Two-stage k-sample designs for the ordered alternative problem

Guogen Shan, Alan D. Hutson, and Gregory E. Wilding*

Department of Biostatistics,University at Buffalo, Buffalo, NY 14214, USA

July 18, 2011

Abstract

In pre-clinical studies and clinical dose-ranging trials, the Jonckheere-Terpstra test is widely used in the assessment of dose-response relationships. Hewett and Spurrier [1] presented a two-stage analog of the test in the context of large sample sizes. In this paper, we propose an exact test based on Simon's minimax and optimal design criteria originally used in one-arm phase II designs based on binary endpoints. The convergence rate of the joint distribution of the first and second stage test statistics to the limiting distribution is studied and design parameters are provided for a variety of assumed alternatives. The behavior of the test is also examined in the presence of ties and the proposed designs are illustrated through application in the planning of a hypercholesterolemia clinical trial. The minimax and optimal two-stage procedures are shown to be preferable as compared to the one-stage procedure due to the associated reduction in required sample size.

Keywords: Clinical Trials, Exact Tests, Jonckheere-Terpstra Test, Minimax Designs, Nonparametric, Optimal Designs, Phase II Designs.

1 Introduction

In pre-clinical studies and clinical dose-ranging trials, investigators are frequently interested in assessing dose-response relationships, the most common conjecture being that the response is a monotonic increasing (decreasing)

*Corresponding author. Department of Biostatistics,University at Buffalo, Buffalo, NY 14214, USA

E-mail address:gwilding@buffalo.edu

function of dose [2]. The assumption of monotonicity is often rooted in historic data, or based on biological reasoning. Consider the clinical trial discussed by Dmitrienko et al. [3] in which patients with hypercholesterolemia are randomized to four groups where patients received one of three doses of a cholesterol-lowering drug or placebo. This four-arm clinical trial is based on a continuous primary endpoint, the reduction in low-density lipoprotein cholesterol after a 12-week treatment. It is hypothesized that the underlying dose response relationship is monotonic increasing.

The null hypothesis to be tested among k ordinal groups is

$$H_0 : \theta_1 = \cdots = \theta_k,$$

and the ordered alternative is

$$H_a : \theta_1 \leq \cdots \leq \theta_k \quad \text{and} \quad \theta_1 < \theta_k,$$

where θ_i is the location parameter associated with group i, $i = 1, 2, \cdots, k$. The above problem has received considerable attention in the literature. Jonckheere [4] and Terpstra [5] were among the first to develop a nonparametric test for the nondecreasing ordered alternative (referred to as the JT test) based on the linear combination of the Mann-Whitney U-statistics [6] associated with the $k(k-1)/2$ possible pairs between the k ordinal groups. Tryon and Hettmansperger [7] proposed a modified procedure based on a weighted version of the original test statistic which was shown to be equivalent to calculating Spearman's rank correlation between the observed data and the corresponding group number. Neuhäuser et al. [8] showed that this modified test can be more powerful than the JT test with smaller sample sizes for reasons related to the less discrete distribution of the test statistic. More recently, McKean et al. [9] implemented a bootstrap Spearman-based approach to the problem. In addition to the p-value, their method produces a measure of association with a range of -1 to 1 along with a corresponding confidence interval. Terpstra and Magel [10] proposed a nonparametric test based on the total number of possible sets of observations (one from each group) which follow the alternative ordering. The receiver operating characteristic (ROC) curve is commonly used for evaluating the ability of a diagnostic marker measured on a continuous scale to correctly classify subjects into two groups. An extension of the ROC curve, namely the ROC surface, has been proposed in the context of three groups [11, 12]. When $k = 3$, the volume under the ROC surface is proportional to the test statistic given by Terpstra and Magel. This is analogous to the

well known relationship observed between the area under the ROC curve and the Mann-Whitney test statistic. Cuzick [13] proposed an extension of the signed rank Wilcoxon test based on a weighted rank for use in testing for trends. Le [14] proposed a test for monotone ordered alternatives similar to the Kruskal-Wallis test which is equivalent to the Cuzick test when the sample sizes are equal for all groups. Mahrer and Magel [15] conducted a study which compared the JT test with those due to Cuzick and Le and found that all three tests were comparable in terms of power.

Two-stage tests have been proposed for a variety of problems with the primary advantage being a reduction in sample size compared to the usual one-stage testing approach. The general procedure for a two-stage test begins with the selection of the first stage subjects used to assess whether continuation of the trial beyond the interim analysis is warranted. Following many two-stage designs applied in clinical trials, the sampling procedure will be stopped only for lack of evidence in favor of the alternative based on the first stage data. If we do not decide in favor of the null in the first stage, additional subjects are enrolled in the second stage. The final decision is based on all the observations from both stages. A number of two-stage procedures have been proposed in the context of clinical research for use with binary outcomes [16, 17, 18]. Wilding et al. [19] executed an exact two-stage Mann-Whitney test for continuous endpoints and showed that the sample size savings are substantial as compared to the one-stage procedure. Generally, it is often true that two-stage designs can substantially reduce the expected sample sizes required to achieve a given power. A two-stage JT test has been proposed by Hewett and Spurrier [20, 1]. The asymptotic null distribution of the first and second stage JT test statistics, denoted by JT_1 and JT_2, respectively, was derived. Although critical values were provided, these values were calculated from the limiting null distribution with pre-specified probabilities of rejecting and accepting the null hypothesis at the first and second stage, and the critical values provided are not claimed to be optimal in any sense. In addition, the authors did not study the rate of convergence of the joint probability distribution of JT_1 and JT_2 to the limiting null distribution. There is no published optimal two-stage k-sample designs for pre-clinical and dose-ranging trials at this time.

In this paper, we consider two-stage k-sample designs for continuous endpoints for use with ordered alternatives. We review results pertaining to the limiting distribution of the two-stage JT test statistics and present a study

of the rate of convergence in Section 2. Section 3 is given to the procedures for obtaining the design parameters of the proposed Monte Carlo exact and asymptotic-based two-stage k-sample designs. The proposed designs are applied in Section 4 to a clinical experiment. Section 5 is given to some recommendations and discussion.

2 The test statistics

The JT test was proposed by Jonckheere and Terpstra independently for testing for monotonically ordered alternatives in the k-sample problem. The underlying distribution functions are assumed absolutely continuous and of the form, $F_i = F(x - \theta_i), i = 1, 2, \cdots, k$, and the sample sizes are $m_1, m_2, \cdots, m_k$ for the k populations. There is no difference among the populations under the null hypothesis, and the distributions under the alternative differ by their location parameters $\theta_i, i = 1, 2, \cdots, k$. Specifically, the hypotheses are

$$H_0 : \theta_1 = \cdots = \theta_k,$$

and

$$H_a : \theta_1 \leq \cdots \leq \theta_k \text{ and } \theta_1 < \theta_k.$$

The JT test statistic is given as

$$JT = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{ij},$$

where the Mann-Whitney test statistic U_{ij} is calculated based on the set of all possible pairs of observations from the $i-$th and $j-$th populations: $U_{ij} = \sum_{l=1}^{m_i} \sum_{v=1}^{m_j} I(X_{il} < X_{jv})$, where X_{il}, $l = 1, 2, \cdots, m_i$, denotes the $l-$th observation from the $i-$th group, and $I(.)$ denotes an indicator function that takes the value 1 if the condition in the parenthesis is satisfied, and takes the value 0, otherwise.

Hewett and Spurrier considered the two-stage analog of the JT test. Samples of sizes $m_1, m_2, \cdots, m_k$ are assigned to k groups, usually via a randomization scheme, the responses of which are used to compute the first stage JT test statistic

$$JT_1 = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{ij}^1. \tag{2.1}$$

If the procedure goes to the second stage, additional $n_1, n_2, \cdots, n_k$ samples are enrolled. The final decision is

based on the second stage test statistic, JT_2, which is calculated using the information from both stages,

$$JT_2 = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{ij}^2. \tag{2.2}$$

The statistics U_{ij}^1 and U_{ij}^2 are Mann-Whitney test statistics based on the samples from the first stage and both stages, respectively. Letting $U^{'} = (U_{12}^1, U_{12}^2, \cdots, U_{(k-1)k}^1, U_{(k-1)k}^2)_{1\times(k-1)k}$, the two-stage test statistics can be expressed as a linear combination of vector U,

$$JT_1 = C_1 U \quad \text{and} \quad JT_2 = C_2 U,$$

where $C_1 = (1, 0, 1, 0, \cdots, 1, 0)_{1\times(k-1)k}$ and $C_2 = (0, 1, 0, 1, \cdots, 0, 1)_{1\times(k-1)k}$. We take a similar approach as Wilding et al. [19] for obtaining designs based upon JT_1 and JT_2. Let corresponding critical values at each stage be denoted as r_1 and r. At the first stage, m_i subjects for the $i-$th treatment, $i = 1, 2, \cdots, k$, are enrolled. If the first stage JT_1 statistic is less than or equal to r_1 then it is deemed there exists no trend, otherwise it continues on to the second stage. In the second stage, more patients (n_i for the $i-$th treatment) are accrued. If at the end of the trial the second stage JT_2 statistic is less than or equal to r, we fail to reject H_0 and conclude that there is no difference among the groups. Otherwise, if $JT_2 > r$, we conclude that the θ_i's follow a nondecreasing ordering.

2.1 Asymptotic distribution of JT_1 and JT_2

We start with an investigation of the asymptotic distribution of the test statistics. The expectation and variance of the Mann-Whitney statistics U_{ij}^1 and U_{ij}^2 directly follow from results in Mann and Whitney [6]. The following lemma from Hewett and Spurrier [20] provides the first and second central moments associated with the two-stage JT test statistics.

Lemma 2.1. (Hewett and Spurrier [20]) *Under H_0, the expectation of the first and second stage JT test statistics, given in (2.1) and (2.2), are*

$$\mu_1 = E(JT_1) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{m_i m_j}{2}, \qquad \mu_2 = E(JT_2) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{(m_i + n_i)(m_j + n_j)}{2}.$$

The corresponding variances and covariance are

$$\sigma_1^2 = Var(JT_1) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{m_i m_j (m_i + m_j + 1)}{12} + \sum_{h=1}^{k-2} \sum_{i=h+1}^{k-1} \sum_{j=i+1}^{k} \frac{m_h m_i m_j}{6},$$

$$\sigma_2^2 = Var(JT_2) = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \frac{(m_i+n_i)(m_j+n_j)(m_i+n_i+m_j+n_j+1)}{12} + \sum_{h=1}^{k-2}\sum_{i=h+1}^{k-1}\sum_{j=i+1}^{k} \frac{(m_h+n_h)(m_i+n_i)(m_j+n_j)}{6},$$

and

$$\sigma_{12} = Cov(JT_1, JT_2) = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \frac{m_i m_j(m_i+n_i+m_j+n_j+1)}{12} + \sum_{h=1}^{k-2}\sum_{i=h+1}^{k-1}\sum_{j=i+1}^{k} \frac{m_h(m_i+n_i)m_j}{6}.$$

The following asymptotic result will allow us to form large sample two-stage JT procedures based on a standardized version of the vector U,

$$U^{st} = \left(\frac{U_{12}^1 - E(U_{12}^1)}{\sqrt{(Var(U_{12}^1))}}, \frac{U_{12}^2 - E(U_{12}^2)}{\sqrt{(Var(U_{12}^2))}}, \cdots, \frac{U_{k-1k}^1 - E(U_{k-1k}^1)}{\sqrt{(Var(U_{k-1k}^1))}}, \frac{U_{k-1k}^2 - E(U_{k-1k}^2)}{\sqrt{(Var(U_{k-1k}^2))}} \right).$$

Theorem 2.1. (Hewett and Spurrier [20]) *If $\rho_i = m_i/(\sum_{i=1}^k m_i) \to b_i$, $\tau_i = n_i/(\sum_{i=1}^k n_i) \to c_i$, and $m_i/(m_i + n_i) \to \lambda$, $i = 1, 2, \cdots, k$, as $m_1, \cdots, m_k$, $n_1, \cdots, n_k$ go to infinity, where b_i, c_i, λ are constant, then the joint limiting distribution of $(JT_1 - \mu_1)/\sigma_1$ and $(JT_2 - \mu_2)/\sigma_2$ is bivariate normal distribution with mean $(0,0)'$, unit variances, and covariance $\sqrt{\lambda}$.*

Hewett and Spurrier [20] proved that U^{st} follows a multivariate normal distribution as sample sizes go to infinity by using Theorem 6(II) from Lehmann [21]. Since there exist matrices A and B, such that $(JT_1, JT_2) = AU^{st} + B$, the limiting distribution of (JT_1, JT_2) is bivariate normal. Large sample two-stage designs may be obtained.

Remark 2.1. *In practice, it is common in two-stage designs to have balanced data across groups, that is,*

$$m_1 = m_2 = \cdots = m_k = m \quad \text{and} \quad n_1 = n_2 = \cdots = n_k = n.$$

In the balanced case

$$\frac{m_1}{m_1+n_1} = \cdots = \frac{m_k}{m_k+n_k} = \frac{m}{m+n}.$$

That is, in each stage, we have the same proportion of observations from each population. In this case, the conditions of Theorem 2.1 are met with $\lambda = \frac{m}{m+n}$, $b_i = c_i = \frac{1}{k}, i = 1, 2, \cdots, k$.

2.2 Convergence rate to the limiting null distribution

It is of interest to know how large the sample sizes must be in order for the limiting null distribution to be adequately used to approximate the finite sample distribution of the first and second stage JT test statistics. We investigated contour plots of the asymptotic and Monte Carlo simulated finite sample joint distributions of

the two-stage 3-sample JT test statistics for $\lambda = 1/4, 2/4, 3/4$ and various values of $N = m + n$. In the finite sample case, the contours were plotted based on a two-dimensional kernel density estimator, as implements by the R function *kde2d* [22]. Although the finite sample distribution does not depend on the underlying distribution assumed for each population, in our Monte Carlo simulation study based on 10,000 iterations, the responses in the 3 groups were assumed to be distributed as standard normal random variables. From the contour plots in the Figure 1, we see that the asymptotic approximation for the finite sample distribution appears adequate even for smaller N for the three values of λ considered.

An additional Monte Carlo simulation study was conducted to assess the asymptotic approximation to the joint distribution of $(JT_1 - \mu_1)/\sigma_1$ and $(JT_2 - \mu_2)/\sigma_2$. In the simulation experiment with $k = 3$, observations were generated from a standard normal distribution for $N = 4, 8, 12, 16, \cdots, 80$. Then JT_1 and JT_2 were calculated based on the first λN and all N observations, respectively, where $\lambda = 1/4, 2/4, 3/4$. This procedure was repeated 3,000 times for each N and λ combination. For each simulated sample, the Shapiro-Wilk test for multivariate normality [23] was performed. Since this test has been seen to perform well in finite sample and is known to be asymptotically exact, we may use the observed Type I error rate to gauge the appropriateness of using the asymptotic approximation. Table 1 shows the relative frequencies of rejecting the null hypothesis of bivariate normality at given nominal levels. As the sample size increases, the relative frequencies converge on the nominal levels as expected. As can be seen, the adequacy of the approximation is highly dependent on the value of λ.

3 Description of designs

In this section we describe the proposed designs, provide design parameters to be used in practice, and study the robustness of the procedures.

3.1 Design criteria

One may employ an upper boundary as considered by Hewett and Spurrier [20] to stop the trial early when substantial evidence in favor of H_a is observed in the first stage, but we consider only early stopping in the case of evidence in favor of H_0 as in [17], [18], and [24]. We consider equal sample sizes in the k groups at each stage for a maximum total sample size of $k(m + n) = kN$. Although unbalanced designs may also be considered, the

balance design is that which is most common in pre-clinical and clinical experiments.

Since the trial is stopped in the first stage if the test statistic JT_1 is less than or equal to r_1, the probability of early termination (PET) is defined as

$$\text{PET} = \sum_{i=0}^{r_1} P(JT_1 = i).$$

Corollary 3.1. *Using the previously presented Theorem 2.1, it follows that*

$$PET \to \Phi\Big(\frac{r_1 - \mu_1}{\sigma_1}\Big),$$

as $m \to \infty$, where $\Phi(.)$ is the cumulative distribution function of a standard normal random variable.

In the event that $JT_1 > r_1$ in the first stage, the trial continues to the second stage with additional subjects enrolled and observed for response. The probability of rejecting the null hypothesis with an experiment based on the design parameters $s = (m, N, r_1, r)$ is defined as

$$\tau_s(\tilde{\theta}) = P(JT_1 > r_1, JT_2 > r), \tag{3.1}$$

where $\tilde{\theta}$ represent the vector of location parameters. The above is equal to the Type I error (TIE) rate of the test under $\tilde{\theta} = \tilde{\theta_0}$, or the power when $\tilde{\theta} = \tilde{\theta_a}$, where $\tilde{\theta_0}$ and $\tilde{\theta_a}$ are the vectors of location parameters under the null and alternative, respectively. The expected sample size (ESS), which is a function of the PET, is given as

$$\text{ESS} = km + (1 - \text{PET})k(N - m).$$

For given first and second stage sample sizes, m and n, ESS is a decreasing function of PET.

There are many solutions of $s = (m, N, r_1, r)$ that satisfy Type I error (α) and power ($1 - \beta$) requirements, and the set of such designs are contained in $\Omega_{\alpha,\beta,\tilde{\theta_0},\tilde{\theta_a}} = \{s|\tau_s(\tilde{\theta_0}) \leq \alpha, \tau_s(\tilde{\theta_a}) \geq (1 - \beta)\}$. The minimax and optimal designs proposed by Simon [18] have both been widely used in two-stage one-arm clinical trials based on binary endpoints, and we may also consider use of the criteria in this context. The optimal design is defined as the design with the minimum ESS under the null, that is,

$$\big\{s|s, t \in \Omega_{\alpha,\beta,\tilde{\theta_0},\tilde{\theta_a}}; \text{ESS}_{s|H_0} = \min_t(\text{ESS}_{t|H_0})\big\},$$

where $\text{ESS}_{s|H_0}$ denotes the expected sample size for design s under the null hypothesis. The minimax design is that which has the smallest maximum sample size, and within this fixed sample size N, the minimum ESS under

the null. Letting N_s denote the maximum sample size for each arm associated with design s, and $\Omega_{\alpha,\beta,\tilde{\theta_0},\tilde{\theta_a}}(N) = \{s|s \in \Omega_{\alpha,\beta,\tilde{\theta_0},\tilde{\theta_a}}, N_s = N\}$, the design criteria may be written as

$$\left\{s|s, t \in \Omega_{\alpha,\beta,\tilde{\theta_0},\tilde{\theta_a}}; q \in \Omega_{\alpha,\beta,\tilde{\theta_0},\tilde{\theta_a}}(N_s); N_s = \min_t(N_t); \text{ESS}_{s|H_0} = \min_q(\text{ESS}_{q|H_0})\right\}.$$

The search algorithm for the optimal and minimax designs begins by evaluation of the set $\Omega_{\alpha,\beta,\tilde{\theta_0},\tilde{\theta_a}}(N)$ for increasing values of N. Reasonable starting values include $\max\{2, N_{os} - 5\}$, where N_{os} is the sample size per group for the one-stage design. For a given N, values of m from 1 to $N-1$ are considered. For fixed m and N, the range for the integer value r_1 (r) is $[0, \frac{k!}{(k-2)!2!}m^2]$ ($[0, \frac{k!}{(k-2)!2!}N^2]$). The search algorithm proceeds by evaluating all possible r values with this range. For a fixed value of r, Equation (3.1) is a non-increasing function of r_1, and therefore the maximum value of r_1, denoted r_1^{max}, can be identified such that the power constraints are satisfied by using the bisection search method. If there exists a set of parameters $(m, N, r_1^*, r) \in \Omega_{\alpha,\beta,\tilde{\theta_0},\tilde{\theta_a}}(N)$ for some r_1^*, then the design based on r_1^{max} is included among them, $r_1^* \leq r_1^{max}$, and the design (m, N, r_1^{max}, r) will have the smallest ESS under the null since it will have the largest PET. If the Type I error rate of (m, N, r_1^{max}, r) is larger than α, there exists no design satisfying the Type I error and power requirements for given m, N, and r. Increasing value of N are considered until the first occurrence of the event that $\Omega_{\alpha,\beta,\tilde{\theta_0},\tilde{\theta_a}}(N) \neq \emptyset$, that is, at least one valid design with adequate power properties is identified. For this set, the design with the smallest ESS under the null is the minimax design. To find the optimal design, values of N must be increased until it is clear that the design within $\Omega_{\alpha,\beta,\tilde{\theta_0},\tilde{\theta_a}}$ with the smallest ESS has been identified. Critical to the algorithm is the calculation of $\tau_s(\tilde{\theta_0})$ and $\tau_s(\tilde{\theta_a})$ for a given design s. Values of $\tau_s(\tilde{\theta_0})$ may be calculated exactly used a recurrence relationship, as in Wilding et al. [19], estimated using the asymptotic distribution provided in Section 2.1, or estimated via Monte Carlo exact methods. Due to complications similar as those in the one-stage design, estimates of $\tau_s(\tilde{\theta_a})$ are limited to Monte Carlo methods.

The minimax design may be more desirable when the difference in expected sample sizes is small, patient accrual is slow and there is a limited source of subjects. Although the null distribution depends only on the sample sizes, the non-null distribution of our test statistics depends on $F_i, i = 1, 2, \cdots, k$. A number of underlying distributions could be considered, but the normal distribution is a continuous probability distribution that often gives a good description of data which cluster around the mean. Therefore, the underlying distributions are assumed to be

that of normal populations.

Remark 3.1. *For the special case of $k = 2$, the two-stage JT designs proposed in this note are equivalent to the two-stage Mann-Whitney designs proposed in Wilding et al. [19].*

3.2 Design comparison

We investigated the proposed minimax and optimal designs under assumed alternatives for the case of $k = 3$ and $k = 4$. In the three sample scenarios the assumed alternatives take the form of different parametric contrasts for various shapes: linear $(\theta_1, \theta_2, \theta_3) = (0, 1, 2)$, convex $(\theta_1, \theta_2, \theta_3) = (0, 0, 3)$, concave $(\theta_1, \theta_2, \theta_3) = (0, 2, 2)$, and increasing $(\theta_1, \theta_2, \theta_3) = (0, 2, 3)$. In the case of $k = 4$, the following are considered: linear $(\theta_1, \theta_2, \theta_3, \theta_4) = (0, 0.5, 1, 1.5)$, convex $(\theta_1, \theta_2, \theta_3, \theta_4) = (0, 0, 0, 2)$, concave $(\theta_1, \theta_2, \theta_3, \theta_4) = (0, 2, 2, 2)$, and increasing $(\theta_1, \theta_2, \theta_3, \theta_4) = (0, 0, 1.5, 1.5)$. The units used for the contrasts are in standard deviations of the outcome of interest. For specified values of the Type I error rate and power, the minimax and optimal designs were determined using both Monte Carlo exact methods and the asymptotic approximation to the null distribution of (JT_1, JT_2). In the case of the non-null distribution, Monte Carlo methods were used in both cases. The Monte Carlo exact methods were based on 10,000 replicates. A Type I error rate of 5% is commonly used in practice and was used throughout our evaluation. The two-stage JT designs for power of 80% are provided in Tables 2 and 4, and for power of 90% see Tables 3 and 5. In addition to the minimax and optimal design parameters obtained by the Monte Carlo exact and asymptotic-based probability calculations, the tables contain the ESS of each design under the null, the PET, the actual Type I error, and the actual power of the test; These quantities were calculated via Monte Carlo exact methods. In addition, the classic one-stage design via Monte Carlo exact methods is also provided for the sake of comparison.

The savings in sample size of the two-stage designs are large in comparison to the single stage versions. For example, for the linear monotone ordering alternative of $\theta_1 = 0, \theta_2 = 1, \theta_3 = 2$, with power of 80% (Table 2), the sample size needed for the one-stage design is 5 per group, for a total of 15. However, the expected total sample size for the two-stage methods is only 8.6, resulting from enrolling 2 subjects at the first stage and another 3 subjects in the second stage if necessary. Under the alternative $\theta_1 = 0, \theta_2 = 0, \theta_3 = 0, \theta_4 = 2$, and power of 90% (Table 5), the total sample size needed for the one-stage design is 32, however the ESS for the two-stage methods

are all less than 20.

When comparing the two-stage designs obtained using Monte Carlo exact and asymptotic methodologies, the ESS difference can be substantial. When using the asymptotic approach to obtain designs for power of 80% (Table 4), the ESS is 12.2 subjects under the alternative $\theta_1 = 0, \theta_2 = 0, \theta_3 = 1.5, \theta_4 = 1.5$; the minimax design based on the Monte Carlo exact method reduces the ESS from 12.2 to 11.2 subjects. Note that we have considered large effect sizes in this section as is often seen in the planning of pre-clinical studies and early phase clinical trials. The same approach may be taken when smaller effect sizes are hypothesized, where difference between the exact and asymptotic-based probability calculations would be negligible.

3.3 Robustness

In the one-stage Mann-Whitney test, Hollander and Wolfe [25] showed that the variance of the Mann-Whitney statistic becomes smaller in the presence of ties. The test statistics JT_1 and JT_2 are linear combinations of Mann-Whitney statistics. Consequently, the variance of the JT test statistic is also affected by the presence of ties in the data, both within and between groups. Wilding et al. [19] showed that exact two-stage Mann-Whitney designs still provide excellent Type I error control and power properties in the presence of different percentages of ties in the data, that is, the procedure was shown to be robust. We investigated the influence of ties in the two-stage 3-sample JT designs by Monte Carlo simulations for Type I error rates of 5% and power of 80% for various alternatives. We rounded simulated observations generated from normal populations to the nearest multiple of 0.01, 0.1, 0.2, and 1 in order to produce data with a portion of the values being tied. When computing the Mann-Whitney test statistic components of the JT test statistics, in the event of ties when evaluating all possible pairs of observations between two samples, a value of 0.5 is assigned for the indicator function. For discussion of this strategy when dealing with rank-based procedures, see Lehmann [26]. The simulated Type I error rates and powers based on 20,000 iterations are given in Table 6. In the presence of tie, the two-stage JT designs based on Monte Carlo exact methods still provide excellent Type I error control and power properties. Similar results are obtained for $k = 4$ in Table 7.

Like other nonparametric approaches, the power of our proposed two-stage design depends on the distribution used and the specified alternative. Although the power of test is a function of the distribution form, the Type I

error rate of the test will be maintained regardless. We conducted a Monte Carlo simulation with 20,000 iterations based on uniform populations with the same mean and standard deviation as the normal distributions specified when obtaining the design parameters. The results showed that the proposed two-stage test still maintains excellent Type I error control and power properties. For example, for the set of parameters ($r_1 = 7, m = 2, r = 52, N = 5$) under the alternative $\theta_1 = 0, \theta_2 = 0, \theta_3 = 3$ (Table 3), the simulated Type I error rate and power are 0.0481 and 0.8437, respectively, with normal populations, and 0.0480 and 0.8281, respectively, with uniform populations.

4 An example

To illustrate the proposed two stage designs based on Monte Carlo exact methods, we revisit the example discussed earlier in the paper. As discussed in Dmitrienko et al. [3], we consider the design of a 4-arm trial of patients with hypercholesterolemia with the objective of assessment of the relationship between a new cholesterol-lowering drug and response. Three doses of the experimental drug are to be included in addition to a group receiving placebo. This clinical trial is to be based on a continuous primary endpoint, the reduction in low-density lipoprotein cholesterol after a 12-week treatment, and it is hypothesized that the underlying dose response relationship will be linear with $\theta_1 = 0\ mg/dL, \theta_2 = 7\ mg/dL, \theta_3 = 14\ mg/dL, \theta_4 = 21\ mg/dL$, where θ_1 is the placebo effect, and θ_2, θ_3 and θ_4 represent the drug effect in the low, medium, and high dose groups, respectively. The common standard deviation (SD) of LDL cholesterol change is expected to be between 20 mg/dL and 25 mg/dL. For a given 5% significance level and 90% power, the one- and two-stage designs were calculated based on the Monte Carlo exact methods proposed in this paper for a range of standard deviation values; see Table 8.

Focusing on the case of a standard deviation of 20, for example, the one stage design requires 64 subjects, while the ESS for the minimax and optimal designs under the null are 44.7 and 41.5, respectively. The minimax design is determined to have $m = 8$ and $n = 7$ per arm as the first and second stage sample sizes. If the first stage statistic $JT_1 \leq r_1 = 195$, the trial stops and fails to reject the null hypothesis, otherwise the trial continues to the second stage. After the trial is finished, the final conclusion is determined by the second stage statistics JT_2. If $JT_2 \leq r = 795$, we fail to reject H_0; otherwise we conclude that there exists a dose response relationship.

5 Discussion

In this article we presented two-stage k-sample designs based on the JT test statistic by adopting Simon's minimax and optimal design criteria. Although Hewett and Spurrier proposed a two-stage procedure and derived the asymptotic limiting distribution of the first and second stage test statistics, tests based on a Monte Carlo exact approach are more desirable for smaller sample sizes as used in early drug development trials. Our results indicate that two-stage designs can save sample size considerably compared to one-stage designs with the same Type I error rate and power constraints.

Two-stage designs are often preferred due to the ability to stop the procedure early in the absence of activity resulting in sample size savings. Although designs which stop early for efficacy may be considered using similar methods as we have used in this article, more often than not investigators choose not to take this approach to intervention evaluation due to the fact that the additional information obtained in the second stage may be used to achieve greater precision in the estimation of effects. We have also not considered designs with three or more stages [27] due to the administrative complexities they introduce into the process, as well as the lack of additional savings in total sample sizes as seen by others with similar designs in the binary endpoint context.

Although the designs have been shown to be robust in the scenarios considered in the note, one weakness of the proposed approach is the inability to directly accommodate the presence of ties in the data. Research into methods which more efficiently accommodate ties in the data are currently being undertaken. We have written an R package to calculate two-stage k-sample designs based on Monte Carlo exact methods for user defined parameters which will be available to readers upon request. Two-stage analogs of Cochran's Q test, Friedman's test, Page's test, and the Kruskal-Wallis test are also under investigation.

References

[1] Hewett JE, Spurrier JD. Some Two-Stage K-Sample Tests. *Journal of the American Statistical Association* 1979; **74**(366):398–404.

[2] Wilson JG. *Environment and Birth Defects.* Academic Press: New York, 1973.

[3] Dmitrienko A., Chuang-Stein C, D'Agostino R. *Pharmaceutical Statistics Using SAS: A Practical Guide (SAS Press)*. SAS Publishing, 2006.

[4] Jonckheere AR. A Distribution-Free k-Sample Test Against Ordered Alternatives. *Biometrika* 1954; **41**(1–2):133–145.

[5] Terpstra TJ. The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indigationes Mathematicae* 1952; **14**:327–333.

[6] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 1947; **18**:50–60.

[7] Tryon PV, Hettmansperger TP. A Class of Non-Parametric Tests for Homogeneity Against Ordered Alternatives. *The Annals of Statistics* 1973; **1**(6):1061–1070.

[8] Neuhäuser M, Liu PY, Hothorn LA. Nonparametric Tests for Trend: Jonckheere's Test, a Modification and a Maximum Test. *Biometrical Journal* 1998; **40**(8):899–909.

[9] McKean JW, Naranjo JD, Huitema BE. A robust method for the analysis of experiments with ordered treatment levels. *Psychological Reports* 2001; **89**(2):267–273.

[10] Terpstra J, Magel R. A new nonparametric test for the ordered alternative problem. *Journal of Nonparametric Statistics* 2003; **15**(3):289–301.

[11] Mossman D. Three-way ROCs. *Medical Decision Making* 1999; **19**(1):78–89.

[12] Nakas CT, Alonzo TA. ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. *Biometrics* 2007; **63**(2):603–609.

[13] Cuzick J. A Wilcoxon-type test for trend. *Statistics in Medicine* 1985; **4**(1):87–90.

[14] Le CT. A New Rank Test Against Ordered Alternatives in K-Sample Problems. *Biometrical Journal* 1988; **30**(1):87–92.

[15] Mahrer JM, Magel RC. A comparison of tests for the k-sample, non-decreasing alternative. *Statistics in Medicine* 1995; **14**(8):863–871.

[16] Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases* 1961; **13**(4):346–353.

[17] Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; **38**(1):143–151.

[18] Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; **10**(1):1–10.

[19] Wilding GE, Shan G, Hutson AD. Exact Two-Stage Designs for Phase II Clinical Trials with Rank-Based Endpoints. Technical Report, Department of Biostatistics, University at Buffalo, 2010.

[20] Hewett JE, Spurrier JD. Some two-stage k-sample tests. Technical Report, Department of statistics, university of Missouri-Columbia, 1978.

[21] Lehmann EL. Robust Estimation in Analysis of Variance. *The Annals of Mathematical Statistics* 1963; **34**(3):957–966.

[22] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2011.

[23] Villasenor-Alva JA, González-Estrada E. A Generalization of ShapiroWilk's Test for Multivariate Normality. *Communications in Statistics - Theory and Methods* 2009; **38**(11):1870–1883.

[24] Jung SH, Carey M, Kim KM. Graphical search for two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 2001; **22**(4):367–372.

[25] Hollander M, Wolfe DA. *Nonparametric Statistical Methods, 2nd Edition.* Wiley-Interscience, 1999.

[26] Lehmann EL. *Nonparametrics: Statistical Methods Based on Ranks.* Prentice Hall: New Jersey, 1998.

[27] Chen TT. Optimal three-stage designs for phase II cancer clinical trials. *Statistics in medicine* 1997; **16**(23):2701–2711.

Table 1: Relative frequency of rejecting the null of bivariate normality at nominal levels of 0.1, 0.05, and 0.01 for two-stage 3-sample JT test.

N	$\lambda = \frac{1}{4}$			$\lambda = \frac{2}{4}$			$\lambda = \frac{3}{4}$		
	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
4	0.073	0.037	0.004	0.099	0.046	0.006	0.108	0.042	0.006
8	0.096	0.044	0.005	0.102	0.046	0.007	0.103	0.048	0.006
12	0.100	0.047	0.008	0.103	0.050	0.009	0.102	0.046	0.008
16	0.105	0.053	0.010	0.101	0.050	0.008	0.100	0.049	0.009
20	0.103	0.051	0.009	0.100	0.049	0.008	0.102	0.051	0.010
24	0.101	0.048	0.009	0.101	0.053	0.011	0.101	0.049	0.009
28	0.106	0.052	0.009	0.101	0.050	0.011	0.107	0.052	0.009

Table 2: Minimax and optimal designs with $k = 3$ for 5% Type I error and 80% power.

θ_1	θ_2	θ_3	Design	r_1	m	r	N	TIE	Power	ESS	PET
0	1	2	exact one-stage			53	5	0.0479	0.8746	15	
			exact minimax	7	2	52	5	0.0481	0.8437	8.6	0.7096
			exact optimal	7	2	52	5	0.0481	0.8437	8.6	0.7096
			asy minimax	7	2	53	5	0.0415	0.8271	8.6	0.7047
			asy optimal	7	2	53	5	0.0415	0.8271	8.6	0.7047
0	0	3	exact one-stage			35	4	0.0448	0.8595	12	
			exact minimax	1	1	35	4	0.0414	0.8448	7.4	0.5095
			exact optimal	1	1	35	4	0.0414	0.8448	7.4	0.5095
			asy minimax	7	2	35	4	0.0422	0.8487	7.7	0.7150
			asy optimal	7	2	35	4	0.0422	0.8487	7.7	0.7150
0	2	2	exact one-stage			53	5	0.0429	0.8041	15	
			exact minimax	17	3	52	5	0.0493	0.8047	10.2	0.7970
			exact optimal	7	2	73	6	0.0462	0.8276	9.5	0.7144
			asy minimax	7	2	74	6	0.0394	0.8202	9.5	0.7117
			asy optimal	7	2	74	6	0.0394	0.8202	9.5	0.7117
0	2	3	exact one-stage			21	3	0.0384	0.8627	9	
			exact minimax	1	1	21	3	0.0378	0.8571	6	0.4957
			exact optimal	1	1	21	3	0.0378	0.8571	6	0.4957
			asy minimax	8	2	21	3	0.0351	0.8483	6.5	0.8355
			asy optimal	8	2	21	3	0.0351	0.8483	6.5	0.8355

Table 3: Minimax and optimal designs with $k = 3$ for 5% Type I error and 90% power.

θ_1	θ_2	θ_3	Design	r_1	m	r	N	TIE	Power	ESS	PET
0	1	2	exact one-stage			74	6	0.0489	0.9385	18	
			exact minimax	6	2	73	6	0.0492	0.9102	11.2	0.5706
			exact optimal	6	2	73	6	0.0492	0.9102	11.2	0.5706
			asy minimax	16	3	74	6	0.0410	0.9027	11.3	0.7396
			asy optimal	16	3	74	6	0.0410	0.9027	11.3	0.7396
0	0	3	exact one-stage			53	5	0.0447	0.9494	15	
			exact minimax	7	2	52	5	0.0454	0.9503	8.6	0.7128
			exact optimal	7	2	52	5	0.0454	0.9503	8.6	0.7128
			asy minimax	7	2	53	5	0.0357	0.9312	8.6	0.7055
			asy optimal	7	2	53	5	0.0357	0.9312	8.6	0.7055
0	2	2	exact one-stage			99	7	0.0461	0.9342	21	
			exact minimax	16	3	97	7	0.0475	0.9071	12.2	0.7460
			exact optimal	16	3	97	7	0.0475	0.9071	12.2	0.7460
			asy minimax	16	3	98	7	0.0414	0.8979	12.2	0.7361
			asy optimal	16	3	98	7	0.0414	0.8979	12.2	0.7361
0	2	3	exact one-stage			35	4	0.0470	0.9721	12	
			exact minimax	8	2	34	4	0.0431	0.9247	7.0	0.8337
			exact optimal	8	2	34	4	0.0431	0.9247	7.0	0.8337
			asy minimax	8	2	35	4	0.0347	0.9199	7.0	0.8266
			asy optimal	8	2	35	4	0.0347	0.9199	7.0	0.8266

Table 4: Minimax and optimal designs with $k = 4$ for 5% Type I error and 80% power.

θ_1	θ_2	θ_3	θ_4	Design	r_1	m	r	N	TIE	Power	ESS	PET
0	0.5	1	1.5	exact one-stage			188	7	0.0463	0.8702	28	
				exact minimax	28	3	140	6	0.0460	0.8008	17.0	0.5798
				exact optimal	13	2	184	7	0.0470	0.8102	15.1	0.6470
				asy minimax	28	3	140	6	0.0460	0.8008	17.0	0.5798
				asy optimal	13	2	185	7	0.0431	0.8045	15.1	0.6470
0	0	0	2	exact one-stage			140	6	0.0457	0.8206	24	
				exact minimax	31	3	138	6	0.0487	0.8144	15.2	0.7335
				exact optimal	13	2	183	7	0.0499	0.8347	15.1	0.6453
				asy minimax	31	3	139	6	0.0451	0.8023	15.2	0.7335
				asy optimal	13	2	185	7	0.0438	0.8215	15.1	0.6453
0	2	2	2	exact one-stage			140	6	0.0469	0.8206	24	
				exact minimax	12	2	139	6	0.0477	0.8085	15.4	0.5405
				exact optimal	13	2	184	7	0.0477	0.8343	15.1	0.6456
				asy minimax	30	3	140	6	0.0440	0.8004	15.8	0.6794
				asy optimal	13	2	185	7	0.0444	0.8272	15.1	0.6456
0	0	1.5	1.5	exact one-stage			99	5	0.0480	0.8496	20	
				exact minimax	14	2	97	5	0.0486	0.8041	11.2	0.7337
				exact optimal	14	2	97	5	0.0486	0.8041	11.2	0.7337
				asy minimax	13	2	99	5	0.0448	0.8137	12.2	0.6500
				asy optimal	13	2	99	5	0.0448	0.8137	12.2	0.6500

Table 5: Minimax and optimal designs with $k = 4$ for 5% Type I error and 90% power.

θ_1	θ_2	θ_3	θ_4	Design	r_1	m	r	N	TIE	Power	ESS	PET
0	0.5	1	1.5	exact one-stage			241	8	0.0486	0.9072	32	
				exact minimax	50	4	240	8	0.0488	0.9023	22.5	0.5929
				exact optimal	56	4	360	10	0.0488	0.9000	21.2	0.7839
				asy minimax	49	4	241	8	0.0461	0.9002	23.1	0.5572
				asy optimal	54	4	299	9	0.0472	0.9038	21.4	0.7283
0	0	0	2	exact one-stage			241	8	0.0476	0.9204	32	
				exact minimax	29	3	239	8	0.0476	0.9053	19.3	0.6352
				exact optimal	29	3	239	8	0.0476	0.9053	19.3	0.6352
				asy minimax	29	3	239	8	0.0476	0.9053	19.3	0.6352
				asy optimal	29	3	239	8	0.0476	0.9053	19.3	0.6352
0	2	2	2	exact one-stage			187	7	0.0479	0.9428	28	
				exact minimax	12	2	237	8	0.0496	0.9005	18.6	0.5579
				exact optimal	12	2	237	8	0.0496	0.9005	18.6	0.5579
				asy minimax	29	3	239	8	0.0452	0.9101	19.1	0.6468
				asy optimal	29	3	239	8	0.0452	0.9101	19.1	0.6468
0	0	1.5	1.5	exact one-stage			242	8	0.0461	0.9161	32	
				exact minimax	31	3	184	7	0.0494	0.9099	16.3	0.7322
				exact optimal	31	3	184	7	0.0494	0.9099	16.3	0.7322
				asy minimax	31	3	185	7	0.0464	0.9074	16.3	0.7322
				asy optimal	31	3	185	7	0.0464	0.9074	16.3	0.7322

Table 6: Type I error and power of designs with $k = 3$ in the presence of tie; 5% Type I error and 80% power.

	No ties	round to			
The alternative $(\theta_1, \theta_2, \theta_3)$		0.01	0.1	0.2	1
$(0, 1, 2)$					
Proportion of tie under the null	0.00	0.04	0.32	0.53	0.93
Proportion of tie under the alternative	0.00	0.03	0.25	0.44	0.90
Actual TIE	0.05	0.05	0.05	0.05	0.04
Power	0.84	0.84	0.85	0.86	0.84
$(0, 0, 3)$					
Proportion of tie under the null	0.00	0.03	0.27	0.45	0.90
Proportion of tie under the alternative	0.00	0.02	0.16	0.30	0.81
Actual TIE	0.04	0.04	0.04	0.04	0.03
Power	0.85	0.85	0.86	0.87	0.87
$(0, 2, 2)$					
Proportion of tie under the null	0.00	0.05	0.37	0.60	0.94
Proportion of tie under the alternative	0.00	0.03	0.28	0.48	0.92
Actual TIE	0.05	0.04	0.05	0.05	0.04
Power	0.83	0.83	0.84	0.85	0.83
$(0, 2, 3)$					
Proportion of tie under the null	0.00	0.02	0.20	0.36	0.85
Proportion of tie under the alternative	0.00	0.01	0.12	0.23	0.72
Actual TIE	0.04	0.04	0.04	0.04	0.03
Power	0.85	0.86	0.87	0.87	0.86

Table 7: Type I error and power of designs with $k = 4$ in the presence of tie; 5% Type I error and 80% power.

	No ties	round to			
The alternative $(\theta_1, \theta_2, \theta_3, \theta_4)$		0.01	0.1	0.2	1
$(0, 0.5, 1, 1.5)$					
Proportion of tie under the null	0.00	0.07	0.52	0.74	0.97
Proportion of tie under the alternative	0.00	0.06	0.47	0.70	0.96
Actual TIE	0.05	0.05	0.05	0.05	0.04
Power	0.80	0.81	0.82	0.82	0.79
$(0, 0, 0, 2)$					
Proportion of tie under the null	0.00	0.07	0.52	0.74	0.97
Proportion of tie under the alternative	0.00	0.06	0.42	0.65	0.96
Actual TIE	0.05	0.05	0.05	0.05	0.05
Power	0.84	0.85	0.85	0.85	0.84
$(0, 2, 2, 2)$					
Proportion of tie under the null	0.00	0.07	0.51	0.74	0.97
Proportion of tie under the alternative	0.00	0.06	0.43	0.65	0.96
Actual TIE	0.05	0.05	0.05	0.05	0.04
Power	0.84	0.84	0.85	0.85	0.84
$(0, 0, 1.5, 1.5)$					
Proportion of tie under the null	0.00	0.05	0.40	0.63	0.95
Proportion of tie under the alternative	0.00	0.04	0.33	0.55	0.94
Actual TIE	0.05	0.05	0.05	0.05	0.04
Power	0.80	0.80	0.81	0.81	0.78

Table 8: Minimax and optimal designs with $k = 4$ for 5% Type I error and 90% power.

SD	Design	r_1	m	r	N	TIE	Power	ESS	PET
20	exact one-stage			901	16	0.0496	0.9193	64	
	exact minimax	195	8	795	15	0.0498	0.9003	44.7	0.5470
	exact optimal	161	7	1245	19	0.0499	0.9005	41.5	0.7192
21	exact one-stage			1017	17	0.0490	0.9064	68	
	exact minimax	198	8	1012	17	0.0495	0.9000	46.7	0.5917
	exact optimal	158	7	1247	19	0.0500	0.9012	43.4	0.6786
22	exact one-stage			1261	19	0.0497	0.9124	76	
	exact minimax	378	11	1134	18	0.0500	0.9001	54.7	0.6196
	exact optimal	207	8	1802	23	0.0496	0.9016	49.4	0.7092
23	exact one-stage			1529	21	0.0498	0.9129	84	
	exact minimax	599	14	1392	20	0.0492	0.9000	66.4	0.5683
	exact optimal	324	10	1802	23	0.0499	0.9007	53.8	0.7342
24	exact one-stage			1672	22	0.0491	0.9078	88	
	exact minimax	536	13	1664	22	0.0496	0.9004	63.4	0.6830
	exact optimal	255	9	1963	24	0.0499	0.9027	57.9	0.6343
25	exact one-stage			1976	24	0.0492	0.9121	96	
	exact minimax	440	12	1815	23	0.0499	0.9000	67.4	0.5589
	exact optimal	312	10	2133	25	0.0495	0.9002	62.8	0.6199

Figure 1: Contour plots for the bivariate normal distribution (solid black line) and finite sample (dashed blue line) joint distribution of two-stage 3-sample JT test statistics . The first, second and third columns are for contour plots with different maximum sample sizes for $\lambda = 1/4, 2/4, 3/4$, respectively.