Published in final edited form as:

Stat Biosci. 2018 April ; 10(1): 217–232. doi:10.1007/s12561-017-9212-1.

Efficient two-stage designs and proper inference for animal studies

Chunyan Cai¹, Jin Piao², Jing Ning³, and Xuelin Huang³

¹Biostatistics/Epidemiology/Research Design Core, Center for Clinical and Translational Sciences, The University of Texas Health Science Center at Houston, Houston, Texas 77030, U.S.A

²Division of Biostatistics, The University of Texas School of Public Health, Houston, Texas 77030, U.S.A

³Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, U.S.A

Abstract

Cost-effective yet efficient designs are critical to the success of animal studies. We propose a two-stage design for cost-effectiveness animal studies with continuous outcomes. Given the data from the two-stage design, we derive the exact distribution of the test statistic under null hypothesis to appropriately adjust for the design's adaptiveness. We further generalize the design and inferential procedure to the K -sample case with multiple comparison adjustment. We conduct simulation studies to evaluate the small sample behavior of the proposed design and test procedure. The results indicate that the proposed test procedure controls the type I error rate for the one-sample design and the family-wise error rate for K -sample design very well; whereas the naive approach that ignores the design's adaptiveness due to the interim look severely inflates the type I error rate or family-wise error rate. Compared with the standard one-stage design, the proposed design generally requires a smaller sample size.

Keywords

Animal studies; cost-effectiveness; family-wise error rate; two-stage designs

1 Introduction

Preclinical studies with animal models play a crucial role in providing the initial evidence of the safety and effectiveness of prospective treatments. They guide the decision of whether these treatments have potential beneficial effects for patients and can be advanced to clinical

Correspondence to: chunyan.cai@uth.tmc.edu, Tel.: 1-713-500-7932; Fax: 1-713-500-0766.

SUPPLEMENTARY MATERIALS

Supplementary materials for "Efficient two-stage designs and proper inference for animal studies". Table S1 shows the results from additional scenarios to evaluate the type I error rate for the one-sample case using the proposed, naive, and standard approaches. Table S2 shows the results from additional scenarios to evaluate the family-wise error rate for the K -sample design using the proposed, naive, and standard approaches.

Author Manuscript

trials (Aban and George, 2015). Unfortunately, a high proportion of the therapeutic effects observed in animal studies do not translate to humans in clinical trials (Henderson et al, 2013). These unsuccessful translations expose patients to unnecessary risk and waste resources (Aguilar-Nascimento, 2005; Perrin, 2014). This issue arises from inappropriate choices of animal models, treatment parameters and conditions for clinical settings (e.g., delivery route, timing of treatment, dose levels, processing method, temperature), and poor statistical designs of animal studies (Aban and George, 2015; Hackam, 2007; Kilkenny et al, 2009; Hess, 2011; Landis et al, 2012; Macleod, 2014).

Author Manuscript

To increase the translation validity of animal research, investigators may test a large number of treatment combinations under various conditions and select the promising ones for clinical studies. However, this can require a huge number of animals, even when a small sample size (e.g., 5 to 10 animals) is used for each combination. For example, a study in mice to test 6 treatments under 10 conditions, with 8 mice per combination needs $6 \times 10 \times 8 = 480$ mice. If 5 different diseases are studied in a lab, the total required sample size can increase to $480 \times 5 = 2440$ mice. Such a large sample size is often unaffordable for investigators. Therefore, it is critically important to have efficient statistical designs for animal studies to reduce the total sample size, but at the same time, maintain the validity of the study. Even when investigators can afford such a large sample size, they can use efficient research designs to reduce their costs and direct the funds that are not used to a new treatment or another disease. Unfortunately, little work has focused on the efficient use of the available resources for animal studies.

Author Manuscript

In contrast, numerous research projects have focused on developing statistical strategies to increase the efficiency of clinical trials. The most popular strategy is the adaptive design, which allows the use of accumulating data to modify aspects of the trial in a pre-planned manner (Gallo et al, 2006; Food et al, 2010). Chow et al (2008), Berry et al (2010), and Chow and Chang (2011) provide comprehensive reviews of adaptive designs in clinical trials. The popular modifications used for clinical trials include adaptive dose finding, modification of randomization schemes, and terminating the trial early for futility or superiority with multi-stage designs or group sequential designs. Some of these adaptive strategies such as multi-stage designs can be adapted for animal studies to increase efficiency, reduce the number of false positive findings, and improve translation validity (Majid et al, 2015). For example, Lewis and Berry (1994) tested the potential applicability of Bayesian decision-theoretic designs for animal studies. Cai et al (2016) proposed a Bayesian multi-stage design to simultaneously identify the optimal dose and determine the therapeutic treatment window for animal studies in stroke research. Although some progress has been made, there is still room for improved development of efficient designs and for establishing methodological standards for animal studies.

Author Manuscript

In this paper, we aim to develop an efficient cost-effectiveness design for animal studies. Our proposed design is motivated by a study in rat models to evaluate whether stem cell transplantation provides functional recovery as measured by behavioral tests (e.g., neurologic severity score, adhesive removal test, limb placement test) after ischemic stroke. There are many factors to consider when evaluating the effect of stem cell transplantation, such as donor cell factors (cell type, safety, and auto vs. allogeneic donation), recipient

factors (stroke subtype and location), and treatment factors (acute, sub-acute, or chronic, delivery route, and cell dose) (Abe et al, 2012). Say that we consider multiple conditions determined by these factors for evaluating a new treatment in one study. The standard non-adaptive design is likely to require a lot of resources. To reduce the cost, we propose a two-stage cost-effectiveness design in order to quickly and accurately identify promising beneficial conditions, while also identifying conditions for which all further development should be stopped.

The two-stage design can efficiently utilize the available resources due to its adaptiveness, but introduces challenges for inferential procedures. The naive method that ignores the adaptive nature of the design will result in inflated type I error rates and misleading results. Accordingly, we propose an inferential procedure in terms of conditional probabilities for the proposed two-stage design. We further generalize the inferential procedure from a one-sample case to a K -sample case with a proper multiple comparison adjustment. There is a large body of literature on methods for multiple comparison adjustment. The simplest adjustment method is Bonferroni correction (Dunn, 1961), which deems the significance for each individual test at the level of α/M , where α is the desired overall significance level, and M is the number of hypotheses. However, this method is conservative and the true probability of observing a false positive effect is much lower than α , resulting in a high chance of missing true positive effects (Blakesley et al, 2009). Several less conservative and more powerful methods based on Bonferroni inequalities have been developed, including Holm's step-down procedure (Holm, 1979), Hochberg's step-up procedure (Hochberg, 1988), and Hommel's procedure (Hommel, 1988). Among these methods, Hommel's procedure is the most powerful one (Blakesley et al, 2009); therefore, we incorporate it into our K -sample design for controlling the family-wise error rate (FWER).

There are some unique features of the proposed two-stage design for animal studies compared to the multi-stage designs used in clinical studies, such as Simon's two-stage design and its extensions (Simon, 1989; Chen, 1997; Lin and Chen, 2000; Lu et al, 2005), as well as group sequential designs (Pocock, 1977; Jennison and Turnbull, 1999). First, different from Simon's two-stage design in which the design parameters need to be determined by achieving the optimization criterion, our proposed design is tailored for animal studies and is flexible in the parameter choice to allow investigators to control the stringency of the efficacy criterion in stage I. Second, our decision in stage I focuses purely on the continuation of the treatment and is quite distinct from the repeated significance test used in group sequential designs. Third, we derive the exact distribution of the test statistic given all collected data for the decision making. This is different from the conditional power approach (Lachin, 2005), which calculates the probability that the final study result will be statistically significant given the data observed thus far for the decision making at interim stages. These unique features enable our design to reflect the exploratory nature of animal studies. The remainder of this paper is organized as follows. In Section 2, we introduce our proposed two-stage design for cost-effectiveness animal studies with continuous outcomes. We derive the exact distribution of the test statistic under null hypothesis to appropriately adjust for the design's adaptiveness. We generalize the one-sample design to a K -sample design with FWER control. In Section 3, we examine the small sample behavior of the proposed design and inferential procedure through extensive simulation studies and compare

it to that of the standard design and test procedures. We provide a brief discussion in Section 4.

2 Methods

2.1 Two-stage design

Suppose we investigate a treatment with continuous responses. We denote the response of the i^{th} animal as Y_i and assume Y_i follows a normal distribution with a mean μ and variance σ^2 . Without loss of generality, we assume that a larger value of the response indicates a better performance of the treatment. The hypothesis compares the mean of the response under the treatment μ with a historical mean μ_0 , i.e., $H_0: \mu = \mu_0$ versus $H_1: \mu > \mu_0$ with a desired type I error rate of α . We propose a two-stage design. In stage I, we administer the treatment to n_1 animals. At the end of stage I, we determine whether the treatment should be advanced to stage II for further evaluation based on the efficacy criterion $\mathcal{Q}: T^{(1)} \geq c_1$, where c_1 is a calibrated boundary and $T^{(1)}$ is the summary statistic used in stage I.

If the criterion $\mathcal{Q}$ is satisfied, we start stage II and administer the treatment to another n_2 animals. At the end of the study, we determine whether the treatment is potentially beneficial or not by testing the hypothesis H_0 using data from both stages. Denoting the test statistic used at the end of the study as $T^{(2)}$, we propose to conclude that the treatment has potential efficacy, i.e., we reject H_0 , if $T^{(2)} \geq c_2$, where c_2 is the corresponding critical value by controlling the type I error rate at α .

2.2 Test and Inference

We denote the sample average of the responses in stage I as $\bar{Y}^{(1)} = \sum_{i=1}^{n_1} Y_i / n_1$ and the sample standard deviation as $S^{(1)} = \sqrt{\sum_{i=1}^{n_1} (Y_i - \bar{Y}^{(1)})^2 / (n_1 - 1)}$. Then we define the summary statistic $T^{(1)} = (\bar{Y}^{(1)} - \mu_0) / S^{(1)}$. Under H_0 , $\sqrt{n_1} T^{(1)}$ follows a student's t -distribution with $n_1 - 1$ degrees of freedom. If we allow the continuation of the treatment evaluation to stage II with a pre-specified probability $0 < \delta < 1$ under H_0 , i.e., $P_{H_0}(\mathcal{Q}) = \delta$, the efficacy criterion $\mathcal{Q}$ can be obtained as

$$\sqrt{n_1} T^{(1)} \geq F_{t_{n_1-1}}^{-1}(1 - \delta), \quad (1)$$

i.e., the boundary $c_1 = F_{t_{n_1-1}}^{-1}(1 - \delta) / \sqrt{n_1}$, where $F_{t_{\nu}}(\cdot)$ denotes the cumulative distribution

function of a student's t -distribution with ν degrees of freedom, and F^{-1} denotes the inverse of function F . The value of δ controls the stringency of the efficacy criterion $\mathcal{Q}$. The lower the value of δ , the more stringent the criterion $\mathcal{Q}$.

Suppose the treatment is selected for continuation and is administered to another n_2 animals in stage II. We denote the sample average of the responses from these n_2 animals as

$\bar{Y}^{(2)} = \sum_{i=n_1+1}^n Y_i/n_2$, where $n = n_1 + n_2$. Then the average response from all n animals is

$\bar{Y}^* = (n_1 \bar{Y}^{(1)} + n_2 \bar{Y}^{(2)})/n$ and the sample standard deviation is $S^* = \sqrt{\sum_{i=1}^n (Y_i - \bar{Y}^*)^2 / (n-1)}$.

To test H_0 at the end of the study, we consider the test statistic $T^{(2)} = (\bar{Y}^* - \mu_0)/S^*$. The naive approach would use the standard t -test with a degree of freedom $n-1$ for $\sqrt{n}T^{(2)}$. Obviously, this approach ignores the two-stage feature of the proposed design and generally leads to an inflated type I error rate.

To take account of this feature of the proposed design, we calculate the conditional probability, denoted as $\psi(t)$, of observing a value of $T^{(2)}$ as extreme as t given that the criterion $\mathcal{Q}$ is met when H_0 is true. This conditional probability is the p-value of testing H_0 at the end of the study with the observed $T^{(2)} = t$. Denoting $T^{(1*)} = (\bar{Y}^{(1)} - \mu_0)/S^*$, we rewrite $\mathcal{Q}$ as $T^{(1*)} \geq \Delta$ and obtain $\psi(t) = P_{H_0}(T^{(2)} = t | T^{(1*)} \geq \Delta)$ with $\Delta = c_1 S^{(1)}/S^*$. To calculate $\psi(t)$, we need to derive the joint distribution of $\mathbf{W} = (T^{(1*)}, T^{(2)})^T$ under H_0 , where A^T indicates the transpose of A . We represent $\mathbf{W} = U\sqrt{\nu/V}$ with $U = (U_1, U_2)^T = ((\bar{Y}^{(1)} - \mu_0)/\sigma, (\bar{Y}^* - \mu_0)/\sigma)^T$, $V = (n-1)S^{*2}/\sigma^2$ and $\nu = n-1$. It can be shown that, under H_0 , U follows a bivariate normal distribution with mean $(0, 0)^T$ and covariance matrix $\Sigma = \begin{bmatrix} 1/n_1 & 1/n \\ 1/n & 1/n \end{bmatrix}$, and V follows a chi-squared distribution with ν degrees of freedom, which is independent with U . Through the representation of multivariate t -distribution (Lin, 1972), under H_0 , $\mathbf{W}$ follows a multivariate t -distribution with parameters Σ, θ, ν , and has the following density function

$$f(\mathbf{w}) = \frac{\Gamma[(\nu+2)/2]}{\Gamma(\nu+2)\nu\pi^{1/2} |\Sigma|^{1/2}} \left[1 + \frac{1}{\nu}(\mathbf{w} - \theta)^T \Sigma^{-1}(\mathbf{w} - \theta)\right]^{-(\nu+2)/2}, \quad (2)$$

where $\theta = (0, 0)^T$, and Γ is the gamma function. Accordingly, the p-value can be derived as

$$\begin{aligned} \psi(t) &= P_{H_0}(T^{(2)} \geq t | T^{(1*)} \geq \Delta) = P_{H_0}(T^{(1*)} \geq \Delta, T^{(2)} \geq t) / \delta \quad (3) \\ &= \int_t^\infty \int_\Delta^\infty f(w_1, w_2) dw_1 dw_2 / \delta, \end{aligned}$$

where $f(w_1, w_2)$ is the probability density function in equation (2) with $\mathbf{w} = (w_1, w_2)^T$. At the significance level of α , we reject H_0 if $\psi(t) \leq \alpha$ and obtain the critical value $c_2 = \psi^{-1}(\alpha)$.

2.3 Determination of sample size

We assume σ is known in the sample size calculation. Under H_0 , $\sqrt{n_1}(U_1, U_2)^T$ follows a bivariate normal distribution with mean $(0, 0)^T$ and covariance matrix $\Sigma_0 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, where $\rho = 1/(1+r)$ and $r = n_2/n_1$. We denote the survival function of this bivariate normal distribution

as $S(\cdot, \cdot)$. At the end of stage I, we continue the evaluation if $\sqrt{n_1}U_1 > Z_\delta$ where Z_δ is the upper δ th quantile of the standard normal distribution. At the end of stage II, we reject H_0 if $\sqrt{n_1}U_2 > c$. To control the type I error rate at α , we have

$$\begin{aligned}\alpha &= P(\text{Reject } H_0 \mid H_0) \\ &= \frac{P(\sqrt{n_1}U_1 > Z_\delta, \sqrt{n_1}U_2 > c \mid H_0)}{P(\sqrt{n_1}U_1 > Z_\delta \mid H_0)} = S(Z_\delta, c)/\delta\end{aligned}\quad (4)$$

Under the alternative hypothesis $H_1: \mu = \mu_1, \sqrt{n_1}(U_1 - u^*, U_2 - u^*)^T$ follows a bivariate normal distribution with mean $(0, 0)$ and covariance matrix Σ_0 , where $u^* = (\mu_1 - \mu_0)/\sigma$. To achieve the desired power of $(1 - \beta)100\%$, we have

$$\begin{aligned}1 - \beta &= P(\text{Reject } H_0 \mid H_1) \\ &= \frac{P[\sqrt{n_1}(U_1 - u^*) > (Z_\delta - \sqrt{n_1}u^*), \sqrt{n_1}(U_2 - u^*) > (c - \sqrt{n_1}u^*) \mid H_1]}{P[\sqrt{n_1}(U_1 - u^*) > (Z_\delta - \sqrt{n_1}u^*) \mid H_1]} \\ &= S(Z_\delta - \sqrt{n_1}u^*, c - \sqrt{n_1}u^*)/[1 - \Phi(Z_\delta - \sqrt{n_1}u^*)],\end{aligned}\quad (5)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. By jointly solving the two equations, we can identify n_1 and then determine the required sample size n under the specifications.

2.4 Extension to a K -sample design

In animal studies, it is increasingly common to evaluate a treatment under varying conditions within one study. Suppose there are K different conditions to be tested and we assign n_1 animals to each of K conditions in stage I. Let Y_{ik} , $k = 1, \dots, K$, denote the response of the i^{th} animal under condition k . We assume Y_{ik} follows a normal distribution with a mean μ_k and variance σ_k^2 . Considering the hypothesis $H_{0k}: \mu_k = \mu_{0k}$ versus $H_{1k}: \mu_k > \mu_{0k}$ with μ_{0k} as the historical mean for condition k , we determine to continue the evaluation of condition k in stage II if the summary statistic in stage I, $T_k^{(1)}$, satisfies the efficacy criterion $\mathcal{Q}_k: T_k^{(1)} \geq c_{1k}$, where c_{1k} is a calibrated boundary for condition k in stage I.

Denoting the average response under condition k collected in stage I as $\bar{Y}_k^{(1)} = \sum_{i=1}^{n_1} Y_{ik}/n_1$ and the sample standard deviation as $S_k^{(1)} = \sqrt{\sum_{i=1}^{n_1} (Y_{ik} - \bar{Y}_k^{(1)})^2 / (n_1 - 1)}$, we define $T_k^{(1)} = (\bar{Y}_k^{(1)} - \mu_{0k})/S_k^{(1)}$. Under H_{0k} , $\sqrt{n_1}T_k^{(1)}$ follows a student's t -distribution with $n_1 - 1$

degrees of freedom. If we allow the continuation of condition k to stage II at a pre-specified probability $0 < \delta_k < 1$ under H_0 , the corresponding criterion $\mathcal{Q}_k$ can be derived as

$$\sqrt{n_1} T_k^{(1)} \geq F_{t_{n_1-1}}^{-1}(1 - \delta_k), \text{ i.e., the boundary for condition } k \text{ in stage I is}$$

$$c_{1k} = F_{t_{n_1-1}}^{-1}(1 - \delta_k) / \sqrt{n_1}. \text{ If we set the same cut-offs for } K \text{ conditions, i.e., } \delta_1 = \dots = \delta_K = \delta,$$

the boundaries for K conditions in stage I would be the same, i.e.,

$$c_{11} = \dots = c_{1K} = c_1 = F_{t_{n_1-1}}^{-1}(1 - \delta) / \sqrt{n_1}.$$

Suppose M conditions are successfully carried forward to stage II and denote a set containing all indexes of these M conditions as A . For condition m , $m \in A$, we assign another n_2 animals to the test conditions. Then, at the end of the study, we determine whether condition m has potential efficacy or not by testing the hypothesis H_{0m} using data collected at both stages. Denoting the test statistic for condition m at the end of the study as $T_m^{(2)}$, we conclude potential efficacy for condition m , i.e., rejecting H_{0m} if $T_m^{(2)} \geq c_{2m}$, where c_{2m} is the calibrated critical value for condition m when controlling the type I error rate at α_m .

Denoting the average response of condition m collected in stage II as

$$\bar{Y}_m^{(2)} = \sum_{i=n_1+1}^n Y_{im} / n_2, \text{ the average response of condition } m \text{ from all } n \text{ animals can be}$$

derived as $\bar{Y}_m^* = (n_1 \bar{Y}_m^{(1)} + n_2 \bar{Y}_m^{(2)}) / n$, and its sample standard deviation is

$$S_m^* = \sqrt{\sum_{i=1}^n (Y_i - \bar{Y}_m^*)^2 / (n - 1)}. \text{ We define the test statistic at the end of the study, i.e.,}$$

$$T_m^{(2)} = (\bar{Y}_m^* - \mu_{0m}) / S_m^*. \text{ The naive approach using the standard } t\text{-test with } n - 1 \text{ degrees of}$$

freedom for $\sqrt{n} T_m^{(2)}$ is problematic and generally inflates the type I error rate. To make

correct inference, we compute the p-value of testing H_{0m} denoted as $\psi_m(t)$, as the conditional probability of observing a value of $T_m^{(2)}$ as extreme as t given $\mathcal{Q}_m$ is met.

Denoting $T_m^{(1*)} = (\bar{Y}_m^{(1)} - \mu_{0m}) / S_m^*$, we rewrite $\mathcal{Q}_m$ as $T_m^{(1*)} \geq \Delta_m$ and obtain

$$\psi_m(t) = P_{H_{0m}}(T_m^{(2)} \geq t \mid T_m^{(1*)} \geq \Delta_m) \text{ with } \Delta_m = c_1 S_m^{(1)} / S_m^*. \text{ Similarly, we show that}$$

$\mathbf{W}_m = (T_m^{(1*)}, T_m^{(2)})^T$ follows a multivariate t -distribution with parameters Σ , θ , and ν and has a density function in equation (2). Then, we can derive the p-value $\psi_m(t)$ as

$$\begin{aligned} \psi_m(t) &= P_{H_{0m}}(T_m^{(2)} \geq t \mid T_m^{(1*)} \geq \Delta_m) = P_{H_{0m}}(T_m^{(1*)} \geq \Delta_m, T_m^{(2)} \geq t) / \delta \quad (6) \\ &= \int_t^\infty \int_{\Delta_m}^\infty f(w_1, w_2) dw_1 dw_2 / \delta, \end{aligned}$$

where $f(w_1, w_2)$ is the probability density function of a multivariate t -distribution in equation (2). We reject H_{0m} at the α_m significance level if $\psi_m(t) > \alpha_m$ and obtain the critical value $c_{2m} = \psi_m^{-1}(\alpha_m)$.

If we tested M hypotheses simultaneously with the significance level α_m for each hypothesis at 0.05, the probability of falsely rejecting at least one true null hypothesis, i.e., the FWER, would be dramatically inflated. Instead, to control the type I error rate per comparison at 0.05, we apply Hommel's procedure to control the FWER q at 0.05 for testing M hypotheses. We obtain the p-value p_m for condition m by plugging in the observed values of $T_m^{(2)}$ for the realization of t in equation (6). After obtaining these M p-values, we rank them and denote the rank-ordered p-values as $p_{(1)}, p_{(2)}, \dots, p_{(M)}$. We find

$$m^* = \max_{m \in \{1, \dots, M\}} \{p_{(M-m+j)} > jq/m \text{ for } j = 1, \dots, m\}. \quad (7)$$

Here, m^* is the largest value of $m \in \{1, \dots, M\}$ such that the condition $p_{(M-m+j)} > jq/m$ is satisfied for all the values of j with $j = 1, \dots, m$. If m^* does not exist, we reject all M hypotheses; otherwise, we reject H_{0m} if $p_m < \alpha_m$ with $\alpha_m = q/m^*, m \in A$.

3 Simulation Study

3.1 Simulation Setting

We conduct simulation studies to evaluate the small sample performance of the proposed two-stage design and the inferential procedure. We also evaluate the performance of a naive approach, which uses the same two-stage design but applies a naive test that ignores the unique features of the design. Specifically, this naive approach, described in Section 2.2, uses the standard t -test with a degree of freedom $n - 1$ for $\sqrt{n}T^{(2)}$ to test H_0 at the end of the study. We also consider a standard approach under which $n = n_1 + n_2$ animals are assigned to the treatment and standard t -tests are used for testing. Although the standard approach is not efficient, it provides a reference standard against which to evaluate the performance of our proposed approach.

We first evaluate the performance of our proposed design and the inferential procedure under the one-sample case, and compare it to that of the naive and standard approaches. For the proposed and naive approaches, we consider $n = 8$, with three different settings: 1) a balanced design with $(n_1 = 4, n_2 = 4)$, 2) an unbalanced design with $(n_1 = 3, n_2 = 5)$, and 3) an unbalanced design with $(n_1 = 2, n_2 = 6)$, for stages I and II, respectively. We choose different values of the parameter δ in the efficacy criterion $\mathcal{Q}$, such as 0.3, 0.4, 0.5, and 0.6. For the standard approach, we assign $n = 8$ animals to the treatment. We study the type I error rates by these three approaches under the null hypothesis. Without loss of generality, we set $\mu = \mu_0 = 0$ and $\sigma = 1$. We simulate 50,000 replicates and evaluate the rejection rates and the average sample sizes. The simulation results are shown in Table 1. Then we assess the statistical power under the alternative scenarios using the proposed and standard approaches. We consider two standardized effect sizes: 1) effect size of 1 with $\mu = 1, \mu_0 = 0$

and $\sigma = 1$, and 2) effect size of $2/3$ with $\mu = 1$, $\mu_0 = 0$ and $\sigma = 1.5$. We simulate 50,000 replicates and list the simulation results in Table 2.

Next we evaluate the FWER control for the K -sample design. FWER is calculated as the proportion of at least one false rejection among all hypotheses testings. Without loss of generality, we set $\mu_k = \mu_{0k} = 0$ and $\sigma_k = 1$ for condition k , $k = 1, \dots, K$. Similarly, for the standard approach, we assign $n = 8$ animals to the treatment, and for the proposed and naive approaches, we consider 1) a balanced design with $(n_1 = 4, n_2 = 4)$, 2) an unbalanced design with $(n_1 = 3, n_2 = 5)$, and 3) an unbalanced design with $(n_1 = 2, n_2 = 6)$. We choose δ at 0.3, 0.4, 0.5, and 0.6. We simulate 50,000 replicates with $K = 5$ and $K = 8$. The results of FWER and the average sample size under these three approaches are shown in Table 3. In addition, we assess the statistical power under the alternative scenarios using the proposed and standard approaches for the K -sample case. We consider $K = 8$ and assume only two conditions are efficacious with $\mu_1 = 1.5$ and $\mu_2 = 1.3$ compared to their historical means $\mu_{10} = \mu_{20} = 0$. The rejection rates of the null hypothesis for these two conditions and the average total sample size are summarized in Table 4.

3.2 Simulation Results

For the one-sample case, as shown in Table 1, both the proposed and standard approaches control the type I error rate very well; the rejection rates are all in a reasonable range around 5%. With a balanced sample size in two stages, the rejection rates of the proposed approach under H_0 are 5.3%, 5.1%, 5.1%, 5.0% for $\delta = 0.3, 0.4, 0.5, 0.6$, respectively. When the sample size of the two stages is unbalanced, its rejection rates under H_0 range from 5.1% to 5.5% for these scenarios. Compared to the standard approach, our proposed approach utilizes much fewer animals for all scenarios. For example, when $\delta = 0.3$, our proposed approach that uses a balanced sample size needs only 5.2 average animals vs. the 8 animals required by the standard approach, and saves 35% of resources in terms of the sample size. With unbalanced settings of $(n_1 = 3, n_2 = 5)$ and $(n_1 = 2, n_2 = 6)$, our proposed approach further saves around 44% and 53% of resources, respectively, compared to the standard approach. When the value of δ increases, the efficacy criterion in stage I becomes less stringent, which results in a relatively higher proportion of the simulated studies continuing to stage II. Therefore, the average sample size utilized by the proposed design increases when δ increases, but is still smaller than that of the standard design. The naive approach uses the same two-stage design and utilizes the same number of animals as the proposed approach; however, under H_0 , the rejection rates of the naive approach range from 8.0% to 15.1% under the balanced setting and from 7.6% to 13.7% under the unbalanced setting, which are much higher than the pre-specified significance level of 5%. These high rejection rates are not surprising because the naive approach ignores the interim look and the fact that the data from stage II are only observed when the efficacy criterion is satisfied by the data from stage I.

We note that in the scenarios presented in Table 1, the proposed design has slightly inflated type I error rates, especially when the sample size of stage I is small, e.g., $n_1 = 2$ or 3. We expand the simulation studies by including scenarios with larger sample sizes ($n=10, 12, 14$) and summarize the results in Table S1 in the supplementary materials. The results show that

under these scenarios, our proposed design controls the type I error rate very well (4.8% to 5.2%), and saves 20%–35% of the sample size. Not surprisingly, the naive approach dramatically inflates the type I error rate under these scenarios (8.0%–15.1%).

Regarding the statistical power under alternative scenarios, Table 2 shows that the standard approach reaches 81.1% power when the effect size is 1 and reaches 51.7% power when the effect size is $2/3$. As expected, our proposed approach is less powerful than the standard design, but utilizes smaller sample sizes. When δ increases from 0.3 to 0.6 with the unbalanced setting of $(n_1 = 2, n_2 = 6)$, the power of our proposed approach increases from 71.4% to 76.2% when the effect size is 1, and increases from 43.5% to 47.0% when the effect size is $2/3$. The simulation results show that with the unbalanced settings of $(n_1 = 3, n_2 = 5)$ and $(n_1 = 2, n_2 = 6)$, our proposed approach has slightly higher power and utilizes fewer animals than those under the balanced setting. Overall, these results demonstrate that our proposed design saves resources, especially when $n_1 < n_2$, but due to adaptation, loses some degree of power compared to the standard design.

When extending to the K -sample case, we evaluate the performance of FWER control under the null hypothesis for these three approaches. The standard approach controls the FWER at 5.0% with both settings of $K = 5$ and $K = 8$. Our proposed approach also controls the FWER well, with rates ranging from 5.2% to 5.4% with the balanced setting and from 5.3% to 5.7% with the unbalanced setting. These results show that our proposed approach is capable of controlling the FWER well. In terms of resources saved, our proposed approach again outperforms the standard approach. When $K = 5$ and δ increases from 0.3 to 0.6, our proposed approach saves 20% to 35% of resources with the balanced setting, saves 25% to 44% of resources with the unbalanced setting of $(n_1 = 3, n_2 = 5)$, and saves 30% to 53% of resources with the unbalanced setting of $(n_1 = 2, n_2 = 6)$. We observe similar findings when $K = 8$. Similar to the one-sample case, the naive approach performs worse in terms of the FWER control and dramatically inflates the FWER, which ranges from 8.0% to 16.2% in the investigated scenarios. These results suggest that, compared to the standard and naive approaches, our proposed K -sample design not only performs well regarding control of the FWER, but also saves resources in terms of the sample size. Compared to the pre-specified significance level of 5%, the FEWR in the K -sample design seems to be slightly inflated. We conduct additional simulations with larger sample sizes and display the results in Table S2 in the supplementary materials. The results show that under these scenarios, our proposed design controls the FWER very well (4.9% to 5.2%), and saves 20%–35% of the sample size. Under these scenarios, the naive approach inflates the FWER to 8.0%–15.6%.

Regarding the statistical power of the K -sample case under alternative scenarios, Table 4 shows that the standard approach reaches 80.0% power for condition 1 and 66.5% power for condition 2. Overall, our proposed approach has power that is similar to that of the standard approach for both conditions. Under the balanced setting, the statistical power of the proposed design ranges from 76.5% to 78.8% for condition 1 and 64.3% to 65.7% for condition 2. In addition, our design tends to terminate the other six inefficacious conditions and saves considerable resources overall compared to the standard approach. Under the balanced setting, our proposed design saves 15% to 26% of the resources. Under the unbalanced settings of $(n_1 = 3, n_2 = 5)$ and $(n_1 = 2, n_2 = 6)$, our proposed approach has more

savings in sample size. When more conditions to be evaluated are nonefficacious, we expect that our proposed design will be more efficient in terms of resource saving.

Our simulation studies also suggest that when the treatment does not have a beneficial effect, the higher the value of δ , the less savings of resources when n_1 and n_2 are fixed. For example, with $n_1 = 3$ and $n_2 = 5$, the percentage of sample size savings reduces from 44% to 25% when δ increases from 0.3 to 0.6. On the other hand, the choice of δ does not have an obvious impact on the type I error control (see Table 1). However, when the treatment has a beneficial effect, with increasing δ , we gain more power when fixing other parameters. For example, when $n_1 = 3$, $n_2 = 5$, and the effect size is 1, the power increases from 67.2% to 74.2% when δ increases from 0.3 to 0.6. Therefore, we recommend using simulation studies to evaluate the performance of our proposed design under the recommended range of 0.3 to 0.6 for δ and then selecting a value with a good balance between sample size savings and robust operating characteristics. For the K -sample design, we recommend using the same value of δ for each condition.

4 Conclusion

We propose a two-stage cost-effectiveness design for animal studies. Given the data from two stages, we derive the exact distribution of the test statistic under null hypothesis to appropriately adjust for the design's adaptiveness. Our design is similar in spirit to multi-stage designs or group sequential designs for clinical trials, but incorporates some unique features. For example, unlike Simon's two-stage design in which the design parameters need to be determined by achieving some optimization criterion, our design is flexible in stage I by allowing the investigators to choose the value of δ to control the stringency of the efficacy criterion and adjust the sample size allocation in each stage. Compared to the repeated significance tests in group sequential designs, our efficacy criterion in stage I mainly focuses on the decision of treatment continuation. These unique features enable our design to reflect the exploratory nature of animal studies.

The simulation studies confirm that, in terms of the sample size, our proposed two-stage design is more efficient than the standard one-stage design, especially when the experimental condition/treatment does not have a promising effect. As noted, the proposed two-stage design may lose some degree of power due to adaptation when the experimental condition/treatment does have a beneficial effect. However, due to the exploratory nature of animal research, investigators may test a large number of treatment combinations under various conditions. Under this scenario, the gain in resource savings may outweigh the power loss. Taking the study discussed in the introduction as an example, the investigators evaluate 6 treatments under 10 conditions, resulting in a total of 60 evaluations. Suppose 80% of the investigated conditions do not have a beneficial treatment effect, and 20% of the investigated conditions have a beneficial treatment effect (effect size=1). If we conduct the proposed design ($n_1 = 3$, $n_2 = 5$ and set $\delta = 0.5$), the average sample size is around $5.5 \times 60 \times 0.8 + 7.8 \times 60 \times 0.2 \approx 358$ mice by using the simulation results in 1 and 2. If the standard design with $n = 8$ is applied, the sample size is $8 \times 60 = 480$ mice, 34% larger than the sample size needed by the two-stage design.

For the choice of a balanced or unbalanced setting for sample size allocation, the results suggest that the unbalanced setting has slightly higher power and utilizes fewer animals than the balanced setting. Therefore, we recommend studying the performance of various sample size settings through simulation studies and selecting a setting with good performance in terms of both sample size and operating characteristics. Our design can be extended to multiple stages, with choices of sample sizes as 3+3+3 or 2+2+2+2, etc. The total sample size can be determined by achieving adequate statistical power when detecting the anticipated effect and the variance of the data.

Motivated by the exploratory nature of animal studies, the main purpose of the proposed design is to identify promising beneficial conditions/treatments for further development. The proposed design can be applied not only to evaluate all or a subset of combinations determined by multiple factors (e.g., the example in the introduction), but also to more general settings, such as the evaluation of a series of parallel conditions. It allows the investigators to evaluate any condition of interest at any time, without the need to wait for data from other conditions for decision making. This is an important feature for preclinical studies in which the investigators may try different conditions and graduate any promising ones to the next stage at any time. When investigators want to compare a series of conditions determined by multiple factors and the objective is to screen out the most promising ones among the combinations, we may consider a factorial design (Whitehead, 1997), which is an efficient way to detect any main effects and potential interactions. This paper focuses on the scenario with continuous outcomes. Given the small sample size in animal studies, the normal approximation for the binomial distribution is inappropriate. Determining how to extend the proposed design using appropriate inference for animal studies with binary outcomes is beyond the scope of this paper and is worthy of future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the editor, the associate editor and two reviewers for their constructive comments that have greatly improved the initial version of this paper. The work was supported in part by the U.S. National Institutes of Health Grants UL1 TR000371, CA193878 and CA016672.

References

- Aban IB, George B. Statistical considerations for preclinical studies. *Exp Neurol*. 2015; 270:82–87. [PubMed: 25725352]
- Abe K, Yamashita T, Takizawa S, Kuroda S, Kinouchi H, Kawahara N. Stem cell therapy for cerebral ischemia: from basic science to clinical applications. *Journal of Cerebral Blood Flow & Metabolism*. 2012; 32(7):1317–1331. [PubMed: 22252239]
- Aguilar-Nascimento JEd. Fundamental steps in experimental design for animal studies. *Acta Cirúrgica Brasileira*. 2005; 20(1):2–3.
- Berry SM, Carlin BP, Lee JJ, Muller P. Bayesian adaptive methods for clinical trials. London: Chapman & Hall; 2010.

- Blakesley RE, Mazumdar S, Dew MA, Houck PR, Tang G, Reynolds CF III, Butters MA. Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology*. 2009; 23(2):255. [PubMed: 19254098]
- Cai C, Ning J, Huang X. A bayesian multi-stage cost-effectiveness design for animal studies in stroke research. *Statistical Methods in Medical Research*. 2016 0962280216657853.
- Chen TT. Optimal three-stage designs for phase ii cancer clinical trials. *Statistics in medicine*. 1997; 16(23):2701–2711. [PubMed: 9421870]
- Chow SC, Chang M. *Adaptive design methods in clinical trials*. London: Chapman & Hall; 2011.
- Chow SC, Chang M, et al. Adaptive design methods in clinical trials-a review. *Orphanet J Rare Dis*. 2008; 3(11):169–90.
- Dunn OJ. Multiple comparisons among means. *Journal of the American Statistical Association*. 1961; 56(293):52–64.
- Food, Administration D, et al. *Guidance for industry: adaptive design clinical trials for drugs and biologics*. Washington DC, USA: Food and Drug Administration; 2010.
- Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive designs in clinical drug developmentan executive summary of the phrma working group. *Journal of biopharmaceutical statistics*. 2006; 16(3):275–283. [PubMed: 16724485]
- Hackam DG. Translating animal research into clinical benefit. *BMJ*. 2007; 7586:163.
- Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. *PLoS Med*. 2013; 10(7):e1001, 489.
- Hess KR. Statistical design considerations in animal studies published recently in cancer research. *Cancer research*. 2011; 71(2):625–625. [PubMed: 21239476]
- Hochberg Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*. 1988; 75(4):800–802.
- Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*. 1979:65–70.
- Hommel G. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*. 1988; 75(2):383–386.
- Jennison C, Turnbull BW. *Group sequential methods with applications to clinical trials*. CRC Press; 1999.
- Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, Hutton J, Altman DG. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PloS one*. 2009; 4(11):e7824. [PubMed: 19956596]
- Lachin JM. A review of methods for futility stopping based on conditional power. *Statistics in medicine*. 2005; 24(18):2747–2764. [PubMed: 16134130]
- Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*. 2012; 490(7419):187–191. [PubMed: 23060188]
- Lewis RJ, Berry DA. Group sequential clinical trials: a classical evaluation of bayesian decision-theoretic designs. *Journal of the American Statistical Association*. 1994; 89(428):1528–1534.
- Lin PE. Some characterizations of the multivariate t distribution. *Journal of Multivariate Analysis*. 1972; 2(3):339–344.
- Lin SP, Chen TT. Optimal two-stage designs for phase ii clinical trials with differentiation of complete and partial responses. *Communications in Statistics-Theory and Methods*. 2000; 29(5–6):923–940.
- Lu Y, Jin H, Lamborn KR. A design of phase ii cancer trials using total and complete response endpoints. *Statistics in medicine*. 2005; 24(20):3155–3170. [PubMed: 16189806]
- Macleod MR. Preclinical research: Design animal studies better. *Nature*. 2014; 510(7503):35–35.
- Majid A, Bae ON, Redgrave J, Teare D, Ali A, Zemke D. The potential of adaptive design in animal studies. *International journal of molecular sciences*. 2015; 16(10):24,048–24,058.
- Perrin S. Preclinical research: Make mouse studies work. *Nature*. 2014; 507(7493):423–425. [PubMed: 24678540]

- Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977; 64(2):191–199.
- Simon R. Optimal two-stage designs for phase ii clinical trials. *Controlled clinical trials*. 1989; 10(1): 1–10. [PubMed: 2702835]
- Whitehead J. The design and analysis of sequential clinical trials. John Wiley & Sons; 1997.

Table 1

Type I error rate under the null hypothesis for the one-sample case using the proposed, naive, and standard approaches with 50,000 simulations

Proportion of rejection (Standard deviation)						Average sample size (Standard deviation)					
μ	μ_0	n_1	n_2	σ	δ	Standard	Proposed	Naive	Standard	Proposed	Naive
Balanced design											
0	0	4	4	1	0.3	4.8 (21.4)	5.3 (22.5)	15.1 (35.8)	8	5.2 (1.8)	5.2 (1.8)
0	0	4	4	1	0.4	4.8 (21.4)	5.1 (22.1)	11.9 (32.4)	8	5.6 (2.0)	5.6 (2.0)
0	0	4	4	1	0.5	4.8 (21.4)	5.1 (21.9)	9.6 (29.5)	8	6.0 (2.0)	6.0 (2.0)
0	0	4	4	1	0.6	4.8 (21.4)	5.0 (21.9)	8.0 (21.2)	8	6.4 (2.0)	6.4 (2.0)
Unbalanced design											
0	0	3	5	1	0.3	4.8 (21.4)	5.4 (22.5)	13.7 (34.4)	8	4.5 (2.3)	4.5 (2.3)
0	0	3	5	1	0.4	4.8 (21.4)	5.3 (22.4)	11.3 (31.7)	8	5.0 (2.4)	5.0 (2.4)
0	0	3	5	1	0.5	4.8 (21.4)	5.2 (22.3)	9.4 (29.2)	8	5.5 (2.5)	5.5 (2.5)
0	0	3	5	1	0.6	4.8 (21.4)	5.1 (22.0)	7.9 (26.9)	8	6.0 (2.4)	6.0 (2.4)
0	0	2	6	1	0.3	4.8 (21.4)	5.4 (22.6)	11.7 (32.2)	8	3.8 (2.7)	3.8 (2.7)
0	0	2	6	1	0.4	4.8 (21.4)	5.5 (22.7)	10.2 (30.3)	8	4.4 (2.9)	4.4 (2.9)
0	0	2	6	1	0.5	4.8 (21.4)	5.4 (22.5)	8.8 (28.3)	8	5.0 (3.0)	5.0 (3.0)
0	0	2	6	1	0.6	4.8 (21.4)	5.2 (22.3)	7.6 (26.5)	8	5.6 (2.9)	5.6 (2.9)

Statistical power under the alternative hypothesis for the one-sample case using the proposed and standard approaches with 50,000 simulations

Table 2

Proportion of rejection (Standard deviation)						Average sample size (Standard deviation)			
μ	μ_0	n_1	n_2	σ	δ	Standard	Proposed	Standard	Proposed
Balanced design									
1	0	4	4	1	0.3	81.1 (39.2)	63.1 (48.3)	8	7.7 (1.1)
1	0	4	4	1	0.4	81.1 (39.2)	67.0 (47.0)	8	7.8 (0.8)
1	0	4	4	1	0.5	81.1 (39.2)	70.1 (45.8)	8	7.9 (0.6)
1	0	4	4	1	0.6	81.1 (39.2)	72.9 (44.5)	8	8.0 (0.4)
1	0	4	4	1.5	0.3	51.7 (50.0)	36.6 (48.2)	8	7.1 (1.7)
1	0	4	4	1.5	0.4	51.7 (50.0)	38.8 (48.7)	8	7.4 (1.4)
1	0	4	4	1.5	0.5	51.7 (50.0)	41.0 (49.2)	8	7.6 (1.2)
1	0	4	4	1.5	0.6	51.7 (50.0)	43.3 (49.5)	8	7.8 (0.9)
Unbalanced design									
1	0	3	5	1	0.3	81.1 (39.2)	67.2 (46.9)	8	7.4 (1.7)
1	0	3	5	1	0.4	81.1 (39.2)	69.8 (45.9)	8	7.6 (1.3)
1	0	3	5	1	0.5	81.1 (39.2)	72.1 (44.9)	8	7.8 (1.0)
1	0	3	5	1	0.6	81.1 (39.2)	74.2 (43.7)	8	7.9 (0.8)
1	0	3	5	1.5	0.3	51.7 (50.0)	40.3 (49.0)	8	6.6 (2.2)
1	0	3	5	1.5	0.4	51.7 (50.0)	41.6 (49.3)	8	7.1 (2.0)
1	0	3	5	1.5	0.5	51.7 (50.0)	43.2 (49.5)	8	7.4 (1.7)
1	0	3	5	1.5	0.6	51.7 (50.0)	44.9 (49.7)	8	7.6 (1.4)
1	0	2	6	1	0.3	81.1 (39.2)	71.4 (45.2)	8	6.7 (2.5)
1	0	2	6	1	0.4	81.1 (39.2)	73.1 (44.3)	8	7.2 (2.0)
1	0	2	6	1	0.5	81.1 (39.2)	74.5 (43.6)	8	7.5 (1.6)
1	0	2	6	1	0.6	81.1 (39.2)	76.2 (42.6)	8	7.7 (1.3)
1	0	2	6	1.5	0.3	51.7 (50.0)	43.5 (49.6)	8	5.8 (2.9)
1	0	2	6	1.5	0.4	51.7 (50.0)	44.7 (49.7)	8	6.5 (2.6)
1	0	2	6	1.5	0.5	51.7 (50.0)	46.0 (49.8)	8	7.0 (2.3)
1	0	2	6	1.5	0.6	51.7 (50.0)	47.0 (50.0)	8	7.3 (1.9)

Table 3

Family-wise error rate for the K-sample case using the proposed, naive, and standard approaches with 50,000 simulations

Family-wise error rate (%)(Standard deviation)					Average sample size (Standard deviation)				
K	n ₁	n ₂	δ	Standard	Proposed	Naive	Standard	Proposed	Naive
Balanced design									
5	4	4	0.3	5.0 (21.8)	5.4 (22.6)	15.9 (36.6)	40	26.0 (4.1)	26.0 (4.1)
5	4	4	0.4	5.0 (21.8)	5.3 (22.4)	12.3 (32.8)	40	28.0 (4.4)	28.0 (4.4)
5	4	4	0.5	5.0 (21.8)	5.2 (22.2)	9.9 (29.8)	40	30.0 (4.5)	30.0 (4.5)
5	4	4	0.6	5.0 (21.8)	5.2 (22.2)	8.3 (27.6)	40	32.0 (4.4)	32.0 (4.4)
8	4	4	0.3	5.0 (21.8)	5.4 (22.5)	16.2 (36.8)	64	41.6 (5.2)	41.6 (5.2)
8	4	4	0.4	5.0 (21.8)	5.3 (22.4)	12.4 (33.0)	64	44.8 (5.5)	44.8 (5.5)
8	4	4	0.5	5.0 (21.8)	5.2 (22.2)	10.0 (30.0)	64	48.0 (5.7)	48.0 (5.7)
8	4	4	0.6	5.0 (21.8)	5.2 (22.1)	8.4 (27.7)	64	51.2 (5.5)	51.2 (5.5)
Unbalanced design									
5	3	5	0.3	5.0 (21.8)	5.6 (23.0)	14.8 (35.5)	40	22.5 (5.1)	22.5 (5.1)
5	3	5	0.4	5.0 (21.8)	5.5 (22.8)	12.0 (32.5)	40	25.0 (5.5)	25.0 (5.5)
5	3	5	0.5	5.0 (21.8)	5.4 (22.6)	9.8 (29.8)	40	27.5 (5.6)	27.5 (5.6)
5	3	5	0.6	5.0 (21.8)	5.3 (22.3)	8.3 (27.5)	40	30.0 (5.5)	30.0 (5.5)
8	3	5	0.3	5.0 (21.8)	5.6 (23.0)	15.4 (36.1)	64	36.0 (6.5)	36.0 (6.5)
8	3	5	0.4	5.0 (21.8)	5.5 (22.8)	12.2 (32.8)	64	40.0 (6.9)	40.0 (6.9)
8	3	5	0.5	5.0 (21.8)	5.4 (22.7)	10.0 (30.0)	64	44.0 (7.1)	44.0 (7.1)
8	3	5	0.6	5.0 (21.8)	5.3 (22.4)	8.3 (27.5)	64	48.0 (6.9)	48.0 (6.9)
5	2	6	0.3	5.0 (21.8)	5.6 (23.0)	12.7 (33.3)	40	19.0 (6.2)	19.0 (6.2)
5	2	6	0.4	5.0 (21.8)	5.6 (23.1)	11.0 (31.2)	40	22.0 (6.6)	22.0 (6.6)
5	2	6	0.5	5.0 (21.8)	5.6 (23.0)	9.4 (29.1)	40	25.0 (6.7)	25.0 (6.7)
5	2	6	0.6	5.0 (21.8)	5.5 (22.8)	8.0 (27.2)	40	28.0 (6.6)	28.0 (6.6)
8	2	6	0.3	5.0 (21.8)	5.7 (23.1)	13.4 (34.1)	64	30.4 (7.8)	30.4 (7.8)
8	2	6	0.4	5.0 (21.8)	5.7 (23.1)	11.4 (31.8)	64	35.2 (8.4)	35.2 (8.4)
8	2	6	0.5	5.0 (21.8)	5.7 (23.1)	9.7 (29.6)	64	39.9 (8.5)	39.9 (8.5)
8	2	6	0.6	5.0 (21.8)	5.5 (22.9)	8.3 (27.5)	64	44.7 (8.3)	44.7 (8.3)

Statistical power under the alternative hypothesis for the K-sample case using the proposed and standard approaches with 50,000 simulations

Table 4

Proportion of rejection (Standard deviation)					Average total sample size (Standard deviation)			
n_1	n_2	δ	Standard		Proposed		Standard	Proposed
			Condition 1	Condition 2	Condition 1	Condition 2		
4	4	0.3	80.0 (40.0)	66.5 (47.2)	Balanced design		64	47.1 (4.5)
	4	0.4	80.0 (40.0)	66.5 (47.2)	76.5 (42.4)	64.3 (47.9)	64	49.5 (4.8)
	4	0.5	80.0 (40.0)	66.5 (47.2)	77.5 (41.8)	64.9 (47.7)	64	52.0 (4.9)
	4	0.6	80.0 (40.0)	66.5 (47.2)	78.3 (41.2)	65.4 (47.6)	64	54.4 (4.8)
3	5	0.3	80.0 (40.0)	66.5 (47.2)	Unbalanced design		64	42.6 (5.8)
	3	0.4	80.0 (40.0)	66.5 (47.2)	78.8 (40.9)	65.7 (47.5)	64	45.8 (6.1)
	3	0.5	80.0 (40.0)	66.5 (47.2)	78.9 (40.8)	67.4 (46.9)	64	48.9 (6.2)
	3	0.6	80.0 (40.0)	66.5 (47.2)	79.1 (40.7)	66.8 (47.1)	64	52.0 (6.0)
2	6	0.3	80.0 (40.0)	66.5 (47.2)	79.1 (40.6)	66.6 (47.2)	64	37.5 (7.2)
	2	0.4	80.0 (40.0)	66.5 (47.2)	79.3 (40.5)	66.4 (47.2)	64	41.8 (7.4)
	2	0.5	80.0 (40.0)	66.5 (47.2)	82.9 (37.7)	72.3 (44.8)	64	45.7 (7.5)
	2	0.6	80.0 (40.0)	66.5 (47.2)	81.7 (38.7)	70.2 (45.8)	64	49.4 (7.3)