Published in final edited form as:

Stat Biopharm Res. 2014 January 2; 6(1): 16–29. doi:10.1080/19466315.2013.826596.

Statistical Methods for Selecting Maximum Effective Dose and Evaluating Treatment Effect When Dose – Response is Monotonic

Maiying Kong [Associate Professor],

Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA

Shesh N. Rai [Professor], and

Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA

Roberto Bolli, M.D. [Professor]

Institute of Molecular Cardiology and the Executive Vice Chair in the Department of Medicine, University of Louisville, Louisville, KY 40202, USA

Maiying Kong: maiying.kong@louisville.edu; Shesh N. Rai: Shesh.Rai@Louisville.edu; Roberto Bolli: rbolli@louisville.edu

Abstract

The maximum effective dose (MaxED) is an important quantity for therapeutic drugs. The MaxED for therapeutic drugs is defined as the dose above which no improvement in efficacy is obtained. In this article, we propose two experimental designs and analytic methods (one single-stage design and one two-stage design) to select the MaxED among several fixed doses and to compare the therapeutic effect of the selected MaxED with a control. The selection of MaxED is based on the isotonic regression under the restriction of monotonicity. In the single-stage design, both the selection of the MaxED and assessing its efficacy are carried out at the end of experiment. In the two-stage design, the selection of the MaxED and assessment of its efficacy are carried out at the interim analysis (first stage), the experiment in the second stage is carried out only at the selected MaxED and control if the first-stage test is not significant. Thus, the two-stage design enables selection of the MaxED at an earlier stage and stopping the trial earlier if the treatment effect at MaxED is extreme. Williams' test (1972) is applied to test whether the selected MaxED is significantly different from control for the single-stage design and the first-stage test of the two-stage design. The sample size calculation for each design is provided. Extensive simulations are carried out to illustrate the performances of the proposed methods.

Keywords

O'Brien and Fleming's repeated significance test; Two-stage design; Williams' test

1. Introduction

The maximum effective dose (MaxED) for therapeutic drugs is defined as the dose “beyond which additional benefit would be unlikely to occur” (ICH E9 1998). In contrast to the MaxED, the minimum effective dose (MinED) is defined as the lowest dose whose effect is significantly different from control (Williams 1971; Ruberg 1995a,b; Ting 2006). Many

statistical methods have been developed for finding a MinED (Ruberg 1989, 1995a,b; Tamhane, Hochberg, Dunnett 1996; Bauer 1997; Bretz, Pinheiro, and Branson 2005; Bretz, Dette, and Pinheiro 2009). However, if a study aims to find treatments with high efficacy, the MaxED is of great interest (Remmenga et al. 1997; Strassburger, Bretz and Finner 2007; Ivanova, Bolognese, and Perevozskaya 2008). A MaxED could be considered as the lowest dose whose effect is within a pre-specified minimum clinically important difference from the effect of the highest dose within the tested dose range. The tested dose range for the MaxED should be within the maximum tolerated dose (MTD), which is the highest dose with toxicity below a certain level (Korn et al. 1994; Garrett-Mayer 2006). With the assumption that the tested dose range is a known prior, the focus of the article is selection of the MaxED from several fixed doses and to test whether the effect of the selected MaxED is significantly different from control.

The current study has been motivated by the ongoing NIH-funded CAESAR study (<http://www.nihcaesar.org/>). The major goal of the CAESAR study is to provide an infrastructure that enables rigorous preclinical evaluation of promising cardioprotective therapies. The CAESAR consortium calls for study designs that enable selection of the MaxED among several fixed doses and to compare the therapeutic effect of the selected MaxED with a control. As recommended by ICH E9 (1998), the hypothesis tests that are used in dose-response studies may need to be tailored to the natural ordering of doses or to particular questions regarding the shape of the dose-response curve (e.g., monotonicity). The current work is developed under the assumption of monotonicity. To be specific, let us assume that R fixed dose levels (e.g., $d_1 < d_2 < \dots < d_R$) and a control group ($d_0 = 0$) are studied in the dose-response study, with μ_i as the mean of the response at dose level d_i ($i=0, 1, \dots, R$). The monotonicity of the form $\mu_0 \leq \mu_1 \leq \dots \leq \mu_R$ is added to the means of the responses if the magnitude of the response increases as the dose level increases. Otherwise, the monotonicity of the form $\mu_0 \geq \mu_1 \geq \dots \geq \mu_R$ is added if the magnitude of the response variable decreases as the dose level increases. For example, studies in subjects with acute myocardial infarction have shown that the size of the infarct is an important prognostic factor, and that reducing infarct size is of paramount importance to alleviate morbidity and mortality (Pfeffer and Braunwald 1990). In the case that the response variable is infarct size, a smaller value of the response variable indicates a stronger therapeutic effect; thus, a constraint of the form $\mu_0 \geq \mu_1 \geq \dots \geq \mu_R$ is appropriate.

Based on our literature survey, there is little work on estimating the MaxED under the minimal assumption that the dose-response is monotonic. Under the constraint that $\mu_0 \leq \mu_1 \leq \dots \leq \mu_R$, Ivanova et al. (2008) suggested that the MaxED be estimated as the lowest dose with mean response at $\mu_R + \delta$ using an adaptive design, where μ_R is the mean effect at the highest dose level and δ is the minimum clinically important difference in response specified by the clinical team. Remmenga et al. (1997) conducted a simulation study to compare several regression model-based procedures for estimating the MaxED. However, when the number of dose levels studied is fixed and small, say $R = 4$, and the waiting time for the outcome is long, the usage of parametric approaches and adaptive designs for estimating the MaxED seems limited. In this article, we propose to estimate the MaxED based on the maximum likelihood estimate (MLE) of μ_i 's ($i = 1, \dots, R$) under the minimal assumption that the

means of the response variable are monotonic as the dose increases (Barlow et al. 1972). In the case that the response variable is normally distributed, the monotonic MLE of the μ_i 's ($i = 1, \dots, R$), and the test statistic and critical values for testing whether the response at dose level d_i is significantly different from control have been developed by Williams (1971, 1972). In Section 2 of this article, we propose to select the MaxED as the lowest dose whose monotonic MLE is within a minimum clinically important difference from the monotonic MLE at the highest dose level, and in Section 3.1, we apply the Williams' procedure (1972) to compare the effect of the selected MaxED with control.

The procedure based on Williams' (1972) is a single-stage procedure, where selecting the MaxED and testing its effectiveness are carried out at the end of the experiment. Motivated by group sequential design (O'Brien and Fleming 1979; Jennison and Turnbull 2000) and multiarmed clinical trials (Follmann, Proschan, and Geller 1994; Proschan, Follmann, and Geller 1994), in Section 3.2, we propose a two-stage design to select the MaxED at an earlier stage (say, stage 1) and stop the trial earlier if the treatment effect at the selected MaxED is extremely effective. The experiment at stage 2 is carried out if the stage 1 test is not significant at a prespecified significance level, and the experiment at stage 2 is carried out only at the selected MaxED and control. With the two-stage design, it is expected that the number of subjects can be reduced while the accuracy to select the MaxED and the power to detect a prespecified effect size can be maintained. The power and sample size analysis for each design method is provided. Section 4 describes extensive simulations carried out to examine whether the proposed designs control the experiment-wise error rate at α and maintain the power at least at $1-\beta$ to detect a specified effect, as well as to examine whether the two-stage designs require less subjects than the single-stage design. Section 5 describes simulations carried out to compare the proposed approaches in this paper with the parametric approach in terms of selecting the MaxED and assessing their efficacy. The last section is devoted to a discussion.

2. Procedure for Selecting MaxED

To select the MaxED, let us assume that the number of subjects assigned to each one of the $R + 1$ groups is the same, denoted by n . Let denote $y_{i1}, y_{i2}, \dots, y_{in}$ as the observations of the n subjects from i th treatment group, where $i = 0, 1, \dots, R$. In this article, we assume that y_{ij} has normal distribution with mean μ_i and a common variance σ^2 . Without any restriction on the mean μ_i 's, the MLE for μ_i is

$$\bar{y}_i = \frac{1}{n} \sum_{j=1}^n y_{ij} \quad (i=0, 1, \dots, R).$$

However, it is generally expected that the therapeutic effect becomes stronger as the dose level becomes larger. Without loss of generality, let assume $\mu_0 \leq \mu_1 \leq \dots \leq \mu_R$. The general theory for the estimation under order restriction is known as isotonic regression (Barlow et al. 1972). The monotonic MLE of μ_i ($i = 1, \dots, R$); that is, the estimator of μ_i under the order restriction, is given as (Barlow et al. 1972; Williams 1972):

$$\hat{\mu}_i = \min_{1 \leq u \leq i} \max_{i \leq v \leq R} \sum_{j=u}^v \bar{y}_j / (v - u + 1). \quad (1)$$

In case that the number of observations for the R dose levels varies, say n_i observations in i th group ($i = 1, \dots, R$), the monotonic MLE of μ_i ($i = 1, \dots, R$) is given by

$$\hat{\mu}_i = \min_{1 \leq u \leq i} \max_{i \leq v \leq R} \sum_{j=u}^v n_j \bar{y}_j / \sum_{j=u}^v n_j. \quad (2)$$

We propose to estimate the MaxED as the lowest dose whose estimated effect is within the minimum clinically important difference, say δ , from $\hat{\mu}_R$, the monotonic MLE of the effect at the highest dose. That is,

$$d_m = \min\{d_i : \hat{\mu}_i \leq \hat{\mu}_R + \delta, i = 1, \dots, R\}. \quad (3)$$

The monotonic MLE in (1) and (2) could be implemented by the “pool-adjacent-violators” algorithm (PAVA) (Ayer et al. 1955). For example, if the sample means for four dose levels are (40, 35, 25, 30), then the monotonic MLE ($\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4$) will be (40, 35, 27.5, 27.5), where the last two adjacent violators are pooled. Based on (3), the estimated MaxED is d_3 as long as $\delta < 7.5$. In the case that the constraint for the means is $\mu_0 \leq \mu_1 \leq \dots \leq \mu_R$, the monotonic MLE of μ_i ($i = 1, \dots, R$) is given by

$\hat{\mu}_i = \max_{1 \leq u \leq i} \min_{i \leq v \leq R} \sum_{j=u}^v n_j \bar{y}_j / \sum_{j=u}^v n_j$ (Williams 1972), and the MaxED could be estimated by the lowest dose whose mean effect is within δ from $\hat{\mu}_R$.

It should be noticed that the selection of MaxED does not depend on a test of statistical significance. Upon the MaxED being selected, the decision on whether the selected MaxED is significantly different from control is made based on the Williams’ test (1972) detailed in the following sections. The test statistics here are constructed under the assumption that the dose–response curve is decreasing, that is, $\mu_0 \geq \mu_1 \geq \dots \geq \mu_R$, which is consistent with our motivating example on designing experiments to test cardioprotective therapies with the infarct size as the primary endpoint. In the cases that the monotonicity of $\mu_0 \geq \mu_1 \geq \dots \geq \mu_R$ is desired, the two terms of the numerator in each following test statistic should have reverse order. All the test procedures stay exactly the same.

3. Experimental Designs and Analytic Methods

3.1 Single-Stage Design Based on Williams’ Test

The single-stage design based on Williams’ test requires randomly assigning n subjects to each one of the $R + 1$ groups. At the end of the experiment, the monotonic MLE of μ_i ($i = 1, \dots, R$), say $\hat{\mu}_i$, is given by (1) or (2), the MaxED is chosen as the lowest dose (say, d_m) whose estimated effect is within the size of δ from $\hat{\mu}_R$, and the significance test for

comparing the effect at the MaxED and control is carried out by the following t_j ($j = m, \dots, R$) statistic (Williams 1972):

$$\bar{t}_j = \frac{\bar{y}_0 - \hat{\mu}_j}{\sqrt{2\text{MSE}/n}}. \quad (4)$$

Here $\text{MSE} = \frac{1}{(n-1)(R+1)} \sum_{i=0}^R \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2$, which is the estimated variance for σ^2 . Let us denote $\text{df} = (n-1)(R+1)$. Note that $\text{df} \times \text{MSE} / \sigma^2$ follows a chi-squared distribution with degrees of freedom df , and is independent of $\bar{y}_i$ ($i = 0, 1, \dots, R$). The test statistic (4) is identical to the Student's t -statistic for comparing treatment at dose d_j with control except that $\hat{\mu}_j$ replaces $\bar{y}_j$ ($j = m, \dots, R$). However, under the null hypothesis that $\mu_0 = \mu_1 = \dots = \mu_R$, the distribution of the test statistic t_j ($j = m, \dots, R$) is different from the central t -distribution. Williams (1972) tabulated the critical values for the distribution of the test statistic t_j for α at 0.005, 0.01, 0.025, and 0.05, j between 2 and 10, and different degrees of freedom df . Let denote the critical values by $\bar{t}_{j,\alpha}(\text{df})$. Based on Williams' approach (1972), the effect at dose level d_m is significantly different from control at significance level α only if $t_j > \bar{t}_{j,\alpha}(\text{df})$ ($j = R, R-1, \dots, m$). That is, if $t_j > \bar{t}_{j,\alpha}(\text{df})$ for some j between m and R , then the effect at d_m is not significantly different from control.

The power of a test statistic is defined as the probability that the null hypothesis is rejected when the alternative hypothesis is correct. Here the null hypothesis is that the means in all groups are the same, which implies that the effect at the MaxED is the same as control, and the alternative hypothesis is that the effect at the MaxED is different from control, say $\mu_0 - \mu_m = \Delta > 0$. Under the alternative hypothesis, the test statistic can be approximated by a t -distribution (Chow, Shao, and Wang 2008). That is,

$$\bar{t}_m = \frac{\bar{y}_0 - \hat{\mu}_m}{\sqrt{2\text{MSE}/n}} \approx \frac{\bar{y}_0 - \bar{y}_m}{\sqrt{2\text{MSE}/n}}, \quad (5)$$

where the last term in (5), under the alternative hypothesis, follows a noncentral t -

distribution with degree of freedom $(n-1)(R+1)$ and a noncentrality parameter $\frac{\mu_0 - \mu_m}{\sigma \sqrt{2/n}}$

denoted by $t(\frac{\mu_0 - \mu_m}{\sigma \sqrt{2/n}}, (n-1)(R+1))$. Therefore, the power can be approximated by the expression

$$\begin{aligned} \text{Power} &= \text{Prob} \left(\frac{\bar{y}_0 - \hat{\mu}_m}{\sqrt{2\text{MSE}/n}} > \bar{t}_{m,\alpha}(\text{df}) \mid \mu_m = \mu_0 - \Delta \right) \\ &\approx \text{Prob} \left(\frac{\bar{y}_0 - \bar{y}_m}{\sqrt{2\text{MSE}/n}} > \bar{t}_{m,\alpha}(\text{df}) \mid \mu_m = \mu_0 - \Delta \right) \\ &= \text{Prob} \left(T > \bar{t}_{m,\alpha}(\text{df}) \mid T \sim t \left(\frac{\mu_0 - \mu_m}{\sigma \sqrt{2/n}}, (\text{df}) \right) \right) \\ &\geq \text{Prob} \left(T > \bar{t}_{R,\alpha}(\text{df}) \mid T \sim t \left(\frac{\mu_0 - \mu_m}{\sigma \sqrt{2/n}}, (\text{df}) \right) \right), \end{aligned} \quad (6)$$

where $\text{df} = (n-1)(R+1)$. The last inequality holds since $\bar{t}_{m,\alpha}(\text{df}) < \bar{t}_{R,\alpha}(\text{df})$. To detect the difference Δ at a significance level α with power $1 - \beta$, the sample size required is obtained

as the smallest integer n such that the last quantity in (6) is greater than $1 - \beta$. The total sample size required for the single-stage design is $n(R + 1)$.

The power and sample size analysis can be carried out in any statistical software package, such as R (<http://www.r-project.org/>) and SAS (<http://www.sas.com/>), using an iterative method. We carried out the power and sample size analysis in R-package. The initial starting value for n can be obtained by using the function *power.t.test*, a function for power and sample size analysis for t -test in the R package. Since the critical value for t -statistic is usually smaller than $t_{R,\alpha}(\text{df})$, the actual power calculated based on *power.t.test* is generally less than $1 - \beta$. We increase n until the actual power based on (6) is at least $1 - \beta$. The R code is available from the first author.

3.2 Two-Stage Design Based on O'Brien and Fleming's Procedure

In the literature, the statistical method for monitoring multiarmed trials, which “drops the loser,” has been proposed under the framework that the multiarms are parallel (Proschan et al. 1994). It seems logical that the total sample size could be decreased if we carry out an interim analysis and drop out the less effective doses and the dose beyond the MaxED. That is, instead of studying n subjects at each dose level, we may study $n/2$ subjects in each dose level first, then select the MaxED and assess its efficacy. Additional $n/2$ subjects would be studied only at the control group and the selected MaxED. The total subjects applied would be $n(R + 1)/2 + n$. Thus, the total sample size may be smaller than single-stage design. Without loss of generality, let us assume the number of subjects at each dose level in the first stage is n_1 . The mean at control group and the monotonic MLE of μ_i ($i = 1, \dots, n_1$) are obtained, the MaxED (d_m) is selected as the lowest dose such that $\hat{\mu}_m$ is within δ from $\hat{\mu}_R$, and the significance test comparing the selected MaxED and control is carried out by the Williams' t_m statistic at significance level α_1 ; in stage 2, n_1 additional subjects per group at the MaxED and control are studied, and the stage 2 significance test is carried out by the following test statistic:

$$T_m^* = \frac{\bar{y}_0^* - \bar{y}_m^*}{\sqrt{2\text{MSE}^*/(n_1 + n_1)}} = \frac{\bar{y}_0^* - \bar{y}_m^*}{\sqrt{\text{MSE}^*/n_1}}, \quad (7)$$

where $\bar{y}_0^*$ and $\bar{y}_m^*$ are, respectively, the means of the $2n_1$ observations in control group and the MaxED group in the entire experiment, that is,

$$\bar{y}_0^* = \frac{1}{2n_1} \sum_{j=1}^{2n_1} y_{0j}$$

and

$$\bar{y}_m^* = \frac{1}{2n_1} \sum_{j=1}^{2n_1} y_{mj},$$

and MSE^* is the mean squares of errors of the two groups in the entire experiment, that is,

$$\text{MSE}^* = \frac{\sum_{j=1}^{2n_1} (y_{0j} - \bar{y}_0^*)^2 + \sum_{j=1}^{2n_1} (y_{mj} - \bar{y}_m^*)^2}{4n_1 - 2}.$$

The experiment in stage 2 is carried out only if the stage 1 significance test indicates that the MaxED is not significantly different from control at significance level α_1 ; that is, $t_j - t_j, \alpha_1$ (df) for some j between m and R . Although the test statistic T_m^* in (7) is formally identical to the Student's t -statistic for comparing treatment effect at dose d_m with control, the distributions are quite different. Since d_m , the selected MaxED, is the lowest dose with its monotonic MLE within δ from $\hat{\mu}_R$, and $\hat{\mu}_1 \dots \hat{\mu}_R$, the sample mean observed at dose d_m at stage 1 is smaller than the sample means at the lower dose groups no matter whether the null hypothesis $\mu_0 = \mu_1 = \dots = \mu_R$ is correct or not. The stage 1 observations at dose d_m and control are applied to construct stage 2 test statistic T_m^* . Thus, under the null hypothesis, the distribution of T_m^* , defined as

$$\frac{\bar{y}_0^* - \bar{y}_m^*}{\sqrt{\text{MSE}^*/n_1}},$$

is larger than the central t -distribution with degree of freedom $4n_1 - 2$. The larger is R (number of doses involved), the larger differentiation of the distribution of T_m^* from the central t -distribution. The exact distribution of the test statistic (7) under the null hypothesis is difficult to derive. To examine the distribution of T_m^* , we calculated 10,000 T_m^* based on 10,000 random samples under the null hypothesis for each combination of R and n_1 . The simulation procedure is given in Section 4.1, and the plot of the density function of T_m^* for each one of the six combinations of R ($=2, 3, 4$) and n_1 ($=5, 10$) is shown in Figure 1 as a solid line in each panel, while the density function of a central t -distribution with a degree of freedom $4n_1 - 2$ is shown as a dashed line in each panel. The vertical line with $T_m^* = 0$ in Figure 1 helps to visualize the property of T_m^* -distribution. From Figure 1, it is clear that the distribution of T_m^* is shifted to the right of the t -distribution, and the shift becomes larger as the number of studied doses R gets larger. The upper tail critical points of T_m^* for different combinations of R ($=2, 3, 4$, and 6) and n_1 , denoted by $t_{\alpha}(n_1, R)$, are tabulated in Table 1, where each combination is based on 100,000 simulation runs. The upper tail critical points tabulated in Table 1 are associated with the two-stage experimental design given below. Each experimental design controls the experiment-wise error rate at α and has a power at least $1 - \beta$ to detect a minimum clinical effect size at the MaxED.

O'Brien and Fleming (1979) proposed a procedure for repeated significance tests that controls the experiment-wise Type I error rate at α . We apply this procedure to the two-stage design, where we consider the selection of the MaxED and the stage 1 significance test as one interim analysis. Let us assume $\alpha = 0.05$. For O'Brien and Fleming's procedure, the significance level for the stage 1 test (i.e., test at the interim analysis) is $2(1 - 2\Phi(1.977\sqrt{2})) = 2(-2\Phi(2.796)) = 0.005$, and the significance level for the stage 2 test (i.e., the test at the final analysis) is $2(1 - 2\Phi(1.977)) = 0.048$ (Jennison and Turnbull, 2000).

The associated null hypothesis is that the effect at MaxED is the same as control. We reject the null hypothesis at stage 1 if the Williams' test statistic t_j is larger than $t_{j,\alpha}(\text{df})$ for $j = R, R-1, \dots, m$, where $\alpha_1 = 0.005$ and $\text{df} = (n_1-1)(R+1)$. The critical values $t_{j,\alpha}(\text{df})$ can be obtained from Table 4 given by Williams (1972). If the null hypothesis is not rejected at stage 1, the trial is continued to stage 2. At the final analysis, the null hypothesis is rejected if $T_m^* \geq \bar{t}_{\alpha_2}^*(n_1, R)$, where $\alpha_2 = 0.048$, and the critical values $\bar{t}_{\alpha_2}^*(n_1, R)$ are tabulated in Table 1.

The power and sample size analysis is carried out in a similar manner as given by Jennison and Turnbull (2000, sec. 3.8), but with critical value $\bar{t}_{\alpha}^*(n_1, R)$ instead of $t_{\alpha}(4n_1 - 2)$, where t_{α} is the upper α point of the central t -distribution with degrees of freedom $4n_1 - 2$. The sample size n_1 required per group is obtained as the smallest integer such that the power defined in the following expression is at least $1 - \beta$:

$$\text{Power} = \text{Prob} \left(T \geq \bar{t}_{\alpha}^*(n_1, R) \mid T \sim t \left(\frac{\mu_0 - \mu_m}{\sigma \sqrt{R_B(2, \alpha, \beta)/n_1}}, 4n_1 - 2 \right) \right), \quad (8)$$

where $R_B(2, \alpha, \beta)$ is the ratio of the maximum sample size to the fixed sample size based on O'Brien and Fleming's group sequential test (Jennison and Turnbull 2000). In particular, when $\alpha = 0.05$ and $\beta = 0.2$, $R_B(2, \alpha, \beta) = 1.008$. For fixed α , $R_B(2, \alpha, \beta)$ may vary as β varies. However, the change is small; for example, $R_B(2, 0.05, \beta)$ varies from 1.008 to 1.007 when β varies from 0.2 to 0.1 (Jennison and Turnbull 2000). In the sample size calculation, we fix $R_B(2, \alpha, \beta)$ with the target α and β .

It should be noticed that the power in (8) is an approximation of the total power, where the sample size is adjusted from the fixed sample size (Jennison and Turnbull 2000). The total power is the rejection rate in both first stage (say $1 - \beta_1$) and second stage, while the conditional power is the rejection rate at second stage provided that the null hypothesis is not rejected at the first stage (say $1 - \beta_2$). The total power, $(1 - \beta_1) + \beta_1(1 - \beta_2) (= 1 - \beta_1\beta_2)$, is always larger than the conditional power (i.e., $1 - \beta_2$), as long as $\beta_1 < 1$. All the power and sample size analyses for the proposed experiment designs are carried out in the R package. The R code is available from the first author. In Section 4.1, we performed simulations to examine the distribution of the stage 2 test statistic T_m^* and to tabulate the critical values required for the stage 2 significance tests and sample size calculation. In Section 4.2, we carried out extensive simulations to examine the performances of the two design methods in terms of power, sample size, and the probability of certain dose being selected as the MaxED.

4. Simulations to Characterize the Proposed Methods

4.1 Simulations for the Distribution of T_m^*

The critical values for testing whether the effect at the selected MaxED is significantly different from control at stage 2 are determined by the distribution of T_m^* under the null hypothesis that the means of the $R + 1$ groups are the same as μ . Recall that T_m^* is defined as

$$\frac{\bar{y}_0^* - \bar{y}_m^*}{\sqrt{\text{MSE}^*/n_1}},$$

where $\bar{y}_0^*$ and $\bar{y}_m^*$ are, respectively, the sample means of the $2n_1$ observations at the control and selected MaxED in the entire experiment. Note that

$$T_m^* = \frac{\bar{y}_0^* - \bar{y}_m^*}{\sqrt{\text{MSE}^*/n_1}} = \frac{\frac{1}{2n_1} \sum_{j=1}^{2n_1} y_{0j} - \frac{1}{2n_1} \sum_{j=1}^{2n_1} y_{mj}}{\sqrt{\text{MSE}^*/n_1}} = \frac{\frac{1}{2n_1} \sum_{j=1}^{2n_1} \left(\frac{y_{0j} - \mu}{\sigma} \right) - \frac{1}{2n_1} \sum_{j=1}^{2n_1} \left(\frac{y_{mj} - \mu}{\sigma} \right)}{\sqrt{\text{MSE}^*/\sigma^2 n_1}}, \quad (9)$$

where $(4n_1 - 2)\text{MSE}^*/\sigma^2$ is a central χ^2 -distribution with degree of freedom $4n_1 - 2$, which is independent of the numerator in (9) (Ravishanker and Dey 2002). Therefore, the distribution of T_m^* stays the same as that obtained from $\mu = 0$ and a common variance $\sigma^2 = 1$.

To examine the distribution of T_m^* under the null hypothesis that $\mu_0 = \mu_1 = \dots = \mu_R = \mu$ for different combinations of n_1 and R , we generated n_1 observations for each one of the $R + 1$ groups with $y_{ij} \sim N(0, 1)$ for $i = 0, 1, \dots, R$ and $j = 1, \dots, n_1$, obtained the monotonic MLE of μ_i ($i = 1, \dots, R$), selected the MaxED as $d_m = \min\{d_i : \hat{\mu}_i - \hat{\mu}_m + \delta, i = 1, \dots, R\}$ with $\delta = 0.1\sigma = 0.1$, generated additional n_1 observations for control group as well as for the selected MaxED, and calculated T_m^* . These steps were repeated 10,000 times to generate the probability density plot for each one of the combinations of R ($=2, 3$, and 4) and n_1 ($=5, 10$). These probability density plots are shown as solid lines in the six panels in Figure 1, along with dashed lines, which are the probability density plots of the central t -distribution with degree of freedom $4n_1 - 2$. From Figure 1, it is clear that the distribution of T_m^* (solid line) is shifted to the right of the central t -distribution (dashed line) with a degree of freedom $4n_1 - 2$, and the shift becomes greater when R gets larger. We repeated 100,000 times to generate the critical values for the upper- α ($=0.048, 0.05$) tail points based on the realization of the 100,000 T_m^* values under each combination of R ($=2, 3, 4$, and 6) and n_1 ($=3$ to 100), which are tabulated in Table 1. The upper 0.048 and 0.05 tail points are used, respectively, for the stage 2 significance test and the power and sample size analysis of the two-stage design.

4.2 Simulations to Examine the Properties of the Proposed Methods

The simulation scenarios are based on our current study on preclinical assessment of cardioprotective therapies (CAESAR), where we propose to test two dose levels for large animals (pigs) and three or four dose levels for small animals (mice and rabbits) (<http://www.nihcaesar.org/>). Based on the preliminary studies of cardioprotective therapies performed by Dr. Bolli's lab (R. Bolli, personal communication) in mice and rabbits, the mean infarct size in the control group is usually about 50% of the risk region with a standard deviation of around 15%. A treatment is considered effective if it decreases infarct size by 20%, which is the minimum clinically significant effect size. We calculate the sample size for each design method with the specification of $\mu_0 = 50$, $\sigma_2 = 15_2$, and the minimal clinical effect at 15. In the simulations, we consider at least three scenarios for each fixed R (i.e., the number of doses to be studied). The simulation results are reported in Tables 2–5 for R being 2, 3, 4, and 6, respectively. For each fixed R , we calculated the sample size required for each

design method to detect the minimum clinical effect at 20 with power at least 80% (i.e., $\beta = 0.2$) and significance level at 0.05 (i.e., $\alpha = 0.05$). The sample sizes required for single-stage design (say n) and two-stage design (n_1 at each stage) is reported under the column titles “ n stage 1,” “ n stage 2” in Tables 2–5. The maximum sample size required for each design is reported in the same tables under the column title “ $N_{\max}$,” which is $n_1(R + 1) + 2n_1$ for the two-stage design and $n(R + 1)$ for the single-stage design. For each fixed R , we examine (i) whether each design method has controlled the experiment-wise error rate by setting the underlying means all equal (i.e., $\mu_0 = \mu_1 = \dots = \mu_R = 50$), (ii) whether the power for each method will be smaller than 80% when the maximum difference (i.e., $\mu_0 - \mu_R$) is less than the minimum clinical effect by setting the underlying means as $50 = \mu_0 \quad \mu_1 \quad \dots \quad \mu_R = 40$, (iii) whether each design method has correct power by setting the underlying means as $50 = \mu_0 \quad \mu_1 \quad \dots \quad \mu_R = 30$; (iv) whether the power for each method will increase when the maximum difference is larger than the minimum clinical effect by setting the underlying means as $50 = \mu_0 \quad \mu_1 \quad \dots \quad \mu_R = 20$. For each fixed R , under each setting for the means ($\mu_0, \mu_1, \dots, \mu_R$), and for each design method, we generated 10,000 simulation runs. Each simulation run is carried out by the following steps:

Step 1. Generate n or n_1 random numbers (n for single stage design and n_1 for the two-stage design) from $N(\mu_i, \sigma^2)$ for the i th dose level ($i = 0, 1, \dots, R$);

Step 2. Obtain the monotonic MLE $\hat{\mu}_i$ of μ_i ($i = 1, \dots, R$), and select the MaxED d_m as the lowest dose whose monotonic MLE is within the size of δ from $\hat{\mu}_R$, where δ is set as 0.1σ , which is 1.5 in this simulation setting;

Step 3. Calculate the test statistic t_R and perform the Williams’s significance test to examine whether the effect at the MaxED is significantly different from control;

Step 4. Go to next run if the single-stage design is applied or the significance test at Step 3 is significant; otherwise, go to next run after stage 2 is performed: where n_1 random numbers from $N(\mu_m, \sigma^2)$ for the selected MaxED d_m and n_1 random numbers from $N(\mu_0, \sigma^2)$ for control are generated, and the test statistic T_m^* is constructed and decision on whether the two groups are the same or different is made.

Among 10,000 runs, we calculated (i) the fraction of dose d_i ($i = 1, \dots, R$) being selected as the MaxED (see the column “% select” under each μ_i in Tables 2–5); (ii) the fraction of the rejection among the 10,000 runs (see the columns “% total reject” in Tables 2–5); (iii) the average sample size among the 10,000 simulation runs (see the column “ N_{ave} ” in Tables 2–5). It should be noted that the average sample size in the single-stage design is the same as the maximum sample size since no earlier stop is applied. For the two-stage designs, particularly, when the maximum effect is large, the stage 1 significance test may show that the MaxED is significantly different from control. In that case, the subjects for stage 2 are saved, and the total sample size will be $n_1(R + 1)$ instead of $n_1(R + 1) + 2n_1$. Thus, the average sample size among the 10,000 simulations will be smaller than the maximum sample size. Based on the simulation results (Tables 2–4), we conclude that (i) the Type I error $\alpha = 0.05$ is controlled for the single-stage as well as the two-stage experimental designs (see the columns “% total reject” under the specification $\mu_0 = \mu_1 = \dots = \mu_R = 50$): the overall fractions of rejection for the single-stage design (see rows “Single-stage”) and the two-stage design (see rows “Two-stage”) are around 0.05; (ii) when the maximum effect (μ_0

– μ_R) is less than the minimum clinical effect 20, the power for the two methods are less than 80% (see the columns “% total reject” under the specification $50 = \mu_0 = \mu_1 = \dots = \mu_R = 40$); (iii) the 80% power has been maintained for all design methods (see the columns “% total reject” under the specification $50 = \mu_0 = \mu_1 = \dots = \mu_R = 30$); (iv) when the maximum effect ($\mu_0 - \mu_R$) is larger than the minimum clinical effect 20, the power for the two methods are greater than 80% (see the columns “% total reject” under the specification $50 = \mu_0 = \mu_1 = \dots = \mu_R = 20$).

Since the stage 2 test statistic has the identical form as the Student’s t -statistic, it is worthwhile to examine what happens if stage 2 significance test is based on the central t -distribution instead of the distribution of T_m^* statistics. We carried out the simulations described in Steps 1–4, except that the stage 2 significance test in Step 4 is based on the central t -distribution with degree of freedom $4n_1 - 2$ instead. The sample sizes were calculated based on the expressions (8) with the critical value $\bar{t}_\alpha^*(n, R)$ replaced by $t_\alpha(4n - 2)$ in (8). The results are reported in Tables 2–5 with the row title “T-test” for the two-stage design. Based on the results, it is clearly that Type I error rate is not controlled at $\alpha = 0.05$ (see the column “% total reject” under the specification $\mu_0 = \mu_1 = \dots = \mu_R = 50$ with row title “T-test”). As R goes from 2 to 6, the actual Type I error rate goes from 0.0796 to 0.1051 for the two-stage design. As we illustrated from Figure 1, the probability density function of T_m^* shifts to the right of the central t -distribution. The critical values based on the central t -distribution are smaller than the critical values based on the distribution of T_m^* under the null hypothesis that all groups have the same mean, thus resulting in more rejections of the null hypothesis if the stage 2 significance test is based on the central t -distribution. Therefore, the stage 2 significance test using critical values based on the central t -distribution is not recommended.

From Tables 2–4, the maximum sample sizes required by the two-stage design are close to the ones required by the single-stage design for $R = 2$ and 3, and smaller than the ones required by the single-stage design for $R = 4$ and 6. The average sample sizes required by the two-stage design are generally smaller than required by the single-stage design. To examine whether these observations are generally true, we calculated the sample size required per group for each design method to detect an effect size at significance level $\alpha = 0.05$ and with power at least 80% (i.e., $\beta = 0.2$). The effect size, say γ , is defined as

$\frac{\mu_0 - \mu_m}{\sigma}$. Panels A1–A3 in Figure 2 show the maximum sample sizes required for each design as the effect size varies from 0.5 to 2. From panels A1–A4, the maximum sample sizes required by the two-stage design are smaller than required by the single-stage design for the effect size less than 1 or for larger number of dose level R (say $R = 4$ and 6), but the maximum sample sizes required by the two-stage design are close to and even slightly larger than the ones required by single-stage design for large effect size (say > 1.0) and small R (say $R = 2$ and 3). The general results from Figure 2 are consistent with the simulation results presented in Tables 2–5, where the effect size is 1.33.

5. Simulations to Compare the Proposed Approaches and the Modeling Approach

The proposed single-stage and two-stage approaches are based on the isotonic regression, where the dose-response is assumed to be monotonic. Studies have shown that the estimate of the effective dose based on the isotonic regression is robust to misspecification of the dose response curves (Bhattacharya and Kong 2006). To compare the proposed approach with the modeling approach in terms of selecting MaxED and testing its efficacy, we considered trial designs with six dose levels, say (1, 2, 4, 8, 16, 32), with the response profile generated from the $E_{\max}$ model:

$$E = E_0 - \frac{E_{\max} \left(\frac{d}{ED_{50}} \right)^m}{1 + \left(\frac{d}{ED_{50}} \right)^m}.$$

Here E_0 is the mean of the response at control, $E_{\max}$ is the maximum reduction compared with control, ED_{50} is the dose achieving 50% $E_{\max}$ reduction in response, and m is the slope of the $E_{\max}$ model. We carried out simulations under the following four response profiles for $(\mu_0, \mu_1, \dots, \mu_6)$:

Truth R6-1: $\mu_0 = \mu_1 = \dots = \mu_6 = 50$;

Truth R6-2: $(\mu_0, \mu_1, \dots, \mu_6) = (50, 46, 40, 34, 31, 30, 30)$, which are based on the $E_{\max}$ model with $E_0 = 50$, $m = 2$, $E_{\max} = 20$, and $ED_{50} = 2$;

Truth R6-3: $(\mu_0, \mu_1, \dots, \mu_6) = (50, 47, 41, 31, 24, 21, 20)$, which are based on the $E_{\max}$ model with $E_0 = 50$, $m = 2$, $E_{\max} = 30$, and $ED_{50} = 3$;

Truth R6-4: $(\mu_0, \mu_1, \dots, \mu_6) = (50, 44, 35, 26, 22, 20, 20)$, which are based on the $E_{\max}$ model with $E_0 = 50$, $m = 2$, $E_{\max} = 30$ and $ED_{50} = 2$.

For each response profile, we consider two modeling approaches: (i) the $E_{\max}$ model, where E_0 and m are fixed at 50 and 2, respectively, and $E_{\max}$ and ED_{50} are treated as model parameters; (ii) the linear model, where $f(d) = \beta_0 + \beta_1 d$ with β_0 and β_1 as model parameters. The $E_{\max}$ model is the one for which the last three response profiles are generated, while linear model is the misspecified one. For each modeling approach, we use the nonlinear least squares method to estimate the model parameters, and obtain the estimated dose-response curve, say $\hat{f}(d)$. To make the selection of the MaxED from modeling approach comparable with the proposed approaches in Section 3, the MaxED is selected as $d_m = \min\{d_i : \hat{f}(d_i) - \hat{f}(d_R) + \delta\}$. The $100(1 - \alpha)\%$ confidence interval for $\hat{f}(d_m)$ is constructed. If the upper limit of the confidence interval is below 50, we reject the null hypothesis and conclude that the selected MaxED is significantly different from control. In addition, to compare the modeling approach with the approaches proposed in Section 3, we take the number of subjects in each group as the rounded integer of the total sample size for two-stage design divided by the number of groups (say, $R + 1$). For each response profile, we calculated the sample size first, and generated random sample from normal distribution with means as specified and with standard deviation $\sigma = 15$. We carried out the simulations with 10,000 runs, and the simulation results are summarized in Table 5. From Table 5, we conclude that (i) the total

power (rejection rates) for the approaches proposed in Section 3 and the modeling approaches are similar; however (ii) when the dose–response curve is specified correctly, the modeling approach is more effective on selecting the true MaxED; (iii) when the parametric model is not specified correctly (such as the linear model here), the performance of the modeling approach is not as good as the approaches proposed in Section 3 in terms of selecting the true MaxED. In addition, we carried out 10,000 simulations for the cases $R=2$, 3, and 4 (see Tables 2–4), where there is no specific underlying model generating the response profiles. From Table 2–4, the overall power for the approaches proposed in Section 3 and modeling approaches are similar; however, the approaches proposed in Section 3, in general, perform better than modeling approaches in terms of selecting the correct MaxED, which are highlighted in Tables 2–5.

6. Conclusion and Discussion

This article proposes one single-stage design and one two-stage design to select the MaxED and assess whether the effect at MaxED is different from control. The simulation results indicate that the Type I error rate for each one of the two designs is controlled, and the power for detecting the minimum clinical effect is maintained. However, the maximum sample sizes required for two designs are different. The two-stage design generally requires a smaller maximum sample size than the single-stage design. Although, in case that the number of dose levels is small (say $R = 2$) or the effect size is large (say $\gamma > 1.0$), the maximum sample sizes required by the two-stage design are quite close to or even slightly larger than those required by the single-stage design, the average sample size of the two-stage design is generally smaller than the single-stage design.

In this article, the number of subjects per group at stage 1 and stage 2 is roughly the same. The equal split in the two-stage design may not be optimal, and how to split the sample sizes between the stages may potentially have some impact on the performance of the proposed two-stage design in terms of the Type I error, power, and distribution of the total sample size. For example, if the total sample size required at the MaxED or control for the final analysis is n , one may consider selecting the MaxED and assessing treatment effect when $n/3$ subjects per group are studied, which may result in less rejection in the first stage, less total sample size, and less accuracy in estimating the MaxED. On the contrary, one may select the MaxED and assess treatment effect when $2n/3$ subjects per group are studied, which may result in larger rejection in the first stage, larger total sample size, and more accuracy in estimating the MaxED. How to obtain an optimal design is challenging yet interesting.

In the two-stage design, we propose to stop the trial early only if the selected dose is extremely effective. It is ideal to set up the boundary to stop the trial for either efficacy or futility at early stage; that is, the trial may stop early if the selected MaxED is either extremely effective (i.e., early stop for efficacy) or the selected MaxED is not effective at all (i.e., early stop for futility). However, it is extremely difficult to derive the exact boundary for futility under the current approach. Thus, we may not be able to stop the trial early due to futility, and the proposed approach will always choose a MaxED, irrespective of the observed dose response appearing to be one for which the MaxED would be observed. To

overcome this shortcoming, an adaptive design targeted at estimating the MaxED could be a possible approach.

As indicated in Tables 2–5, there will be a high probability of declaring a wrong MaxED based on the proposed procedure. It should be noted that the test performed is the one that conditions on the estimated MaxED, instead of the true MaxED. That is, the question the test addresses is that whether the estimated MaxED, which may well be different from the true MaxED, has higher efficacy than control. Therefore, the overall power is calculated based on whether the selected MaxED is different from control, and the power cannot be carried over to the true MaxED. It is also noticed that the proposed approaches, along with modeling approaches, consistently select the highest dose with a relatively large probability, regardless of whether the highest dose is the true MaxED. The proposed approaches in general behave better than the modeling approaches. Although the proposed approaches in this study have some limitations, the importance of MaxED and the difficulty of the study is clearly indicated by the ICH E4 guideline (1994), which states that the purpose of dose-response information is “to find the smallest dose with a discernible useful effect or to find a maximum dose beyond which no further beneficial effects is seen, but practical study designs do not exist to allow for precise determination of these doses.” The experimental design and analytic and testing methods for MaxED deserve further investigation.

Acknowledgments

The authors thank the reviewers and the Associate Editor for their constructive comments that have improved the article significantly. The current work for M. Kong and R. Bolli is partially supported by the NHLBI-funded CAESAR study (U24 HL094373). S.N. Rai was partially supported by Dr. Miller (Director, J. G. Brown Cancer Center at the University of Louisville) and the Wendell Cherry Chair in Clinical Trial Research Fund.

References

- Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E. An Empirical Distribution Function for Sampling With Incomplete Information. *The Annals of Mathematical Statistics*. 1955; 26:641–647.
- Barlow, RE.; Bartholomew, DJ.; Bremner, JM.; Brunk, HD. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. London: Wiley; 1972.
- Bauer P. A Note on Multiple Testing Procedures in Dose Finding. *Biometrics*. 1997; 53:1125–1128. [PubMed: 9333343]
- Bhattacharya R, Kong M. Consistency and Asymptotic Normal Property of Estimated Effective Doses in Bioassay. *Journal of Statistical Planning and Inference*. 2006; 137:643–658.
- Bretz F, Dette H, Pinheiro JC. Practical Considerations for Optimal Designs in Clinical Dose Finding Studies. *Statistics in Medicine*. 2009; 29:731–742. [PubMed: 20213708]
- Bretz F, Pinheiro JC, Branson M. Combining Multiple Comparisons and Modeling Techniques in Dose-Response Studies. *Biometrics*. 2005; 61:738–748. [PubMed: 16135025]
- Chow, SC.; Shao, J.; Wang, H. *Sample Size Calculations in Clinical Research*. 2nd ed.. New York: CRC Press; 2008.
- Follmann DA, Proschan MA, and Geller NL. Monitoring Pairwise Comparisons in Multi-Armed Clinical Trials. *Biometrics*. 1994; 50:325–336. [PubMed: 8068834]
- Garrett-Mayer E. The Continual Reassessment Method for Dose-Finding Studies: A Tutorial. *Clinical Trials*. 2006; 3:57–71. [PubMed: 16539090]
- ICH E4. “Dose-Response Information to Support Drug Registration,” International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use. 1994. Available at http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E4/Step4/E4_Guideline.pdf

- ICH E9. "Statistical Principles for Clinical Trials," International Conference on Harmonization of Technical Requirements for Human Use. 1998. Available at http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf
- Ivanova A, Bolognese JA, Perevozskaya I. Adaptive Dose Finding Based on t -Statistic for Dose-Response Trials. *Statistics in Medicine*. 2008; 27:1581–1592. [PubMed: 18241082]
- Jennison, C.; Turnbull, BW. *Group Sequential Methods With Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC; 2000.
- Korn E, Midthune D, Chen T, Rubinstein L, Christian MC, Simon R. A Comparison of Two Phase I Trial Designs. *Statistics in Medicine*. 1994; 13:1799–1806. [PubMed: 7997713]
- O'Brien PC, Fleming TR. A Multiple Testing Procedure for Clinical Trials. *Biometrics*. 1979; 35:549–556. [PubMed: 497341]
- Pfeffer MA, Braunwald E. Ventricular Remodeling After Myocardial Infarction. Experimental Observations and Clinical Implications. *Circulation*. 1990; 81:1161–1172. [PubMed: 2138525]
- Proschan MA, Follmann DA, and Geller NL. Monitoring Multi-Armed Trials. *Statistics in Medicine*. 1994; 13:1441–1452. [PubMed: 7973223]
- Ravishanker, N.; Dey, DR. *A First Course in Linear Model Theory*. Boca Raton, FL: Chapman & Hall/CRC; 2002.
- Remmenga MD, Milliken GA, Kratzer D, Schwenke JR, Rolka HR. Estimating the Maximum Effective Dose in a Quantitative Dose-Response Experiment. *Journal of Animal Science*. 1997; 75:2174–2183. [PubMed: 9263066]
- Ruberg SJ. Contrasts for Identifying the Minimum Effective Dose. *Journal of the American Statistical Association*. 1989; 84:816–822.
- Ruberg SJ. Dose-Response Studies I, Some Design Considerations. *Journal of Biopharmaceutical Statistics*. 1995a; 5:1–14. [PubMed: 7613556]
- Ruberg SJ. Dose-Response Studies II, Analysis and Interpretation. *Journal of Biopharmaceutical Statistics*. 1995b; 5:15–42. [PubMed: 7613559]
- Strassburger K, Bretz F, Finner H. Ordered Multiple Comparisons With the Best and Their Applications to Dose-Response Studies. *Biometrics*. 2007; 63:1143–1151. [PubMed: 17489969]
- Tamhane AC, Hochberg Y, Dunnett CW. Multiple Test Procedures for Dose Finding. *Biometrics*. 1996; 52:21–37. [PubMed: 8934584]
- Ting, N. Introduction and New Drug Development Process. In: Ting, N., editor. *Dose Finding in Drug Development*. New York: Springer; 2006. p. 1-16.
- Williams DA. A Test for Differences Between Treatment Means When Several Dose Levels are Compared With a Zero Control. *Biometrics*. 1971; 27:103–117. [PubMed: 5547548]
- Williams DA. The Comparison of Several Dose Levels With a Zero Dose Control. *Biometrics*. 1972; 28:519–531. [PubMed: 5037867]

Figure 1.

The probability density function plot for the stage 2 test statistic T_m^* (solid line) under the null hypothesis that the means of all groups are the same, and the probability density function plot of central t -distribution with degrees of freedom $4n_1 - 2$ (dashed lines). Here R is the number of doses studied, and n_1 is the number of subjects per group. It is clear that the test statistic T_m^* shifts to the right of the central t -distribution.

Figure 2. Panels A1–A4 show the maximum sample size required for the entire experiment for the two different designs with R fixed at 2, 3, 4, and 6, respectively.

Table 1

The upper-tail critical points for T_m^* based on 100,000 simulation runs for each combination of n_1 and R , where n_1 is the number of subjects studied per group in stage 1, as well as the number of subjects per group for the selected MaxED and control in stage 2. R is the number of dose levels studied excluding control.

n_1 Per group	$R = 2$		$R = 3$		$R = 4$		$R = 6$	
	$\bar{t}_{0.048}^*$	$t_{0.05}^*$	$\bar{t}_{0.048}^*$	$t_{0.05}^*$	$\bar{t}_{0.048}^*$	$t_{0.05}^*$	$\bar{t}_{0.048}^*$	$t_{0.05}^*$
3	2.102	2.076	2.207	2.182	2.285	2.263	2.304	2.278
4	2.040	2.017	2.134	2.108	2.202	2.178	2.241	2.219
5	1.998	1.977	2.097	2.074	2.143	2.121	2.184	2.165
6	1.979	1.958	2.069	2.045	2.113	2.092	2.166	2.143
7	1.965	1.945	2.050	2.029	2.099	2.079	2.140	2.120
8	1.947	1.924	2.035	2.012	2.090	2.069	2.133	2.111
9	1.930	1.909	2.041	2.022	2.065	2.048	2.123	2.102
10	1.923	1.906	2.042	2.019	2.074	2.052	2.125	2.103
12	1.918	1.898	2.010	1.989	2.060	2.039	2.096	2.079
14	1.923	1.902	1.990	1.970	2.045	2.024	2.086	2.067
16	1.912	1.892	1.995	1.976	2.034	2.015	2.082	2.064
18	1.911	1.889	1.985	1.967	2.026	2.005	2.074	2.055
20	1.900	1.878	1.978	1.959	2.03	2.013	2.069	2.050
25	1.895	1.875	1.975	1.956	2.009	1.987	2.060	2.042
30	1.887	1.869	1.970	1.951	2.005	1.985	2.057	2.039
50	1.871	1.853	1.943	1.924	1.971	1.952	2.018	1.999
100	1.847	1.829	1.923	1.902	1.943	1.923	1.961	1.942

Simulation results based on 10,000 simulations when the number of doses studied is 2 ($R = 2$) and $\delta = 1.5$, where the highlighted dose are the true MaxED.

Table 2

	μ_0	μ_1	% Select	μ_2	% Select	% Total reject	n -Stage 1	n -Stage 2	$N_{\max}$	N_{ave}
Truth R2-1	50		50		50					
Single-stage		0.5828		0.4172		0.0510	8	0	24	24
Two-stage		0.5642		0.4358		0.0492	5	5	25	24.96
T-test		0.5596		0.4404		0.0796	5	5	25	24.95
$E_{\max}$		0.6361		0.3639		0.0892	9	0	27	27
Linear		0.6694		0.3306		0.0755	9	0	27	27
Truth R2-2	50		40		40					
Single-stage		0.5871		0.4129		0.3870	8	0	24	24
Two-stage		0.5649		0.4351		0.4303	5	5	25	24.50
T-test		0.5601		0.4399		0.5345	5	5	25	24.48
$E_{\max}$		0.1960		0.8040		0.8297	9	0	27	27
Linear		0.1546		0.8454		0.8520	9	0	27	27
Truth R2-3	50		40		30					
Single-stage		0.1329		0.8671		0.8035	8	0	24	24
Two-stage		0.1901		0.8099		0.7946	5	5	25	22.84
T-test		0.1819		0.8181		0.8635	5	5	25	22.87
$E_{\max}$		0.0454		0.9546		0.9948	9	0	27	27
Linear		0.0074		0.9926		0.9967	9	0	27	27
Truth R2-4	50		30		30					
Single-stage		0.5813		0.4187		0.8739	8	0	24	24
Two-stage		0.5616		0.4384		0.8995	5	5	25	22.47
T-test		0.5671		0.4329		0.9407	5	5	25	22.45
$E_{\max}$		0.1696		0.8304		0.9994	9	0	27	27
Linear		0.0074		0.9926		0.9967	9	0	27	27
Truth R2-5	50		35		20	0	0	0	0	0
Single-stage		0.0397		0.9603		0.9809	8	0	24	24
Two-stage		0.0787		0.9213		0.9786	5	5	25	19.56

	μ_0	μ_1	% Select	μ_2	% Select	% Total reject	n -Stage 1	n -Stage 2	$N_{\max}$	N_{ave}
T-test			0.0773		0.9227	0.9894	5	5	25	19.57
$E_{\max}$			0.0122		0.9878	1	9	0	27	27
Linear			0		1	1	9	0	27	27
Truth R2-6	50		20		20					
Single-stage			0.5872		0.4128	0.9948	8	0	24	24
Two-stage			0.5648		0.4352	0.997	5	5	25	18.94
T-test			0.5637		0.4363	0.9984	5	5	25	18.91
$E_{\max}$			0.0001		0.9999	1	9	0	27	27
Linear			0		1	1	9	0	27	27

Simulation results based on 10,000 simulations when the number of doses studied is 3 ($R = 3$) and $\delta = 1.5$, where the highlighted dose are the true MaxED.

Table 3

	μ_0	μ_1	% Select	μ_2	% Select	μ_3	% Select	% Total reject	n-Stage 1	n-Stage 2	$N_{\max}$	N_{ave}
Truth R3-1	50	50	50	50	50	50	50					
Single-stage		0.4364		0.2375		0.3261		0.0477	8	0	32	32
Two-stage		0.4098		0.2301		0.3601		0.0507	5	5	30	29.95
T-test		0.4087		0.2332		0.3581		0.0897	5	5	30	29.95
$E_{\max}$		0.5865		0.1588		0.2546		0.0796	8	0	32	32
Linear		0.6142		0.0512		0.3346		0.0931	8	0	32	32
Truth R3-2	50	50	40	40								
Single-stage		0.062		0.5145		0.4235		0.3910	8	0	32	32
Two-stage		0.1004		0.4818		0.4178		0.3767	5	5	30	29.47
T-test		0.0968		0.4744		0.4288		0.4995	5	5	30	29.46
$E_{\max}$		0.0473		0.1466		0.8061		0.7771	8	0	32	32
Linear		0.0955		0.0275		0.877		0.7871	8	0	32	32
Truth R3-3	50	40	35	35		30						
Single-stage		0.0738		0.2661		0.6601		0.8250	8	0	32	32
Two-stage		0.1094		0.2814		0.6092		0.8004	5	5	30	27.55
T-test		0.1011		0.2791		0.6198		0.8763	5	5	30	27.53
$E_{\max}$		0.0441		0.1670		0.7889		0.9978	8	0	32	32
Linear		0.008		0.0051		0.9869		0.9990	8	0	32	32
Truth R3-4	50	40	30	30								
Single-stage		0.0602		0.5310		0.4088		0.8707	8	0	32	32
Two-stage		0.1010		0.4746		0.4244		0.8602	5	5	30	27.21
T-test		0.0944		0.4762		0.4294		0.9268	5	5	30	27.25
$E_{\max}$		0.0352		0.1745		0.7904		0.9991	8	0	32	32
Linear		0.0086		0.0025		0.9889		0.9998	8	0	32	32
Truth R3-5	50	44	35	35		20						
Single-stage		3.00E-04		0.0372		0.9625		0.9837	8	0	32	32
Two-stage		0.0045		0.0772		0.9183		0.9747	5	5	30	24.21

	μ_0	μ_1	% Select	μ_2	% Select	μ_3	% Select	% Total reject	n -Stage 1	n -Stage 2	$N_{\max}$	N_{ave}
T-test			0.003	0.0747	0.9223	0.9223		0.9882	5	5	30	24.13
$E_{\max}$			0.0016	0.0078	0.9906	0.9906		1	8	0	32	32
Linear			0	0	1	1		1	8	0	32	32
Truth R3-6	50	44	20	20	20							
Single-stage		1.00E-04		0.5755	0.4244			0.9943	8	0	32	32
Two-stage		0.001		0.5619	0.4371			0.9949	5	5	30	23.49
T-test		0.0022		0.5690	0.4288			0.9973	5	5	30	23.49
$E_{\max}$		0		0.0033	0.9967			1	8	0	32	32
Linear		0		0	1			1	8	0	32	32

Table 4

Simulation results based on 10,000 simulations when the number of doses studied is 4 ($R = 4$) and $\delta = 1.5$, where the highlighted dose are the true MaxED.

	μ_0	μ_1	% Select	μ_2	% Select	μ_3	% Select	μ_4	% Select	% Total reject	n-Stage 1	n-Stage 2	$N_{\max}$	N_{ave}
Truth R4-1	50		50	50	50	50	50	50						
Single-stage		0.3514		0.1766		0.1682		0.3038		0.0493	8	0	40	40
Two-stage		0.3305		0.1673		0.1849		0.3173		0.0507	5	5	35	34.96
T-test		0.3279		0.1722		0.1848		0.3151		0.1002	5	5	35	34.95
$E_{\max}$		0.5992		0.1028		0.1281		0.1699		0.0747	7	0	35	35
Linear		0.6045		0.0147		0.0492		0.3316		0.0934	7	0	35	35
Truth R4-2	50		50	45		40		40						
Single-stage		0.0353		0.1743		0.4055		0.3849		0.3863	8	0	40	40
Two-stage		0.0615		0.1829		0.3546		0.401		0.3516	5	5	35	34.42
T-test		0.0624		0.1898		0.3641		0.3837		0.4933	5	5	35	34.46
$E_{\max}$		0.0387		0.0754		0.2562		0.6296		0.8045	7	0	35	35
Linear		0.0946		0.0081		0.0274		0.8699		0.7795	7	0	35	35
Truth R4-3	50		46	40		34		30						
Single-stage		0.0048		0.0694		0.3147		0.6111		0.8414	8	0	40	40
Two-stage		0.018		0.0939		0.3149		0.5732		0.7888	5	5	35	32.35
T-test		0.0178		0.0905		0.314		0.5777		0.8859	5	5	35	32.40
$E_{\max}$		0.0122		0.0422		0.1858		0.7598		0.9972	7	0	35	35
Linear		0.006		0.0009		0.0037		0.9894		0.9979	7	0	35	35
Truth R4-4	50		46	40		30		30						
Single-stage		0.0028		0.0618		0.5189		0.4165		0.8750	8	0	40	40
Two-stage		0.0104		0.0895		0.4748		0.4253		0.8477	5	5	35	32.11
T-test		0.0108		0.0865		0.4764		0.4263		0.9176	5	5	35	32.09
$E_{\max}$		0.0073		0.0294		0.1702		0.7931		0.9987	7	0	35	35
Linear		0.0053		0.0005		0.0031		0.9911		0.9975	7	0	35	35
Truth R4-5	50		44	35		26		20						
Single-stage		3.00E-04		0.0177		0.2579		0.7241		0.9904	8	0	40	40
Two-stage		0.0013		0.0402		0.2845		0.6740		0.9796	5	5	35	28.66

	μ_0	μ_1	% Select	μ_2	% Select	μ_3	% Select	μ_4	% Select	% Total reject	n -Stage 1	n -Stage 2	$N_{\max}$	N_{ave}
T-test			0.0013	0.0421	0.2813	0.6753	0.9908	5	5	5	5	5	35	28.64
$E_{\max}$			0.0017	0.0058	0.0821	0.9103	1	7	0	35	35	35	35	35
Linear			0	0	0.0003	0.9997	1	7	0	35	35	35	35	35
Truth R4-6	50	44	35	20	0.5665	0.4202	0.9954	8	0	40	40	40	40	40
Single-stage		1.00E-04	0.0132	0.0289	0.5303	0.4403	0.9913	5	5	35	35	35	35	28.19
Two-stage		5.00E-04	0.0336	0.0036	0.5301	0.4357	0.9950	5	5	35	35	35	35	28.21
T-test			0.0006	0.0036	0.0492	0.9465	1	7	0	35	35	35	35	35
$E_{\max}$			0	0	0	1	1	7	0	35	35	35	35	35
Linear			0	0	0	1	1	7	0	35	35	35	35	35

Table 5

Simulation results based on 10,000 simulations when the number of doses studied is 6 ($R = 6$) and $\delta = 1.5$, where the highlighted dose are the true MaxED.

	μ_0	μ_1 % Select	μ_2 % Select	μ_3 % Select	μ_4 % Select	μ_5 % Select	μ_6 % Select	% Total reject	n-Stage 1	n-Stage 2	$N_{\max}$	N_{ave}
Truth R6-1		50	50	50	50	50	50					
Single-stage	0.2675	0.1246	0.1011	0.1021	0.1425	0.2622	0.0477	0.0477	8	0	56	56
Two-stage	0.2564	0.1237	0.0997	0.1016	0.1434	0.2752	0.0509	0.0509	6	6	54	53.93
T-test	0.2632	0.1158	0.1050	0.1057	0.1386	0.2717	0.1051	0.1051	6	6	54	53.94
$E_{\max}$	0.6112	0.0630	0.0741	0.0780	0.0816	0.0920	0.0657	0.0657	8	0	56	56
Linear	0.6033	0.0050	0.0053	0.0193	0.0630	0.3041	0.0985	0.0985	8	0	56	56
Truth R6-2		46	40	34	31	30	30					
Single-stage	0.0008	0.0252	0.1624	0.2336	0.2419	0.3361	0.8908	0.8908	8	0	56	56
Two-stage	0.0030	0.0381	0.1615	0.2239	0.2326	0.3409	0.8901	0.8901	6	6	54	49.02
T-test	0.0024	0.0369	0.1635	0.2246	0.2329	0.3397	0.9403	0.9403	6	6	54	49.11
$E_{\max}$	0.0042	0.0181	0.1241	0.4620	0.3547	0.0370	1	1	8	0	56	56
Linear	0.0069	0.0003	0.0003	0.0013	0.0056	0.9856	1	1	8	0	56	56
Truth R6-3		47	41	31	24	21	20					
Single-stage	0	0.0003	0.0323	0.216	0.3261	0.4253	0.9963	0.9963	8	0	56	56
Two-stage	0	0.0007	0.0435	0.2242	0.3123	0.4193	0.9968	0.9968	6	6	54	44.10
T-test	0.0001	0.0008	0.0435	0.2225	0.3053	0.4278	0.9984	0.9984	6	6	54	44.11
$E_{\max}$	0.0003	0.0002	0.0038	0.0984	0.7234	0.1738	1	1	8	0	56	56
Linear	0	0	0	0	0	1	1	1	8	0	56	56
Truth R6-4		44	35	26	22	20	20					
Single-stage	0	0.0048	0.1089	0.2461	0.2913	0.3489	0.9972	0.9972	8	0	56	56
Two-stage	0.0001	0.0087	0.1240	0.2356	0.2789	0.3527	0.9978	0.9978	6	6	54	43.91
T-test	0.0001	0.0082	0.1224	0.2375	0.2768	0.3550	0.9996	0.9996	6	6	54	43.91
$E_{\max}$	0.0001	0.0010	0.0220	0.4031	0.5583	0.0155	1.0000	1.0000	8	0	56	56
Linear	0	0	0	0	0	1	1	1	8	0	56	56