Published in final edited form as:

Clin Trials. 2012 June ; 9(3): 293–302. doi:10.1177/1740774512443429.

A Bayesian approach for unplanned sample sizes in phase II cancer clinical trials

Yimei Li^a, Rosemarie Mick^b, and Daniel F Heitjan^b

^aDivision of Oncology, The Children's Hospital of Philadelphia & University of Pennsylvania, Philadelphia, PA, USA

^bDepartment of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Background—Phase II cancer clinical trials commonly employ two-stage designs that incorporate a single interim analysis for lack of efficacy and are designed to achieve specified frequentist properties. The requirement to examine the outcome at a prespecified sample size (SS) can be problematic, because the attained SS often differs from the planned SS.

Purpose—We propose to address unplanned SSs achieved at either stage by means of a Bayesian approach that approximately preserves the original design's properties.

Methods—Our approach translates the rejection rule of the original frequentist design into equivalent statements about the posterior distribution of the response rate and applies this Bayesian criterion to the analysis with any realized SS.

Results—The results demonstrate that our approach approximately maintains operating characteristics of the original frequentist design including type I and type II error rates, probability of early termination, and expected SS under the null hypothesis.

Limitations—Designs attained under this approach may not satisfy target limits for type I error rate and power.

Conclusions—Our method offers a coherent analysis plan when the attained SS at either stage deviates from that specified in the original design. The price of its flexibility, in terms of erosion of the desired frequentist properties, is modest.

Introduction

A major purpose of a phase II cancer clinical trial is to determine whether the experimental therapy has sufficient antitumor effect to warrant further investigation. Commonly used

© The Author(s), 2012

Reprints and permission: <http://www.sagepub.co.uk/journalsPermissions.nav>

Author for correspondence: Yimei Li, Division of Oncology, The Children's Hospital of Philadelphia, 10207 Colket Translational Research Building, 3614 Civic Center Boulevard, Philadelphia, PA 19104, USA. liy3@email.chop.edu.

Conflict of interest

All authors declare that they have no conflict of interest.

primary endpoints include tumor response, time to progression, progression-free survival, and overall survival [1]. Despite growing interest in the use of randomized controls [2], open-label, single-arm studies remain common. Thus, we henceforth focus on the typical setting of a single-arm phase II trial with clinical response as the outcome.

Ethical and practical considerations militate for early termination when interim results strongly suggest that the new regimen is inactive. Thus, many designs incorporate a single, preplanned interim analysis for lack of efficacy [3–9]. That is, one conducts an interim analysis when outcomes are available on the first n_1 patients, proceeding to enroll a further $n_2 = n - n_1$ patients only if the number of responses in the first stage exceeds a critical value r_1 . When all n patients have been evaluated, one conducts a final analysis that rejects the treatment if the total response count falls below a critical value r . The various designs choose n_1 , r_1 , n , and r to satisfy particular statistical criteria. For example, Simon's optimal design [6] minimizes the expected total sample size (SS) under a specified null value of the response rate, subject to constraints on type I and type II error probabilities. Most such designs eschew early termination for efficacy, because if the new treatment is effective, there is no reason to stop the trial early.

In practice, the requirement to examine the outcome at a prespecified SS in each stage can be problematic when, as commonly occurs, the attained SS differs from the planned SS [10–13]. This can cause the trial to forfeit its desired statistical properties and to create ambiguity about how to analyze the data. For example, consider Simon's design with $n_1 = 17$, $r_1 = 3$, $n = 37$, and $r = 10$. Suppose that at the end of the study one has actually enrolled 38 patients, which could occur, for example, if there is an uptick in the acceptance of participation by patients toward the end. What then is the appropriate criterion to use in determining whether the new treatment should be rejected? The situation becomes more complex if only 10 of the first 37 patients respond (implying rejection), but the extra patient achieves a response. Thus, the overall response rate is now 29% and exceeds the 10/37 (27%) specified by the design as the maximal rate for rejection. The opposite situation could occur as well. Imagine that 11 of the first 37 patients respond, but that the trial inadvertently enrolls 5 extra patients, none of whom responds. Strict adherence to the design suggests acceptance based on the first 37 patients, whereas the overall response rate ($11/42 = 26\%$) is lower than the 27% critical rate and suggests rejection.

There have been some attempts to modify classical frequentist two-stage designs to address this issue. Green and Dahlberg [10] proposed an extension of the standard phase II approach used by the Southwest Oncology Group that allows for inexact SSs in both stages. They suggested testing $p = p_1$ against $p < p_1$ at the 0.02 level at the first stage (p_1 is the desirable response rate), terminating the study if the test is significant. They then test $p = p_0$ against $p > p_0$ at the 0.05 level (p_0 is the undesirable response rate), accepting the drug as promising if the test is significant. Although this flexible design works well (with realized type I error below 0.05 and power close to the target level), the choice of the 0.02 level at the first stage and 0.05 level at the second stage is arbitrary and lacks theoretical justification. Moreover, the choice of significance levels is specific for the type I error rate of 0.05, and it is unclear what levels one should apply if one desires a different type I error rate (e.g., 0.1).

Chen and Ng [11] defined a flexible design as a collection of two-stage designs with a range of SSs in both stages. They explicitly considered a range of possible interim SSs $\{n^{(1)}_1, n^{(2)}_1, \dots, n^{(k)}_1\}$ and total SSs $\{n^{(1)}, n^{(2)}, \dots, n^{(k)}\}$, assuming that each would occur with equal probability, $1/k$, for any of these values. Within this scenario, they search for the ‘optimal’ and ‘minimax’ designs with designated limits on type I and type II error rates. Masaki *et al.* [12] further extended this method by allowing unequal probabilities of achieving the k possible SSs. While such approaches strictly control type I and type II error rates, the variability margin (the difference between planned and achieved SSs) must be prespecified (e.g., Masaki *et al.* used a margin of 2, giving $k = 5$). The actual attained SS could of course fall outside of this range.

Koyama and Chen [13] presented a method for inference from Simon’s design, using both the planned and the actual SSs. Accounting for the adaptive nature of the design, this method computes the proper p value and confidence interval of the response rate (i.e., correct based on the realized SS). However, the method can only accommodate unplanned SS in stage 2 and assumes that in stage 1, the actual SS is always the same as that planned.

In this article, we propose a flexible Bayesian approach that can handle any unplanned SS in both stages and can be applied with any frequentist two-stage design. Our proposal is to translate the cut points (i.e., n_1 , r_1 , n , and r) in a frequentist design to a Bayesian criterion relative to the posterior distribution of the response rate and to apply the same criterion to the data when the attained SS differs from the planned SS. We illustrate our Bayesian approach for a range of popular frequentist designs: Simon’s optimal and minimax designs [6], Ye and Shyr’s balanced design [7], and Jung *et al.*’s admissible design [8,9]. We demonstrate empirically that the method has only modest effect on the designs’ original frequentist properties.

Methods

The framework

We assume that the trial seeks to test the null hypothesis $H_0: p = p_0$ against the alternative $H_1: p = p_1$, where p is the regimen’s unknown true response rate, p_0 is the maximum uninteresting response rate, and p_1 is a target response rate at which we seek a designated level of power. As discussed previously, in the first stage, one enrolls n_1 patients and terminates if no more than r_1 respond. In stage 2, one enrolls a further $n_2 = n - n_1$ patients and deems the treatment a failure if the total number of responses is r or less. Henceforth, we denote n_1 and n as the planned first stage and total SSs, respectively, and N_1 and N as the attained first stage and total SSs, respectively. Similarly, we denote r_1 and r as the decision cut points at planned interim and final SSs and denote R_1 and R as the new decision cut points at attained SSs. Moreover, we denote M as a random variable for the number of evaluated outcomes (not specifically for the end of a stage) and m as a realized value of that random variable. Similarly, we denote X as the random variable for number of observed responses and x as a realized value of that random variable. Furthermore, we will use ‘reject’ to refer to rejection of the regimen (i.e., failing to reject the null hypothesis) and ‘accept’ to refer to acceptance of the regimen (i.e., rejecting the null).

In such designs, a type I error constitutes falsely accepting a regimen whose response rate does not exceed p_0 , and a type II error constitutes falsely rejecting a regimen whose response rate exceeds p_1 . The probability of early termination (PET) is the probability that the trial is terminated at the first stage, that is, the probability of r_1 or fewer responses in the first n_1 patients. The expected SS (ESS) is the average number of patients to be enrolled in a hypothetical series of replications of the design: $ESS = n_1 + (1 - PET) \times (n - n_1)$. Generally, we prefer designs that hold type I and type II error probabilities below specified rates α and β , respectively, and have a high PET under p_0 , a low PET under p_1 , and a low ESS under p_0 .

The Bayesian approach

Our strategy for adapting a frequentist design to accommodate unplanned SSs is to translate the frequentist rejection criteria into equivalent statements about the posterior probability that the response rate is definitely acceptable or definitely unacceptable. In this way, the decision rule is the same as the original frequentist design when the attained SS at either stage is the same as the planned SS. When there are deviations in SSs, one can evaluate the data under the Bayesian decision rule, which is coherent with the original design in that the same Bayesian rule is applied at any SS. This provides guidance for handling inexact SSs with, as we shall show, only modest effects on operating characteristics.

Our Bayesian approach assumes a $\text{Beta}(a_0, b_0)$ prior for the response rate, so that with $X = x$ observed responses in $M = m$ evaluated outcomes, the posterior of the response rate is $\text{Beta}(a_0 + x, b_0 + m - x)$. In the interim analysis, we stop for futility if the current posterior assigns a small probability (less than a specified value τ_l) to the hypothesis that the true response rate exceeds p_1 , that is $\Pr[p > p_1 \mid \text{data}] < \tau_l$. At the end of the trial, we accept the regimen if the posterior indicates a large probability (greater than a specified τ_u) that the true response rate exceeds p_0 , that is $\Pr[p > p_0 \mid \text{data}] > \tau_u$.

To ensure that the Bayesian criterion is consistent with the original frequentist design, we calculate the target posterior probabilities under the planned SSs and cut points. Specifically, under a designated $(p_0, p_1, \alpha, \beta)$, a two-stage frequentist design yields rejection criteria (n_1, r_1, n, r) . We denote $\Pr[p > p_1 \mid X = r_1, M = n_1]$ as the posterior probability of $p > p_1$ for potential observed data $X = r_1$ and $M = n_1$, that is, the probability of $p > p_1$ under the $\text{Beta}(a_0 + r_1, b_0 + n_1 - r_1)$ distribution. Because the frequentist design implies stopping for futility when observing the data $X = r_1$ and $M = n_1$, the Bayesian cut point τ_l has to be greater than $\Pr[p > p_1 \mid X = r_1, M = n_1]$. And because the frequentist design also implies proceeding to the second stage when observing the data $X = r_1 + 1$ and $M = n_1$, the Bayesian cut point τ_l has to be less than $\Pr[p > p_1 \mid X = r_1 + 1, M = n_1]$. Because the data are discrete, τ_l can take any value between these two posterior probabilities; to avoid arbitrariness, we will take the average. Similarly, for end-of-trial decision making, we determine τ_u from n and r . That is, we designate

$$\begin{aligned} \theta_l &= \frac{\Pr[p > p_1 \mid X = r_1, M = n_1] + \Pr[p > p_1 \mid X = r_1 + 1, M = n_1]}{2} \\ \theta_u &= \frac{\Pr[p > p_0 \mid X = r, M = n] + \Pr[p > p_0 \mid X = r + 1, M = n]}{2} \end{aligned} \quad (1)$$

Thus, the rejection region implied by the Bayesian criterion will be the same as for the original design at the planned SSs n_1 and n .

We revisit Simon's optimal design with $\alpha = \beta = 0.1$, $p_0 = 0.2$, and $p_1 = 0.4$, which gives $n_1 = 17$, $r_1 = 3$, $n = 37$, and $r = 10$ and results in $\text{PET}(p_0) = 0.55$ and $\text{ESS}(p_0) = 26.0$ with a realized type I error rate of 0.095 and type II error rate of 0.097. Taking p to have a Beta(1, 1) (uniform) prior, Equation (1) gives $\tau_l = 0.063$ and $\tau_u = 0.908$. That is, we stop the trial for futility at the first stage if $\Pr[p > 0.4 \mid \text{data}] < 0.063$, and we accept the regimen at the second stage if $\Pr[p > 0.2 \mid \text{data}] > 0.908$. Table 1 presents the implied rejection region for various possibly interim and final SSs. We can see that the rejection region is the same as Simon's design at $n_1 = 17$ and $n = 37$ and is also reasonable for other attained SSs. For example, with attained final SS of 41, one should reject the regimen if there are 11 or fewer responses.

Potential SS deviation

We evaluated how application of our approach would affect the frequentist characteristics of the original design. For comparison, we included Simon's optimal and minimax designs, Ye and Shyr's balanced design, and Jung *et al.*'s admissible design. For each design, we considered various scenarios for p_0 and p_1 (p_0 ranges from 0.05 to 0.7; the difference between p_1 and p_0 is 0.15 or 0.2) and three combinations of type I and type II error rates ($\alpha = \beta = 0.1$; $\alpha = 0.05$, $\beta = 0.2$; and $\alpha = 0.05$, $\beta = 0.1$).

To reflect potential inexact SS, we assume that the attained SS can deviate from the planned SS with a range of ± 2 for the interim analysis and with a range of ± 5 for the final analysis. The probability of a given SS decreases as it deviates farther from the target SS. Specifically, we assume that at the end of the first stage, $P(N_1 = n_1 \pm 2) = 0.05$, $P(N_1 = n_1 \pm 1) = 0.2$, and $P(N_1 = n_1) = 0.5$; at the end of the second stage, we set $P(N = n \pm 5) = 0.01$, $P(N = n \pm 4) = 0.02$, $P(N = n \pm 3) = 0.04$, $P(N = n \pm 2) = 0.08$, $P(N = n \pm 1) = 0.15$, and $P(N = n) = 0.4$. These assumptions reflect the SS deviations identified by Masaki *et al.* [12] in a review of the clinical literature. For all possible realized SSs N_1 and N , we derive the new decision cut points R_1 and R (implied by our Bayesian criteria τ_l and τ_u) and calculate the size, power, PET, and so forth (as in the original Simon's article [6]); we then average these values across the probabilities of deviation to obtain the overall operating characteristics. We performed all computations in R 2.14.0; code is available from the first author.

Results

Simon's optimal design

Simon's optimal design chooses the combination of (n_1, r_1, n, r) to minimize $\text{ESS}(p_0)$ among those two-stage designs satisfying designated type I and type II error constraints. We summarize selected results in Table 2 for $\alpha = \beta = 0.1$. The first three columns show the null and alternative response rates and their difference, and columns 4–7 show the decision criteria for the optimal design and the corresponding Bayesian approach. The remaining columns show the properties of two methods, where the columns labeled with F refer to the

theoretical properties of the original frequentist design, and columns labeled with B refer to the properties of the Bayesian approach under the possibility of unplanned SSs.

Among all the designs we evaluated (including but not limited to those shown in Table 2), the Bayesian approach on average yields a slightly higher PET under both p_0 and p_1 compared to the original optimal design; the difference in $PET(p_0)$ ranges from -0.002 to 0.036 with an average of 0.009 , and the difference in $PET(p_1)$ ranges from 0.001 to 0.013 with an average of 0.005 . The $ESS(p_0)$ with the Bayesian approach is slightly smaller than the theoretical value (on average 0.3 smaller). Also, the flexible Bayesian approach sometimes inflates type I and type II error rates slightly, so that they occasionally exceed their targets. In the scenarios we evaluated, the realized type I error rate can exceed the target by as much as 0.003 , and the power can fall below the target level by as much as -0.011 .

The results are similar for other combinations of type I and type II error rates, that is, $\alpha = 0.05$ and $\beta = 0.2$ and $\alpha = 0.05$ and $\beta = 0.1$. Figure 1 presents the results for a 0.15 difference between p_0 and p_1 . In general, the type I error and power are close to the target levels. The type I error rate can exceed the target level by as much as 0.002 and 0.002 , and the power can fall short by as much as -0.010 and -0.006 , for these error rate combinations, respectively. Other frequentist properties ($PET(p_0)$, $PET(p_1)$, and $ESS(p_0)$) are also similar to the theoretical values of the original design (not shown in Figure 1).

Simon's minimax design

Simon's minimax design minimizes the total SS among all two-stage designs that satisfy the designated type I and type II error constraints. This may be preferable to the optimal design when there is concern that patient accrual will be slow. Table 3 summarizes the selected results when $\alpha = \beta = 0.1$. The results for the other two combinations ($\alpha = 0.05$, $\beta = 0.2$; $\alpha = 0.05$, $\beta = 0.1$) are similar and are not shown here.

The Bayesian approach leads to a slightly higher PET under both p_0 and p_1 , compared to the original minimax design: the difference in $PET(p_0)$ ranges from -0.001 to 0.030 with an average of 0.008 ; the difference in $PET(p_1)$ ranges from 0 to 0.005 with an average of 0.002 . The $ESS(p_0)$ under the Bayesian approach is slightly smaller than the theoretical value (on average 0.2 smaller). Also with the flexible Bayesian approach, the type I error rate can exceed the target limit by as much as 0.003 , and the power can decrease from the target level by as much as -0.004 .

Ye and Shyr's balanced design

In both the optimal and minimax designs, SSs in the two stages can be highly unbalanced. Conducting the interim analysis at a small percentage of the target total SS may be undesirable if there is concern that early patients are not representative of the target population. On the other hand, if the ratio of the SS of the first stage to the second stage ($n_1:n_2$) is too high, the potential savings in SS from the two-stage design is minimal. To address this issue, Ye and Shyr proposed the balanced design [7]. Here we have applied our approach to this design and evaluated its properties assuming potentially inexact SSs.

The results are shown in Table 4 for $\alpha = \beta = 0.1$ (the results for the other combinations of error rates are similar). Again, on average, the PET values under both p_0 and p_1 are somewhat higher with our approach: the difference in $\text{PET}(p_0)$ ranges from -0.005 to 0.036 with an average of 0.006 ; the difference in $\text{PET}(p_1)$ ranges from 0.001 to 0.010 with an average of 0.003 . The Bayesian approach has smaller $\text{ESS}(p_0)$ (on average 0.2 smaller). The type I and type II error rates exceed nominal levels by no more than 0.004 and 0.007 , respectively.

Jung *et al.*'s admissible design

In some situations, the optimal and minimax designs can diverge substantially, in that the optimal design may have a much larger total SS n compared to the minimax design, or the minimax design may have a much larger $\text{ESS}(p_0)$ than the optimal design. In practice, one may prefer a compromise design, neither optimal nor minimax, that has $\text{ESS}(p_0)$ comparable to the optimal design and n comparable to the minimax design. Jung *et al.* termed such designs 'admissible' according to a Bayesian decision-theoretic criterion based on an ethically justifiable loss function [9]. For a given (p_0, p_1, a, b) , the admissible designs are not unique and include as special cases the optimal and minimax designs. But sometimes, the optimal and minimax are the only two admissible designs, and no compromise design exists. In our comparisons, we selected the admissible design (other than optimal and minimax) with the smallest n (indicated as NA if nonexistent). We developed the corresponding Bayesian approach and evaluated it under unplanned SSs.

The results are shown in Table 5 for $\alpha = \beta = 0.1$ (the results for the other two combinations of error rates are similar). Like the previous results, the Bayesian approach leads to a slightly higher PET under both p_0 and p_1 (the difference in $\text{PET}(p_0)$ ranges from -0.001 to 0.038 with an average of 0.007 ; the difference in $\text{PET}(p_1)$ ranges from 0.001 to 0.011 with an average of 0.002) and a smaller $\text{ESS}(p_0)$ (on average 0.2 smaller). Under the Bayesian approach, the type I error rate can exceed the target limit by as much as 0.003 , and the power can vary from the target level as much as -0.006 .

Sensitivity to the prior distribution

In all the above results, we used the uniform prior $\text{Beta}(1, 1)$. To investigate the sensitivity of our approach to the prior, we considered two other priors: an optimistic prior and a pessimistic prior. The optimistic prior is defined as $\text{Beta}(a_0 = nprior\ p_1, b_0 = nprior\ (1 - p_1))$, which implies that the prior mean of the response rate is p_1 based on the information from a prior trial with a SS of $nprior$. Similarly, the pessimistic prior is $\text{Beta}(a_0 = nprior\ p_0, b_0 = nprior\ (1 - p_0))$, implying a prior mean of p_0 based on the information from $nprior$ patients. As typically we have efficacy outcomes from at least 6 patients enrolled at the recommended phase II dose in a preceding phase I trial, we set $nprior = 6$ in all scenarios. For example, for $p_0 = 0.1$ and $p_1 = 0.25$, the optimistic prior is $\text{Beta}(1.5, 4.5)$ and the pessimistic prior is $\text{Beta}(0.6, 5.4)$. For each scenario of each frequentist design discussed above, we repeated the calculations using the optimistic and pessimistic priors.

Figure 2 displays the Bayesian decision criteria and selected frequentist properties under uniform, optimistic, and pessimistic priors for Simon's optimal design with $\alpha = \beta = 0.1$. The

results for the other combinations of (α, β) and the other frequentist designs are similar. We observed some differences in the Bayesian decision criteria (τ_l and τ_u) across priors (top row of Figure 2). However, the frequentist properties ($\text{PET}(p_0)$, $\text{PET}(p_1)$, $\text{ESS}(p_0)$, size, and power) are nearly the same for all scenarios (size and power are displayed in the bottom row of Figure 2). This result is expected because the prior distribution influenced the posterior distribution, which in turn affected the Bayesian decision criteria, to force the rejection rule to be consistent with the original frequentist design. But since the Bayesian decision rule remained coherent with the original design, regardless of priors, the frequentist properties are hardly changed. We note that the Bayesian decision criteria (values of τ_l and τ_u) for the uniform prior do not necessarily fall between those for the optimistic and pessimistic priors, especially when p_0 is low. This is because the uniform prior $\text{Beta}(a_0 = 1, b_0 = 1)$ can be informative when p_0 is small, that is, the prior information $a_0 = 1$ is relatively more influential when the observed responses are few (e.g., $X = 1$ when $p_0 = 0.05$ and $M = 20$).

Sensitivity to the deviation in SS

In the previous examples, we calculated the sampling properties averaged across a range of differences from the planned SS. We refer to this model, which we took from the literature review of Masaki *et al.* [12], as the ‘random deviation’ model. Recall that it allows the attained SSs to diverge by as much as ± 2 for the interim analysis and ± 5 for the final analysis. One might also wish to consider the sampling properties conditional on the attained SSs. Thus, we evaluated the design at each possible combination of attained SSs (N_1, N) (a total of $5 \times 11 = 55$ combinations).

Figure 3 illustrates the results for Simon’s optimal design with $p_0 = 0.3$, $p_1 = 0.45$, and target $\alpha = \beta = 0.1$, for all the possible combinations of attained SSs. In this example, the planned SSs are $n_1 = 30$ and $n = 82$, so that the possible attained SSs are 28 N_1 32 and 77 N 87. The results suggested that both type I error and power can depart substantially from the target levels (Table 2, second row). In all combinations, the type I error ranges from 0.082 to 0.119 and power ranges from 0.871 to 0.924. We note that the power does not necessarily increase with the attained SS (interim or final). For example, when $N_1 = 30$, power decreases from 0.904 to 0.893 when final N increases from 80 to 81; this is because the final rejection cut point (R) also increases (from 28 to 29). Similar patterns are observed in type I error.

Summary

For the frequentist designs that we evaluated, the corresponding Bayesian approach yields slightly higher PET and smaller $\text{ESS}(p_0)$ than the original under a model of random divergence from planned SSs. The type I and type II error rates under the Bayesian approach are typically close to the nominal levels, if slightly higher in some scenarios. The results averaged across symmetric deviations from the target SSs smooth out larger deviations from the nominal levels that occur at particular attained SSs.

Discussion

We have presented a novel Bayesian approach that merges a frequentist two-stage design with a Bayesian data analysis. Our method offers a coherent analysis plan when the attained

SS deviates from that specified under the original design, while approximately maintaining the frequentist properties of the original design.

Unlike recently proposed Bayesian designs [14–18], our approach is not a formal Bayesian design per se; instead, it is a strategy to adapt the frequentist design when the attained SS differs from the planned SS. In this way, the investigator can choose any frequentist design with the desired properties when planning the trial, and at the interim analysis or at the end of the trial, apply the corresponding Bayesian approach in the setting of under- and over-accrual.

Our approach is preferable to other frequentist methods for unplanned SSs [10–13], in that it has a Bayesian justification and it can accommodate enrollment deviations of *any* size in *both* stages. Our results show that for four popular frequentist designs (Simon's optimal and minimax designs, Ye and Shyr's balanced design, and Jung *et al.*'s admissible design), the corresponding Bayesian approach essentially preserves the original frequentist properties for the sizes of deviations that are realized in practice. We expect similar results for other two-stage frequentist designs.

We have considered the use of uniform, optimistic, and pessimistic priors. The uniform prior is common in practice, as it has small potential impact on the posterior distribution [17]. Our selection of optimistic and pessimistic priors also mimics those proposed previously in the literature [15,19]. With different priors, our Bayesian approach adjusts the Bayesian cut points accordingly but remains consistent with the original frequentist design and maintains the desirable frequentist properties.

Under specific deviations from planned values, the error rates can differ substantially from the target values, reflecting the discreteness of the sample space. Thus, to some extent, the good performance of the method overall reflects our averaging of its error probabilities over a symmetric range of possible divergences. Investigators who anticipate a particular pattern of over- or under-accrual may wish to investigate the frequentist properties of the approach averaged across the likely pattern of attained SSs.

To implement a two-stage design precisely, one may need to suspend accrual as the trial approaches the end of the first stage. Suspension is not desirable because it can both complicate and lengthen the trial. When suspension adds substantial complexity (e.g., in multicenter studies), the continuous enrollment design is an alternative [20]. Trials with continuous enrollment are unlikely to meet the original SS requirement; however, in which case, our approach would be helpful. When there is serious concern that the test regimen poses a safety risk (that would only be justified by the chance of a substantially improved outcome), suspension should be mandatory. Yet even careful suspension may not guarantee in achieving the planned SS. In any event, we recommend that investigators follow the original design as closely as they can, applying our approach to analyze the data as needed.

We note that one should not misuse the flexibility of our or alternative approaches to fish for a desirable result. For example, an unscrupulous investigator who analyzes 17 patients but observes only 3 responses might decide to enroll an additional (18th) subject in the hope of getting a positive outcome that would allow the trial to continue (Table 1). This is equivalent

to taking an unplanned look at the data and will cause the forfeit of any targeted frequentist properties. Like all two-stage designs, our approach achieves appropriate frequentist characteristics (as shown in the tables) only when the data are examined exactly twice (at the interim and final analyses). At each analysis, one should use all the available data to make a decision, without arbitrarily adding or removing some patients' outcomes.

Although phase II designs are commonly rooted in hypothesis testing criteria, the estimation of response rates is also important. The imposition of early stopping induces a bias in the estimated response rate, for which numerous correction methods exist [21–24]. With the proposed Bayesian approach, one can also apply a bias correction as needed [24]. Nevertheless, a decision is needed at the end of phase II trials on whether to move the drug forward, and if the realized accrual is different from the planned accrual, such a decision could be ad hoc and subjective. Our approach eliminates ambiguity in the decision making, permitting trials to be conducted in an objective and coherent manner.

One could construct an alternative approach by matching frequentist significance levels rather than Bayesian posterior probabilities. That is, one would compute the type I error rate of the interim test, at the planned SS, of the null hypothesis that $p = p_1$, and use this as the nominal level for the same test at the attained interim SS. Likewise, one would select the nominal size for the final analysis from the nominal size of the test of $p = p_0$ at the planned SS. Because the standard frequentist tests of these hypotheses are likelihood-ratio tests, this procedure would be similar to our Bayesian procedure and would enjoy similar operating characteristics.

For example, for $p_0 = 0.3$, $p_1 = 0.45$, $\alpha = 0.1$, and $\beta = 0.1$, Simon's optimal design gives $n_1 = 30$, $r_1 = 9$, $n = 82$, and $r = 29$. If we define the p value as $\Pr[X \leq x | p_1]$ at stage 1 and as $\Pr[X \leq x | p_0]$ at stage 2, then the original optimal design implies that we should reject the drug if $p < 0.102$ at stage 1 and accept the drug if $p < 0.146$ at stage 2. Using such a frequentist rule in the situation of 'random deviation' for the SS, the realized type I error and power is 0.098 and 0.898, respectively, which are very close to the nominal levels and quite similar to the results under the Bayesian approach (size = 0.099 and power = 0.899 in Table 2).

Acknowledgments

Funding

The National Institutes of Health supported the research of all three authors under USPHS grant P30-CA016520.

References

- Schlesselman JJ, Reis IM. Phase II clinical trials in oncology: Strengths and limitations of two-stage designs. *Cancer Invest.* 2006; 24:404–12. [PubMed: 16777694]
- Rubinstein LV, Korn EL, Freidlin B, et al. Design issues of randomized phase II trials and a proposal for phase II screening trials. *J Clin Oncol.* 2005; 23:7199–206. [PubMed: 16192604]
- Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *J Chron Dis.* 1961; 13:364–53.
- Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics.* 1982; 38:143–51. [PubMed: 7082756]

5. Lee YJ. Phase II trials in cancer: Present status and analysis methods. *Drug Exp Clin Res.* 1986; 12:57–71.
6. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials.* 1989; 10:1–10. [PubMed: 2702835]
7. Ye F, Shyr Y. Balanced two-stage designs for phase II clinical trials. *Clin Trials.* 2007; 4:514–24. [PubMed: 17942467]
8. Jung SH, Carey M, Kim KM. Graphical search for two-stage designs for phase II clinical trials. *Clin Trials.* 2001; 22:367–72.
9. Jung SH, Lee T, Kim KM, George SL. Admissible two-stage designs for phase II cancer clinical trials. *Stat Med.* 2004; 23:561–69. [PubMed: 14755389]
10. Green SJ, Dahlberg S. Planned versus attained design in phase II clinical trials. *Stat Med.* 1992; 11:853–62. [PubMed: 1604065]
11. Chen TT, Ng T-H. Optimal flexible designs in phase II clinical trials. *Stat Med.* 1998; 17(20): 2301–12. [PubMed: 9819829]
12. Masaki N, Koyama T, Yoshimura I, Hamada C. Optimal two-stage designs allowing flexibility in number of subjects for phase II clinical trials. *J Biopharm Stat.* 2009; 17:721–31. [PubMed: 20183436]
13. Koyama T, Chen H. Proper inference from Simon's two-stage designs. *Stat Med.* 2008; 27:3145–54. [PubMed: 17960777]
14. Thall PF, Simon R. A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. *Control Clin Trials.* 1994; 15:463–81. [PubMed: 7851108]
15. Heitjan DF. Bayesian interim analysis of phase II cancer clinical trials. *Stat Med.* 1997; 16:1791–802. [PubMed: 9280033]
16. Sylvester RJ. A Bayesian approach to the design of phase II clinical trials. *Biometrics.* 1998; 44:823–36. [PubMed: 3203131]
17. Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. *Clin Trials.* 2008; 5:93–106. [PubMed: 18375647]
18. Johnson VE, Cook JD. Bayesian design of single-arm phase II clinical trials with continuous monitoring. *Clin Trials.* 2009; 6:217–26. [PubMed: 19528131]
19. Mayo MS, Gajewski BJ. Bayesian sample size calculations in phase II clinical trials using informative conjugate priors. *Control Clin Trials.* 2004; 25:157–67. [PubMed: 15020034]
20. Herndon JE. A design alternative for two-stage, phase II, multicenter cancer clinical trials. *Control Clin Trials.* 1998; 19:440–50. [PubMed: 9741865]
21. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika.* 1986; 73:573–81.
22. Chang MN, Wieand HS, Chang VT. The bias of the sample proportion following a group sequential phase II clinical trial. *Stat Med.* 1989; 8:563–70. [PubMed: 2727475]
23. Jung SH, Kim KM. On the estimation of the binomial probability in multistage clinical trials. *Stat Med.* 2004; 23:881–96. [PubMed: 15027078]
24. Guo HY, Liu A. A simple and efficient bias-reduced estimator of response probability following a group sequential phase II trial. *J Biopharm Stat.* 2005; 15:773–81. [PubMed: 16078384]

Figure 1.

Type I and type II error rates under the Bayesian adaptation of Simon's optimal design, with the target levels of $\alpha = 0.05$ and $\beta = 0.2$ and $\alpha = 0.05$ and $\beta = 0.1$. Each point/triangle represents a design with a specific value of p_0 (assuming $\pi = 0.15$). The dots show the realized type I error, the triangles show the realized type II error, and solid lines show the target levels

Figure 2.

Properties of the Bayesian approach for Simon's optimal design under uniform, optimistic, and pessimistic priors (target $\alpha = \beta = 0.1$). Each point/square/triangle represents a design with a specific value of p_0 (assuming $\gamma = 0.15$).

Figure 3.

Properties of the Bayesian approach for Simon's optimal design with $p_0 = 0.3$, $p_1 = 0.45$, and target $\alpha = \beta = 0.1$, conditional on the attained SSs. The gray flat lines represent the targeted levels of type I error and power.

SS: sample size.

Table 1

The corresponding Bayesian approach for Simon's optimal design with target $\alpha = \beta = 0.1$; uniform prior for p ; $p_0 = 0.2$; $p_1 = 0.4$

Interim analysis		Final analysis	
Attained SS	Rejection region^a	Attained SS	Rejection region^b
11	1	28–31	8
12–15	2	32–35	9
16	3	36	10
17	3	37	10
18	3	38–39	10
19–21	4	40–43	11
22–23	5	44–46	12

SS: sample size.

The values given in boldface type designate the planned SSs and cut points in the original design.

^aThe rejection region is implied by the Bayesian decision criterion $\theta_I = 0.063$.

^bThe rejection region is implied by the Bayesian decision criterion $\theta_U = 0.908$.

Properties of Simon's optimal design with planned SSs (F) and the corresponding Bayesian approach with random SSs (B), with target $\alpha = \beta = 0.1$

Table 2

p_0	p_1	Decision criteria				PET(p_0)			PET(p_1)			ESS(p_0)			Size			Power		
		r_1/n_1	r/n	θ_I	θ_u	F	B		F	B		F	B		F	B		F	B	
0.15	0.10	0.25	2/21	7/50	0.112	0.901	0.648	0.649	0.075	0.076	31.2	31.2	0.098	0.099	0.901	0.899				
	0.30	0.45	9/30	29/82	0.079	0.886	0.589	0.589	0.069	0.070	51.4	51.4	0.099	0.099	0.901	0.899				
	0.50	0.65	18/35	47/84	0.067	0.882	0.632	0.631	0.068	0.069	53.0	53.1	0.095	0.100	0.900	0.902				
	0.70	0.85	14/20	45/59	0.056	0.869	0.584	0.592	0.067	0.072	36.2	35.9	0.095	0.098	0.901	0.898				
0.20	0.10	0.30	1/12	5/35	0.133	0.896	0.659	0.672	0.085	0.093	19.8	19.5	0.098	0.097	0.901	0.893				
	0.30	0.50	7/22	17/46	0.076	0.888	0.671	0.701	0.067	0.080	29.9	29.2	0.097	0.094	0.905	0.894				
	0.50	0.70	11/21	26/45	0.065	0.879	0.668	0.666	0.068	0.069	29.0	29.0	0.096	0.103	0.902	0.904				
	0.70	0.90	6/9	22/28	0.041	0.859	0.537	0.550	0.053	0.060	17.8	17.6	0.099	0.103	0.910	0.906				

SS: sample size; PET: probability of early termination; ESS: expected sample size.

Table 3

Properties of Simon's minimax design with planned SSs (F) and the corresponding Bayesian approach with random SSs (B), with target $\alpha = \beta = 0.1$

p_0	p_1	Decision criteria				PET(p_0)			PET(p_1)			ESS(p_0)			Size			Power		
		r/n_1	r/n	θ_1	θ_u	F	B	θ_u	F	B	θ_u	F	B	θ_u	F	B	F	B	F	B
0.15	0.10	2/27	6/40	0.036	0.921	0.485	0.485	0.021	0.021	0.021	0.021	33.7	33.7	0.098	0.100	0.100	0.900	0.899		
	0.30	16/50	25/69	0.047	0.900	0.684	0.684	0.043	0.043	0.043	0.043	56.0	56.0	0.100	0.100	0.100	0.902	0.901		
	0.50	19/40	41/72	0.017	0.900	0.437	0.462	0.017	0.021	0.021	0.021	58.0	57.2	0.096	0.101	0.101	0.900	0.904		
	0.70	15/22	40/52	0.031	0.878	0.506	0.515	0.037	0.040	0.040	0.040	36.6	36.6	0.098	0.102	0.102	0.903	0.904		
0.2	0.10	1/16	4/25	0.048	0.924	0.515	0.530	0.026	0.029	0.029	0.029	20.4	20.2	0.095	0.101	0.101	0.903	0.903		
	0.30	7/28	15/39	0.008	0.911	0.365	0.365	0.006	0.007	0.007	0.007	35.0	35.0	0.094	0.095	0.095	0.900	0.899		
	0.50	11/23	23/39	0.021	0.895	0.500	0.499	0.021	0.022	0.022	0.022	31.0	31.0	0.098	0.098	0.098	0.902	0.900		
	0.70	11/16	20/25	0.013	0.882	0.550	0.560	0.017	0.019	0.019	0.019	20.1	20.0	0.090	0.097	0.097	0.902	0.905		

SS: sample size; PET: probability of early termination; ESS: expected sample size.

Properties of Ye and Shyr’s balanced design with planned SSs (F) and the corresponding Bayesian approach with random SSs (B), with target $\alpha = \beta = 0.1$

Table 4

p_0	p_1	Decision criteria				PET(p_0)		PET(p_1)		ESS(p_0)		Size		Power	
		r/n_1	r/n	θ_I	θ_u	F	B	F	B	F	B	F	B	F	B
0.15	0.10	2/24	7/48	0.064	0.918	0.560	0.565	0.040	0.040	34.5	34.4	0.093	0.094	0.918	0.917
	0.30	12/39	28/78	0.057	0.900	0.620	0.618	0.050	0.051	53.9	53.9	0.093	0.093	0.908	0.907
	0.50	20/39	44/78	0.053	0.891	0.630	0.625	0.054	0.055	53.6	53.7	0.094	0.099	0.901	0.903
	0.70	18/26	40/52	0.027	0.878	0.540	0.548	0.032	0.034	38.0	37.8	0.099	0.104	0.908	0.909
0.20	0.10	2/17	5/34	0.112	0.907	0.760	0.771	0.077	0.084	21.1	20.9	0.088	0.087	0.909	0.903
	0.30	6/21	16/42	0.047	0.909	0.550	0.586	0.039	0.048	30.4	29.7	0.090	0.088	0.901	0.894
	0.50	12/23	27/46	0.053	0.903	0.660	0.659	0.055	0.055	30.8	30.9	0.081	0.081	0.904	0.902
	0.70	11/15	24/30	0.043	0.901	0.700	0.700	0.056	0.056	19.5	19.5	0.068	0.072	0.901	0.903

SS: sample size; PET: probability of early termination; ESS: expected sample size.

Properties of Jung *et al.*'s admissible design with planned SSs (F) and the corresponding Bayesian approach with random SSs (B), with target $\alpha = \beta = 0.1$

Table 5

p_0	p_1	Decision criteria				PET(p_0)			PET(p_1)			ESS(p_0)			Size			Power		
		r/n_1	r/n	θ_1	θ_u	F	B	θ_u	F	B	θ_u	F	B	θ_u	F	B	F	B	F	B
0.15	0.10	0.25	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	0.30	0.45	11/37	26/72	0.049	0.900	0.566	0.566	0.043	0.043	0.043	52.2	52.2	0.097	0.097	0.097	0.901	0.900	0.901	0.900
	0.50	0.65	18/37	42/74	0.030	0.897	0.500	0.526	0.030	0.035	0.035	55.5	54.5	0.096	0.095	0.095	0.902	0.898	0.902	0.898
	0.70	0.85	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.20	0.10	0.30	2/18	4/26	0.090	0.913	0.734	0.743	0.060	0.065	0.065	20.1	20.1	0.100	0.100	0.103	0.904	0.901	0.904	0.901
	0.30	0.50	6/21	16/42	0.047	0.909	0.551	0.586	0.039	0.048	0.048	30.4	29.7	0.090	0.088	0.088	0.901	0.894	0.901	0.894
	0.50	0.70	8/17	24/41	0.040	0.889	0.500	0.538	0.040	0.051	0.051	29.0	28.1	0.098	0.103	0.103	0.901	0.898	0.901	0.898
	0.70	0.90	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

SS: sample size; PET: probability of early termination; ESS: expected sample size.