Exact Two-Stage Designs for Phase II Clinical Trials with Rank-Based Endpoints

Gregory E. Wilding[1*], Guogen Shan[1] and Alan D. Hutson[1]

[1] *Department of Biostatistics,University at Buffalo, 3435 Main Street, Buffalo, NY, 14214, USA*

August 6, 2010

Abstract

Features common to Phase II clinical trials include limited knowledge of the experimental treatment being evaluated, design components reflecting ethical considerations, and small to moderate sample sizes as a result of resource constraints. It is for these reasons that there exist many two-stage designs proposed in the literature for use in this context. The majority of these designs are for binary endpoints and based on exact probability calculations, or are for continuous endpoints and rooted in asymptotic approximations to the null distribution. We present exact two-stage Mann-Whitney designs in the context of Phase II clinical trials. In addition to describing the designs, we present tables of decision rules under a variety of assumed realities for use in trial planning.

Keywords: Clinical Trials, Exact Test, Mann-Whitney Test, Minimax Design, Nonparametric, Optimal Design, Phase II Design.

*Corresponding author. Department of Biostatistics,University at Buffalo, 3435 Main Street, Buffalo, NY, 14214, USA.

E-mail address:gwilding@buffalo.edu

1 Introduction

One important goal of the Phase II clinical trial is to determine whether a new treatment has sufficient activity for further investigation. If activity is identified in Phase II, the intervention typically proceeds to Phase III for further evaluation. The primary endpoint for Phase II evaluation is disease- and treatment- type specific and includes categorical and numeric outcomes. Design constraints and complications encountered in this context include ethical concerns, resource limitations, and the use of surrogate endpoints.

Currently accepted Phase II trial designs in the area such as oncology include a wide range of single arm designs with a focus on binary endpoints. Due to ethical concerns and resource management, common designs used in Phase II trials include those which allow for early stoppage in the presence of excessive toxicity or lack of efficacy. These designs often utilize historical data so to gage the efficacy of the experimental intervention. A single stage design incorporating sequential early stopping rules for adverse events was proposed by Fleming [4]. Simon's optimal and minimax designs [20] have been widely applied in Phase II clinical trials and allow for early stoppage for futility. Other two-stage designs which incorporated early stoppage for futility include those due to Gehan [5], Green and Dahlberg [7], Chen and Ng [3], and Storer [24]. Designs which allow for stopping early for either futility and/or efficacy are discussed by Kepner and Chang [12]. While frequentist design have predominated Phase II clinical trials, Bayesian Phase II trials have become more visible after extensive development by Thall, Simon and Estey [27], and Mayo and Gajewski [17]. A review of Phase II designs may be found in Mariani and Marubini [16], Thall [26], and in the specific setting of clinical oncology in Green, Benedetti and Crowley [6].

When single arm Phase II studies are not appropriate, randomized designs may be considered; see Simon, Wittes and Ellenberg [21] or Lee and Feng [13] for discussion on the merits of such an approach to intervention evaluation. Taylor, Braun and Li [25] pointed out that a two-stage two-arm design for binary endpoints may be preferable under the condition of uncertainty of the historical response rates. Due to the type of endpoint and the sample size limitations experienced in the Phase II setting, many of the above mentioned designs used in conjunction with binary endpoints are exact. Solutions for non-binary endpoints are currently restricted in the literature to those which are mainly reliant on asymptotic approximations to the null distribution of the test statistic.

Mann and Whitney [15] presented the two-sample one-stage rank-sum test which is widely used in many areas of application, including randomized Phase II trials, and its statistical properties are well studied. The usual normal approximation may be used for larger sample sizes in that calculation of the exact distribution may be computational difficult for even moderate sample size. Mann and Whitney computed the null distribution of the test statistics based on sample sizes m and n, denoted $U_{m,n}$, for sample sizes up to $m = n = 8$ by using the recurrence relation tables. They pointed out that the distribution of $U_{m,n}$ is almost normal distributed for large values of m and n. Kurt Hornik [9] recently developed a function for computing the distribution of the Mann-Whitney statistic for given sample sizes for the statistical software R based on the algorithm presented by Mann and Whitney. A faster algorithm for computation of the exact distribution of the Mann-Whitney statistic $U_{m,n}$ under the null hypothesis was given by Castagliola [1]. He suggested use of the normal approximation when $m + n > 20$. We note that for studies planned to detect larger effects, sample sizes would be small and therefore the asymptotic approximation might not be adequately applied. Recently, Shuster et al. [19] proposed a Mann-Whitney test for group sequential clinical trials with ordinal data. The test is based on use of asymptotic results thereby making the design adequate in the Phase III context. A two-stage test based on sample sizes m_1, n_1 in the first stage and another m_2, n_2 added at the second stage, has been proposed by Spurrier and Hewett [23]. The asymptotic null distribution of first and second stage test statistics, denoted by U_{m_1,n_1} and U_{m_1,n_1,m_2,n_2}, respectively, are derived and recommended for use in practice when the sample sizes are as small as $m_1 = n_1 = m_2 = n_2 = 9$. Although an exact recurrence relation is discussed, the authors had not been able to find an efficient method for computing the joint probability distribution of U_{m_1,n_1} and U_{m_1,n_1,m_2,n_2} under null for small sample sizes at the time of publication. There is no published exact two-stage Mann-Whitney design for Phase II clinical trials at this time.

In this paper, we consider parallel-arm two-stage exact designs for continuous endpoints [11]. We present methodology for computation of the distribution of the Mann-Whitney statistics in section 2 and a probability recurrence relation to get the exact two-stage Mann-Whitney probability distribution is reviewed. Section 3 is given to the description of the proposed exact two-stage designs for different shift values of the two treatments populations. Section 4 is given to a discussion.

2 Two-stage procedure

The Mann-Whitney test provides a simple and effective method for comparing a new treatment with a standard-of-care in regards to some continuous endpoint. Let the responses of subjects corresponding to the new intervention be denoted by $Y_1, Y_2, \cdots, Y_n$ and let $X_1, X_2, \cdots, X_m$ denote the responses for the control group. Furthermore, let higher values of the response be an indication of greater health in the subject. In the randomization model [14], the $m+n$ subjects are assigned to treatment at random. The observations are assumed to be deterministic under the null, and a probabilistic basis is created by assigning subjects at random. The inference can not be extended beyond the subjects at hand without further assumptions. Under the invoked population model, the $m+n$ subjects are assumed to be randomly drawn in some specified manner from populations of users of the two treatments, n from the treatment and m from the control. The subjects are compared with the end goal of making inferences regarding the populations. Either of the discussed models may be considered, but for trial planning purposed the population model must be invoked. The null hypothesis of no treatment effect under the population model can be stated as

$$H_0 : F = G,$$

where F and G represent the cumulative distribution functions corresponding to the random variables X and Y, respectively. Generally, one purpose of a phase II trial is to determine whether or not the new treatment is more efficiency. Therefore the alternative of interest is the case where the random variable Y is stochastically larger than X, implying $G(u) \leq F(u)$ for all u with strict inequality for some point u.

2.1 One-Stage Mann-Whitney test

The test statistic of the Mann-Whitney test is formulated by considering the $m \times n$ pairs $(X_i, Y_j), i = 1, 2, \cdots, m,\ j = 1, 2, \cdots, n$. Assuming no ties in the data, if the distribution from which the observations come are equal, we would expect the number of pairs such that $X_i > Y_j$ to be approximately equal to the number of pairs where $X_i < Y_j$. Therefore, a measure of distributional differences and our test statistic is,

$$U_{m,n} = \sum_{j=1}^{n} \sum_{i=1}^{m} I(X_i < Y_j),$$

where

$$I(X_i < Y_j) = \begin{cases} 1, & X_i < Y_j, \\ 0, & X_i \geq Y_j. \end{cases}$$

Larger values of the above defined test statistic would signify evidence in favor of the alternative. Under null hypothesis, Mann and Whitney [15] showed the mean and variance of $U_{m,n}$ to be,

$$E(U_{m,n}) = \frac{mn}{2}$$

and

$$Var(U_{m,n}) = \frac{mn(m+n+1)}{12},$$

respectively. The Mann-Whitney test is well known to be equivalent to the Wilcoxon rank sum test [28] and it may be seen that $U_{m,n} = R_{m,n} - n(n+1)/2$, where $R_{m,n}$ is the Wilcoxon rank sum test statistic. Due to the rank based nature of the test, the null distribution only depends on the sample sizes making it convenient for statistical practice. See Figure 1 for the probability distribution for $U_{2,2}$ under the null hypothesis. We also note a finite number of type I error rates is available when planning the trial due to the discreteness of the test statistic.

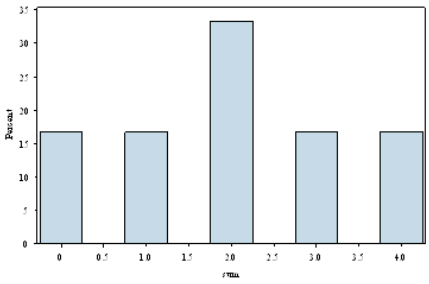

Figure 1: Exact distribution of $U_{2,2}$

2.2 Two-Stage Mann-Whitney test

In the two-stage Mann-Whitney procedure, we first obtain the samples $X_1, X_2, \cdots, X_{m_1}$ from F and $Y_1, Y_2, \cdots, Y_{n_1}$ from G, used in the calculation of the first stage test statistic U_{m_1,n_1}. In the second stage, m_2 observations

$X_{m_1+1}, X_{m_1+2}, \cdots, X_{m_1+m_2}$ from F and n_2 observations $Y_{n_1+1}, Y_{n_1+2}, \cdots, Y_{n_1+n_2}$ from G are collected. The test statistics after the second stage,

$$U_{m_1,n_1,m_2,n_2} = \sum_{j=1}^{n_1+n_2} \sum_{i=1}^{m_1+m_2} I(X_i < Y_j),$$

which includes the subjects from both stages, is then used to make a final decision in regards to the hypotheses. The statistical dependence of U_{m_1,n_1} and U_{m_1,n_1,m_2,n_2} is obvious and is illustrated through examination of the distribution of $U_{2,2,1,1}$ under null, see Figure 2. The distribution of U_{m_1,n_1} is that of the one-stage Mann-Whitney test statistic $U_{2,2}$, which has five possible outcomes with probability greater than zero. The conditional distribution of U_{m_1,n_1,m_2,n_2} is reliant on the information from stage 1. Note the conditional distribution of U_{m_1,n_1,m_2,n_2} shifts from left to right with the increase value of U_{m_1,n_1}. See Figure 3 for the unconditional distribution of $U_{2,2,1,1}$.

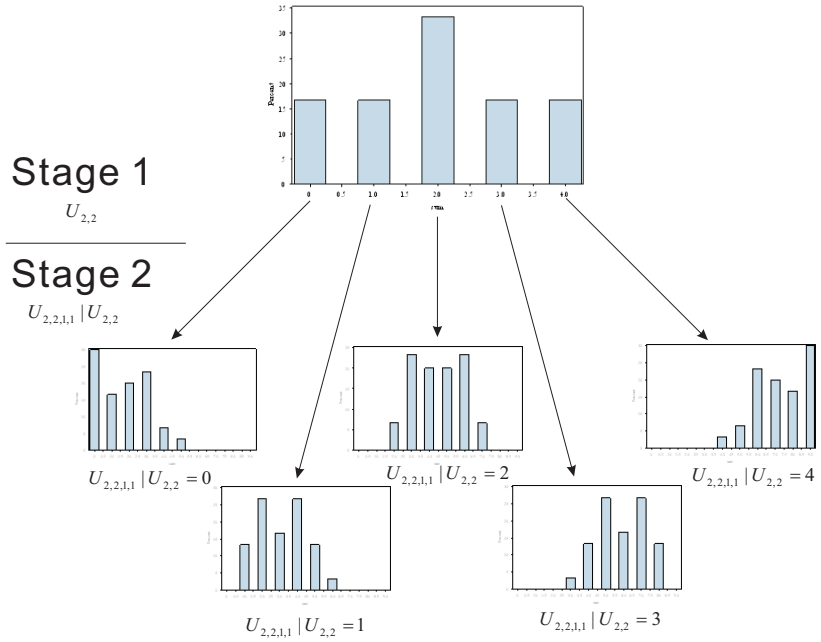

Figure 2: Exact conditional distribution of $U_{2,2,1,1}$

It has been shown in Spurrier and Hewett[23] that, under null hypothesis,

$$\mu_1 = E(U_{m_1,n_1}) = \frac{m_1 n_1}{2},$$

$$\sigma_1^2 = Var(U_{m_1,n_1}) = \frac{m_1 n_1 (m_1 + n_1 + 1)}{12},$$

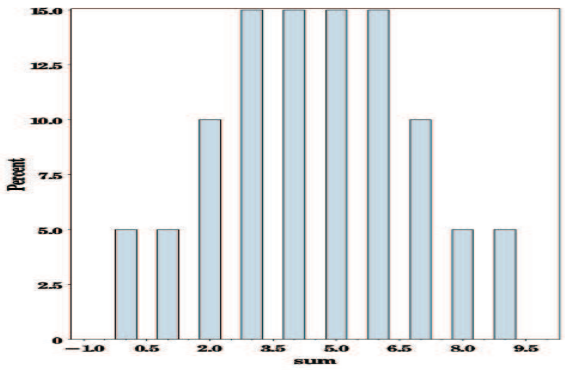

Figure 3: unconditional distribution of $U_{2,2,1,1}$

$$\mu_2 = E(U_{m_1,n_1,m_2,n_2}) = \frac{MN}{2},$$

and

$$\sigma_2^2 = Var(U_{m_1,n_1,m_2,n_2}) = \frac{MN(M+N+1)}{12},$$

where $N = n_1 + n_2$ and $M = m_1 + m_2$. The bivariate relationship between U_{m_1,n_1} and U_{m_1,n_1,m_2,n_2} may be described by using the result,

$$\rho_{U_{m_1,n_1},U_{m_1,n_1,m_2,n_2}} = \big[\frac{m_1 n_1 (N+1)}{MN(m_1+n_1+1)}\big]^{\frac{1}{2}},$$

which demonstrates that the Pearson correlation decrease as the ratios m_2/m_1 or n_2/n_1 increase. The following theorem is the basis for asymptotic approximation of the joint distribution of U_{m_1,n_1} and U_{m_1,n_1,m_2,n_2}.

Theorem 2.1. (Spurrier and Hewett [23]) *If $m_1, n_1, m_2, n_2 \to \infty$ such that $\frac{m_1}{M+N} \to b_1$ and $\frac{n_1}{M+N} \to b_2$, where $b_1, b_2 > 0$ with $b_1 + b_2 < 1$, and $\frac{m_1}{m_2} = \frac{n_1}{n_2} = a$, for some $a > 0$, then the joint limiting distribution of V is bivariate normal, where $V = (V_1, V_2)'$,*

$$V_1 = \frac{U_{m_1,n_1} - \mu_1}{\sigma_1},$$

and

$$V_2 = \frac{U_{m_1,n_1,m_2,n_2} - \mu_2}{\sigma_2},$$

with mean vector $\mu = (0,0)'$ and covariance matrix

$$\sum = \begin{pmatrix} 1 & (a+1)^{-1/2} \\ (a+1)^{-1/2} & 1. \end{pmatrix}.$$

The above results was used by Spurrier and Hewett [23] to develop a two-stage procedure which allow early acceptance of the new treatment under the alternative $G(u) \leq F(u)$ for all u with strict inequality for some point u. They employ an upper boundary to stop the trial early when a significantly high efficacy is observed from the first stage. Monte Carlo simulation demonstrated their two-stage Mann-Whitney test is reasonable when the sample sizes as low as $m_1 = n_1 = m_2 = n_2 = 9$. Tables of critical values for in practice were provided in Spurrier and Hewett [22]. They compared the two-stage and one-stage tests under a range of parametric families and found that both maintain equivalent type I and II errors. In addition they demonstrated the two-stage tests can save $10 - 30\%$ in expected sample sizes as compared to the one-stage tests for given type I and II error rates.

With small sample sizes, the asymptotic approximation is questionable and a reasonably efficient method to calculate $P_{m_1,n_1,m_2,n_2}(U_{m_1,n_1}, U_{m_1,n_1,m_2,n_2})$, the joint probability under null hypothesis, is needed. The largest value of the pooled data,

$$Z = \max(X_1, X_2, \cdots, X_{m_1+m_2}, Y_1, Y_2, \cdots, Y_{n_1+n_2}),$$

belongs to one of the four samples of the two-stage design. If Z belongs to sample X, it does not contribute to the observed values of U_{m_1,n_1} or U_{m_1,n_1,m_2,n_2}. If Z belongs to sample Y and stage 1, it contributes m_1 and $m_1 + m_2$ to the observed values of U_{m_1,n_1} and U_{m_1,n_1,m_2,n_2}, respectively. If Z belongs to sample Y and stage 2, it does not contribute to the observed values of U_{m_1,n_1}, but contributes $m_1 + m_2$ to U_{m_1,n_1,m_2,n_2}.

The probability under the null that Z comes from stage i of sample X(Y) is $m_i/(M+N)(n_i/(M+N))$. By conditioning on the observation with the largest value in either sample and on whether this observation is in the first or the second sample, we get the following recurrence relation:

$$\begin{aligned} P_{m_1,n_1,m_2,n_2}(U_{m_1,n_1}, U_{m_1,n_1,m_2,n_2}) &= \tfrac{m_1}{M+N} P_{m_1-1,n_1,m_2,n_2}(U_{m_1,n_1}, U_{m_1,n_1,m_2,n_2}) \\ &+ \tfrac{n_1}{M+N} P_{m_1,n_1-1,m_2,n_2}(U_{m_1,n_1} - m_1, U_{m_1,n_1,m_2,n_2} - M) \\ &+ \tfrac{m_2}{M+N} P_{m_1,n_1,m_2-1,n_2}(U_{m_1,n_1}, U_{m_1,n_1,m_2,n_2}) \\ &+ \tfrac{n_2}{M+N} P_{m_1,n_1,m_2,n_2-1}(U_{m_1,n_1}, U_{m_1,n_1,m_2,n_2} - M), \end{aligned}$$

where $P_{0,0,0,0}(0,0) = 1$, and $P_{m_1,n_1,m_2,n_2}(U_{m_1,n_1}, U_{m_1,n_1,m_2,n_2}) = 0$ if any of the following occur:

1. $m_1 < 0$ or $n_1 < 0$ or $m_2 < 0$ or $n_2 < 0$,

2. $U_{m_1,n_1} < 0$ or $U_{m_1,n_1} > m_1 n_1$ or $U_{m_1,n_1,m_2,n_2} < 0$ or $U_{m_1,n_1,m_2,n_2} > MN$ or $U_{m_1,n_1} > U_{m_1,n_1,m_2,n_2}$,

3. $m_1 = 0$ or $n_1 = 0$, but $U_{m_1,n_1} \neq 0$,

4. $m_2 = 0$ or $n_2 = 0$, but $U_{m_1,n_1,m_2,n_2} \neq U_{m_1,n_1}$.

Although an algorithm had been proposed to obtain joint probabilities, Spurrier and Hewett [23] stated in their paper, "The authors have been unable to find an efficient method for computing the joint probability distribution of U_{m_1,n_1} and U_p under H for small sample sizes." With the improvement of computer technology it is now possible to calculate the joint probability distribution of U_{m_1,n_1} and U_{m_1,n_1,m_2,n_2} for small to median sample sizes so to implement exact two-stage designs.

3 Minimax and optimal two-stage Mann-Whitney exact designs

The proposed design is as follows. For simplicity, we assume that the sample sizes of the two groups are equal at each stage, that is, $m_1 = n_1$ and $m_2 = n_2$, for a total sample size of $S = 2n_1 + 2n_2$. Although unbalance design may also be considered, the balance design is that which is most common in practice. Corresponding critical values at each stage are denoted as r_1 and r. A total of n_1 subjects are assigned to each treatment at stage 1; if the statistic U_{n_1,n_1} less than or equal to r_1, we terminate the trial and reject the new treatment. Otherwise, we enter n_2 additional patients to each treatment at the second stage; if the test statistic U_{n_1,n_1,n_2,n_2} is less than or equal to r, the trial concludes that the new treatment lacks the efficacy to proceed to Phase III. Alternatively, if $U_{n_1,n_1,n_2,n_2} > r$, we conclude that the treatment is promising and merits further testing. One may also employ an upper boundary to stop the trial early when a significantly high efficacy is observed in stage 1 as considered by Spurrier and Hewett [23], but we consider only early stopping in case of lack of efficacy as in Simon [20], Fleming [4] and Jung et al. [10].

The probability of early termination (PET) is defined as the observed value of U_{n_1,n_1} falling in the stopping region at stage 1, that is,

$$PET = \sum_{i=0}^{r_1} P_{n_1,n_1}(U_{n_1,n_1} = i).$$

Using the previous presented theorem, the asymptotic PET can be written as

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

where Φ is the cumulative distribution function of standard normal. The expected sample size (ESS), which is a function of the PET, is given as

$$ESS = 2n_1 + (1 - PET) * 2n_2.$$

There are many solutions of n_1, n_2, r_1 and r that satisfy α and power requirements. The minimax and optimal designs by Simon [20] have both been widely used in the two-stage one-arm clinical trials based on binary endpoints. We may also consider the optimal design in this context which is defined as the design with the minimum ESS under the null. In addition, we also consider the minimax design which is the design which has the minimum total sample size S under the null, and within this fixed sample size S, the minimum ESS. 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 patients. Although the null distribution depends only on the sample sizes, the non-null distribution of our test statistics depends on F and G. 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 normal with common variance, and the distributions under the alternative are assumed to be differ only by a shift value δ. We considered only median to larger shift values as is generally seen in phase II trials.

To investigate the performance of minimax and optimal exact two-stage design, Monte Carlo studies were performed. For specified values of α, power and δ, we can determine the minimax and optimal designs by using the exact or asymptotic distributions of V. For example, Figure 4 shows the plot of ESS against the maximum sample size S for our proposed exact designs under $\alpha = 0.05$, power=0.85 and alternative $\delta = 1.5$. Table 1 provides the minimax and optimal designs by the exact and asymptotic methodologies under the alternative that $\delta = 2$, the ESS under the null, the PET for the stage 1, actual type I error, and actual power of the test. Table 2 and Table 3 provide the minimax and optimal designs for $\delta = 1.5$ and $\delta = 1$, respectively. Small values of δ would correspond to relative greater sample sizes where the asymptotic approximation can be applied with adequately.

When comparing the minimax designs obtained using exact and asymptotic methodologies, the ESS difference

Table 1: Minimax and Optimal Designs for $\delta = 2$.

				Reject Drug if U					
α	power	Type	Design	$\leq r_1/n_1$	$\leq r/n$	ESS	PET	TIE	POWER
0.05	0.8	Asy	Minimax	6 / 3	28 / 6	7.5	0.74	0.046	0.87
			Optimal	6 / 3	28 / 6	7.5	0.74	0.046	0.87
		Exact	Minimax	0 / 1	20 / 5	6.0	0.50	0.042	0.82
			Optimal	0 / 1	20 / 5	6.0	0.50	0.042	0.82
0.05	0.85	Asy	Minimax	6 / 3	28 / 6	7.5	0.74	0.046	0.87
			Optimal	6 / 3	28 / 6	7.5	0.74	0.046	0.87
		Exact	Minimax	5 / 3	20 / 5	7.4	0.65	0.047	0.87
			Optimal	2 / 2	28 / 6	6.7	0.67	0.039	0.87
0.05	0.9	Asy	Minimax	11 / 4	28 / 6	8.8	0.81	0.049	0.90
			Optimal	6 / 3	47 / 8	8.6	0.74	0.044	0.90
		Exact	Minimax	5 / 3	28 / 6	8.1	0.65	0.044	0.91
			Optimal	2 / 2	36 / 7	7.3	0.67	0.049	0.91
0.1	0.8	Asy	Minimax	2 / 2	19 / 5	7.0	0.50	0.080	0.87
			Optimal	7 / 3	23 / 6	6.8	0.86	0.091	0.80
		Exact	Minimax	0 / 1	12 / 4	5.0	0.50	0.088	0.85
			Optimal	0 / 1	12 / 4	5.0	0.50	0.088	0.85
0.1	0.85	Asy	Minimax	2 / 2	19 / 5	7.0	0.50	0.080	0.87
			Optimal	2 / 2	19 / 5	7.0	0.50	0.080	0.87
		Exact	Minimax	2 / 2	12 / 4	5.3	0.67	0.088	0.86
			Optimal	2 / 2	12 / 4	5.3	0.67	0.088	0.86
0.1	0.9	Asy	Minimax	5 / 3	19 / 5	7.7	0.59	0.084	0.91
			Optimal	6 / 3	25 / 6	7.5	0.74	0.097	0.91
		Exact	Minimax	5 / 3	19 / 5	7.4	0.65	0.073	0.91
			Optimal	2 / 2	33 / 7	7.3	0.67	0.089	0.92

For each design, ESS and PET denote the expected sample size and the probability of early termination under null hypotheses. n_1,n are the sample sizes per treatment for first stage and overall, respectively.

Table 2: Minimax and Optimal Designs for $\delta = 1.5$.

				Reject Drug if U					
α	power	Type	Design	$\leq r_1/n_1$	$\leq r/n$	ESS	PET	TIE	POWER
0.05	0.8	Asy	Minimax	5 / 3	48 / 8	10.1	0.59	0.042	0.81
			Optimal	11/ 4	58 / 9	9.9	0.81	0.044	0.80
		Exact	Minimax	4 / 3	37 / 7	10.0	0.50	0.047	0.81
			Optimal	2/ 2	69 / 10	9.3	0.66	0.048	0.81
0.05	0.85	Asy	Minimax	35 / 7	47 / 8	14.2	0.91	0.050	0.85
			Optimal	10 / 4	71 / 10	11.4	0.72	0.045	0.86
		Exact	Minimax	9 / 4	47 / 8	10.7	0.66	0.049	0.86
			Optimal	5 / 3	58 / 9	10.2	0.65	0.047	0.86
0.05	0.9	Asy	Minimax	9 / 4	71 / 10	12.6	0.61	0.050	0.90
			Optimal	9 / 4	71 / 10	12.6	0.61	0.050	0.90
		Exact	Minimax	15 / 5	71 / 10	12.7	0.73	0.046	0.90
			Optimal	16 / 5	83 / 11	12.5	0.79	0.049	0.91
0.1	0.8	Asy	Minimax	5 / 3	26 / 6	8.5	0.59	0.091	0.81
			Optimal	5 / 3	26 / 6	8.5	0.59	0.091	0.81
		Exact	Minimax	5 / 3	26 / 6	8.1	0.65	0.080	0.82
			Optimal	0 / 1	33 / 7	8.0	0.50	0.098	0.81
0.1	0.85	Asy	Minimax	8 / 4	26 / 6	10.0	0.50	0.099	0.85
			Optimal	5 / 3	34 / 7	9.3	0.59	0.096	0.85
		Exact	Minimax	4 / 3	26 / 6	9.0	0.50	0.087	0.85
			Optimal	4 / 3	26 / 6	9.0	0.50	0.087	0.85
0.1	0.9	Asy	Minimax	9 / 4	44 / 8	11.1	0.61	0.093	0.90
			Optimal	9 / 4	44 / 8	11.1	0.61	0.093	0.90
		Exact	Minimax	8 / 4	34 / 7	10.7	0.56	0.100	0.90
			Optimal	8 / 4	34 / 7	10.7	0.56	0.100	0.90

For each design, ESS and PET denote the expected sample size and the probability of early termination under null hypotheses. n_1,n are the sample sizes per treatment for first stage and overall, respectively.

Table 3: Minimax and Optimal Designs for $\delta = 1$.

				Reject Drug if U					
α	power	Type	Design	$\leq r_1/n_1$	$\leq r/n$	ESS	PET	TIE	POWER
0.05	0.8	Asy	Minimax	37 / 8	151 / 15	20.2	0.70	0.050	0.80
			Optimal	21 / 6	189 / 17	18.9	0.68	0.050	0.81
		Exact	Minimax	20 / 6	150 / 15	18.3	0.65	0.050	0.80
			Optimal	14 / 5	188 / 17	18.3	0.66	0.050	0.81
0.05	0.85	Asy	Minimax	46 / 9	191 / 17	23.0	0.69	0.050	0.85
			Optimal	29 / 7	280 / 21	21.9	0.72	0.050	0.85
		Exact	Minimax	26 / 7	191 / 17	22.0	0.60	0.048	0.85
			Optimal	20 / 6	256 / 20	21.8	0.65	0.050	0.86
0.05	0.9	Asy	Minimax	56 / 10	259 / 20	26.5	0.67	0.050	0.90
			Optimal	47 / 9	334 / 23	25.9	0.72	0.050	0.90
		Exact	Minimax	52 / 10	236 / 19	27.7	0.57	0.050	0.90
			Optimal	43 / 9	259 / 20	26.8	0.60	0.050	0.90
0.1	0.8	Asy	Minimax	29 / 7	79 / 11	16.3	0.72	0.098	0.80
			Optimal	14 / 5	107 / 13	16.0	0.62	0.099	0.81
		Exact	Minimax	13 / 5	79 / 11	15.0	0.58	0.096	0.80
			Optimal	13 / 5	79 / 11	15.0	0.58	0.096	0.80
0.1	0.85	Asy	Minimax	74 / 11	94 / 12	22.4	0.81	0.099	0.85
			Optimal	28 / 7	124 / 14	18.6	0.67	0.097	0.85
		Exact	Minimax	35 / 8	94 / 12	18.9	0.64	0.095	0.85
			Optimal	35 / 8	94 / 12	18.9	0.64	0.095	0.85
0.1	0.9	Asy	Minimax	43 / 9	143 / 15	23.0	0.59	0.098	0.90
			Optimal	35 / 8	160 / 16	22.0	0.62	0.099	0.90
		Exact	Minimax	42 / 9	143 / 15	23.2	0.57	0.097	0.90
			Optimal	25 / 7	160 / 16	22.1	0.55	0.100	0.90

For each design, ESS and PET denote the expected sample size and the probability of early termination under null hypotheses. n_1,n are the sample sizes per treatment for first stage and overall, respectively.

Figure 4: Two-stage Mann-Whitney exact designs for $\alpha = 0.05$, power=0.85, under alternative $\delta = 1.5$

can be substantial. If we use the asymptotic method for obtaining the minimax design, $n_1 = 2$ and $n_2 = 4$ for each treatment are called for so to satisfy the constraints $\alpha = 0.1$ and power=0.8, and the ESS is 7 patients for the design. The exact minimax design reduces the ESS from 7 to 5 with the same PET=0.5. Although the total sample size may be equal when comparing the two methods obtained via minimax criteria, often the exact minimax design realizes savings in the number of patients used for the first stage. In order to satisfy the constraints $\alpha = 0.05$ and power=0.85 under the alternative $\delta = 1.5$ (Table 2), the total sample sizes for both methods using the minimax criteria are $2 \times 8 = 16$, and the same decision rule is used at the second stage. Note, though the exact design reduces the first stage's sample size from $2 \times 7 = 14$ to $2 \times 4 = 8$, and further reduces the ESS from 14.2 to 10.7. The difference in the ESS between the minimax and optimal designs by exact calculation is very small. However, for the asymptotic method, the differences can be larger. For small δ, the total sample sizes are large, and computation of exact designs may be found to be computational infeasible. Rather than computing exact probabilities under the null, each quantities may be computed via Monte Carlo methods.

In the one-stage Mann-Whitney test, Hollander and Wolfe [8, 18] showed that the variance of the Mann-Whitney statistic becomes smaller in the presence of ties. The variance is affected by both the ties within and

Table 4: Type I error control and power of the design in the presence of tie; $\alpha = 0.05$,power=0.8 under the alternative $\delta = 2$.

	no tie	round to 0.01	round to 0.1	round to 0.2	round to 1
Percentage of tie under the null	0	0.03	0.22	0.40	0.87
Percentage of tie under the alternative	0	0.02	0.15	0.28	0.80
Actual α	0.042	0.0432	0.0440	0.0456	0.0414
Power	0.82	0.83	0.83	0.84	0.83

Table 5: Type I error control and power of the design in the presence of tie; $\alpha = 0.05$,power=0.8 under the alternative $\delta = 1.5$.

	no tie	round to 0.01	round to 0.1	round to 0.2	round to 1
Percentage of tie under the null	0	0.05	0.41	0.63	0.95
Percentage of tie under the alternative	0	0.04	0.34	0.55	0.94
Actual α	0.048	0.047	0.052	0.050	0.046
Power	0.81	0.81	0.82	0.83	0.84

between groups. We investigated the influence of ties in the exact two-stage Mann-Whitney designs by a Monte Carlo simulation for $\alpha = 0.05$ and power= 0.8 with various shift value $\delta = 2, 1.5$ and 1. We rounded simulated observations from the normal distributions to nearest multiple of 0.01,0.1,0.2,and 1 in order to produce data with a portion of the values being tied. The simulated type I error and power for $\delta = 2, 1.5$ and 1 based on 20,000 simulations are given in table 4, 5, and 6, respectively. In the presence of tie, exact two-stage Mann-Whitney designs still provide excellent type I error control and power properties.

Table 6: Type I error control and power of the design in the presence of tie; $\alpha = 0.05$,power=0.8 under the alternative $\delta = 1$.

	no tie	round to 0.01	round to 0.1	round to 0.2	round to 1
Percentage of tie under the null	0	0.09	0.58	0.79	0.97
Percentage of tie under the alternative	0	0.08	0.54	0.76	0.97
Actual α	0.050	0.052	0.053	0.051	0.046
Power	0.80	0.81	0.81	0.82	0.79

4 Discussion

In this article, we presented two-stage Phase II clinical trial designs for use with continuous endpoints. In Phase II clinical trials, 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 addition patient information obtained in stage 2 may be used to achieve greater precision in the estimates of effect. We have also not considered designs with three or more stages [2] 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 design have been shown to be robust in the scenarios considered in the note, one weakness of the proposed design 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 a R package to calculate the density and cumulative distribution function of the two-stage Mann-Whitney test statistic for finite sample sizes. The program is available upon request. In this R package, the exact two-stage designs for continuous endpoints are made available for user defined parameters.

References

[1] P. Castagliola. Optimized algorithms for computing wilcoxon's Tn, wilcoxon's Wm,n and ansari-bradley's Am,n statistics when m and n are small. *Journal of Applied Statistics*, pages 41–58, February 1996.

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

[3] T. T. Chen and T. H. Ng. Optimal flexible designs in phase II clinical trials. *Statistics in medicine*, 17(20):2301–2312, October 1998.

[4] T. R. Fleming. One-sample multiple testing procedure for phase II clinical trials. *Biometrics*, 38(1):143–151, 1982.

[5] E. A. Gehan. 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*, 13(4):346–353, April 1961.

[6] S. Green, J. Crowley, J. Benedetti, and A. Smith. *Clinical Trials in Oncology, Second Edition.* Chapman and Hall/CRC, 2nd edition, July 2002.

[7] S. J. Green and S. Dahlberg. Planned versus attained design in phase II clinical trials. *Statistics in Medicine*, 11(7):853–862, 1992.

[8] M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods, 2nd Edition.* Wiley-Interscience, 2 edition, January 1999.

[9] K. Hornik. *Wilcoxon rank sum statistic in R package version 2.10-1.*, 2010.

[10] S. H. Jung, M. Carey, and K. M. Kim. Graphical search for two-stage designs for phase II clinical trials. *Controlled clinical trials*, 22(4):367–372, August 2001.

[11] T. G. Karrison, M. L. Maitland, W. M. Stadler, and M. J. Ratain. Design of phase II cancer trials using a continuous endpoint of change in tumor size: Application to a study of sorafenib and erlotinib in non small-cell lung cancer. *J. Natl. Cancer Inst.*, 99(19):1455–1461, October 2007.

[12] J. Kepner and M. N. Chang. On the maximum total sample size of a group sequential test about binomial proportions. *Statistics & Probability Letters*, 62(1):87–92, March 2003.

[13] J. J. Lee and L. Feng. Randomized phase II designs in cancer clinical trials: Current status and future directions. *J Clin Oncol*, 23(19):4450–4457, July 2005.

[14] E. L. Lehmann. Nonparametrics statistical methods based on ranks. 1975.

[15] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 1947.

[16] L. Mariani and E. Marubini. Design and analysis of phase II cancer trials: A review of statistical methods and guidelines for medical researchers. *International Statistical Review / Revue Internationale de Statistique*, 64(1):61–88, 1996.

[17] M. S. Mayo and B. J. Gajewski. Bayesian sample size calculations in phase II clinical trials using informative conjugate priors. *Controlled clinical trials*, 25(2):157–167, April 2004.

[18] R. L. Ott and M. T. Longnecker. *An Introduction to Statistical Methods and Data Analysis*. Duxbury Press, 5 edition, December 2000.

[19] J. J. Shuster, M. N. Chang, and L. Tian. Design of group sequential clinical trials with ordinal categorical data based on the mannwhitneywilcoxon test. *Sequential Analysis: Design Methods and Applications*, 23(3):413–426, 2004.

[20] R. Simon. Optimal two-stage designs for phase II clinical trials. *Controlled clinical trials*, 10(1):1–10, March 1989.

[21] R. Simon, R. E. Wittes, and S. S. Ellenberg. Randomized phase II clinical trials. *Cancer treatment reports*, 69(12):1375–1381, December 1985.

[22] J. D. Spurrier and J. E. Hewett. Wilcoxon rank sum statistic. *Technical Report No. 62G10-1, Department of Mathematics and Computer Science, University of South Carolina*, 1975.

[23] J. D. Spurrier and J. E. Hewett. Two-stage wilcoxon tests of hypotheses. *Journal of the American Statistical Association*, 71(356):982–987, 1976.

[24] B. E. Storer. A class of phase II designs with three possible outcomes. *Biometrics*, 48(1):55–60, 1992.

[25] J. M. G. Taylor, T. M. Braun, and Z. Li. Comparing an experimental agent to a standard agent: relative merits of a one-arm or randomized two-arm phase II design. *Clin Trials*, 3(4):335–348, August 2006.

[26] P. F. Thall. A review of phase 2-3 clinical trial designs. *Lifetime Data Analysis*, 14(1):37–53, March 2008.

[27] P. F. Thall, R. M. Simon, and E. H. Estey. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine*, 14(4):357–379, 1995.

[28] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.