Bayesian Analysis (2017) **12**, Number 1, pp. 31–51

Bayesian Two-Stage Design for Phase II Clinical Trials with Switching Hypothesis Tests

Haolun Shi* and Guosheng Yin†

Abstract. Conventional phase II clinical trials use either a single-arm or a double-arm scheme to examine the treatment effect of an investigational drug. The hypotheses tests under these two schemes are different, as a single-arm study usually tests the response rate of the new drug against a set of fixed reference rates and a double-arm randomized trial compares the new drug with the standard treatment or placebo. To bridge the single- and double-arm schemes in one phase II clinical trial, we propose a Bayesian two-stage design with changing hypothesis tests. Stage 1 enrolls patients solely to the experimental arm to make a comparison with the reference rates, and stage 2 imposes a double-arm comparison of the experimental arm with the control arm. The design is calibrated with respect to error rates from both the frequentist and Bayesian perspectives. Moreover, we control the "type III error rate", defined as the probability of prematurely stopping the trial at stage 1 when the trial is supposed to move on to stage 2. We conduct extensive simulations on the calculations of these error rates to examine the operational characteristics of our proposed method, and illustrate it with a non-small cell lung cancer trial.

MSC 2010 subject classifications: Primary 62C10; secondary 62P10.

Keywords: Bayesian error rates, expected sample size, phase II clinical trial, single-to-double arm design, two-stage procedure, type I error, type II error.

1 Introduction

As the proof-of-concept stage of drug development, phase II trials focus on the evaluation of the new agent's therapeutic effects, screening out unpromising drugs and carrying the promising ones forward to confirmative phase III trials. Many statistical methods have been developed for phase II trial designs. Gehan (1961) suggested a two-stage design with the provision of stopping the trial early for futility if there is no response observed in the first stage. Fleming (1982) proposed multiple testing procedures for phase II clinical trials. Simon et al. (1985) discussed sample sizes for selection designs with response endpoints in randomized phase II trials. Chang et al. (1987) studied group sequential methods and suggested minimizing the average expected sample size under the null and alternative hypotheses in phase II trials. Sylvester (1988) introduced a Bayesian approach to phase II trial designs on he basis of loss functions. Simon (1989) proposed an optimal and a minimax two-stage design by controlling the type I and

*Department of Statistics and Actuarial Science, The University of Hong Kong, 91 Pokfulam Road, Hong Kong

†Department of Statistics and Actuarial Science, The University of Hong Kong, 91 Pokfulam Road, Hong Kong, gyin@hku.hk

 DOI: 10.1214/15-BA988

type II error rates under the frequentist hypothesis testing framework. To address the patient accrual problem, Green and Dahlberg (1992) developed phase II designs to allow for variable attained sample sizes. In the Bayesian paradigm, Thall and Simon (1994) provided useful framework for continuously assessing the trial outcomes based on posterior probabilities in single-arm phase II trials. Chen and Ng (1998) proposed a flexible design by optimizing the expected sample size under an uninteresting response rate.

Recent years have witnessed vast development in the statistical theories with applications to phase II clinical trial designs. In particular, Lee and Zelen (2000) introduced the concept of Bayesian posterior error rates and recommended that the control of error rates should be conditional on the trial outcomes. Steinberg and Venzon (2002) proposed an early selection approach to randomized phase II trials. Tan and Machin (2002) proposed two Bayesian two-stage designs for phase II clinical trials where the decisions are based on the posterior distribution of the true response proportion. As extensions, Mayo and Gajewski (2004) considered the cases where the prior distribution is informative and provided methods for sample size calculation; and Sambucini (2008) proposed a predictive version of the Bayesian two-stage phase II design. Wang et al. (2005) introduced a Bayesian single-arm design by considering both frequentist and Bayesian error rates. Similar to the continuously monitoring scheme proposed by Thall and Simon (1994) which adopts decision boundaries on posterior probabilities, Lee and Liu (2008) studied posterior predictive probability monitoring rules for single-arm phase II trials. Liu et al. (2010) modified Simon's two-stage design (Simon, 1989) using beta–binomial distributions and presented some asymptotic conditions. Sambucini (2010) suggested a Bayesian predictive strategy in a two-stage phase II trial to adapt the sample size based on the data in the first stage. In an extension to randomized phase II studies, Yin et al. (2011) bridged predictive probability monitoring and adaptive randomization, and provided a detailed comparison with group sequential methods in a two-arm trial. Dong et al. (2012) proposed a two-stage design with control of both frequentist and Bayesian error rates. Inoue et al. (2002) developed a seamless phase II/III design where the discrete outcomes in phase II and the survival times in phase III can be combined together. Lai et al. (2012) studied cancer trial designs based on modeling the bivariate endpoints of tumor response and survival, and developed likelihood ratio statistics for such a model under the group sequential framework. Posch et al. (2005) proposed a design that seamlessly integrates a selection phase and a confirmation phase into a single trial.

All the aforementioned designs use either a single-arm or two-arm comparison to examine the drug's therapeutic effects. Both single-arm and multi-arm evaluations have their own merits, hence they are implemented according to the practical situations. When there is no standard therapy and placebo cannot serve as a control due to ethical considerations, it is rational to conduct a single-arm trial. Moreover, since there is no randomization in a single-arm trial, it is easier to establish hypothesis testing under a one-sample case, and is more convenient to conduct such a study. Nevertheless, in reality, many seemingly promising drugs eventually fail in phase III trials even though they have shown potential efficacious effects in phase II trials. Apart from the fact that the endpoints used in phase II trials are typically different from those of phase III

trials, one of the main reasons for such failures is that the experimental drug is merely compared with the standard response rate or historical data in a single-arm setting. Although single-arm trials are inherently comparative, they are less objective and can be biased due to many differences between the current and previous studies, such as patient populations, study criteria, and medical facilities. To overcome these problems, a randomized two-arm phase II trial is often preferred when a standard treatment is available.

It is a common situation that patients enrolled in a multi-arm trial tend to be more willing to take the experimental drug instead of the control, particularly for those with advanced or refractory diseases, because no standard treatment had worked and the trial's experimental drug could be the last hope. The current practice is to conduct a single-arm phase IIa trial and a randomized phase IIb trial separately without any information borrowing across the two studies. For time saving and information sharing, we propose a Bayesian two-stage single-to-double arm design with a single-arm comparison of the experimental drug with the standard response rate (no concurrent treatment) in stage 1 and a two-arm comparison of the experimental drug with the standard of care in stage 2. Not only does such a design eliminate the gap between the conventional phase IIa and phase IIb trials, it also help to pool patients together from separate trials for better decision making.

The rest of this article is organized as follows. In Section 2, we describe the Bayesian single-to-double arm transition design and derive the frequentist and Bayesian error rates. Calibration of design parameters and several simulation studies are presented in Section 3. Section 4 illustrates the proposed design with a lung cancer trial, and Section 5 concludes with some remarks. The R code of our proposed design is available upon request.

2 Bayesian Two-Stage Design

2.1 Single-to-Double Arm Transition

We are interested in testing the response rate on binary outcomes of an experimental treatment versus the standard treatment. The outcome takes a value of 1 when a response is observed and 0 otherwise; for example, whether there is tumor shrinkage after treatment. In stage 1, which is a single-arm trial, we compare the experimental drug with a standard response rate. The null and alternative hypotheses are formulated as

$$H_0 : \theta_E \leq \theta_0 \quad \text{versus} \quad H_1 : \theta_E \geq \theta_1,$$

where θ_E is the response rate of the experimental drug, θ_0 is the maximum uninteresting response rate, and θ_1 is the minimum response rate of clinical interest. This hypothesis is commonly adopted in single-arm trial designs, which sets the boundary for a clinically uninteresting rate; see Simon (1989), Thall and Simon (1995), Mariani and Marubini (1996), and Yin (2012) for a more detailed exposition of the subject. In the first stage, n_1 patients are enrolled, and suppose there are x_1 responses, then $x_1|\theta_E \sim \text{Bin}(n_1, \theta_E)$, where $\text{Bin}(n, \theta)$ denotes the binomial distribution with the success probability θ. Similar

to the two-stage design proposed by Dong et al. (2012), early stopping for efficacy or futility is allowed at the end of stage 1 and the probability of early termination (PET) equals to the sum of the probability for efficacy stopping and that for futility stopping. Let l_1 denote the lower bound for accepting the null hypothesis, and u_1 denote the upper bound for rejecting the null hypothesis, then the decision rules at the end of stage 1 are described as follows:

(i) If $x_1 \leq l_1$, stop the trial and claim the experimental drug unpromising.

(ii) If $x_1 \geq u_1$, stop the trial and claim the experimental drug promising.

(iii) Otherwise, the trial proceeds to stage 2 where a total number of $2n_2$ patients are equally allocated to the experimental and standard arms.

The expected sample size (ESS) of our proposed design is

$$\text{ESS} = n_1 + 2n_2(1 - \text{PET}).$$

In stage 2, which is a two-arm trial, we examine the superiority of the experimental drug compared with a standard treatment. The testing hypotheses change to

$$H_0 : \theta_E \leq \theta_S \quad \text{versus} \quad H_1 : \theta_E > \theta_S,$$

where θ_E is the same response rate of the experimental drug in stage 1, and θ_S is that of the standard treatment. It is possible to formulate the hypotheses as $H_0 : \theta_E \leq \theta_S + \delta \quad \text{versus} \quad H_1 : \theta_E > \theta_S + \delta$, where δ is the minimal clinically meaningful difference in the response rate. When $\delta > 0$, the design usually requires a larger sample size as the test becomes more stringent on the experimental response rate. For simplicity, we set $\delta = 0$. Suppose that in stage 2, x_2 responses are observed in the experimental arm and y_2 responses in the standard arm, then $x_2|\theta_E \sim \text{Bin}(n_2, \theta_E)$ and $y_2|\theta_S \sim \text{Bin}(n_2, \theta_S)$. If we set the prior distributions as $\theta_E \sim \text{Beta}(a, b)$ and $\theta_S \sim \text{Beta}(c, d)$, the posterior distributions are

$$\begin{aligned} \theta_E|(x_1, x_2) &\sim \text{Beta}(a + x_1 + x_2, b + n_1 - x_1 + n_2 - x_2), \\ \theta_S|y_2 &\sim \text{Beta}(c + y_2, d + n_2 - y_2). \end{aligned}$$

As a result, the posterior probability (PoP) of $\theta_E > \theta_S$ can be written as

$$\begin{aligned} \text{PoP} \equiv P(\theta_E > \theta_S|x_1, x_2, y_2) &= \int_0^1 \int_{\theta_S}^1 P(\theta_E|x_1, x_2) P(\theta_S|y_2) d\theta_E d\theta_S \\ &= \int_0^1 \int_{\theta_S}^1 \frac{\theta_E^{a+x_1+x_2-1}(1-\theta_E)^{b+n_1-x_1+n_2-x_2-1}\theta_S^{c+y_2-1}(1-\theta_S)^{d+n_2-y_2-1}}{B(a + x_1 + x_2, b + n_1 - x_1 + n_2 - x_2)B(c + y_2, d + n_2 - y_2)} d\theta_E d\theta_S, \end{aligned} \tag{1}$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ represents the beta function with parameters a and b. Let c_T denote the cutoff probability for decision making, then the decision rules at the end of stage 2 are given as follows:

(i) If PoP $\geq c_T$, we reject the null hypothesis and claim the experimental drug promising.

(ii) Otherwise, we fail to reject the null and claim the experimental drug unpromising.

In summary, at the end of the first stage, the proposed design uses decision boundaries (u_1, l_1) to determine whether the trial should be continued or stopped for efficacy/futility. At the end of the second stage, to reach a final decision, the trial uses c_T as the cutoff value on a clinically meaningful quantity, PoP, calculated from the trial outcomes of both stages.

The design parameters of our proposed method, $(n_1, n_2, u_1, l_1, c_T)$, are determined by controlling error rates, in conjunction with the aim of minimizing the ESS of the trial. The error rates are calculated from both the frequentist and Bayesian perspectives. In the frequentist framework, we control the commonly used type I and II error rates (Yin, 2012). In addition, to fine-tune the transition between the two hypothesis tests across two stages, we control the frequentist type III error rate, defined as the probability of prematurely stopping the trial at stage 1 when the decision is supposed to move on to stage 2. Under the Bayesian framework, the error rates are derived using two approaches. One is based on Bayesian marginal probabilities conditional on an assumed truth, whereas the other is based on Bayesian posterior probabilities conditional on the action of rejecting or accepting the null hypothesis.

2.2 Frequentist Error Rates

Frequentist type I and II error rates refer to the probability of declaring efficacy given the null is true and the probability of declaring futility given the alternative is true, respectively. We declare efficacy when the number of responders reaches the efficacy stop u_1 in stage 1 or when the calculated PoP is larger than c_T in stage 2, and declare futility when the number of responders reaches the futility stop l_1 in stage 1 or when the calculated PoP is smaller than c_T in stage 2.

Frequentist stage 1 error rates. Let R and A represent "rejecting the null hypothesis" and "accepting the null hypothesis", and let α_1^f and β_1^f denote frequentist type I and type II error rates at stage 1, respectively. Given $x_1|\theta_E \sim \text{Bin}(n_1, \theta_E)$ and the lower and upper bounds l_1 and u_1, we have

$$\alpha_1^f = P(\text{R at stage 1}|H_0) = \sum_{x_1=u_1}^{n_1} P(x_1|\theta_0) = 1 - F_{\text{Bin}}(u_1 - 1; n_1, \theta_0),$$

$$\beta_1^f = P(\text{A at stage 1}|H_1) = \sum_{x_1=0}^{l_1} P(x_1|\theta_1) = F_{\text{Bin}}(l_1; n_1, \theta_1),$$

where F_{Bin} denotes the binomial cumulative distribution function (CDF).

Frequentist stage 2 error rates. If x_1 lies between l_1 and u_1, the trial proceeds to stage 2, where a total number of $2n_2$ patients are equally allocated to the experimental

and standard arms. Suppose that we observe x_2 responses among n_2 patients in the experimental arm, and y_2 responses in the standard arm, then $x_2|\theta_E \sim \text{Bin}(n_2, \theta_E)$ and $y_2|\theta_S \sim \text{Bin}(n_2, \theta_S)$. In consistence with the settings at stage 1, we specify $\theta_E = \theta_S = \theta_0$ under the null hypothesis, and $\theta_E = \theta_1$ and $\theta_S = \theta_0$ under the alternative hypothesis. The frequentist type I and type II error rates at stage 2 are respectively given by

$$\alpha_2^f = \sum_{x_1=l_1+1}^{u_1-1} \sum_{x_2=0}^{n_2} \sum_{y_2=0}^{n_2} P(x_1|\theta_0)P(x_2|\theta_0)P(y_2|\theta_0)I(\text{PoP} \geq c_T),$$

$$\beta_2^f = \sum_{x_1=l_1+1}^{u_1-1} \sum_{x_2=0}^{n_2} \sum_{y_2=0}^{n_2} P(x_1|\theta_1)P(x_2|\theta_1)P(y_2|\theta_0)I(\text{PoP} < c_T),$$

where $I(\cdot)$ is the indicator function, and $I(\text{PoP} \geq c_T)$ serves as a "filter" to select all the possible values of x_1, x_2 and y_2 at stage 2 that lead to rejection of the null hypothesis. The overall frequentist type I and type II error rates are $\alpha^f = \alpha_1^f + \alpha_2^f$ and $\beta^f = \beta_1^f + \beta_2^f$, respectively.

Frequentist type III error rates. When $\theta_E < \theta_0$ or $\theta_E > \theta_1$, the trial should be terminated with a conclusive decision at the end of stage 1, instead of further continuing into the randomization stage. When $\theta_0 < \theta_E < \theta_1$, we expect the number of responses in stage 1 to lie between l_1 and u_1. Due to the uncertainty about the superiority of the new treatment, we prefer further confirmation through a randomized study. To accommodate this unconventional phenomenon when combining two stages, we define the failure to move on to the double-arm stage when $\theta_0 < \theta_E < \theta_1$ as the type III error (Storer, 1992). As a result, the frequentist type III error rates can be formulated as $\gamma^f = P(\text{R} \cup \text{A at stage 1}|\theta_0 < \theta_E < \theta_1)$, and if we specify $\theta_E = \theta_m = (\theta_0 + \theta_1)/2$, then $\gamma^f = \gamma_R^f + \gamma_A^f$, where

$$\gamma_R^f = P(\text{R at stage 1}|\theta_E = \theta_m) = \sum_{x_1=u_1}^{n_1} P(x_1|\theta_m) = 1 - F_{\text{Bin}}(u_1 - 1; n_1, \theta_m),$$

$$\gamma_A^f = P(\text{A at stage 1}|\theta_E = \theta_m) = \sum_{x_1=0}^{l_1} P(x_1|\theta_m) = F_{\text{Bin}}(l_1; n_1, \theta_m).$$

2.3 Bayesian Error Rates

Motivated by the work of Lee and Zelen (2000) and Dong et al. (2012), we define the Bayesian type I and type II error rates as

$$\alpha^B = P(\text{R}|H_0) = \frac{P(\text{R} \cap H_0)}{P(H_0)},$$

$$\beta^B = P(\text{A}|H_1) = \frac{P(\text{A} \cap H_1)}{P(H_1)},$$

which correspond to the prior probabilities of rejecting and accepting the null hypothesis given the null and alternative hypotheses being true, respectively. Lee and Zelen (2000)

estimated $P(H_0)$ and $P(H_1)$ from historical data, while we calculate them based on the prior distributions of parameters at each stage.

Lee and Zelen (2000) pointed out that conventional type I and type II error rates may be inadequate to quantify trial results in practice. They suggested the Bayesian posterior false positive and false negative rates by conditioning on trial outcomes,

$$\alpha^* = P(H_0|\mathrm{R}) = \frac{P(\mathrm{R} \cap H_0)}{P(\mathrm{R})},$$

$$\beta^* = P(H_1|\mathrm{A}) = \frac{P(\mathrm{A} \cap H_1)}{P(\mathrm{A})}.$$

Bayesian stage 1 settings. Let α_1^B, β_1^B, α_1^* and β_1^* denote the Bayesian type I and type II error rates, the Bayesian posterior false positive and false negative rates at stage 1, respectively; that is,

$$\begin{aligned}
\alpha_1^B &= P(\mathrm{R} \text{ at stage } 1|H_0 \text{ at stage } 1),\\
\beta_1^B &= P(\mathrm{A} \text{ at stage } 1|H_1 \text{ at stage } 1),\\
\alpha_1^* &= P(H_0 \text{ at stage } 1|\mathrm{R} \text{ at stage } 1),\\
\beta_1^* &= P(H_1 \text{ at stage } 1|\mathrm{A} \text{ at stage } 1).
\end{aligned}$$

In the first stage, $x_1|\theta_E \sim \mathrm{Bin}(\theta_E, n_1)$, and under a beta prior distribution, $\theta_E \sim \mathrm{Beta}(a, b)$, the marginal distribution of x_1 is a beta–binomial distribution,

$$\begin{aligned}
P(x_1) &= \int_0^1 \left(\begin{array}{c} n_1 \\ x_1 \end{array} \right) {\theta_E}^{x_1} (1-\theta_E)^{n_1 - x_1} \frac{\theta_E^{a-1}(1-\theta_E)^{b-1}}{B(a,b)} d\theta_E \\
&= \left(\begin{array}{c} n_1 \\ x_1 \end{array} \right) \frac{B(a + x_1, b + n_1 - x_1)}{B(a,b)}.
\end{aligned}$$

Based on the marginal distribution of x_1, we can derive the formulae of α_1^B, β_1^B, α_1^* and β_1^*, which are given in the Supplementary Material (Shi and Yin, 2015).

Bayesian stage 2 settings. Let α_2^B, β_2^B, α_2^* and β_2^* denote the Bayesian type I and type II error rates, Bayesian posterior false positive and false negative rates at stage 2, respectively; that is,

$$\begin{aligned}
\alpha_2^B &= P(\mathrm{R} \text{ at stage } 2|H_0 \text{ at stage } 2),\\
\beta_2^B &= P(\mathrm{A} \text{ at stage } 2|H_1 \text{ at stage } 2),\\
\alpha_2^* &= P(H_0 \text{ at stage } 2|\mathrm{R} \text{ at stage } 2),\\
\beta_2^* &= P(H_1 \text{ at stage } 2|\mathrm{A} \text{ at stage } 2).
\end{aligned}$$

For the experimental arm, the posterior predictive distribution of x_2 conditional on x_1 is given by

$$\begin{aligned}
P(x_2|x_1) &= \int_0^1 P(x_2|\theta_E) P(\theta_E|x_1) d\theta_E \\
&= \left(\begin{array}{c} n_2 \\ x_2 \end{array} \right) \frac{B(a + x_1 + x_2, b + n_1 - x_1 + n_2 - x_2)}{B(a + x_1, b + n_1 - x_1)}.
\end{aligned}$$

For the standard arm, the marginal distribution of y_2 is also beta–binomial,

$$P(y_2) = \left(\begin{array}{c} n_2 \\ y_2 \end{array} \right) \frac{B(c + y_2, d + n_2 - y_2)}{B(c, d)}.$$

Detailed derivation of α_2^B, β_2^B, α_2^* and β_2^* are given in the Supplementary Material.

Calibration of Bayesian error rates. As stage 1 and stage 2 are integrated in one trial, α_2^B, β_2^B, α_2^* and β_2^* need to be calibrated with respect to the stage 1 condition. For instance, if we assume that "H_0 is true", then the entire trial must satisfy

$$\begin{aligned} &P(\text{R at stage 1}|H_0 \text{ at stage 1}) + P(\text{A at stage 1}|H_0 \text{ at stage 1}) \\ &+ P(\text{R at stage 2}|H_0 \text{ at stage 2}) + P(\text{A at stage 2}|H_0 \text{ at stage 2}) = 1, \end{aligned}$$

which means that, if H_0 is true, the probabilities of rejecting and accepting the null hypothesis at stage 1 and stage 2 should be summed up to 1. However, stage 2 has been treated as a separate trial in the derivation of Bayesian error rates. Therefore, the Bayesian stage 2 type I error rate should be recalibrated by multiplying

$$\begin{aligned} &P(\text{trial proceeds to stage 2}|H_0 \text{ at stage 1}) \\ &= P(l_1 + 1 \le x_1 \le u_1 - 1 \mid \theta_E \le \theta_0) \\ &= \frac{\sum_{x_1=l_1+1}^{u_1-1} \left\{ \left(\begin{array}{c} n_1 \\ x_1 \end{array} \right) \frac{B(a + x_1, b + n_1 - x_1)}{B(a, b)} F_{\text{Beta}}(\theta_0; a + x_1, b + n_1 - x_1) \right\}}{F_{\text{Beta}}(\theta_0; a, b)}, \end{aligned}$$

where F_{Beta} represents the CDF of a beta distribution. As a result, the calibrated Bayesian type I error rate at stage 2 is given by

$$\alpha_{2c}^B = P(l_1 + 1 \le x_1 \le u_1 - 1 \mid \theta_E \le \theta_0)\alpha_2^B.$$

As shown in the Supplementary Material, the calibrated Bayesian type II error rate, the Bayesian posterior false positive and false negative rates at stage 2, β_{2c}^B, α_{2c}^* and β_{2c}^*, can be derived similarly. The overall Bayesian error rates and Bayesian posterior false rates of the trial are given by

$$\begin{cases} \alpha^B = \alpha_1^B + \alpha_{2c}^B, & \beta^B = \beta_1^B + \beta_{2c}^B; \\ \alpha^* = \alpha_1^* + \alpha_{2c}^*, & \beta^* = \beta_1^* + \beta_{2c}^*. \end{cases}$$

2.4 Determination of Design Parameters

In the design stage, we need to specify θ_0, θ_1, and the type I, II and III error rates α, β and γ. As defaults, we use a noninformative $\text{Beta}(\theta_0, 1 - \theta_0)$ prior for θ_E and θ_S and set $c_T = 1 - \alpha$. To avoid the ambiguity on minimizing the expected sample size under the null or the alternative hypothesis, we minimize the Bayesian expected sample size,

$$\text{ESS}^B = n_1 + 2n_2(1 - \text{PET}^B),$$

subject to the error rate constraints on $(\alpha^f, \beta^f, \gamma^f, \alpha^B, \beta^B, \alpha^*, \beta^*)$, where PET^B denotes the Bayesian probability of early termination, and is given by

$$\text{PET}^B = \sum_{x_1=u_1}^{n_1} P(x_1) + \sum_{x_1=0}^{l_1} P(x_1).$$

We develop an enumeration algorithm to find the design parameters n_1, n_2, u_1, and l_1, as described below.

(i) Set n_1 from 10 to $[N/2]$, the integer part of $N/2$, where N is the required sample size in each arm for the standard two-arm randomized design with the type I error rate α and power $1-\beta$.

(ii) Given n_1, find the combinations of (l_1, u_1) satisfying $F_{\text{Bin}}(l_1; n_1, \theta_1) < \beta$.

(iii) Given n_1, l_1 and u_1, find n_2 that satisfies the constraints on the type I error rate and power.

(iv) Enumerate all possible values of the aforementioned (n_1, n_2, u_1, l_1) and choose the set that satisfies the specified constraints on error rates, and finally select the one that minimizes ESS^B.

We can speed up the numerical search by utilizing the approximately monotonic relationship between n_2 and power to formulate a bisectional search, and only perform enumeration in the neighborhood of the solution to the bisectional search to find the smallest n_2.

2.5 Commensurate Prior

To effectively control the type I error rate, typically we assume noninformative prior for θ_S. Hobbs et al. (2011, 2012) proposed a class of commensurate prior distributions that can borrow strength from historical trials depending on the exchangeability between historical and current data. When there exist historical data containing information on the efficacy of the standard drug, we may consider using the commensurate prior for θ_S. Let θ_{S_0} denote the response rate for the control arm in the historical trial. We adopt the probit link function $\theta_S = \Phi(\eta)$ and $\theta_{S_0} = \Phi(\eta_0)$, where $\Phi(\cdot)$ denotes the CDF of the standard normal distribution, such that the transformed parameters η and η_0 have support on the real line. Let x_H denote the number of responders among n_H subjects in the historical trial. Based on the historical data, we formulate the joint commensurate prior of (η, η_0, τ) as

$$\pi(\eta, \eta_0, \tau | x_H) \propto \Phi(\eta_0)^{x_H} \left\{1 - \Phi(\eta_0)\right\}^{n_H - x_H} \times \phi(\eta | \eta_0, \tau^{-1}) \times p(\tau),$$

where $\phi(\cdot|\mu, \sigma^2)$ denotes a normal density with mean μ and variance σ^2 and $p(\tau)$ a gamma density with mean $\tilde{\tau}$ and variance $\tilde{\tau}/c$, i.e., $\tau \sim \text{Gamma}(c\tilde{\tau}, c)$.

Let x_C denote the number of responders among n_C subjects in the control arm, and x_E that among n_E subjects in the experimental arm of the current trial. Under the commensurate prior, the joint posterior distribution of (η, η_0, τ) is given by

$$q(\eta, \eta_0, \tau | x_C, x_H) \propto \Phi(\eta)^{x_C} \{1 - \Phi(\eta)\}^{n_C - x_C} \times \pi(\eta, \eta_0, \tau | x_H).$$

The marginal posterior distribution of η can be obtained by integrating out η_0 and τ,

$$q(\eta | x_C, x_H) \propto \Phi(\eta)^{x_C} \big\{1 - \Phi(\eta)\big\}^{n_C - x_C} \times \int_{-\infty}^{+\infty} \left\{ \frac{(\eta - \eta_0)^2}{2} + c \right\}^{-c\tilde{\tau} - 1/2} \Phi(\eta_0)^{x_H} \big\{1 - \Phi(\eta_0)\big\}^{n_H - x_H} d\eta_0.$$

Under a beta prior for the experimental response rate, $\theta_E \sim \text{Beta}(a, b)$, the posterior probability of $\theta_E > \theta_S$ is given by

$$\begin{aligned} \text{PoP} &\equiv P(\theta_E > \theta_S | x_E, x_C, x_H) \\ &= \int_{-\infty}^{+\infty} \big\{1 - F_{\text{Beta}}(\Phi(\eta); a + x_E, b + n_E - x_E)\big\} q(\eta | x_C, x_H) d\eta. \end{aligned}$$

where F_{Beta} denotes the CDF of a beta distribution.

For computational simplicity, we can formulate the commensurate prior by plugging in the historical mean $\hat{\eta}_0 = \Phi^{-1}(x_H / n_H)$, so that it solely depends on η and τ,

$$\pi(\eta, \tau | x_H) \propto \phi\big(\eta | \hat{\eta}_0, \tau^{-1}\big) \times p(\tau).$$

Thus, the marginal prior distribution of η has an explicit expression after integrating out τ,

$$\pi(\eta | x_H) \propto \left\{ \frac{(\eta - \hat{\eta}_0)^2}{2} + c \right\}^{-c\tilde{\tau} - 1/2} \tag{2}$$

where we constrain $c\tilde{\tau} \geq 1/2$ to attain a proper prior distribution. It is worth noting that when $c\tilde{\tau} = 1/2$, (2) reduces to a Cauchy distribution. As a result, the posterior distribution of η is given by

$$q(\eta | x_C, x_H) \propto \Phi(\eta)^{x_C} \big\{1 - \Phi(\eta)\big\}^{n_C - x_C} \left\{ \frac{(\eta - \hat{\eta}_0)^2}{2} + c \right\}^{-c\tilde{\tau} - 1/2}.$$

In a single-to-double arm design, we set $x_C = y_2$, $x_E = x_1 + x_2$, $n_C = n_2$ and $n_E = n_1 + n_2$.

3 Simulation Studies

In the simulation study, we examine three paired values of $(\theta_0, \theta_1) = (0.2, 0.4), (0.3, 0.5)$ and $(0.4, 0.6)$, and we set the type I and type II error rates as $\alpha = 0.05$ and $\beta = 0.2$. For each pair of (θ_0, θ_1), we calibrate the optimal design parameters under different

Table 1: The optimal single-to-double arm design for different values of (θ_0, θ_1) and γ, assuming a noninformative prior Beta$(\theta_0, 1-\theta_0)$ for θ_S and θ_E, with all error rates in percentage.

θ_0	θ_1	γ	n_1	n_2	l_1	u_1	α^f	β^f	γ^f	α^B	β^B	α^*	β^*	PET^B	ESS^B
0.2	0.4	0.2	15	55	2	8	4.88	19.42	17.68	0.04	2.49	0.06	0.73	0.83	33.5
		0.3	13	59	2	7	4.49	19.92	26.49	0.05	2.78	0.10	1.11	0.85	30.5
		0.4	17	51	3	8	4.76	19.76	30.65	0.05	1.87	0.13	1.10	0.88	29.1
		0.5	15	59	3	7	4.66	19.93	42.80	0.08	2.23	0.23	1.52	0.90	26.5
0.3	0.5	0.2	14	65	3	9	4.97	19.92	18.26	0.06	2.76	0.10	0.74	0.80	39.6
		0.3	19	60	5	11	4.78	19.56	25.14	0.06	1.86	0.11	0.90	0.85	36.5
		0.4	20	59	6	11	4.97	19.91	37.75	0.08	1.67	0.17	1.31	0.89	32.9
		0.5	20	59	6	11	4.97	19.91	37.75	0.08	1.67	0.17	1.31	0.89	32.9
0.4	0.6	0.2	20	63	7	14	4.66	19.96	18.92	0.05	2.25	0.06	0.78	0.82	42.5
		0.3	21	61	8	14	4.69	19.72	28.63	0.07	1.80	0.11	1.08	0.86	38.2
		0.4	21	61	8	14	4.69	19.72	28.63	0.07	1.80	0.11	1.08	0.86	38.2
		0.5	26	52	11	16	4.93	19.69	44.21	0.10	1.22	0.17	1.49	0.91	35.5

Note that (θ_0, θ_1) are hypothesis testing parameters, γ is the specified type III error rate, PET^B is the Bayesian probability of early termination, and ESS^B is the Bayesian expected sample size.

specifications of the type III error rate γ, ranging from 0.2 to 0.5. We use noninformative prior distributions for θ_E, Beta$(\theta_0, 1-\theta_0)$, and set $c_T = 1-\alpha$.

Tables 1 and 2 present the simulation results under different design settings assuming noninformative Beta$(\theta_0, 1-\theta_0)$ priors and commensurate priors for θ_S, respectively. Under each combination of (θ_0, θ_1), as the constraint on γ becomes more stringent, the expected sample size of the optimal design becomes larger. A larger probability of early termination corresponds to a smaller expected sample size. The designs that assume commensurate priors have smaller Bayesian expected sample sizes than the designs that assume noninformative priors, as the historic information incorporated in the commensurate prior may contribute to a saving in sample size and provide us more confidence to terminate the trial early when the number of responses is unfavorable. It is observed that the Bayesian error rates are much smaller than their frequentist counterparts. This is certainly expected because the Bayesian error rate control is essentially equivalent to sampling θ_E and θ_S from their prior distributions and then average the type I and type II error rates over the entire parameter space. On the other hand, frequentist approaches ensure that the suprema of the type I and type II error rates do not exceed the respective cutoffs in an asymptotic sense. As a numerical example, consider the case with $\theta_S = 0.2$; the frequentist procedure aims to control the probability of rejection when $\theta_E = \theta_S$ below $\alpha = 0.05$, so that when $\theta_E < \theta_S$ the rejection probability would be even smaller than α. In fact, as θ_E becomes smaller than θ_S, the type I error rate diminishes rapidly; for example, fixing $\theta_S = 0.2$, when θ_E takes the value of 0.15, 0.1, and 0.05, the corresponding type I error rates are 0.01, 0.001 and

Table 2: The optimal single-to-double arm design for different values of (θ_0, θ_1) and γ, assuming a commensurate prior for θ_S and a noninformative Beta$(\theta_0, 1-\theta_0)$ prior for θ_E, with all error rates in percentage.

θ_0	θ_1	γ	n_1	n_2	l_1	u_1	α^f	β^f	γ^f	PET^B	ESS^B
0.2	0.4	0.2	11	55	1	6	4.85	19.79	19.12	0.81	32.0
		0.3	14	54	2	7	4.97	19.78	25.41	0.85	29.7
		0.4	12	61	2	6	4.92	19.99	37.07	0.88	26.9
		0.5	19	46	4	8	4.79	19.95	46.42	0.92	26.3
0.3	0.5	0.2	14	66	3	9	4.48	19.70	18.26	0.85	33.7
		0.3	12	76	3	8	4.19	19.75	28.26	0.87	32.0
		0.4	11	84	3	7	4.76	19.95	39.56	0.89	29.0
		0.5	16	74	5	9	4.85	19.83	47.11	0.92	27.2
0.4	0.6	0.2	13	70	4	10	4.79	19.66	17.96	0.86	32.4
		0.3	17	66	6	12	4.70	19.83	23.79	0.89	31.0
		0.4	17	78	7	12	3.84	19.88	38.63	0.92	29.8
		0.5	14	105	6	10	3.93	19.90	48.50	0.93	29.4

The parameters for the prior distribution of τ are assumed as $c = \tilde{\tau} = 1$. For simplicity, computation of the Bayesian error rates is omitted and we adopt the plug-in method and center the historical mean at θ_0 in the formulation of the commensurate prior.

0.00001. Figure 1 shows the rejection probability as a function of θ_E given $\theta_S = 0.2$ for a standard two-arm design with $\alpha = 0.05$.

Figure 2 shows the operating characteristics of the optimal design with $(\theta_0, \theta_1) = (0.2, 0.4)$ and $\gamma = 0.2$. When the true θ_E is below θ_0, the probability of rejecting the stage 1 null hypothesis is α_1^f (solid line). When θ_E is between θ_0 and θ_1, the rejection probability is the type III error rate γ_R^f (dashed line) and its value is below γ when $\theta_E = (\theta_0 + \theta_1)/2$. The lower panel of Figure 2 shows that when the true θ_E is above θ_1, the probability of accepting the stage 1 null hypothesis is β_1^f (solid line). When θ_E is between θ_0 and θ_1, the probability of accepting the null hypothesis is the type III error rate γ_A^f (dashed line) and its value is below γ when $\theta_E = (\theta_0 + \theta_1)/2$.

Figure 3 shows the overall probabilities of rejection when $\theta_E = \theta_S$ and the overall probabilities of acceptance when $\theta_E = \theta_S + 0.2$ by comparing our design to Simon's two-stage design and a standard two-arm design. The probability of rejection and that of acceptance under our proposed design lie in between those of Simon's two-stage design and the standard two-arm design. As the true standard response rate exceeds θ_0, the probability of rejection increases immensely beyond α for Simon's two-stage design and the single-to-double arm design. This is because the hypothesis tests in Simon's design and the single-arm stage of our design compare the experimental response rate with θ_0, instead of θ_S, which may be different from θ_0. As discussed in Viele et al. (2014), this phenomenon is fundamental and inevitable because compared to the standard two-arm design, we essentially test a different set of hypotheses during the single-arm stage.

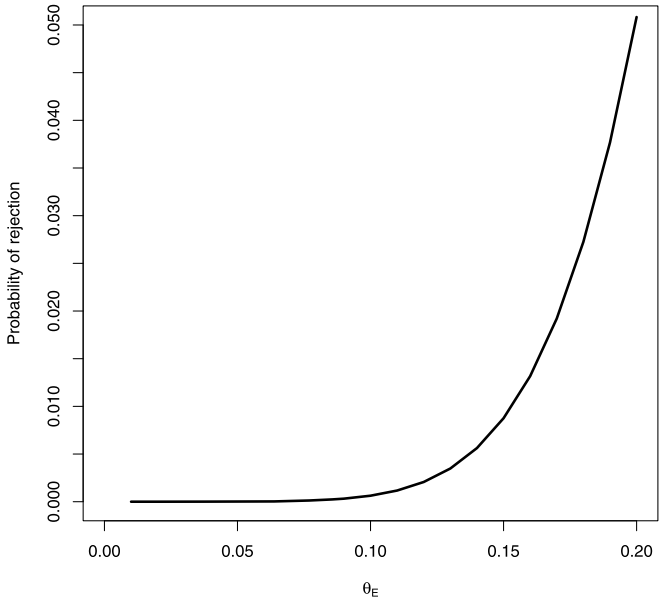
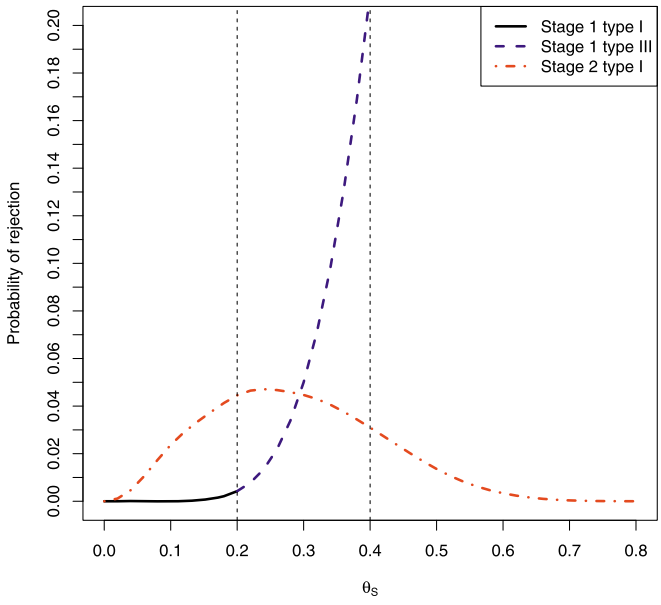


Figure 1: Rejection probability as a function of θ_E given $\theta_S = 0.2$ for a standard two-arm design with $\alpha = 0.05$.

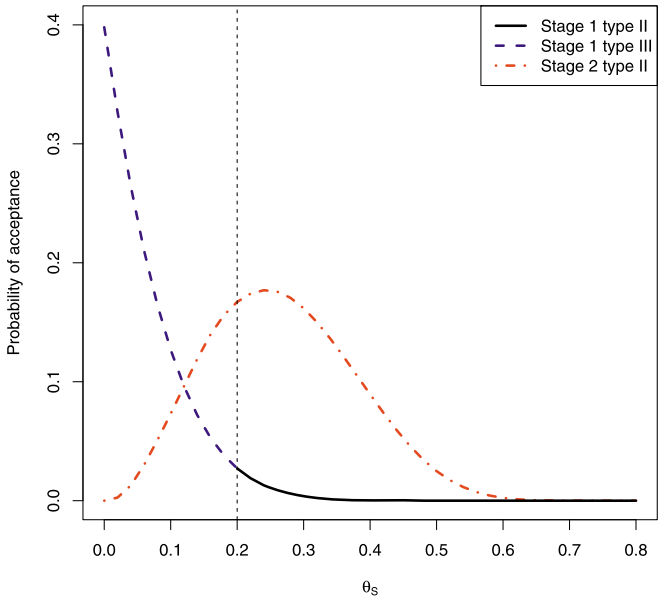
Throughout the simulation studies, we assume a noninformative prior for θ_S, if an informative commensurate prior is assumed, the operating characteristics of our design would be similar to those of a two-arm trial that incorporates historical data in the control arm (Viele et al., 2014).

Figure 4 shows the expected sample size as a function of θ_E for the three designs under comparison. It can be seen that the expected sample size of the single-to-double arm design lies in between that of Simon's two-stage design and the standard two-arm design. Since we prefer the trial to proceed into stage 2 when θ_E is in between θ_0 and θ_1, the expected sample size of the single-to-double arm design attains its maximum within this interval (around 0.3). When θ_E is very large, a single-to-double arm design can achieve a smaller expected sample size than Simon's two-stage design because of a high probability of early stopping.

Figure 5 shows the power, which is defined as the probability of declaring the experimental treatment promising in either the single-arm stage or the double-arm stage, given that the response rate of the experimental treatment is higher than that of the control by a required margin of 0.2. The sample size of the standard two-arm design is calculated using the standard formula to achieve power of 0.8 when $\theta_S = 0.2$ and $\theta_E = 0.4$. The design parameters for the single-to-double arm trial are calibrated similarly, so that the two power curves intersect at the point around $\theta_E = 0.4$ with power 0.8. The power curve of the standard two-arm design has a symmetric shape, whereas that of the single-to-double arm design exhibits an asymmetric pattern. For the stan-

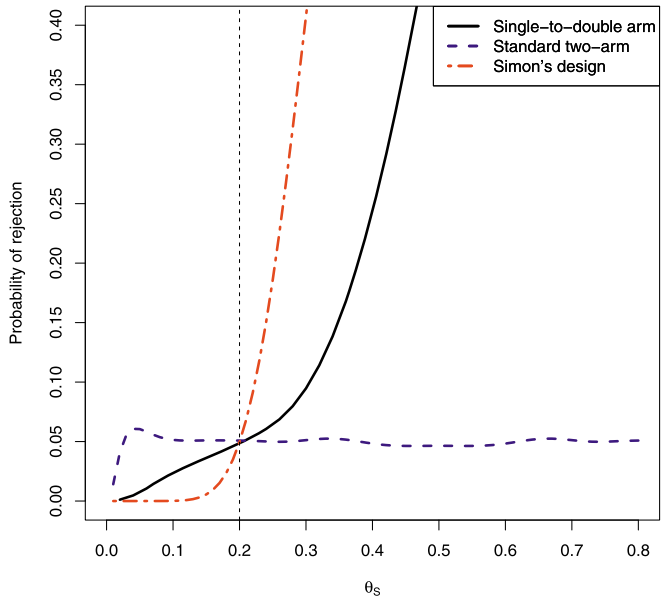


(a)

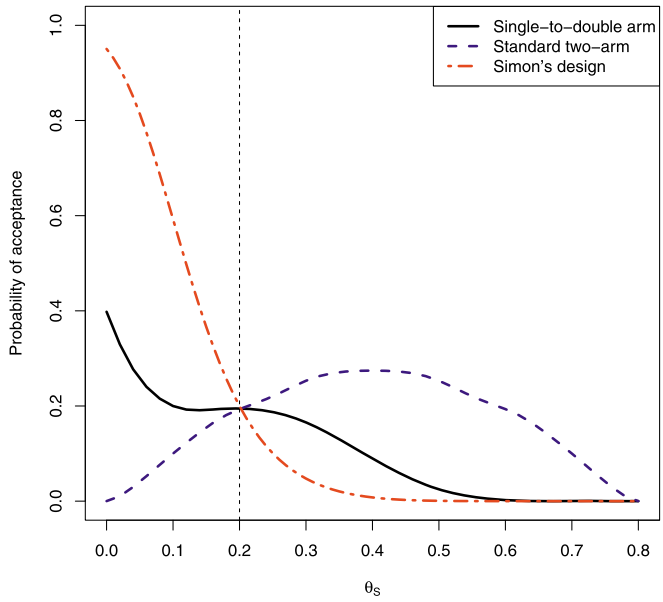


(b)

Figure 2: Breakdown into type I, II and III error rates in the two stages of the single-to-double arm design of (a) probability of rejection with $\theta_E = \theta_S$, and (b) probability of acceptance with $\theta_E = \theta_S + 0.2$, where $(\theta_0, \theta_1) = (0.2, 0.4)$ and $\gamma = 0.2$.



(a)



(b)

Figure 3: Comparison among the single-to-double arm design, the standard two-arm design, and Simon's two-stage design of (a) probability of rejection with $\theta_E = \theta_S$, and (b) probability of acceptance with $\theta_E = \theta_S + 0.2$, where $(\theta_0, \theta_1) = (0.2, 0.4)$ and $\gamma = 0.2$.

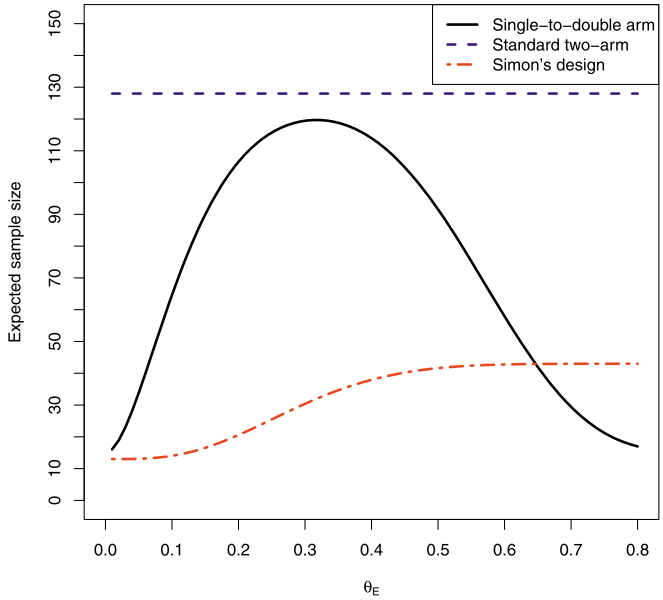


Figure 4: Expected sample size comparison among the single-to-double arm design, the standard two-arm design, and Simon's two-stage design.

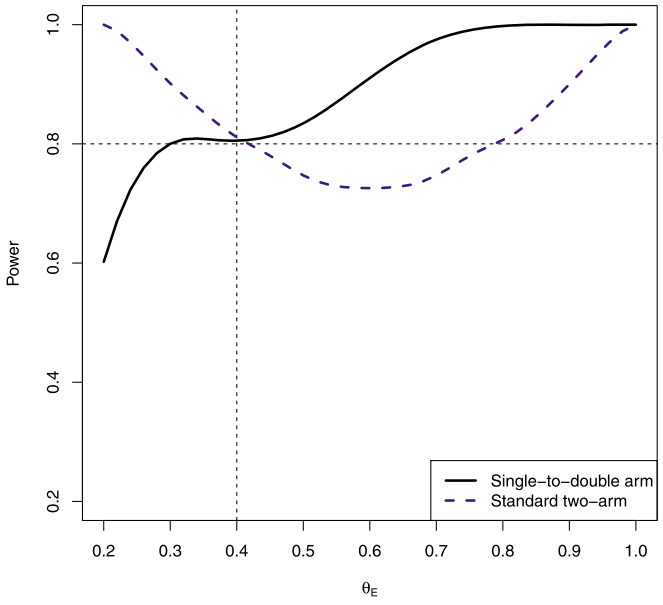


Figure 5: Power comparison between the single-to-double arm design and the standard two-arm design, with $\theta_E = \theta_S + 0.2$, $(\theta_0, \theta_1) = (0.2, 0.4)$ and the target power of 80%.

dard two-arm design, power is large when θ_E is close to 0 or 1, while that under the single-to-double arm design is small when θ_E is close to 0 and increase as θ_E becomes closer to 1. The asymmetric shape of the power curve of the single-to-double arm design can be explained by the fact that the lower the response rate of the experimental drug, the less likely it is to pass the threshold l_1 in stage 1.

4 Trial Example

Our research is motivated by an open-label two-stage phase II trial of a new regimen in patients with advanced non-small cell lung cancer. To successfully recruit a sufficient number of patients in a reasonable period of time, a single-agent trial is intended to be conducted such that all the eligible patients will receive the new regimen at the first stage. If none of the early stopping criteria for futility and efficacy are met by the end of stage 1, subsequent patients will be enrolled into the second stage. At stage 2, eligible patients are randomized to receive either the new regimen or the single-agent chemotherapy (docetaxel or pemetrexed). The final analysis is to be performed to examine the superiority of the new regimen compared with the single-agent chemotherapy.

In order to meet the aforementioned requirements as well as to fully utilize the advantages of single- and double-arm comparisons, we apply the proposed Bayesian two-stage single-to-double arm design for such a phase II clinical trial. The single-arm comparison of the experimental drug with the standard response rate is carried out in stage 1, and the two-arm comparison of the experimental drug with the standard of care is conducted in stage 2. Our goal for this phase II trial is two-fold. First, to fulfill the expectation from the patients of receiving the new drug, we conduct a single-arm trial at stage 1 with all the patients treated with the experimental drug, which can be implemented in a straightforward way. Second, owing to the availability of the standard treatment, we can further use a two-arm comparison to examine the new drug's superiority relative to the standard of care for more objective assessment. To enhance the objectivity, it is indispensable to incorporate a two-arm comparison in the trial. We aim to control the frequentist type I, II and III error rates at 5%, 20% and 20%, respectively. We assume a noninformative Beta$(0.2, 0.8)$ prior for θ_E and θ_S. Based on these trial specifications, the sample size for stage 1 is 15, that for stage 2 is 55 per arm, the lower and upper bounds for the number of responders in stage 1 are 2 and 8, respectively. Our interpretation of the results is that if we only observe 2 or fewer responders in stage 1, we stop the trial for futility; if we observe 8 or more responders in stage 1, we terminate the trial for superiority. On the other hand, if we adopt a commensurate prior for θ_S with a historical mean centered at 0.2 and set $c = \tilde{\tau} = 1$, the required sample sizes for the first and second stages are 11 and 55 in each arm, and the lower and upper bounds for the first stage are 1 and 6, respectively.

For comparison, we explore the possibility of using Simon's two-stage design for this trial. Under the hypotheses

$$H_0 : \theta_E \leq 0.2 \quad \text{versus} \quad H_1 : \theta_E \geq 0.4,$$

the experimental treatment would be considered unpromising if its response rate θ_E is below 0.2, and promising if θ_E is above 0.4. By minimizing the expected sample

size under the null hypothesis, we obtain the parameters for Simon's optimal two-stage design: after enrolling 13 patients in stage 1, if there are 3 or fewer responders, the trial would be terminated for futility; otherwise, the trial moves on to stage 2 with a total of 30 patients, and if there are 12 or fewer responders among the 43 patients from the two stages, we declare the treatment unpromising. In addition to Simon's approach, we also make a comparison with a standard two-arm design, which requires 64 patients per arm. Our proposed design requires 125 patients if we assume a noninformative prior for the standard response rate, and requires 121 patients if we instead assume a commensurate prior. In both versions of our proposed design, 55 patients are randomized to each arm during the second stage. Among the three types of designs under our comparison, Simon's two-stage design has the smallest sample size. It is most efficient if we can have an accurate guess of θ_0. On the other hand, a standard two-arm design has the largest sample size as it seeks the most objective comparison between the standard arm and the experimental arm. As discussed in the simulation study, the expected sample size and error rates of our design lie in between those of Simon's two-stage design and the standard two-arm design, and thus our design serves as a middle-ground between these two designs.

5 Discussions

We have proposed to combine the single-arm and double-arm hypothesis tests into one phase II trial for more comprehensive decision making. Through selecting the appropriate constraint on the type III error rate, we allow the flexibility to switch from a straightforward single-arm study to a more objective two-arm study based on the degree of conservativeness toward the type III error. The proposed Bayesian error rate control extends the work of Dong et al. (2012) to the single-to-double arm setting, and similarly the values of Bayesian error rates are rather small. Contrary to the frequentist approach to controlling the type I error rate, which seeks to maintain the supremum of the rejection probability below α, the Bayesian counterparts represent the probability of committing a type I error averaged over the prior distributions of θ_E and θ_S.

The value of a phase II design should be assessed under the context of the entire program. Instead of looking at a phase II design in isolation, it is important to consider its implication to the ultimate success in the phase III trial. A potential direction of future work is to refine the calibration of our design specifications in connection to the subsequent phase III trial, especially when the phase III trial is using the same endpoints as phase II. Moreover, our proposed design uses an equal allocation of subjects to the experimental and the standard arm in stage 2. One natural extension of our design is to use an unequal allocation of subjects with an allocation ratio r, which might result in higher power when the numbers of subjects in the two arms are balanced. More specifically, in stage 2, if n_2 subjects are assigned to the experimental arm and rn_2 subjects to the standard arm, we can then enumerate r in addition to n_2 in the searching algorithm and choose the design with the smallest ESS while maintaining the frequentist and Bayesian error rates.

Supplementary Material

Supplementary Material of "Bayesian Two-Stage Design for Phase II Clinical Trials with Switching Hypothesis Tests" (DOI: 10.1214/15-BA988SUPP; .pdf).

References

Chang, M. N., Therneau, T. M., Wieand, H. S., and Cha, S. S. (1987). "Designs for group sequential phase II clinical trials." *Biometrics*, 43(4): 865–874. 31

Chen, T. and Ng, T.-H. (1998). "Optimal flexible designs in phase II clinical trials." *Statistics in Medicine*, 17(20): 2301–2312. 32

Dong, G., Shih, W. J., Moore, D., Quan, H., and Marcella, S. (2012). "A Bayesian-frequentist two-stage single-arm phase II clinical trial design." *Statistics in Medicine*, 31(19): 2055–2067. MR2956061. doi: http://dx.doi.org/10.1002/sim.5330. 32, 34, 36, 48

Fleming, T. R. (1982). "One-sample multiple testing procedures for phase II clinical trials." *Biometrics*, 38(1): 143–151. 31

Gehan, E. A. (1961). "The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent." *Journal of Chronic Diseases*, 13(4): 346–353. 31

Green, S. J. and Dahlberg, S. (1992). "Planned versus attained design in phase II clinical trials." *Statistics in Medicine*, 11(7): 853–862. 32

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., and Sargent, D. J. (2011). "Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials." *Biometrics*, 67(3): 1047–1056. MR2829239. doi: http://dx.doi.org/10.1111/j.1541-0420.2011.01564.x. 39

Hobbs, B. P., Sargent, D. J., and Carlin, B. P. (2012). "Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models." *Bayesian Analysis*, 7(3): 639–674. MR2981631. doi: http://dx.doi.org/10.1214/12-BA722. 39

Inoue, L. Y., Thall, P. F., and Berry, D. A. (2002). "Seamlessly expanding a randomized phase II trial to phase III." *Biometrics*, 58(4): 823–831. MR1945019. doi: http://dx.doi.org/10.1111/j.0006-341X.2002.00823.x. 32

Lai, T. L., Lavori, P. W., and Shih, M. C. (2012). "Sequential design of phase II–III cancer trials." *Statistics in Medicine*, 31(18): 1944–1960. MR2956028. doi: http://dx.doi.org/10.1002/sim.5346. 32

Lee, J. J. and Liu, D. D. (2008). "A predictive probability design for phase II cancer clinical trials." *Clinical Trials*, 5(2): 93–106. 32

Lee, S. J. and Zelen, M. (2000). "Clinical Trials and sample size considerations: another perspective (with discussion)." *Statistical Science*, 15(2): 95–100. 32, 36, 37

Liu, J. F., Lin, Y., and Shih, W. J. (2010). "On Simon's two-stage design for single-arm phase IIA cancer clinical trials under beta–binomial distribution." *Statistics in Medicine*, 29(10): 1084–1095. MR2756813. doi: http://dx.doi.org/10.1002/sim.3805. 32

Mariani, L. and Marubini, E. (1996). "Design and analysis of phase II cancer trials: a review of statistical methods and guidelines for medical researchers." *International Statistical Review*, 64(1): 61–88. 33

Mayo, M. S. and Gajewski, B. J. (2004). "Bayesian sample size calculations in phase II clinical trials using informative conjugate priors." *Controlled Clinical Trials*, 25(2): 157–167. 32

Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C., and Bauer, P. (2005). "Testing and estimation in flexible group sequential designs with adaptive treatment selection." *Statistics in Medicine*, 24(24): 3697–3714. MR2221962. doi: http://dx.doi.org/10.1002/sim.2389. 32

Sambucini, V. (2008). "A Bayesian predictive two-stage design for phase II clinical trials." *Statistics in Medicine*, 27(8): 1199–1224. MR2420154. doi: http://dx.doi.org/10.1002/sim.3021. 32

Sambucini, V. (2010). "A Bayesian predictive strategy for an adaptive two-stage design in phase II clinical trials." *Statistics in Medicine*, 29(13): 1430–1442. MR2758126. doi: http://dx.doi.org/10.1002/sim.3800. 32

Shi, H. and Yin, G. (2015). "Supplementary Material of Bayesian Two-Stage Design for Phase II Clinical Trials with Switching Hypothesis Tests" *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/15-BA988SUPP. 37

Simon, R. (1989). "Optimal two-stage designs for phase II clinical trials." *Controlled Clinical Trials*, 10(1): 1–10. 31, 32, 33

Simon, R., Wittes, R. E., and Ellenberg, S. S. (1985). "Randomized phase II clinical trials." *Cancer Treatment Reports*, 69(12): 1375–1381. 31

Steinberg, S. M. and Venzon, D. J. (2002). "Early selection in a randomized phase II clinical trial." *Statistics in Medicine*, 21(12): 1711–1726. 32

Storer, B. E. (1992). "A class of phase II designs with three possible outcomes." *Biometrics*, 48(1): 55–60. 36

Sylvester, R. J. (1988). "A Bayesian approach to the design of phase II clinical trials." *Biometrics*, 44(3): 823–836. MR0963916. doi: http://dx.doi.org/10.2307/2531594. 31

Tan, S. B. and Machin, D. (2002). "Bayesian two-stage designs for phase II clinical trials." *Statistics in Medicine*, 21(14): 1991–2012. 32

Thall, P. F. and Simon, R. (1994). "Practical Bayesian guidelines for phase IIB clinical trials." *Biometrics*, 50(2): 337–349. MR1294683. doi: http://dx.doi.org/10.2307/2533377. 32

Thall, P. F. and Simon, R. (1995). "Recent developments in the design of phase II clinical trials." *Cancer Treatment and Research*, 75(2): 49–71. 33

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., and Thompson, L. (2014). "Use of historical control data for assessing treatment effects in clinical trials." *Pharmaceutical Statistics*, 13(1): 41–54. 42, 43

Wang, Y. G., Leung, D. H., Li, M., and Tan, S. B. (2005). "Bayesian designs with frequentist and Bayesian error rate considerations." *Statistical Methods in Medical Research*, 14(1): 445–456. MR2196274. doi: http://dx.doi.org/10.1191/0962280205sm410oa. 32

Yin, G. (2012). *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*. New York: John Willey & Sons, Inc. 33, 35

Yin, G., Chen, N., and Lee, J. J. (2011). "Phase II trial design with Bayesian adaptive randomization and predictive probability." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(5): 219–235. MR2905060. doi: http://dx.doi.org/10.1111/j.1467-9876.2011.01006.x. 32

Acknowledgments

We thank the two referees, the associate editor, and the editor for their constructive and insightful comments that led to significant improvements in the article, and also thank Y. Wang for many discussions and some preliminary work. The research was supported in part by a grant (grant number 17125814) from the Research Grants Council of Hong Kong.