Title
Phase II two-stage single-arm clinical trials for testing toxicity levels

Authors
Kim, Seongho
Chak, Lorin

Publication Date
2019-03-31

DOI
10.29220/CSAM.2019.26.2.163

Peer reviewed

Communications for Statistical Applications and Methods
2019, Vol. 26, No. 2, 163–173

https://doi.org/10.29220/CSAM.2019.26.2.163
Print ISSN 2287-7843 / Online ISSN 2383-4757

Phase II two-stage single-arm clinical trials for testing toxicity levels

Seongho Kim[1,a], Weng Kee Wong[b]

[a]Biostatistics Core, Karmanos Cancer Institute/Wayne State University, USA;
[b]Department of Statistics, UCLA, USA

Abstract

Simon's two-stage designs are frequently used in phase II single-arm trials for efficacy studies. A concern of safety studies is too many patients who experience an adverse event. We show that Simon's two-stage designs for efficacy studies can be similarly used to design a two-stage safety study by modifying some of the design parameters. Given the type I and II error rates and the proportion of adverse events experienced in the first stage cohort, we prescribe a procedure whether to terminate the trial or proceed with a stage 2 trial by recruiting additional patients. We study the relationship between a two-stage design with a safety endpoint and an efficacy endpoint as well as use simulation studies to ascertain their properties. We provide a real-life application and a free R package `gen2stage` to facilitate direct use of two-stage designs in a safety study.

Keywords: safety, single-stage design, tolerability, toxicity, two-stage design

1. Introduction

Phase II single-arm two-stage designs are typically used to determine if a drug is sufficiently efficacious to move on a randomized phase III trial. Sometimes, we would like to design a trial to examine whether the drug is safe due to a significant adverse event or toxicity. These are often called toxicity or tolerability studies instead of safety studies and they occur frequently. Results from a sample of subjects are then compared to an adverse event rate from historical controls. If the drug appears to have a fewer number of patients with a toxic response than expected, the trial proceeds to stage 2 for further research; otherwise, the study terminates at stage 1.

Drug trials estimate the true success rate p of a drug. These are efficacy studies and are conducted in one stage or two stages. In Simon's two-stage efficacy studies (Simon, 1989), a minimal efficacy response rate p_0 is pre-specified in the null hypothesis versus a user-specified higher rate p_1 in the alternative hypothesis. Given type I and II error rates, there are four variables to optimize the design problem: the patient number n_1 to recruit in stage 1, the total number of patients n required in the entire trial, the rejection number r_1 in stage 1 and the cumulative rejection number r in the entire trial. If the estimated response rate is smaller than p_0, the treatment is deemed not efficacious and the trial stops in the first stage. Otherwise, it proceeds to the second stage, where the number of additional patients and the number of additional responders have to be determined to arrive at a decision for the trial.

[1] Corresponding author: Biostatistics Core, Karmanos Cancer Institute, Department of Oncology, Wayne State University School of Medicine, Detroit, MI, USA. E-mail: kimse@karmanos.org

Published 31 March 2019 / journal homepage: http://csam.or.kr

These optimization problems are typically solved by a greedy search over the set of constrained positive integers (Kim and Wong, 2017). There are two optimality criteria for Simon's two-stage designs: a minimax criterion that minimizes the maximum sample size and an optimal design that minimizes the expected sample (Simon, 1989). Since Simon's landmark paper was proposed, many variations of the strategies have been proposed for phase II designs. Green and Dahlberg (1992) investigated a two-stage design for multicenter trials when the attained sample size is not the planned one. Mander and Thompson (2010) and Mander *et al.* (2012) proposed novel designs that are optimal under the alternative hypothesis. Wason *et al.* (2011) proposed reducing the sample size for phase II trials with a continuous outcome, and Kwak and Jung (2014) considered a two-stage adaptive optimal design for single arm trials with right-censored survival time. Ensign *et al.* (1994), Chen (1997), and Chen and Shan (2007) proposed phase II designs with three stages. Under the Bayesian framework, Thall and Simon (1994) proposed posterior probability to account for uncertainty regarding the response rates of the null and alternative hypotheses. Lee and Liu (2008) developed an efficient and flexible design using Bayesian predictive probability and the minimax criterion. Cai *et al.* (2014) constructed a phase II trial design for continuous monitoring when delayed responses are present using multiple imputation. Zhou *et al.* (2017) recently proposed a flexible Bayesian optimal phase II (BOP2) design to handle simple and complex endpoints such as binary, ordinal, nested, and co-primary endpoints.

Several phase II single-arm two-stage designs are available for toxicity studies, but those studies are designed to monitor toxicity as a secondary or a co-primary endpoint to determine if a drug or treatment is efficacious; see, Bryant and Day (1995), Conaway and Petroni (1996), Ray and Rai (2011), and references. Therefore, those designs are not directly applicable for safety studies where the primary endpoint is the proportion of adverse events or toxicity. The motivation for this work is that there are no phase II single-arm safety studies that utilize Simon's two-stage designs. There are two possible reasons for this: (i) there are allusions that this is possible but the theoretical justifications have not been worked out, and (ii) there is no software package to generate a two-stage design for a safety study. PASS 15 (NCSS, LLC), a widely used power and sample size calculation commercial software package, has an option for Simon's two-stage designs under the 'Proportions: One Proportion: Group-Sequential: Two-Stage Phase II Clinical Trials' category, but cannot handle safety cases where a null proportion p_0 is larger than an alternative proportion p_1 (i.e., $p_0 > p_1$). Similarly, STATA 15.0 (StataCorp, LLC) and nQuery 8.0 (StatSols, Ltd.) only have a code to generate a Simon's two-stage design for an efficacy trial but not for a safety trial.

As an example of a safety study, Rugo *et al.* (2017) used a one-sample proportion test to estimate the required sample size for a phase II single-arm study to assess the incidence of grade 2 or worse everolimus-related stomatitis in women when dexamethasone mouthwash (SWISH) is used. On the basis of historical controls, the incidence rate of grade 2 or worse stomatitis without SWISH was considered to be 0.33 (i.e., the null hypothesis is $p_0 = 0.33$). The alternative hypothesis is that the use of SWISH would lead to an absolute reduction of 0.13 (i.e., $p_1 = 0.20$). A direct calculation shows that the estimated required sample size was 73 for a one-sided 5% sized test with 80% power using a one-sample proportion test. If two-stage designs are applied to this trial using our proposed approaches, the estimated sample sizes are $(r_1, n_1, r, n) = (18, 26, 63, 85)$ and $(50, 67, 54, 72)$ for the optimal and minimax designs, respectively. Therefore, they could have an opportunity to evaluate the study early at the end of the first stage using either minimax or optimal designs that could also save 3 patients compared to their original design if the minimax design was employed. We provide details on how to apply a two-stage design to this SWISH trial in Section 3.2.

We now develop a theory to construct a two-stage safety trial following Simon's original two-

stage design. In particular, we provide analytical formulas parallel to those for an efficacy study and show how a two-stage safety study can be found directly from Simon's two-stage design. We also compare single-stage and two-stage designs for safety studies through simulation studies and apply our results to construct a phase II two-stage design for a real application (Rugo *et al.*, 2017). The user can implement our phase II single- and two-stage designs for a single-arm trial using our R package called `gen2stage` available at http://cansur.wayne.edu/.

2. Designs for one and two-stage safety studies

Throughout, let $b(x, n, p)$ and $B(x, n, p)$ be the probability mass function (pmf) and the cumulative distribution function (cdf), respectively, of a random variable X that follows the binomial distribution with parameters n and p, i.e., $X \sim \text{Binom}(n, p)$. We now describe the theoretical relationships between designs for safety studies and those for efficacy studies.

2.1. Single-stage design

A single-stage design is estimated using Fleming's single-stage procedure with the exact binomial distribution (Fleming, 1982; A'Hern, 2001). The true rate (or proportion) is denoted by p. The standard rate is assumed to be p_0 under the null hypothesis (i.e., $H_0 : p = p_0$) and the new treatment can use further research if the null hypothesis is rejected in favor of the alternative hypothesis, where the rate is p_1 (i.e., $H_1 : p = p_1$). For a safety study, the toxicity (or adverse event) rate is the primary endpoint. We wish to test the toxicity rate is p_0, which is the maximum allowed; consequently, we reject the null hypothesis and terminate the study if there are too many toxic responses among the first cohort of patients. In the alternative hypothesis, the target toxicity rate is p_1 with $p_0 > p_1$. Then the single-stage design aims to find a good design by testing $H_0 : p = p_0$ vs. $H_1 : p = p_1$ with predetermined p_0, p_1, type I error rate α and type II error rate β, respectively. The single-stage design for the efficacy or safety studies is performed as:

1. Begin by recruiting n patients of the single-arm phase II design and observe the number of the responses or toxicity, x, out of n patients.

2. When $p_0 < p_1$ (for efficacy), the null hypothesis is $H_0 : p = p_0$ and the alternative hypothesis $H_1 : p = p_1$ with $p_1 > p_0$. The decision rule is to fail to reject the null hypothesis H_0 if the total number x of responses is less than r (i.e., $x \le r$).

3. When $p_0 > p_1$ (for safety), the null hypothesis is $H_0 : p = p_0$ and the alternative hypothesis $H_1 : p = p_1$ with $p_1 < p_0$. The decision rule is to fail to reject the null hypothesis H_0 if the total number x of subjects experiencing toxicity is larger than r (i.e., $x \ge r$).

The single-stage design has two parameters $\theta = (r, n)$ to be optimized. Given error rates α and β, their optimal values are found by a greedy search. If the true rate is p, the probability of failing to reject H_0 is given by

$$\begin{cases} P_1(\theta|p) = B(r, n, p), & \text{if } p_0 < p_1, \\ S_1(\theta|p) = 1 - B(r-1, n, p), & \text{if } p_0 > p_1, \end{cases} \tag{2.1}$$

where B is the cdf of a binomial distribution with parameters n and p. Note that we hereafter use the subscripts of 1 and 2 to indicate the probability of failing to reject H_0 for the single-stage and the

two-stage designs, respectively. To find an appropriate design, we first identify a set $\widetilde{\Theta}$ that contains all θ that satisfies the two user-specified error constraints:

$$\begin{cases} P_1(\theta|p_0) \geq 1-\alpha; P_1(\theta|p_1) \leq \beta, & \text{if } p_0 < p_1, \\ S_1(\theta|p_0) \geq 1-\alpha; S_1(\theta|p_1) \leq \beta, & \text{if } p_0 > p_1. \end{cases} \tag{2.2}$$

The optimal choice of $\hat{\theta}$ in $\widetilde{\Theta}$ may be determined by the following optimality criterion for both efficacy and safety cases:

$$\hat{\theta} = \arg\min_{\theta \in \widetilde{\Theta}} n_\theta, \tag{2.3}$$

where n_θ is the required sample size corresponding to $\theta = (r, n)$.

2.2. Two-stage design

The two-stage design evaluates the trial endpoint at each stage and allows the trial to proceed to the second stage only if one rejects the null hypothesis at the end of the first stage. The required sample sizes for the first and second stages are denoted by n_1 and n_2, respectively, where the total required sample size $n = n_1 + n_2$. Using similar notation used in the single-stage design, the two-stage design for efficacy or toxicity tests $H_0 : p = p_0$ versus $H_1 : p = p_1$ with $p_0 < p_1$ or $H_1 : p = p_1$ with $p_1 < p_0$, respectively, as:

- Step I: Begin by recruiting n_1 patients in the first stage of the trial and observe the number of the responses or toxicity, x, out of n_1 patients.
- Step II:
  1. When $p_0 < p_1$ (for efficacy), if $x \leq r_1$, stop the trial and fail to reject H_0 (i.e., $p \leq p_0$); if $x > r_1$, power the study at $(1-\beta)$ for $p = p_1$ and enter $n_2 = n - n_1$ additional patients into the study. Reject the alternative hypothesis H_1 (i.e., $p \geq p_1$) if the total number of responses $\leq r$ out of n patients.
  2. When $p_0 > p_1$ (for safety), if $x \geq r_1$, stop the trial and fail to reject H_0 (i.e., $p \geq p_0$); if $x < r_1$, power the study at $(1-\beta)$ for $p = p_1$ and enter $n_2 = n - n_1$ additional patients into the study. Reject the alternative hypothesis H_1 (i.e., $p \leq p_1$) if the total number of toxicity $\geq r$ out of n patients.

The two-stage designs for efficacy or safety have four parameters $\theta = (r_1, n_1, r, n)$ to optimize given error rate constraints. The probability of failing to reject H_0 when the true rate is p is given by

$$\begin{cases} P_2(\theta|p) = B(r_1, n_1, p) + \sum_{x=r_1+1}^{\min(n_1, r)} b(x, n_1, p) B(r-x, n_2, p), & \text{if } p_0 < p_1, \\ S_2(\theta|p) = 1 - B(r_1 - 1, n_1, p) + \sum_{x=r_1-1}^{\max(0, r-n_2)} b(x, n_1, p)(1 - B(r-x-1, n_2, p)), & \text{if } p_0 > p_1, \end{cases} \tag{2.4}$$

where b and B are the pmf and cdf of a binomial distribution with $n = n_1 + n_2$. Furthermore, the probability of early termination (PET) at the end of the first stage is

$$\text{PET}(p|\theta) = \begin{cases} \text{PET}_P(p|\theta) = B(r_1, n_1, p), & \text{if } p_0 < p_1, \\ \text{PET}_S(p|\theta) = 1 - B(r_1 - 1, n_1, p), & \text{if } p_0 > p_1, \end{cases} \tag{2.5}$$

where the subscripts P and S hereafter indicate the designs for testing $H_0 : p = p_0$ versus $H_1 : p = p_1$ with $p_0 < p_1$ for efficacy and designs for testing $H_0 : p = p_0$ versus $H_1 : p = p_1$ with $p_0 > p_1$ for safety, respectively. The expected sample size with p as the response probability is then

$$E(N|p,\theta) = \begin{cases} E_P(N|p,\theta) = n_1 + (1 - B(r_1, n_1, p))n_2, & \text{if } p_0 < p_1, \\ E_S(N|p,\theta) = n_1 + B(r_1 - 1, n_1, p)n_2, & \text{if } p_0 > p_1. \end{cases} \tag{2.6}$$

The sought two-stage design for evaluating toxicity is to find a design $\hat{\theta} \in \widetilde{\Theta}$, where $\widetilde{\Theta}$ contains all θ that satisfies two natural error constraints:

$$\begin{cases} P_2(\theta|p_0) \geq 1 - \alpha; P_2(\theta|p_1) \leq \beta, & \text{if } p_0 < p_1, \\ S_2(\theta|p_0) \geq 1 - \alpha; S_2(\theta|p_1) \leq \beta, & \text{if } p_0 > p_1. \end{cases} \tag{2.7}$$

The goodness of $\hat{\theta}$ may be determined by one of the following two optimality criteria:

$$\text{Optimal: } \hat{\theta} = \arg\min_{\theta \in \widetilde{\Theta}} E(N|p_0, \theta), \tag{2.8}$$

$$\text{Minimax: } \hat{\theta} = \arg\min_{\theta \in \widetilde{\Theta}} n_\theta, \tag{2.9}$$

where n_θ is the required sample size corresponding to $\theta = (r_1, n_1, r, n)$.

2.3. Theoretical results

We describe several relationships between the designs for efficacy, $p_0 < p_1$, and those for safety, $p_0 > p_1$. These relationships enable us to find one or two stage optimal designs for toxicity studies for testing $p = p_0(> p_1)$ versus $H_1 : p = p_1$ from optimal designs for efficacy studies to test $p = p_0(< p_1)$ versus $H_1 : p = p_1$. Suppose a random variable X follows the binomial distribution with parameters n and p, i.e., $X \sim \text{Binom}(n, p)$. An important result which we will use repeatedly and without further mention is $b(x, n, p) = b(n - x, n, 1 - p)$, where $x, n \in N$, $x \leq n$, and $p \in [0, 1]$.

We now discuss several useful relationships between decision rules to test toxicity and efficacy rates with justifications.

Theorem 1. *Consider a single-stage design with $\theta = (r, n)$. Then $P_1(\theta|p) = S_1(\phi|1 - p)$, where $\phi = (n - r, n)$.*

Proof:

$$\begin{aligned} P_1(\theta|p) &= B(r, n, p) \quad \text{(by Equation (2.1))} \\ &= \sum_{x=0}^{r} b(n - x, n, 1 - p) \\ &= \sum_{y=n}^{n-r} b(y, n, 1 - p) \quad \text{(by substituting } y \text{ for } n - x) \\ &= 1 - B(n - r - 1, n, 1 - p) \\ &= S_1(\phi|1 - p) \quad \text{(by Equation (2.1))} \end{aligned}$$

□

Corollary 1. *For a single-stage design, if $\hat{\theta} = (r, n)$ is the optimal single-stage design for $p_0 < p_1$, then $\hat{\phi} = (n - r, n)$ is the optimal single-stage design for $1 - p_0 > 1 - p_1$ and vice versa.*

Proof: By Theorem 1, finding the set of feasible solutions $\hat{\Theta}$ that satisfies $P_1(\theta|p_0) \geq 1 - \alpha$ and $P_1(\theta|p_1) \leq \beta$ is equivalent to finding the set of feasible solutions $\hat{\Phi}$ that satisfies $S_1(\theta|p_0) \geq 1 - \alpha$ and $S_1(\theta|p_1) \leq \beta$. Therefore, $\theta = (r, n) \in \hat{\Theta}$ if and only if $\phi = (n - r, n) \in \hat{\Phi}$. Thus, $\min_\theta n_\theta = \min_\phi n_\phi$ and so $\hat{\phi} = (n - r, n)$ is the optimal design for $1 - p_0 > 1 - p_1$. □

Corollary 1 implies that finding the optimal single-stage design for $p_0 > p_1$ is identical to finding the design for $1 - p_0 < 1 - p_1$. A similar relationship on the probability of failing to reject H_0 when the true rate is p can be established.

Here is an example for a single-stage design for testing toxicity.

Example 1. Consider a standard therapy (historical control) whose incidence rate of adverse events is 0.5 (i.e., $p_0 = 0.5$) and suppose that an investigator wants to see if experimental therapy can reduce the incidence rate of adverse events down to 0.3 (i.e., $p_1 = 0.3$). We want to design a phase II single-stage single-arm trial with 90% power at one-sided 10% level (i.e., $\alpha = 0.1$ and $\beta = 0.1$). Then, by Corollary 1, it is sufficient to find the optimal design, $\hat{\theta} = (r, n)$, for $0.5 = p_0 < p_1 = 0.7$ in order to find the optimal single-stage design, $\hat{\phi} = (n - r, n)$, for $0.5 = p_0 > p_1 = 0.3$. Therefore, $\hat{\phi} = (16, 39)$ because $\hat{\theta} = (23, 39)$, implying that, if 16 or more adverse events occur in the 39 patients, the experimental therapy is rejected.

Theorem 2. *Consider a two-stage design with $\theta = (r_1, n_1, r, n)$. Then $P_2(\theta|p) = S_2(\phi|1 - p)$, where $\phi = (n_1 - r_1, n_1, n - r, n)$.*

Proof:

$$
\begin{aligned}
P_2(\theta|p) &= B(r_1, n_1, p) + \sum_{x=r_1+1}^{\min(n_1,r)} b(x, n_1, p) B(r - x, n_2, p) \quad \text{(by Equation (2.4))} \\
&= 1 - B(n_1 - r_1 - 1, n_1, 1 - p) + \sum_{x=r_1+1}^{\min(n_1,r)} b(n_1 - x, n_1, 1 - p)\big(1 - B(n_2 - r + x - 1, n_2, 1 - p)\big) \\
&\quad \text{(by Theorem 1)} \\
&= 1 - B(n_1 - r_1 - 1, n_1, 1 - p) + \sum_{y=n_1-r_1-1}^{\max(0,n-r-n_2)} b(y, n_1, 1 - p)(1 - B(n - r - y - 1, n_2, 1 - p)) \\
&\quad \text{(by substituting } y \text{ for } n_1 - x) \\
&= S_2(\phi|1 - p) \quad \text{(by Equation (2.4)).}
\end{aligned}
$$

□

Using Theorem 1, calculations of the PET and the expected sample size for toxicity studies are similar to those for efficacy studies. It can be shown that for a two-stage design with $\theta = (r_1, n_1, r, n)$, we have $\text{PET}(p|\theta) = \text{PET}(1 - p|\phi)$ and $E_P(N|p, \theta) = E_S(N|1 - p, \phi)$, where PET is the probability of early termination after the first stage and $\phi = (n_1 - r_1, n_1, n - r, n)$. Therefore, on the basis of Theorem 2 and Corollary 2, two-stage designs for $p_0 > p_1$ can be estimated by finding the two-stage designs for $p_0 < p_1$, respectively.

Corollary 2. *If $\hat{\theta} = (r_1, n_1, r, n)$ is the optimal two-stage design for $p_0 < p_1$, then $\hat{\phi} = (n_1 - r_1, n_1, n - r, n)$ is the optimal two-stage design for $1 - p_0 > 1 - p_1$ and vice versa. This relationship also holds for the minimax two-stage design.*

Proof: Since $\hat{\theta}$ is the optimal two-stage design,

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta\in\hat{\Theta}} E_P(N|p_0,\theta) \quad \text{(by Equation (2.8))}\\ &= \arg\max_{\theta\in\hat{\Theta}}\{n_1 + (1 - B(r_1, n_1, p))n_2\} \quad \text{(by Equation (2.6))}\\ &= \arg\max_{\theta\in\hat{\Theta}}\{n_1 + B(n_1 - r_1 - 1, n_1, 1 - p)n_2\} \quad \text{(by Theorem 1).}\end{aligned}$$

However, by Theorem 2, finding the set of feasible solutions $\hat{\Theta}$ that satisfies $P_2(\theta|p_0) \geq 1 - \alpha$ and $P_2(\theta|p_1) \leq \beta$ is equivalent to finding a set of feasible solutions $\hat{\Phi}$ that satisfies $S_2(\theta|p_0) \geq 1 - \alpha$ and $S_2(\theta|p_1) \leq \beta$. That is, $\theta = (r_1, n_1, r, n) \in \hat{\Theta}$ if and only if $\phi = (n_1 - r_1, n_1, n - r, n) \in \hat{\Phi}$, resulting that $\max_{\theta\in\hat{\Theta}}\{n_1 + B(n_1 - r_1 - 1, n_1, 1 - p)n_2\} = \max_{\phi\in\hat{\Phi}}\{n_1 + B(n_1 - r_1 - 1, n_1, 1 - p)n_2\}$. Therefore, by Equation (2.8), $\hat{\phi} = \arg\max_{\phi\in\hat{\Phi}} E_S(N|1 - p_0, \phi)$, showing that $\hat{\phi} = (n_1 - r_1, n_1, n - r, n)$ is the optimal design for $1 - p_0 > 1 - p_1$.

When $\hat{\theta}$ is the minimax two-stage design, by Theorem 2, we have $\min_{\theta\in\hat{\Theta}} n_\theta = \min_{\phi\in\hat{\Phi}} n_\phi$; therefore, $\hat{\phi} = \arg\min_{\phi\in\hat{\Phi}} n_\phi$ by Equation (2.9), showing that $\hat{\phi} = (n_1 - r_1, n_1, n - r, n)$ is the minimax design for $1 - p_0 > 1 - p_1$ □

It follows from Theorem 2 and Corollary 2. The two-stage design for testing toxicity when $p_0 > p_1$ can be obtained by finding the two-stage designs for $p_0 < p_1$ and conversely.

Here is an example of a two-stage design for testing toxicity.

Example 2. Consider the same design parameter values in Example 1 for the single-stage design, $(p_0, p_1, \alpha, \beta) = (0.5, 0.3, 0.1, 0.1)$. By Corollary 2, it is sufficient to find an optimal design, $\hat{\theta} = (r_1, n_1, r, n)$, for $0.5 = p_0 < p_1 = 0.7$ in order to find an optimal two-stage design, $\hat{\phi} = (n_1 - r_1, n_1, n - r, n)$, for $0.5 = p_0 > p_1 = 0.3$. The resulting optimal two-stage design is $\hat{\phi} = (10, 21, 19, 45)$ because $\hat{\theta} = (11, 21, 26, 45)$ and the minimax two-stage design is $\hat{\phi} = (12, 23, 16, 39)$ because $\hat{\theta} = (11, 23, 23, 39)$. The expected null sample sizes $E(N|p_0)$ for the optimal and minimax designs are 28.96 and 31, respectively. For the optimal design, this means that if 10 or more adverse events occur among the first 21 patients at the end of stage 1, we terminate the trial for excess adverse events and concluded that experimental therapy does not warrant further investigation. Otherwise, the trial proceeds to stage 2 and enrolls additional 24 patients. If 19 or more of the 45 patients experience adverse events at the end of stage 2, we conclude that the experimental therapy results in excess adverse events not requiring further investigation. Otherwise, we consider the experimental therapy for further investigation in subsequent trials. A similar interpretation applies for the minimax design.

3. Applications

3.1. Simulation studies

We use simulation studies to show optimal single- and two-stage designs for toxicity studies using various design parameters where $p_0 > p_1$. Our simulation covers two sets of type I (α) and type II (β) error rates with $(\alpha, \beta) \in \{(0.05, 0.20), (0.10, 0.20)\}$ and seven sets of target toxicity rates with

Table 1: Single-stage and two-stage (optimal and minimax) designs with $\alpha = 0.05$ and $\beta = 0.2$

p_0	p_1		r_1	n_1	r	n	$E(N\|p_0)$	PET(p_0)	$\hat{\alpha}$	$\hat{\beta}$
		Single			5	28			0.047	0.142
0.3	0.1	Optimal	2	6	5	27	14.82	0.58	0.049	0.196
		Minimax	4	23	5	26	23.16	0.95	0.045	0.199
		Single			10	36			0.045	0.168
0.4	0.2	Optimal	4	11	13	43	20.48	0.70	0.049	0.198
		Minimax	5	13	10	35	20.77	0.65	0.050	0.192
		Single			14	37			0.049	0.193
0.5	0.3	Optimal	7	15	17	43	23.50	0.70	0.050	0.196
		Minimax	11	23	14	37	27.74	0.66	0.048	0.199
		Single			20	42			0.038	0.197
0.6	0.4	Optimal	9	16	23	46	24.52	0.72	0.049	0.199
		Minimax	17	34	19	39	34.44	0.91	0.049	0.198
		Single			23	39			0.050	0.168
0.7	0.5	Optimal	10	15	28	46	23.63	0.72	0.050	0.197
		Minimax	13	19	23	39	25.69	0.67	0.046	0.196
		Single			24	35			0.034	0.195
0.8	0.6	Optimal	10	13	31	43	20.58	0.75	0.050	0.200
		Minimax	14	18	23	33	22.25	0.72	0.046	0.199
		Single			20	25			0.033	0.194
0.9	0.7	Optimal	9	10	24	29	15.01	0.74	0.047	0.195
		Minimax	14	15	20	25	19.51	0.55	0.033	0.198

PET = probability of early termination.

$0.3 \leq p_0 \leq 0.9$ and $p_0 - p_1 = 0.2$. The estimation procedure is the same as the original Simon's two-stage design (Simon, 1989).

Table 1 shows single-stage and two-stage designs when $\alpha = 0.05$ and $\beta = 0.2$. We observe that the required total sample sizes (n) of the single-stage design are larger than the minimax two-stage design; however, the single-stage design has the smaller required total sample sizes than the optimal two-stage design except when $p_0 = 0.3$. As expected, the optimal two-stage design has a smaller expected total sample size under the null hypothesis $E(N|p_0)$ than the minimax two-stage design. Interestingly, the single-stage design has smaller estimates of β than the two-stage designs, implying that β is underestimated under the single-stage design. Two-stage designs have advantages over single-stage designs because they allow the early evaluation of the plausibility of the null hypothesis and can save up to 3 patients compared to the single-stage design. It is noteworthy that the critical values, r_1 and r, are the minimum number of adverse events or toxicity leading to a rejection of a null hypothesis different from the efficacy designs. As an example, suppose $0.4 = p_0 > p_1 = 0.2$ and the minimax design is used. We then terminate the study at the end of stage 1 if 5 or more toxicities occur out of 13 patients for excess toxicities. Otherwise, the study proceeds to the second stage and enrolls additional 22 patients. If 10 or more toxicities occur out of 35 patients at the end of the second stage, we conclude that the experimental therapy has excess toxicities. Otherwise, the experimental therapy is considered for further investigation in subsequent trials.

Table 2 displays the simulated single-stage and two-stage designs when $\alpha = 0.10$ and $\beta = 0.2$. Similarly, as observed when $\alpha = 0.05$ in Table 1, the required total sample sizes (n) for the minimax two-stage designs are always smaller than the single-stage designs. However, the optimal two-stage designs require the same or larger total sample sizes than single-stage designs except when $p_0 = 0.3$. One interesting case is when $p_0 = 0.9$ and all designs have the same required total sample size ($n = 18$) and the same rejection boundary ($r = 15$). Moreover, the rejection boundary ($r_1 = 7$) and the required sample size ($n_1 = 7$) at stage 1 are the same across the two two-stage designs, meaning that,

Table 2: Single-stage and two-stage (optimal and minimax) designs with $\alpha = 0.10$ and $\beta = 0.2$

p_0	p_1		r_1	n_1	r	n	$E(N\|p_0)$	PET(p_0)	$\hat{\alpha}$	$\hat{\beta}$
		Single			4	21			0.086	0.152
0.3	0.1	Optimal	2	6	4	20	11.88	0.58	0.090	0.194
		Minimax	3	15	4	19	15.51	0.87	0.092	0.199
		Single			7	24			0.096	0.189
0.4	0.2	Optimal	4	11	10	31	16.93	0.70	0.100	0.192
		Minimax	5	11	7	24	17.93	0.47	0.093	0.199
		Single			11	28			0.092	0.191
0.5	0.3	Optimal	6	12	13	32	19.74	0.61	0.090	0.195
		Minimax	8	15	11	28	21.50	0.50	0.090	0.199
		Single			15	30			0.097	0.175
0.6	0.4	Optimal	7	12	20	38	20.70	0.67	0.098	0.195
		Minimax	10	16	14	28	21.67	0.53	0.099	0.197
		Single			18	30			0.084	0.181
0.7	0.5	Optimal	10	15	20	32	19.73	0.72	0.100	0.196
		Minimax	9	12	17	28	20.12	0.49	0.095	0.198
		Single			17	24			0.089	0.192
0.8	0.6	Optimal	10	12	18	25	17.74	0.56	0.099	0.185
		Minimax	12	14	17	24	19.52	0.45	0.087	0.198
		Single			15	18			0.098	0.165
0.9	0.7	Optimal	7	7	15	18	12.74	0.48	0.089	0.200
		Minimax	7	7	15	18	12.74	0.48	0.089	0.200

PET = probability of early termination.

if all 7 patients experience toxicities at the end of the first stage, we terminate the study for excessive toxicities.

3.2. Stomatitis study

We revisit the stomatitis study conducted by Rugo *et al.* (2017). The purpose of that study was to assess the incidence of grade 2 or worse everolimus-related stomatitis in women who used dexamethasone mouthwash (SWISH). The study assumed that the incidence rate of grade 2 or worse stomatitis without SWISH was 0.33 (i.e., $p_0 = 0.33$) based on historical controls and then proposed that the use of SWISH would lead to an absolute reduction of 0.13 (i.e., $p_1 = 0.20$). We now use Corollaries 1 and 2 to construct a single-stage design and a two-stage design under the hypotheses that $0.67 = 1 - 0.33 = p_0 < p_1 = 1 - 0.20 = 0.80$ for specified type I and II error rates.

Suppose $\alpha = 0.05$ and $\beta = 0.20$. Since the optimal exact single-stage design with $0.67 = p_0 < p_1 = 0.80$ is $(r, n) = (55, 73)$, the optimal exact single-stage design with $0.33 = p_0 > p_1 = 0.20$ becomes $(r, n) = (18, 73)$ and the estimated $\hat{\alpha} = 0.047$ and $\hat{\beta} = 0.196$. By Corollary 1, The minimax and optimal two-stage designs with $0.67 = p_0 < p_1 = 0.80$ are $(r_1, n_1, r, n) = (50, 67, 54, 72)$ and $(18, 26, 63, 85)$, respectively. It follows that, by Corollary 2, the minimax and optimal two-stage designs with $0.33 = p_0 > p_1 = 0.20$ become $(r_1, n_1, r, n) = (50, 67, 54, 72)$ and $(18, 26, 63, 85)$, respectively. The estimated $\hat{\alpha}$ and $\hat{\beta}$ are 0.049 and 0.200 for the minimax design and 0.050 and 0.196 for the optimal design. Compared to the single-stage design, the two-stage designs have the flexibility to stop the trial early when there is evidence of excess adverse events. Consequently, patients can benefit from the two-stage designs by not having them exposed unnecessarily to treatments with undue adverse events. Therefore, employing the optimal two-stage design can assess the incidence of adverse events early as a form of an interim analysis at the end of stage 1 and protect patients. For example, if 18 patients experienced grade 2 or worse everolimus-related stomatitis out of 26 patients at the end of

stage 1, we can terminate the study for excessive adverse events and do not have to wait till the trial concludes.

4. Discussion

The primary objective of a phase II single-arm study is often to assess safety and/or tolerability of a certain drug or treatment by the incidence of adverse events or toxicity. A phase II single-arm safety study aims to show that the rate of an adverse event is lower in the experimental therapy than that in the historical control (i.e., $p_0 > p_1$). In contrast, a phase II single-arm efficacy study aims to show that the rate of having a responder is higher in experimental therapy than in the historical control (i.e., $p_0 < p_1$). This is one of the major differences between safety and efficacy studies.

Two-stage designs are widely used in phase II single-arm efficacy studies because of the flexibility to stop early due to futility and avoid the unnecessary exposure of patients to ineffective therapies. However, there appears to be no phase II single-arm safety studies that employ two-stage designs. This means that current safety studies have no opportunity to stop a trial early due to futility. One possible explanation may be a lack of theoretical justification and dedicated software. Our work shows that the traditional Simon's two-stage designs to evaluate efficacy in a single one-arm trial can also be used for a one or two-stage safety trial with $p_0 > p_1$. We presented their analytical formulas and established relationships between the two types of designs and investigators can now easily find single-stage or two-stage designs for a safety study using available designs for an efficacy study.

Jung *et al.* (2001) proposed a graphical approach to find a suboptimal two-stage design that is a trade-off between optimal and the minimax designs. This approach graphically searches for the suboptimal design solely based on the required total sample size (n) and the expected null sample size $E(N|p_0)$. It is not difficult to observe that Corollary 2 shows that optimal and minimax two-stage designs for safety hold for the same n and $E(N|p_0)$ as those for efficacy, implying that Corollary 2 is valid for suboptimal two-stage designs for safety and efficacy. Therefore, the graphical approach can be applied to find two-stage safety designs to find a suboptimal two-stage design.

It is noteworthy that the Bayesian optimal phase II (BOP2) design (Zhou *et al.*, 2017) can monitor the toxicity endpoint through the binary toxicity endpoint at http://www.trialdesign.org/. However, in order to implement the BOP2 design, a user needs to provide preselected sample sizes for the first and second stages, while the proposed approach requires a desirable power. The BOP2 design computes the achieved power based on fixed sample sizes (i.e., a post-hoc power calculation), while our proposed approach is to estimate the sample sizes based on a pre-specified power (i.e., the prior sample size estimation).

To facilitate practitioners to use our single-stage and two-stage designs for a phase II single-arm study, including Jung *et al.*'s graphical approaches, we implement our computer codes in an R package `gen2stage`. Using the `gen2stage` package, investigators can generate single-stage or two-stage designs for an efficacy study or a toxicity study whether we are concerned with $p_0 < p_1$ or $p_0 > p_1$.

Acknowledgements

The Biostatistics Core is supported, in part, by NIH Center Grant P30 CA022453 to the Karmanos Cancer Institute at Wayne State University. WKK was partially supported by NIH Grant R01GM1076 39. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

A'Hern RP (2001). Sample size tables for exact single-stage phase II designs, *Statistics in Medicine*, **20**, 859–866.

Bryant J and Day R (1995). Incorporating toxicity considerations into the design of two-stage phase ii clinical trials, *Biometrics*, **51**, 1372–1383.

Cai C, Liu S, and Yuan Y (2014). A Bayesian design for phase II clinical trials with delayed responses based on multiple imputation, *Statistics in Medicine*, **33**, 4017–4028.

Chen K and Shan M (2007). Optimal and minimax three-stage designs for phase II oncology clinical trials, *Contemporary Clinical Trials*, **28**, 32–41.

Chen TT (1997). Optimal three-stage designs for phase II cancer clinical trials, *Statistics in Medicine*, **16**, 2701–2711.

Conaway MR and Petroni GR (1996). Designs for phase ii trials allowing for a trade-off between response and toxicity, *Biometrics*, **52**, 1375–1386.

Ensign LG, Gehan EA, Kamen DS, and Thall PF (1994). An optimal three-stage design for phase II clinical trials, *Statistics in Medicine*, **13**, 1727–1736.

Fleming TR (1982). One-sample multiple testing procedure for phase ii clinical trials, *Biometrics*, **38**, 143–151.

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

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

Kim S and Wong WK (2017). Extended two-stage adaptive designs with three target responses for phase II clinical trials, *Statistical Methods in Medical Research*, **27**, 3628–3642.

Kwak M and Jung SH (2014). Phase II clinical trials with time-to-event endpoints: optimal two-stage designs with one-sample log-rank test, *Statistics in Medicine*, **33**, 2004–2016.

Lee JJ and Liu DD (2008). A predictive probability design for phase II cancer clinical trials, *Clinical Trials*, **5**, 93–106.

Mander AP and Thompson SG (2010). Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials, *Contemporary Clinical Trials*, **31**, 572–578.

Mander AP, Wason JM, Sweeting MJ, and Thompson SG (2012). Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis, *Pharmaceutical Statistics*, **11**, 91–96.

Ray HE and Rai SN (2011). An evaluation of a Simon 2-stage phase II clinical trial design incorporating toxicity monitoring, *Contemporary Clinical Trials*, **32**, 428–436.

Rugo HS, Seneviratne L, Beck JT, *et al.* (2017). Prevention of everolimus-related stomatitis in women with hormone receptor-positive, her2-negative metastatic breast cancer using dexamethasone mouthwash (SWISH): a single-arm, phase 2 trial, *Lancet Oncol*, **18**, 654–662.

Simon R (1989). Optimal two-stage designs for phase II clinical trials, *Controlled Clinical Trials*, **10**, 1–10.

Thall PF and Simon R (1994). Practical Bayesian guidelines for phase IIB clinical trials, *Biometrics*, **50**, 337–349.

Wason JM, Mander AP, and Eisen TG (2011). Reducing sample sizes in two-stage phase II cancer trials by using continuous tumour shrinkage end-points, *European Journal of Cancer*, **47**, 983–989.

Zhou H, Lee JJ, and Yuan Y (2017). BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints, *Statistics in Medicine*, **36**, 3302–3314.

Received August 21, 2018; Revised December 16, 2018; Accepted January 9, 2019