.8$	0.2763	10	0.10	42	0.0522	0.0638	0.9528	0.9578
				0.20	31	0.0492	0.0650	0.8822	0.8952
			5	0.10	65	0.0542	0.0568	0.9792	0.9800
				0.20	47	0.0506	0.0616	0.9270	0.9296
Exponential	Equal $\rho=0.5$	0.2900	2	0.10	51	0.0504	0.0517	0.9184	0.9066
				0.20	37	0.0466	0.0522	0.8332	0.8174
			5	0.10	41	0.0492	0.0538	0.9260	0.9216
				0.20	30	0.0514	0.0566	0.8370	0.8388
	Unequal $\rho=0.5$	0.2900	10	0.10	38	0.0522	0.0536	0.9330	0.9374
				0.20	27	0.0498	0.0506	0.8388	0.8488
			5	0.10	41	0.0530	0.0610	0.9596	0.9606
				0.20	30	0.0506	0.0601	0.8904	0.8958
	Unequal $\rho=0.8$	0.2900	10	0.10	38	0.0520	0.0576	0.9484	0.9538
				0.20	27	0.0508	0.0600	0.8656	0.8776
			5	0.10	58	0.0518	0.0538	0.9680	0.9672
				0.20	42	0.0498	0.0564	0.9018	0.9014
			10	0.10	56	0.0484	0.0510	0.9492	0.9490

Distribution	Correlation	θ_a	K	β	$n_1=n_2$	Simulated type I error		Simulated power	
						$T_1=T_{1a}$	Z	$T_1=T_{1a}$	Z
Beta	Equal	0.1974	2	0.20	41	0.0528	0.0584	0.8840	0.8868
				0.10	110	0.0534	0.0528	0.9212	0.9214
			5	0.20	80	0.0516	0.0534	0.8414	0.8548
				0.10	88	0.0504	0.0488	0.9264	0.9346
	Unequal	0.1974	10	0.20	64	0.0546	0.0558	0.8514	0.8688
				0.10	81	0.0474	0.0466	0.9318	0.9436
			5	0.20	59	0.0538	0.0544	0.8600	0.8830
				0.10	88	0.0530	0.0541	0.9676	0.9740
	Unequal	0.5	10	0.20	64	0.0506	0.0518	0.8990	0.9138
				0.10	81	0.0466	0.0482	0.9540	0.9622
			5	0.20	59	0.0512	0.0510	0.8918	0.9148
				0.10	124	0.0522	0.0508	0.9544	0.9596
	Unequal	0.8	10	0.20	89	0.0472	0.0484	0.8878	0.9050
				0.10	121	0.0520	0.0496	0.9446	0.9552
			5	0.20	87	0.0474	0.0468	0.8598	0.8836
				0.10	121	0.0520	0.0496	0.9446	0.9552

Note: Type I error and power are computed from 5000 simulations with sample size $n_1=n_2$ in both groups.