PLOS ONE

RESEARCH ARTICLE

Optimal two-stage group sequential designs based on Mann-Whitney-Wilcoxon test

Yeonhee Park*

Department of Statistics, Sungkyunkwan University, Seoul, South Korea

* yeonheepark@skku.edu

OPEN ACCESS

Citation: Park Y. (2025) Optimal two-stage group sequential designs based on Mann-Whitney-Wilcoxon test. PLoS ONE 20(2): e0318211. https://doi.org/10.1371/journal.pone.0318211

Editor: Asli Suner Karakulah, Ege University, Faculty of Medicine, TÜRKIYE

Received: August 19, 2024

Accepted: January 10, 2025

Published: February 20, 2025

Data availability statement: No data was used for the research described in the article.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The Mann-Whitney-Wilcoxon test, often referred to as the Mann-Whitney U test or Wilcoxon rank-sum test, is a non-parametric statistical test used to compare two independent groups when the dependent variable is ordinal or continuous but not normally distributed. It's particularly useful for small sample sizes or when the assumptions of parametric tests, such as the t-test, are violated, including cases where the data is skewed. This study focuses on the Mann-Whitney-Wilcoxon test for ordinal data, which frequently arises in biomedical research when the proportional odds assumption does not hold. Currently, there are no optimal two-stage randomized clinical trial designs utilizing the Mann-Whitney-Wilcoxon test. To address this research gap, our study proposes optimal two-stage designs based on the Mann-Whitney-Wilcoxon test. We demonstrate the application of these designs through illustrative examples and evaluate their operating characteristics.

Introduction

Outcomes in clinical trials are often ordinal. In 2020, the World Health Organization (WHO) developed a nine-point clinical progression scale to track COVID-19 severity over time [1]. The scale ranges from 0 (no infection) to 8 (death), covering the full spectrum of clinical outcomes. This comprehensive scale has been widely used in COVID-19 trials and cohort studies [2–4]. Also, the modified Rankin Score (mRS) is a measure of disability ranging from 0 (no symptoms) to 6 (death). This is a widely used outcome measure in stroke clinical trials [5–7]. For example, randomized clinical trials such as NCT05199194, NCT05046106, and NCT00059332 primarily measure the mRS to evaluate the effectiveness of the intervention.

In practice, a dichotomized approach to ordered categorical variables is often used for comparative tests. For example, dichotomized mRS (0–1/2–6, 0–2/3–6, and 0–4/5–6) is commonly used [8–10]. However, as Altman and Royston (2006) [11] pointed out, dichotomizing leads to a reduction in statistical power to detect relevant treatment effects [12,13]. To address the issue of analyzing ordinal endpoints, proportional odds logistic regression has been recommended. The proportional odds assumption states that the relationship between the independent variables and the log-odds of the cumulative probabilities of the ordinal categories is the same across all levels or categories of the dependent variable. Using the proportional

odds model, Whitehead (1993) [14] proposes a statistical method to calculate the sample size based on the score test statistic for two-arm randomized clinical trials. The method by Whitehead (1993) [14] was improved by calculating higher order moments of the test statistic using the Cornish-Fisher series and the Edgeworth series [15]. These proportional odds methods preserve the ordinal nature of outcome measures, but the proportional odds assumption can be violated in clinical trials. As an alternative to avoid the proportional odds assumption, the Mann-Whitney-Wilcoxon test or other linear rank tests and win odds test are utilized. Shuster et al. (2002) [16] illustrates a minimax two-stage design using the calculation for a single-stage design and variances of the test statistics at the first and second stages. Shuster et al. (2004) [17] and Nowak et al. (2022) [18] investigate the Mann-Whitney-Wilcoxon test in the context of group sequential trials using the asymptotic canonical joint distribution. Hilton and Mehta (1993) [19] and Rabbee et al. (2003) [20] propose methods of sample size calculation applicable to any linear rank test under a general stochastically ordered alternative hypothesis for single-stage clinical trials. Moreover, Gasparyan et al. (2021) [21] proposes a sample size calculation based on the single-stage win odds test, which is used to analyze hierarchical composite endpoints, for ordinal variables.

Currently, no optimal two-stage randomized clinical trial designs utilize the Mann-Whitney-Wilcoxon test. Optimal two-stage designs based on the Mann-Whitney-Wilcoxon test can offer several advantages, particularly in terms of efficiency and flexibility [22–25]. These designs can reduce the total sample size needed to reach conclusive results, as unpromising trials can be stopped early, conserving resources and time. This flexibility not only protects participants from ineffective treatments but also minimizes patient exposure to potentially harmful interventions.

To fill this gap, we propose new methods to determine the sample size and monitoring rule for the optimal two-stage randomized clinical trial designs based on the Mann-Whitney-Wilcoxon test. We provide two types of the two-stage designs called F design and FS design allowing early stopping of the trial due to futility or superiority of the experimental treatment. In addition, we provide three criteria to select the optimal designs that minimize the expected total sample size. These methodologies offer clinical practitioners user-friendly study design options, enabling them to easily identify appropriate designs tailored to their specific study objectives.

The rest of this paper is organized as follows. We first propose optimal two-stage designs using the Mann-Whitney-Wilcoxon test. We provide explicit formulas for calculating sample sizes and identifying monitoring rules based on the optimal design parameters. We illustrate the methods with two examples and compare the designs. Additionally, we apply the proposed methods to redesign a trial and provide concluding remarks.

Optimal two-stage designs using Mann-Whitney-Wilcoxon test

We consider two-arm clinical trials in which patients are enrolled and randomized with a 1:1 ratio to either a control C or an experimental treatment E. Let X be a treatment assignment having 1 or 2 for the patient receiving treatment C or E, respectively, and Y be the ordinal outcome with J ordered levels, i.e., $y \in \{1, 2, \dots, J\}$. Without loss of generality, we regard the smallest outcome $Y = 1$ as the best clinical outcome and the largest outcome $Y = J$ as the worst clinical outcome. For $i = 1, 2$ and $j = 1, 2, \dots, J$, let $p_{ij} = \Pr(Y = j|X = i)$ be the probability that an ordinal outcome on treatment i falls in the jth level. Let F and G be the distribution function of Y on treatment C and E, respectively. Suppose that we are interested in testing the null hypothesis $H_0 : F(y) = G(y)$ versus the alternative hypothesis $H_a : F(y) < G(y)$. The Mann-Whitney-Wilcoxon test is used to compare two independent samples based on the

ranks of the observed ordinal outcome data. Specifically, let $Y_1 \sim F$ and $Y_2 \sim G$ denote the two independent random variables. Since the smallest value indicates the best outcome, we can compare the outcomes between two arms: for any $l, m = 1, 2, \ldots$, the event of $Y_{1l} > Y_{2m}$ indicates that E wins while C loses; the event $Y_{1l} < Y_{2m}$ indicates that C wins while E loses; the event $Y_{1l} = Y_{2m}$ indicates that C and E are equivalent. This constructs the test statistics as follows. Based on the samples $Y_{11}, \ldots, Y_{1n_1}$ from F and $Y_{21}, \ldots, Y_{2n_2}$ from G, the Mann-Whitney-Wilcoxon test statistic is defined in [26] and [17] by

$$W = \sum_{l=1}^{n_1} \sum_{m=1}^{n_2} \{I(Y_{1l} > Y_{2m}) - I(Y_{1l} < Y_{2m})\}/(n_1 n_2). \tag{1}$$

This accounts for the proportion of all comparisons leading to a win for arm E, i.e., $F(y) < G(y)$. In the context of the clinical trials investigating the treatment efficacy of arms C and E, ordinal data measure the treatment efficacy, and the event that E wins corresponds to the superiority of the experimental drug E while the event that C wins corresponds to the futility of the experimental drug E. We use the Mann-Whitney-Wilcoxon test based on (1) to propose an optimal two-stage design for ordinal outcomes.

Before we see the optimal two-stage design for ordinal outcomes based on the Mann-Whitney-Wilcoxon test (1), we first review the Mann-Whitney-Wilcoxon test for the design. For $i = 1, 2$, $j = 1, 2, \ldots, J$, and $k = 1, 2$, let n_{ijk} denote the corresponding observed frequencies for the jth level at the kth analysis when treatment i is received. Then, $n_{\cdot jk} = n_{1jk} + n_{2jk}$ denotes the marginal frequency for the jth level at the kth analysis. Based on the accumulating data at the kth analysis, (1) becomes

$$W_k = \left[\sum_{j=2}^{J} n_{1jk}\left(\sum_{l=1}^{j-1} n_{2lk}\right) - \sum_{j=2}^{J} n_{2jk}\left(\sum_{l=1}^{j-1} n_{1lk}\right)\right] \Big/ \left(\sum_{j=1}^{J} n_{1jk} \sum_{j=1}^{J} n_{2jk}\right), \tag{2}$$

which is the Mann-Whitney-Wilcoxon test statistic used at the kth analysis. The test statistic W_k is standardized to

$$T_k = W_k \Big/ \sqrt{\frac{\sum_j n_{\cdot jk} + 1}{3 \sum_j n_{1jk} \sum_j n_{2jk}} \left\{1 - \frac{\sum_j (n_{\cdot jk}^3 - n_{\cdot jk})}{(\sum_j n_{\cdot jk})^3 - \sum_j n_{\cdot jk}}\right\}}, \tag{3}$$

where the denominator of (3) is the estimate of standard deviation of W_k under the null hypothesis [17]. For a large sample, Shuster et al. (2004) [17] provides the variance of W_k approximated by $(Q + R)/n_k$, where

$$p^+ = \sum_{j=1}^{J-1} p_{2j} \sum_{l=j+1}^{J} p_{1l}, \qquad p^- = \sum_{j=1}^{J-1} p_{1j} \sum_{l=j+1}^{J} p_{2l}, \qquad D = p^+ - p^-,$$

$$Q = \sum_{j=1}^{J-1} p_{1j}\left(\sum_{l=j+1}^{J} p_{2l}\right)^2 + \sum_{j=2}^{J} p_{1j}\left(\sum_{l=1}^{j-1} p_{2l}\right)^2 - 2\sum_{j=2}^{J-1} p_{1j}\left(\sum_{l=1}^{j-1} p_{2l}\right)\left(\sum_{l=j+1}^{J} p_{2l}\right) - D^2,$$

and

$$R = \sum_{j=1}^{J-1} p_{2j}\left(\sum_{l=j+1}^{J} p_{1l}\right)^2 + \sum_{j=2}^{J} p_{2j}\left(\sum_{l=1}^{j-1} p_{1l}\right)^2 - 2\sum_{j=2}^{J-1} p_{2j}\left(\sum_{l=1}^{j-1} p_{1l}\right)\left(\sum_{l=j+1}^{J} p_{1l}\right) - D^2.$$

We notice that D_0, which indicates the value of D under the null hypothesis, is zero. The quantity of D under the alternative hypothesis is denoted by D_a. To distinguish the statistics under the null and alternative hypotheses, hereafter we will use notations Q_0 and R_0 for Q and R, respectively, under the null hypothesis and Q_a and R_a for Q and R, respectively, under the alternative hypothesis.

Sample size calculation for two-stage design using Mann-Whitney-Wilcoxon test

The two-stage design consists of an interim analysis planned when the total accrual reaches n_1 patients per group, i.e., $n_1 = \sum_{j=1}^{J} n_{1j1} = \sum_{j=1}^{J} n_{2j1}$, and the final analysis is performed after the evaluation of total planned number of n_2 patients per group, i.e., $n_2 = \sum_{j=1}^{J} n_{1j2} = \sum_{j=1}^{J} n_{2j2}$. At the interim, we monitor the treatment efficacy based on the accumulating data and determine whether to continue or stop the trial.

Futility monitoring design (F design). When the total accrual reaches $2n_1$ patients, an interim analysis is performed to test for futility, such that the trial stops if the experimental treatment E is more futile than the control C, i.e., $W_1 \leq t_1$ for a prespecified constant t_1. If $W_1 > t_1$, we continue to enroll $2(n_2 - n_1)$ more patients for the second stage, and the test proceeds to the final analysis. Based on all enrolled patients in the trial, if $W_2 > t_2$ for a prespecified constant t_2, the test terminates with the rejection of the null hypothesis.

Let α_1 and β_1 be given such that $1 - \alpha_1$ and $1 - \beta_1$ denote the probability of making a correct decision at the interim analysis under the null hypothesis and the alternative hypothesis, respectively. By the definition of α_1, i.e., $\Pr(W_1 \leq t_1 | H_0) = 1 - \alpha_1$, we obtain

$$t_1 = z_{\alpha_1} \sqrt{\frac{Q_0 + R_0}{n_1}} \tag{4}$$

where z_{α_1} denotes the critical value of the standard normal distribution at α_1. With the Eq (4), the definition of β_1, i.e., $\Pr(W_1 > t_1 | H_a) = 1 - \beta_1$, gives

$$n_1 = \left(\frac{z_{\alpha_1} \sqrt{Q_0 + R_0} + z_{\beta_1} \sqrt{Q_a + R_a}}{D_a} \right)^2. \tag{5}$$

Assuming that the type I error rate is at most α and the expected power is at least $1 - \beta$, we set $\Pr(W_2 > t_2 | H_0) = \alpha$ and $\Pr(W_2 > t_2 | H_a) - \Pr(W_1 \leq t_1 | H_a) = 1 - \beta$, because $\Pr(W_2 > t_2, W_1 > t_1 | H_0) \leq \Pr(W_2 > t_2 | H_0)$ and $\Pr(W_2 > t_2 | H_a) - \Pr(W_1 \leq t_1 | H_a) \leq \Pr(W_2 > t_2, W_1 > t_1 | H_a)$. Therefore, we obtain

$$t_2 = z_{\alpha} \sqrt{\frac{Q_0 + R_0}{n_2}} \tag{6}$$

and

$$n_2 = \left(\frac{z_{\alpha} \sqrt{Q_0 + R_0} + z_{\beta_2} \sqrt{Q_a + R_a}}{D_a} \right)^2, \tag{7}$$

where $\beta_2 = \beta - \beta_1$.

Futility and superiority monitoring design (FS design). The interim analysis can be conducted to evaluate the efficacy of the intervention and to stop the trial if an experimental treatment is ineffective or is found to be significantly more effective than the control. Let v_1, v_2, and v_3 be the prespecified thresholds for the analyses. Specifically, we stop the trial for

futility if $W_1 \leq v_1$, and we stop the trial for superiority if $W_1 > v_2$. Otherwise, i.e., $v_1 < W_1 \leq v_2$, we continue to enroll $2(n_2 - n_1)$ more patients for the second stage. At the final analysis based on all $2n_2$ enrolled patients, we reject the null hypothesis if $W_2 > v_3$.

Let α_1 and β_2 be given such that $1 - \alpha_1$ and $1 - \beta_2$ denote the probability of making a correct decision at the interim analysis under the null hypothesis and the alternative hypothesis, respectively, i.e., $\Pr(W_1 \leq v_1|H_0) = 1-\alpha_1$ and $\Pr(W_1 > v_2|H_a) = 1-\beta_2$. Let $\alpha_2 = \Pr(W_1 > v_2|H_0)$ and $\beta_1 = \Pr(W_1 \leq v_1|H_a)$. By the definition of α_1, α_2, and β_1, we have

$$v_1 = z_{\alpha_1}\sqrt{\frac{Q_0 + R_0}{n_1}}, \tag{8}$$

$$v_2 = z_{\alpha_2}\sqrt{\frac{Q_0 + R_0}{n_1}}, \tag{9}$$

and

$$n_1 = \left(\frac{z_{\alpha_1}\sqrt{Q_0 + R_0} - z_{1-\beta_1}\sqrt{Q_a + R_a}}{D_a}\right)^2. \tag{10}$$

We notice that the overall type I error rate is $\Pr(W_1 > v_2|H_0) + \Pr(v_1 < W_1 \leq v_2, W_2 > v_3|H_0)$, which is less than or equal to $\Pr(W_1 > v_2|H_0) + \Pr(W_2 > v_3|H_0)$. Since we want the type I error rate to be at most α, we set $\Pr(W_2 > v_3|H_0)$ to be $\alpha - \alpha_2$. In addition, we observe that

$$\begin{aligned} \text{power} &= \Pr(W_1 > v_2|H_a) + \Pr(v_1 < W_1 \leq v_2, W_2 > v_3|H_a) \\ &\geq 1 - \beta_2 + \Pr(v_1 < W_1 \leq v_2|H_a) + \Pr(W_2 > v_3|H_a) - 1 = \Pr(W_2 > v_3|H_a) - \beta_1. \end{aligned}$$

Since we want the expected power to be at least $1 - \beta$, we set $\Pr(W_2 > v_3|H_a)$ to be $1 - (\beta - \beta_1)$. Therefore, we obtain

$$v_3 = z_{\alpha-\alpha_2}\sqrt{\frac{Q_0 + R_0}{n_2}} \tag{11}$$

and

$$n_2 = \left(\frac{z_{\alpha-\alpha_2}\sqrt{Q_0 + R_0} - z_{1-(\beta-\beta_1)}\sqrt{Q_a + R_a}}{D_a}\right)^2. \tag{12}$$

Optimal choice of design parameters

Determination of the thresholds and sample sizes for a two-stage design requires the specification of the design parameters such as α_1 and β_1 for F design and α_1, α_2, and β_1 for FS design. We consider three criteria to specify the design parameters: (1) minimizing the expected total sample size under the null hypothesis, (2) minimizing the expected total sample size under the alternative hypothesis, and (3) minimizing the expected overall total sample size. The overall total sample size is defined as the total sample size, averaged equally across the null and alternative hypotheses.

In two-stage designs with futility monitoring (i.e., F design), if $W_1 \geq t_1$, we stop the trial for futility, and the total sample size is $2n_1$. Otherwise, i.e., $W_1 < t_1$, the trial continues to enroll $2(n_2 - n_1)$ patients for the second stage, and the total sample size is $2n_2$. Therefore, the expected total sample size under the null hypothesis is $2n_1 + 2(n_2 - n_1)\alpha_1$, the expected total

sample size under the alternative hypothesis is $2n_1 + 2(n_2 - n_1)(1 - \beta_1)$, and the expected overall total sample size is $2n_1 + (n_2 - n_1)(\alpha_1 + 1 - \beta_1)$. Of note, n_1 and n_2 are functions of α_1 and β_1. We choose the optimal values of α_1 and β_1 according to the criteria over $\alpha \leq \alpha_1$ and $\beta/2 \leq \beta_1 < \beta$. The constraint of $\beta/2 \leq \beta_1$ guarantees $n_1 \leq n_2$.

Similarly, in two-stage designs with futility and superiority monitoring (i.e., FS design), the expected total sample size under the null hypothesis is $2n_1 + 2(n_2 - n_1)(\alpha_1 - \alpha_2)$, the expected total sample size under the alternative hypothesis is $2n_1 + 2(n_2 - n_1)(\beta_2 - \beta_1)$, and the expected overall total sample size is $2n_1 + (n_2 - n_1)(\alpha_1 - \alpha_2 + \beta_2 - \beta_1)$. The values of β_2, n_1, and n_2 depend on α_1, α_2, and β_1, and the optimal values of α_1, α_2, and β_1 are chosen according to the criteria over $\alpha \leq \alpha_1 < 1$, $0 < \alpha_2 < \alpha$, and $0 < \beta_1 < \beta$. Any pairs of α_1, α_2, and β_1 are excluded unless $0 < n_1 < n_2$.

We find that the minimax criterion is not applicable to determine the design parameters of the two-stage designs. In F design, minimizing the maximum sample size is equivalent to maximizing β_1 over $\beta/2 \leq \beta_1 < \beta$. Thus, it chooses the lower boundary of β_1 as the optimal value to attain the minimum value of n_2 regardless of α_1. Thus, the value of α_1 cannot be uniquely determined for F design. Similarly, in FS design, the minimum value of n_2 is attained over $0 < \alpha_2 < \alpha$ and $0 < \beta_1 < \beta$ satisfying $0 < n_1 < n_2$, but the value of α_1 is not uniquely determined.

Decision rule for single-stage design

The decision rule for a single-stage design can be straightforwardly derived from that of a two-stage design. By setting $t_1 = -\infty$, it becomes impossible to stop the trial for futility at stage 1. Furthermore, by setting $n_1 = n_2$, the second stage sample size is effectively eliminated, as the entire sample is enrolled in a single stage. This simplifies the design, removing the need for interim analysis or additional patient accrual after the initial enrollment. Specifically, the single-stage design employs the Mann-Whitney-Wilcoxon test statistic denoted by W, calculated using data from all $2n$ patients at the end of the trial. The null hypothesis is rejected if $W > t$, where t is a prespecified threshold.

The decision rule (t, n) of the single–stage design is determined to ensure that the design satisfies $\Pr(W_2 > t_2|H_0) = \alpha$ and $\Pr(W_2 > t_2|H_a) = 1 - \beta$, meeting the specified type I error rate and power requirements. Similarly to the derivation of Eqs (4)–(12), by using the property of the Mann-Whitney-Wilcoxon test statistic [17], we obtain

$$t = z_\alpha \sqrt{\frac{Q_0 + R_0}{n}} \quad \text{and} \quad n = \left(\frac{z_\alpha \sqrt{Q_0 + R_0} + z_\beta \sqrt{Q_a + R_a}}{D_a} \right)^2.$$

Examples

Example 1. We consider a clinical trial whose overall clinical outcome is ordinal with 3 categories and a descending order of desirability, e.g., 1=clinical benefit without adverse effects (AEs), 2=clinical benefit with some AEs, and 3=survival without clinical benefit but with AEs. Suppose that the distribution of the desirability score for the standard treatment is uniform, i.e., the probabilities of each category for the standard treatment are 1/3, while investigators wish to have the probabilities 1/2, 1/3, and 1/6 for each category for the experimental treatment. Target error rates are set to be $\alpha = 0.05$ and $\beta = 0.2$ for the test comparing two treatments.

First, we consider a two-stage design with futility monitoring. By applying the Eqs (4)–(7), the optimal decision rule (t_1, n_1, t_2, n_2) is determined according to the criterion. Let

EN0 denote the expected total sample size under the null hypothesis, and ENa denote the expected total sample size under the alternative hypothesis. As described, Criterion 1 is to minimize EN0, Criterion 2 is to minimize ENa, and Criterion 3 is to minimize (1/2)EN0 + (1/2)ENa. Second, we consider a two-stage design monitoring both futility and superiority. Similarly to the design with futility monitoring, by applying the Eqs (8)–(12), the decision rule $(v_1, v_2, n_1, v_3, n_2)$ is determined according to the criterion. Lastly, we include a single-stage design, referred to as *1 design*, to provide a basis for comparison. For consistency in notation, the decision rule (t, n) is represented in the two–stage format as $(t_1, n_1, t_2, n_2) = (-\infty, n, t, n)$, where $t_1 = -\infty$ and $n_1 = n_2 = n$. Table 1 summarizes the decision rules of the designs.

To investigate the operating characteristics of the proposed designs, we conduct a simulation study with 10000 replications. We report four performance metrics: overall type I error rate denoted by $\hat{\alpha}$, power (i.e., one minus overall type II error rate), and the expected sample size under null and alternative hypotheses denoted by EN0 and ENa, respectively. For all three criteria, all designs preserve overall type I and II error rates at the target rates. The overall type I error rates are slightly smaller than the nominal level, which yet does not sacrifice the power, i.e., overall type II error rates are also smaller than the nominal level. Criterion 1 yields a smaller size of the first stage for both F design and FS design compared to other criteria, which leads to a smaller expected sample size under the null hypothesis. Specifically, both the F design and FS design with Criterion 1 spend only 30% of information of the trial to make the decision earlier than others. The FS design with Criterion 2 or Criterion 3 requires n_2 =83 patients per group, which is the smallest maximum sample size of the trial among all six decision rules. We also notice that the the F design with Criterion 2 uses extremely small threshold for futility monitoring, and $n_1 = 1$ is calculated. With $n_1 = 1$ being calculated, implementing a two-stage design becomes impractical and is therefore not recommended. While 1 design requires a smaller maximum sample size compared to the two-stage designs (F and FS designs), we observe that two-stage designs—except for the F design with Criterion 2—tend to use a smaller expected sample size under the null hypothesis. This is an advantage, as it reduces the number of patients enrolled when the treatment shows no effect, thus saving resources and minimizing patient exposure to potentially ineffective treatments. Moreover, most two-stage designs utilize a larger expected sample size under the alternative hypothesis, which is advantageous. Enrolling more patients in the effective treatment not only enhances

Table 1. Decision rule and its operating characteristics such as overall type I error rate denoted by $\hat{\alpha}$, power, expected sample size under null hypothesis denoted by EN0, and expected sample size under alternative hypothesis denoted by ENa. The sample sizes n_1 and n_2 are obtained by rounding up the results.

		Decision rule					Operating characteristics			
		t_1		n_1	t_2	n_2	$\hat{\alpha}$	Power	EN0	ENa
F design	Criterion 1	0.060		32	0.125	104	0.040	0.836	113.64	191.47
F design	Criterion 2	-36.75		1	0.127	100	0.049	0.899	200	200
F design	Criterion 3	0.047		30	0.127	100	0.040	0.833	110.51	185.66
		Decision rule					Operating characteristics			
		v_1	v_2	n_1	v_3	n_2	$\hat{\alpha}$	Power	EN0	ENa
FS design	Criterion 1	0.061	0.423	32	0.125	105	0.042	0.834	112.22	184.68
FS design	Criterion 2	-0.046	0.270	42	0.151	83	0.043	0.817	135.95	137.64
FS design	Criterion 3	-0.004	0.314	38	0.144	83	0.043	0.825	121.44	143.73
		Decision rule					Operating characteristics			
		t_1		n_1	t_2	n_2	$\hat{\alpha}$	Power	EN0	ENa
1 design		-∞		73	0.149	73	0.048	0.799	146	146

https://doi.org/10.1371/journal.pone.0318211.t001

the precision of estimating the treatment effect but also maximizes the potential benefits for participants receiving the treatment.

Example 2. We consider an example trial illustrated in [14], whose primary endpoint is the patients' response evaluating the progress after the treatment, which is categorized as *very good* (=1), *good* (=2), *moderate* (=3), or *poor* (=4). Let $p_{ij}, i = 1, 2, j = 1, 2, 3, 4$ denote the probability that an ordinal outcome on treatment i falls in the jth level. Suppose that we have $p_{11} = 0.2, p_{12} = 0.5, p_{13} = 0.2$, and $p_{14} = 0.1$ for the control group. Then, the control probability of a *good* or *very good* outcomes is expected to be 0.7. Investigators assume the common odds ratio of the cumulative distributions and wish to achieve the probability 0.85 of a *good* or *very good* outcomes to test the effect of the experimental treatment. This provides the reference improvement 0.887 for superiority of the experimental treatment and the response probability for the experimental treatment [14]. Table 2 describes the null and alternative probabilities that patients' response falls in each category. We consider target error rates $\alpha = 0.05$ and $\beta = 0.1$ for the test.

Table 3 summarizes the decision rules of F design, FS design, and 1 design. We investigate the operating characteristics of the proposed design through simulations with 10000 replications. We observe that all designs preserve the overall type I and II error rates at target levels. Similarly to Example 1, Criterion 1 yields a smaller size for the first stage and results in the minimum value of EN0 among the criteria. In all decision rules for two-stage designs in Table 3, the maximum sample size of the trial is almost the same.

The common odds ratio assumption for the cumulative distributions holds in this example. Using the proportional odds model, Whitehead (1993) [14] proposes a method for determining the sample size of the trial based on the score test. The method of Whitehead (1993) [14] requires a total sample size of 188 (i.e., 94 patients per group) to achieve the improvement of the experimental treatment for the trial with a type I error rate of 0.05 and power 90%. This is obtained based on a single-stage test. We notice that the difference in the required total sample size of the trial is very small, but the proposed two-stage designs use, on average, 52.65 fewer patients across the six decision rules compared to the single-stage design in the method of Whitehead (1993) [14], when the experimental treatment is not effective. Moreover, the proposed single-stage design using the Mann-Whitney-Wilcoxon test requires 36 fewer patients in total compared to using the score test.

Trial illustration

The trial NCT03183167 is a longitudinal cohort study for patients with intracerebral hemorrhage (ICH), which is a devastating sub-type of stroke with a high mortality rate. Hagen et al. (2019) [27] analyzes the study data collected from 2008 to 2015 in order to investigate whether systematic inflammatory response syndrome (SIRS) is predictive of a long-term functional outcome in patients with spontaneous ICH. Functional outcome is evaluated using the modified Rankin Scale (mRS) at 3 and 12 months. The mRS is a widely used measure for assessing functional disability and dependence in individuals who have experienced stroke

Table 2. Response probability $p_{ij}, i = 1, 2, j = 1, 2, 3, 4$, which denotes the probability that an ordinal outcome on treatment i falls in the jth level.

	Very good ($j = 1$)	Good ($j = 2$)	Moderate ($j = 3$)	Poor ($j = 4$)
p_{1j}	0.2	0.5	0.2	0.1
p_{2j}	0.378	0.472	0.106	0.044

https://doi.org/10.1371/journal.pone.0318211.t002

Table 3. Decision rule and its operating characteristics such as overall type I error rate denoted by $\hat{\alpha}$, power, expected sample size under null hypothesis denoted by EN0, and expected sample size under alternative hypothesis denoted by ENa. The sample size n_1 and n_2 are obtained by rounding up the results.

		Decision rule					Operating characteristics			
		t_1		n_1	t_2	n_2	$\hat{\alpha}$	Power	EN0	ENa
F design	Criterion 1	0.048		37	0.126	98	0.041	0.915	116.21	189.88
F design	Criterion 2	-5.711		1	0.126	98	0.047	0.948	196	196
F design	Criterion 3	0.043		35	0.126	98	0.040	0.917	115.70	190.29
		Decision rule					Operating characteristics			
		v_1	v_2	n_1	v_3	n_2	$\hat{\alpha}$	Power	EN0	ENa
FS design	Criterion 1	0.046	0.393	36	0.126	99	0.040	0.917	116.84	177.12
FS design	Criterion 2	-0.012	0.234	43	0.147	97	0.043	0.917	142.83	133.02
FS design	Criterion 3	0.035	0.255	42	0.138	98	0.040	0.918	124.50	139.31
		Decision rule					Operating characteristics			
		t_1		n_1	t_2	n_2	$\hat{\alpha}$	Power	EN0	ENa
1 design		-∞		78	0.141	78	0.048	0.897	156	156

https://doi.org/10.1371/journal.pone.0318211.t003

or other neurological conditions. The ordinal form of mRS is more strongly related to the 5-year mortality rate than the dichotomous mRS, which splits the mRS into favorable or unfavorable outcomes (e.g., using 0-1/2-6 or 0-2/3-6 dichotomies) [13]. To compare patients with noninfectious SIRS to a control group of patients without systematic inflammatory response, Hagen et al. (2019) [27] use propensity score matching to have a well-balanced cohort of 104 patients per group [27]. The mRS at 3 months for patients with SIRS falls in categories 0 to 6 with probabilities 0.048, 0.077, 0.173, 0.125, 0.202, 0.125, 0.25 while the mRS at 3 months for patients in the control group falls in categories 0 to 6 with probabilities 0.058, 0.154, 0.25, 0.163, 0.154, 0.106, 0.115.

Using the information, we apply the proposed designs to determine the monitoring rule and sample size for the trial. Since investigators argue that SIRS is associated with poor mRS, the alternative hypothesis has the opposite inequality sign to our description, i.e., $H_a : F(y) > G(y)$. Suppose that target error rates are $\alpha = 0.05$ and $\beta = 0.2$. By applying Eqs (4)–(7) and (8)–(12), decision rules are obtained for F design and FS design. We also add the decision rule for 1 design. The performance is evaluated based on 10000 replications through simulation. The results are summarized in Table 4. The F design using Criterion 1 determines

Table 4. Decision rule and its operating characteristics such as overall type I error rate denoted by $\hat{\alpha}$, power, expected sample size under null hypothesis denoted by EN0, and expected sample size under alternative hypothesis denoted by ENa.

		Decision rule					Operating characteristics			
		t_1		n_1	t_2	n_2	$\hat{\alpha}$	Power	EN0	ENa
F design	Criterion 1	0.064		30	0.133	98	0.039	0.830	105.17	180.65
F design	Criterion 2	-17.45		1	0.136	95	0.049	0.902	190	190
F design	Criterion 3	0.050		29	0.136	95	0.038	0.836	106.21	176.92
		Decision rule					Operating characteristics			
		v_1	v_2	n_1	v_3	n_2	$\hat{\alpha}$	Power	EN0	ENa
FS design	Criterion 1	0.065	0.452	31	0.134	99	0.038	0.838	106.95	174.99
FS design	Criterion 2	-0.051	0.290	39	0.161	78	0.044	0.818	128.36	128.27
FS design	Criterion 3	-0.006	0.337	36	0.155	79	0.043	0.822	115.83	136.19
		Decision rule					Operating characteristics			
		t_1		n_1	t_2	n_2	$\hat{\alpha}$	Power	EN0	ENa
1 design		-∞		69	0.160	69	0.050	0.802	138	138

https://doi.org/10.1371/journal.pone.0318211.t004

$t_1 = 0.064, n_1 = 30, t_2 = 0.133, n_2 = 98$. The decision rule yields a power of 83% and is expected to use 105.17 patients when SIRS is not predictive of poorer long-term functional outcome. The FS design using Criterion 1 determines $v_1 = 0.065, v_2 = 0.452, n_1 = 31, v_3 = 0.134, n_2 = 99$. The decision rule yields a power of 83.8% and is expected to use 106.95 patients when SIRS is not predictive of poorer long-term functional outcome. The F and FS designs with Criterion 1 show that overall error rates are controlled at target levels and spend only 30% of the trial's information to make a decision earlier than others. Notably, the FS design using Criterion 2 or 3 requires a total sample size of less than 80 patients per group, which is smaller than other decision rules for two-stage designs. This decision rule would be preferable if investigators want to minimize the maximum sample size for the trial.

Discussion

We have proposed optimal two-stage clinical trial designs for ordinal data using the Mann-Whitney-Wilcoxon test. The proposed designs enable monitoring treatment efficacy during the trial and making an efficient decision earlier due to futility or superiority based on interim data. The proposed design works with three criteria depending on the objective of optimality. The optimal two-stage design with Criterion 1 minimizes the expected total sample size when the experimental treatment is not efficacious compared to the control. The optimal two-stage design with Criterion 2 minimizes the expected total sample size when the experimental treatment is efficacious compared to the control. The optimal two-stage design with Criterion 3 minimizes the expected overall total sample size. We provide explicit formulas for calculating sample sizes and identifying stopping boundaries for two-stage designs. The operating characteristics of the proposed designs are investigated through simulations. The simulation results show that overall type I and II error rates are well controlled at target rates. The proposed designs do not require a proportional odds assumption for the ordinal data and work well regardless of the proportional odds assumption. Thus, the proposed designs offer more options for clinical practitioners to consider when designing randomized clinical trials whose primary endpoint is an ordinal outcome.

In this study, we do not employ sample size adaptation. This choice aligns with our focus on a fixed sample size framework, which facilitates a more straightforward analysis and application of the proposed method. Optimizing the adaptation rule introduces additional complexities that are beyond the scope of this work. However, this is an area of interest for future research, where re-estimating the sample size at interim analysis could enhance the flexibility of clinical trial designs.

We have assumed that the primary ordinal data would be evaluated at fixed timelines and fully observed before the interim analysis. However, missing data in clinical trials is almost unavoidable due to factors such as patient dropout, loss of follow-up, or technical issues during data collection. For example, a COVID-19 patient discharged early may not have complete follow-up data for subsequent ordinal outcomes, such as disease progression. Additionally, clinical trials for immunotherapies, where responses can take weeks or months to manifest, and chronic conditions, where outcomes like organ failure or survival may occur long after the intervention, often encounter late-onset outcomes. These delayed events may remain unobserved by the end of the trial, resulting in incomplete ordinal outcome data. This presents opportunities to enhance the proposed method by incorporating advanced statistical approaches, such as multiple imputation or joint modeling, to address missing ordinal data effectively.

Author contributions

Conceptualization: Yeonhee Park.

Formal analysis: Yeonhee Park.

Investigation: Yeonhee Park.

Methodology: Yeonhee Park.

Validation: Yeonhee Park.

Writing – original draft: Yeonhee Park.

Writing – review & editing: Yeonhee Park.

References

1. WHO Working Group on the Clinical Characterisation and Management of COVID-19 infection. A minimal common outcome measure set for COVID-19 clinical research. Lancet Infect Dis. 2020;20(8):e192–7. https://doi.org/10.1016/S1473-3099(20)30483-7 PMID: 32539990
2. Brown SM, Peltan ID, Webb B, Kumar N, Starr N, Grissom C, et al. Hydroxychloroquine versus Azithromycin for Hospitalized Patients with Suspected or Confirmed COVID-19 (HAHPS). Protocol for a pragmatic, open-label, active comparator trial. Ann Am Thorac Soc. 2020;17(8):1008–15. https://doi.org/10.1513/AnnalsATS.202004-309SD PMID: 32425051
3. Wang Y, Zhang D, Du G, Du R, Zhao J, Jin Y, et al. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. Lancet. 2020;395(10236):1569–78. https://doi.org/10.1016/S0140-6736(20)31022-9 PMID: 32423584
4. Ivashchenko AA, Dmitriev KA, Vostokova NV, Azarova VN, Blinow AA, Egorova AN, et al. AVIFAVIR for treatment of patients with moderate Coronavirus Disease 2019 (COVID-19): interim results of a phase II/III multicenter randomized clinical trial. Clin Infect Dis. 2021;73(3):531–4. https://doi.org/10.1093/cid/ciaa1176 PMID: 32770240
5. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. Stroke. 1988;19(5):604–7. https://doi.org/10.1161/01.str.19.5.604 PMID: 3363593
6. de Haan R, Limburg M, Bossuyt P, van der Meulen J, Aaronson N. The clinical meaning of Rankin "handicap" grades after stroke. Stroke. 1995;26(11):2027–30. https://doi.org/10.1161/01.str.26.11.2027 PMID: 7482643
7. Kwon S, Hartzema AG, Duncan PW, Min-Lai S. Disability measures in stroke: relationship among the Barthel index, the functional independence measure, and the modified rankin scale. Stroke. 2004;35(4):918–23. https://doi.org/10.1161/01.STR.0000119385.56094.32 PMID: 14976324
8. Hacke W, Bluhmki E, Steiner T, Tatlisumak T, Mahagne MH, Sacchetti ML. Dichotomized efficacy end points and global end-point analysis applied to the ECASS intention-to-treat data set: post hoc analysis of ECASS I.. Stroke. 1998;29(10):2073–5.
9. O'Donnell MJ, Fang J, D'Uva C, Saposnik G, Gould L, McGrath E, et al. The PLAN score: a bedside prediction rule for death and severe disability following acute ischemic stroke. Arch Intern Med. 2012;172(20):1548–56. https://doi.org/10.1001/2013.jamainternmed.30 PMID: 23147454
10. Flint AC, Xiang B, Gupta R, Nogueira RG, Lutsep HL, Jovin TG, et al. THRIVE score predicts outcomes with a third-generation endovascular stroke treatment device in the TREVO-2 trial. Stroke. 2013;44(12):3370–5. https://doi.org/10.1161/STROKEAHA.113.002796 PMID: 24072003
11. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332(7549):1080. https://doi.org/10.1136/bmj.332.7549.1080 PMID: 16675816
12. Dijkland SA, Voormolen DC, Venema E, Roozenbeek B, Polinder S, Haagsma JA, et al. Utility-weighted modified rankin scale as primary outcome in stroke trials: a simulation study. Stroke. 2018;49(4):965–71. https://doi.org/10.1161/STROKEAHA.117.020194 PMID: 29535271
13. Ganesh A, Luengo-Fernandez R, Wharton RM, Rothwell PM, Oxford vascular study. Ordinal vs dichotomous analyses of modified rankin scale, 5-year outcome, and cost of stroke. Neurology. 2018;91(21):e1951–60. https://doi.org/10.1212/WNL.0000000000006554 PMID: 30341155
14. Whitehead J. Sample size calculations for ordered categorical data. Stat Med. 1993;12(24):2257–71. https://doi.org/10.1002/sim.4780122404 PMID: 8134732

15. Kolassa JE. A comparison of size and power calculations for the Wilcoxon statistic for ordered categorical data. Stat Med. 1995;14(14):1577–81. https://doi.org/10.1002/sim.4780141408 PMID: 7481194

16. Shuster J, Link M, Camitta B, Pullen J, Behm F. Minimax two-stage-designs with applications to tissue banking case-control studies. Stat Med. 2002;21(17):2479–93. https://doi.org/10.1002/sim.1198 PMID: 12205694

17. Shuster JJ, Chang MN, Tian L. Design of group sequential clinical trials with ordinal categorical data based on the Mann–Whitney–Wilcoxon test. Sequent Anal. 2004;23(3):413–26. https://doi.org/10.1081/sqa-200027053

18. Nowak CP, Mütze T, Konietschke F. Group sequential methods for the Mann-Whitney parameter. Stat Methods Med Res. 2022;31(10):2004–20. https://doi.org/10.1177/09622802221107103 PMID: 35698787

19. Hilton JF, Mehta CR. Power and sample size calculations for exact conditional tests with ordered categorical data. Biometrics. 1993;49(2):609–16. PMID: 8369392

20. Rabbee N, Coull BA, Mehta C, Patel N, Senchaudhuri P. Power and sample size for ordered categorical data. Stat Methods Med Res. 2003;12(1):73–84. https://doi.org/10.1191/0962280203sm317ra PMID: 12617509

21. Gasparyan SB, Kowalewski EK, Folkvaljon F, Bengtsson O, Buenconsejo J, Adler J, et al. Power and sample size calculation for the win odds test: application to an ordinal endpoint in COVID-19 trials. J Biopharm Stat. 2021;31(6):765–87. https://doi.org/10.1080/10543406.2021.1968893 PMID: 34551682

22. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989;10(1):1–10. https://doi.org/10.1016/0197-2456(89)90015-9 PMID: 2702835

23. Jung S-H, Lee T, Kim K, George SL. Admissible two-stage designs for phase II cancer clinical trials. Stat Med. 2004;23(4):561–9. https://doi.org/10.1002/sim.1600 PMID: 14755389

24. Jennison C, Turnbull BW. Adaptive and nonadaptive group sequential tests. Biometrika. 2006;93(1):1–21. https://doi.org/10.1093/biomet/93.1.1

25. Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. Stat Med. 2006;25(6):917–32. https://doi.org/10.1002/sim.2251 PMID: 16220524

26. Lehmann EL, et al. Statistical methods based on ranks. Nonparametrics. San Francisco, CA: Holden-Day; 1975.

27. Hagen M, Sembill JA, Sprügel MI, Gerner ST, Madžar D, Lücking H, et al. Systemic inflammatory response syndrome and long-term outcome after intracerebral hemorrhage. Neurol Neuroimmunol Neuroinflamm. 2019;6(5):e588. https://doi.org/10.1212/NXI.0000000000000588 PMID: 31355322