Statistics
in Medicine

Research Article

Received 5 January 2011, Accepted 8 December 2011 Published online 23 February 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4511

Bayesian decision theoretic two-stage design in phase II clinical trials with survival endpoint

Lili Zhao,[a*†] Jeremy M. G. Taylor[b] and Scott M. Schuetze[c]

In this paper, we consider two-stage designs with failure-time endpoints in single-arm phase II trials. We propose designs in which stopping rules are constructed by comparing the Bayes risk of stopping at stage I with the expected Bayes risk of continuing to stage II using both the observed data in stage I and the predicted survival data in stage II. Terminal decision rules are constructed by comparing the posterior expected loss of a rejection decision versus an acceptance decision. Simple threshold loss functions are applied to time-to-event data modeled either parametrically or nonparametrically, and the cost parameters in the loss structure are calibrated to obtain desired type I error and power. We ran simulation studies to evaluate design properties including types I and II errors, probability of early stopping, expected sample size, and expected trial duration and compared them with the Simon two-stage designs and a design, which is an extension of the Simon's designs with time-to-event endpoints. An example based on a recently conducted phase II sarcoma trial illustrates the method. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: Bayesian; decision theory; time to event; phase II clinical trial; two-stage design

1. Introduction

In phase II cancer clinical trials, the standard approach consists of a single-arm design where a single binary endpoint is compared with a specified target value. The sample sizes are typically small, maybe 30∼70 patients. Improvements to the study could be made by increasing the sample size, using randomization and using an endpoint that is more informative than a binary one. Limitations on the available number of patients frequently limits power for a randomized study. Our focus in this paper will be on enhancing the trials using nonbinary endpoints. We will consider designs with failure-time endpoints, measuring time until some event, such as a device-related complication, disease progression, relapse, or death. Progression-free survival (PFS) is increasingly used as an endpoint for cancer clinical trials, and a recent review suggests that PFS is being utilized more commonly and may predict for greater success in the phase III setting [1]. Using PFS as an endpoint in single-arm studies does raise some issues concerning the possibility of bias in the comparison with the historical control group [2–4]. Although there is always the possibility of differences in the populations and in methods for assessing the endpoint between the trial and the control group, when using PFS, a consistent surveillance strategy for assessing progression is also needed.

A single-arm phase II trial is typically designed to accrue patients in two stages [5–7], with the Simon design being very popular. It will stop at stage I if a preset level of futility has been demonstrated, thereby reducing the number of patients exposed to an ineffective therapy. Similar to classical phase II studies, a study based on a time-to-event endpoint (such as median PFS or PFS rate at a specific clinical landmark point, t_0) will be deemed a success if there is sufficient statistical evidence to conclude that the endpoint

[a]*Biostatistics Unit, University of Michigan Comprehensive Cancer Center, Ann Arbor, MI 48109, U.S.A.*
[b]*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[c]*Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
**Correspondence to: Lili Zhao, Biostatistics Unit, University of Michigan Comprehensive Cancer Center, Ann Arbor, MI 48109, U.S.A.*
[†]*E-mail: zhaolili@umich.edu*

exceeds, at a clinically relevant level, that of a relevant historical control. Many statistical designs are based on the probability that the patient survives to specific time t_0 without suffering the event. The most severe problem created by this approach is that a patient has to be followed-up for t_0 time to ensure that event has not occurred, and waiting until all patients in stage I complete the follow-up of t_0 may cause long recruitment suspension especially when t_0 is large. The impact of study suspension on accrual momentum and timeliness of the studies completion is often negative. A number of authors have formulated underlying statistical models and interim decision rules directly in terms of time-to-event variables to overcome this problem. Herndon [8] proposed a frequentist *ad hoc* approach to conducting two-stage phase II studies to avoid study suspension. Case and Morgan [9] proposed frequentist two-stage phase II designs using the estimator developed by Lin *et al.* [10] to minimize the expected sample size or expected total study length (ETSL) under H_0. Huang *et al.* [11] modified their approach to protect type I error rate and improve robustness of the design. These designs [9, 11] are essentially an extension of the Simon designs with failure time endpoints using nonparametric statistics.

Researchers developed several Bayesian approaches to continuously monitor survival endpoints. Follman and Albert [12] used a Dirichlet process prior for the probabilities of the event on a large set of potential discretized event times. They compute an approximate posterior distribution that is a mixture of Dirichlet processes by using a data augmentation algorithm. Rosner [13] took a similar approach but used Gibbs sampling to generate posteriors. Cheung and Thall [14] constructed futility monitoring rules on the basis of an approximate posterior through a weighted average of beta distributions for one or more event times in phase II trials. These approaches incorporate the censored data into the posterior estimation in a nonparametric fashion. Thall *et al.* [15] developed model-based approaches to monitor time-to-event endpoints, assuming exponentially distributed failure times with an inverse gamma prior on the mean. They also examined the robustness of the method by assuming that the survival data follows a generalized gamma distribution. In the aforementioned Bayesian designs (e.g., [14, 15]), decisions are made using the posterior distribution of the clinically relevant survival endpoints, and the futility monitoring rule is typically based on $P(p_E > p_S + \delta|\text{data}) < p_L$. Thus, the trial is stopped early if the posterior probability that p_E (such as median survival or survival rate at a specific time point of the experimental treatment) exceeds a clinical meaningful threshold (δ) over the traditional treatment p_S by less than a prespecified cutoff, p_L. The p_L can be calibrated by simulations to obtain good trial properties such as probability of early termination, type I error, and power [16].

The property of Bayesian procedures to accumulate evidence based on updated data is very attractive in clinical trial designs. However, at the end of stage I, most investigators are interested in knowing what is the probability that the study would yield a significant result in favor of the new treatment if the study were to be continued to stage II, given what has been observed to date. Lachin [17] reviewed many frequentist approaches that construct stopping rules on the basis of an assessment of this idea using (predictive) conditional power. Pepe and Anderson [18] presented expressions for the types I and II error probabilities for use with time-to-event endpoints assuming current trends continue for survival data. Berry [19] is also a strong advocate for the use of predictive probabilities in making decisions. However, the aforementioned approaches (frequentist or Bayesian) use only the predictive distribution to make interim decisions. In this paper, we will weight the evidence for stopping or continuing on the basis of posterior and predictive distributions. Once the trial is stopped, the posterior distribution is used for making terminal decisions.

In the literature, researchers have developed decision-based Bayesian methods for binary endpoints [20–24]. Zhao and Woodworth [25] proposed a decision-based Bayesian approach for continually monitoring survival endpoints in single-arm phase II trials with medical devices, but the time-to-event data was assumed to be exponentially distributed.

In practice, the decision theoretic clinical trial designs remain relatively uncommon. The main reasons for this lack of application are fundamental concerns with the decision theoretic setup and practical difficulty of specifying a good loss function [26]. In decision-theoretic approaches, costs need to be specified, and in most situations, it is very hard to relate the costs to tangible quantities. Our strategy to make this approach feasible is to treat the cost parameters, defined in a simple loss structure, as tuning parameters, which are calibrated to achieve desired operating characteristics such as type I error and power. This approach alleviate concerns about difficulties in specifying the loss function, and as we will show, the properties of the decision theoretic designs appear to be very attractive.

Section 2 presents the general framework and methodology. Section 3 contains simulation studies to evaluate the properties of the proposed methods and compares the results to two frequentist designs. Section 4 contains results from a phase II sarcoma trial. Section 5 is the concluding discussion.

2. Method

The primary endpoint is a time-to-event outcome at some clinically meaningful landmark point t_0, such as 6 months or 1 year from the start of the treatment. Let $S(t_0)$ be the survival rate at t_0. Similar to most other designs with a binary endpoint of tumor response, decisions are made based on two hypotheses H_0: $S(t_0) \leqslant p_1$ that the true survival rate at t_0 is less than some uninteresting level p_1 and H_1: $S(t_0) \geqslant p_2$ that the survival rate is at least some desirable target level p_2.

With right-censored data, the likelihood for n subjects is

$$L(\Theta|D) = \prod_{i=1}^{n} \{f(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i}$$

and

$$S(t_0) = e^{-H(t_0)}, \quad \text{where} \quad H(t_0) = \int_0^{t_0} h(u)\,\mathrm{d}u,$$

where Θ are the parameters that determine the distribution $f(t)$ and D is the observed data, which include survival times, $t = (t_1, t_2, \ldots, t_n)$ and corresponding indicators of censoring, $\delta = (\delta_1, \delta_2, \ldots, \delta_n)$.

In this paper, we have chosen a very simple loss function as specified in Table I. There are two possible wrong decisions: (1) false rejection (type I error) and (2) false acceptance (type II error). c_2 is the penalty you are willing to pay for a false rejection decision over a false acceptance decision. In decision theory, c_2 could be a function of the sample size or could differ from stage I to stage II. In this paper, not to overcomplicate things, we consider c_2 as fixed and treat c_2 as a tuning parameter to control type I error and power. A rejection decision will be made if the posterior expected loss of making a rejection decision (e.g., $c_2 P(S(t_0) <= p_1|D)$) is less than that of making an acceptance decision (e.g., $P(S(t_0) >= p_2|D)$), and an acceptance decision will be made otherwise. The higher the c_2, the less likely we are going to reject H_0 to avoid the high penalty of a false rejection decision. But before we decide to stop and make a terminal decision (reject or accept H_0), we should compare the risk of stopping at stage I to the 'expected' risk if the trial will be continued to stage II. The (posterior) Bayes risk of immediate stopping at stage I, denoted by $\rho_0(\pi^1)$, is defined as the minimum of the (posterior) expected losses under two decisions,

$$\rho_0(\pi^1) = \min\{P(S(t_0) \geqslant p_2|D_1), c_2 P(S(t_0) \leqslant p_1|D_1)\}$$

where D_1 are the observed data up to stage I; π^1 is the posterior distribution of $S(t_0)$ given the observed data up to stage I, denoted by $f(S(t_0)|D_1)$.

The expected (posterior) Bayes risk of continuing to stage II, denoted by $\rho_c(\pi^1)$, is defined as,

$$\rho_c(\pi^1) = E_{D_2|D_1}[\rho_0(\pi^2)] + c_3 \tag{1}$$

$$= E_{D_2|D_1}[\min\{P(S(t_0) \geqslant p_2|D_2), c_2 P(S(t_0) \leqslant p_1|D_2)\}] + c_3 \tag{2}$$

Table I. Threshold loss structure.

		True status of the treatment		
			Noninferior	
		Inferior	Neither	Superior
		$(S(t_0) \leqslant p_1)$	$(p_1 < S(t_0) < p_2)$	$(S(t_0) \geqslant p_2)$
Decision	Accept H_0	0	0	1
	Reject H_0	c_2	0	0

where π^2 is the posterior distribution of $S(t_0)$ given the observed data up to stage II, denoted by $f(S(t_0)|D_2)$. c_3 is the cost of running the trial to the final stage, which could be a function of the sample size or study length in stage II. In this paper, c_3 is also fixed to control the probability of stopping at the end of stage I. The higher the c_3, the more likely the trial will be halted at stage I to avoid the high cost of continuing the trial. $E_{D_2|D_1}[\rho_0(\pi^2)]$ defined in (1) can be approximated by

$$E_{D_2|D_1}[\rho_0(\pi^2)] \approx \frac{1}{B}\sum_{l=1}^{B}[\rho_0(\pi^{2,l})] \quad \text{when } B \text{ is large.} \tag{3}$$

where $\pi^{2,l}$ is $f(S(t_0)|D_{2,l})$ and $D_{2,l}$ $(l = 1, 2, \ldots, B)$ are the random samples of potential datasets D_2 generated given D_1.

In brief, the decision rule is as follows:

- At the end of stage I
  - if $\rho_0(\pi^1) \leqslant \rho_c(\pi^1)$, then stop and
    - Reject H_0 if $\frac{P(S(t_0)>p_2|D_1)}{P(S(t_0)<p_1|D_1)} > c_2$
    - Accept otherwise
  - else if $\rho_0(\pi^1) > \rho_c(\pi^1)$, then continue to stage II
- At the end of stage II, stop and
  - Reject H_0 if $\frac{P(S(t_0)>p_2|D_2)}{P(S(t_0)<p_1|D_2)} > c_2$
  - Accept otherwise

The aforementioned decision rules are attractive for two reasons. First, the terminal decision rule is constructed using the ratio of two posterior probabilities under two hypotheses (e.g., similar to posterior odds). Second, the decision to 'stop' or 'continue' is the result of weighting the evidence between the observed data and the data that will be observed if the trial would continue.

Once the posterior distribution of Θ is known, the distribution of $S(t_0)$, as a function of Θ, can be computed using Monte Carlo methods. Therefore, it is easy to calculate the posterior probabilities under the two hypotheses. It is, however, much harder to calculate the Bayes risk of continuation, which involves the posterior predictive distribution of D_2 given D_1 in the presence of censoring. In the following section, we derive algorithms to estimate $E_{D_2|D_1}[\rho_0(\pi^2)]$ for exponential failure time distributions, Weibull distributions, and time-to-event data that do not follow any parametric distribution.

2.1. *Method assuming exponential distribution*

Assume the time-to-event data follows an exponential distribution, $f(t) = \lambda e^{-\lambda t}$ for $t > 0$, and the primary endpoint $S(t_0)$ is defined as $S(t_0) = e^{-\lambda t_0}$.

The likelihood function for λ at stage k is $L(\lambda|D_k) = \lambda^{f^k} e^{-\lambda e^k}$, where f^k is the total number of failures and e^k is the total exposure time up to stage k $(k = 1, 2)$. These are sufficient statistics to estimate λ. D then can be simplified to $D_1 = (f^1, e^1)$ and $D_2 = (f^2, e^2)$. For mathematical convenience, we use the conjugate prior for λ such that $\lambda \sim \Gamma(\alpha_0, \beta_0)$. Then, the posterior distribution is easily determined to be a Gamma distribution, $f(\lambda|D_k) \sim \Gamma(\alpha_0 + f^k, \beta_0 + e^k)$.

Given the data up to stage I (D_1), we can simulate the patients' survival data in stage II. The number of patients in stage II (denoted by N_2) has two parts: (1) the number of event-free patients that have not reached the final study point at the end of stage I; and (2) the number of new patients recruited in stage II. Because of the 'lack of memory" property of the exponential distribution, those patients that are event free at the end of stage I are conditionally exchangeable with patients who are recruited in stage II, and their survival times in stage II will also be distributed as an exponential distribution with rate $\lambda \sim \Gamma(\alpha_0 + f^1, \beta_0 + e^1)$. Zhao and Woodworth [25] defined the algorithm to obtain samples of D_2 given D_1.

2.2. *Method assuming Weibull distribution*

The Weibull distribution, denoted by $\mathcal{W}(\alpha, \gamma)$, is $f(t) = \alpha\gamma t^{\alpha-1} e^{-\gamma t^\alpha}$ for $t > 0, \alpha > 0$, and $\gamma > 0$. Let $\Theta = (\alpha, \gamma)$ and the primary endpoint is $S(t_0) = e^{-\gamma t_0^\alpha}$.

In this case, $D_k = (t^k, \delta^k)$, $k = 1, 2$. t^k includes all survival times for patients enrolled up to stage k, and δ^k are the corresponding censoring indicators. Unlike the exponential distribution, there are no

simple sufficient statistics and no simple posterior distributions for α and γ. MCMC methods will be used to estimate these two parameters.

This Weibull model can be expressed as a log-linear model as $\log t_i = \mu + \sigma\epsilon_i$ [27], where $\mu = -\log(\gamma)/\alpha$ and $\sigma = 1/\alpha$. The density of the log time, $y_i = \log t_i$, is given by

$$f(y_i) = \frac{1}{\sigma}\exp(z_i - e^{z_i}), \quad \text{where } z_i = \frac{y_i - \mu}{\sigma}$$

If we assign μ a uniform prior and for σ the usual noninformative prior proportional to $1/\sigma$, then the posterior density is given by

$$f(\mu, \sigma|D_k) \propto \frac{1}{\sigma}L(\mu, \sigma|D_k), \qquad k = 1, 2 \tag{4}$$

Metropolis–Hasting algorithms can easily be applied to estimate μ and σ.

Given the data up to stage I ($D_1 = (t^1, \delta^1)$), survival times for patients enrolled in stage II will be generated from a Weibull distribution with the updated Θ from stage I. For any patient i that is censored at the end of stage I and has not reached the final study point, the probability that the patient will survive additional time x_i is expressed as

$$S(x_i) = P(t > t_i^1 + x_i | t > t_i^1) = e^{-\gamma\left((t_i^1 + x_i)^\alpha - {t_i^1}^\alpha\right)} \tag{5}$$

To simulate this additional time, we simulate $S(x_{i,l})$ from $U(0, 1)$, and solving for x gives

$$x_{i,l} = \left\{\frac{-\log(S(x_{i,l}))}{\gamma_l} + {t_i^1}^{\alpha_l}\right\}^{\frac{1}{\alpha_l}} - t_i^1,$$

where α_l and γ_l are from the lth MCMC iteration. Then, $x_{i,l}$, $l = 1, \ldots, B$ are random samples from the posterior predictive distribution of the remaining survival time of patient i in stage II.

Let $t_{i,l}^2 = \min\{t_i^1 + x_{i,l}, M_i\}$, $\delta_{i,l}^2 = 1$ if $t_{i,l}^2 < M_i$ and $\delta_{i,l}^2 = 0$ otherwise, where M_i is the maximum follow-up for patient i defined from the enrollment to the final study time point. We repeat this process for all patients ($i = 1, \ldots, N_2$) in stage II, and together with the observed survival data of patients that had event, we obtain $D_{2,l}$. Then, $D_{2,1}, \ldots, D_{2,B}$ form a random sample from the posterior predictive distribution of $f(D_2|D_1)$.

Similar methods can be developed assuming Gamma or log-normal distributions.

2.3. *Grouped-data method*

In practice, interval-censored survival data is common in medical settings where the patient's disease status is evaluated periodically by tests such as magnetic resonance imaging or computed axial tomography scan. The actual time of any patients disease progression is not available, rather, it is only known whether progression occurred during each time interval between successive examinations. To account for this type of interval censoring, we propose a nonparametric method in this section, which handles the interval-censored data for various time-to-event data distributions.

Given the data up to stage I (D_1), we construct a finite partition of the time axis, $0 < s_1 < s_2 < \cdots < s_J$, with $s_J = t_0$. Thus, we have the J disjoint intervals $(0, s_1], (s_1, s_2], \ldots, (s_{J-1}, s_J]$, and $I_j = (s_{j-1}, s_j]$. The value of s_j ($j = 1, \ldots, J$), in the interval-censored case, should be determined by the intended gap between two consecutive scheduled appointments. For example, the scheduled appointment is every 2 months for the first half year, then the intervals can be set up as $s_1 = 2, s_2 = 4, s_3 = 6$, and $J = 3$. For studies where the evaluation is more frequent, the value of J would be much larger.

The observed data D is assumed to be available as grouped within these intervals such that $D_k = (\mathcal{R}_j^k, \mathcal{D}_j^k : j = 1, 2, \ldots, J)$, where $\mathcal{R}_j^k$ is the risk set and $\mathcal{D}_j^k$ is the failure set of the jth interval I_j up to stage k. Let h_j denote the increment in the cumulative baseline hazard in the jth interval, that is,

$$h_j = H_0(s_j) - H_0(s_{j-1}), \qquad j = 1, 2, \ldots, J \tag{6}$$

and

$$S(t_0) = \exp\left(-\sum_{j=1}^{J} h_j\right)$$

The grouped data likelihood is

$$L(\mathbf{h}|D_k) \propto \prod_{j=1}^{J} G_j^k, \qquad \mathbf{h} = (h_1, h_2, \ldots, h_J)$$

and

$$G_j^k = \exp\left\{-h_j^k\left(r_j^k - d_j^k\right)\right\}\left\{1 - \exp\left(-h_j^k\right)\right\}^{d_j^k} \tag{7}$$

where r_j^k and d_j^k are the number of subjects in the sets $\mathcal{R}_i^k$ and $\mathcal{D}_j^k$ up to stage k, respectively.

The Gamma process is used as a prior for the cumulative baseline hazard function $H_0(t)$ [28] such that

$$H_0 \sim \mathcal{GP}(\tau_0 H^*, \tau_0), \tag{8}$$

where $H^*(t)$ is an increasing function with $H^*(0) = 0$. H^* is assumed to be a Weibull distribution with hyperparameter η_0 and κ_0, such that $H^*(t) = \eta_0 t^{\kappa_0}$. τ_0 is a positive scalar quantifying the degree of prior confidence in $H^*(t)$. If $\kappa_0 = 1$, this simplifies as an exponential distribution with rate η_0.

The Gamma process prior in (8) implies that h_j's are independent and

$$h_j \sim \Gamma(\tau_0(H^*(s_j) - H^*(s_{j-1})), \tau_0). \tag{9}$$

With the likelihood and priors set up as that previously discussed, we can write the posterior distribution of $\mathbf{h}$ as specified in [29]

$$f(\mathbf{h}|D_k) \propto \prod_{j=1}^{J} G_j^k h_j^{\tau_0(H^*(s_j)-H^*(s_{j-1}))-1} e^{\tau_0 h_j} \tag{10}$$

We can carry out the following Gibbs sampling scheme sampling h_j from

$$\pi(h_j|\mathbf{h}^{-j}, D_k) \propto G_j^k h_j^{\tau_0(H^*(s_j)-H^*(s_{j-1}))-1} e^{\tau_0 h_j} \tag{11}$$

where $\mathbf{h}^{-j}$ denote the $\mathbf{h}$ vector without jth component. Plugging in the form for G_j^k from Equation (7), we have

$$\pi(h_j|\mathbf{h}^{-j}, D_k) \propto h_j^{\tau_0(H^*(s_j)-H^*(s_{j-1}))-1}(1 - e^{-h_j})^{d_j^k} e^{-(\tau_0 + r_j^k - d_j^k)h_j}. \tag{12}$$

Survival times are approximated by piecewise constant hazard survival model. Thus, the memoryless property of the exponential distribution holds in each interval I_j, $j = 1, 2, \ldots, J$. Given data observed up to stage I (D_1), survival times for patients in stage II ($i = 1, \ldots, N_2$) will be generated on the basis of the updated $h_1, \ldots, h_J$. For any patient i that is censored in interval I_j at the end of stage I and has not reached the final study point, the remaining survival time, $x_{i,j,l}$, will first be generated from an exponential distribution with rate $\frac{h_j}{s_j - s_{j-1}}$, then moving from left to right along the time line until the event occurs or the final study time point is reached, which will result in generated survival times, $x_{i,j,l}, x_{i,j+1,l}, \ldots$ in intervals $I_j, I_{j+1}, \ldots$, respectively. Let $t_{i,j,l} = \min\{x_{i,j,l}, (\min\{s_j, M_i\} - t_i^1)\}$; if $t_{i,j,l} = s_j - t_i^1$ (no event occurred in I_j and the final study point has not been reached), we will move to I_{j+1}. If no event occurred in I_{j+1} and $M_i \geqslant s_{j+1}$, then $t_{i,j+1,l} = s_{j+1} - s_j$. Finally,

$t^2_{i,l} = t^1_i + t_{i,j,l} + t_{i,j+1,l} + \ldots$. For patients enrolled in stage II, this process starts at I_1. We repeat this process for all patients ($i = 1, \ldots, N_2$) in stage II to obtain a random sample from the posterior predictive distribution of $f(D_2|D_1)$.

We note that with the use of a large number of intervals, this grouped-data method can be viewed as a nonparametric method for general distributions that would be applicable even if the data were not interval censored.

3. Simulation studies

In this section, we ran simulations to investigate properties of the proposed designs (Bayes designs) including type I error, power, probability of early stopping (PET), expected sample size, and ETSL under both null and alternative hypotheses. In the simulation setup, we assumed the landmark time point, t_0, to be 6 months; we took the interim analysis 1 day before the first patient of stage II was enrolled such that there is no trial suspension and defined the final study point as 6 months after the enrollment of the last patient; the maximum follow-up per patient is 12 months; we sampled the patient arrival time from a Poisson distribution with a rate of 0.1 so that one patient arrives an average of every 10 days.

In Simon's designs, we dichotomize the survival time into a binary variable $I(t \leqslant t_0)$. With such designs, the trial will continue immediately to stage II once the number of successes reaches the threshold rather than waiting for all enrolled patients to complete the follow-up and stops immediately if there is no hope that the threshold will be met. For example, consider the situation where three successes of 10 is needed for the trial to continue to stage II. If only one success is observed for the first nine patients that had completed the follow-up, the trial should stop even though the last patient has not completed the follow-up because we would observe at most two successes in this case. For the Simon designs, we will consider both the MiniMax and the optimal designs. To be comparable with the Simon's designs, which only allow early stopping for futility, the stopping rule in the Bayes design is modified as

$$\rho_c \geqslant \rho_0 \quad \text{and} \quad \frac{P(S(t_0) > p_2|D_1)}{P(S(t_0) < p_1|D_1)} \leqslant c_2.$$

This 'Stop only for futility' rule is used throughout the simulations unless specified otherwise.

As well as comparing with the Simon designs, we will also make comparisons with the method described in [11]. This method prevents trial suspension by using Nelson–Aalen estimates of the survival calculated at different calendar times to account for the information available from those with partial follow-up. We implemented this method using the R program *OptimPhase2*.

3.1. *Effect of the cost parameters*

In this simple loss structure, cost parameter c_2 controls the trade-off between types I and II errors and cost parameter c_3 controls the PET. In this section, we investigated the effect of c_2 and c_3 while fixing the other. We calculated total sample size ($n = 47$) and the sample size in stage I ($n_1 = 24$) from the Simon MiniMax design by restricting type I error to be 0.05 and type II error to be 0.15 with the two decision thresholds $p_1 = 0.1$ and $p_2 = 0.25$. We generated survival data from an exponential distribution with rate $-\frac{\log S(t_0)}{t_0}$, where $S(t_0) = 0.1$ or 0.25. We used an uninformative Gamma prior distribution, with mean 0.0001 and rate 0.0001, for λ.

Figure 1 demonstrates the effect of c_3 when c_2 is fixed to be 5. The highest c_3 (e.g., $c_3 = 0.1$) gives the highest PET across different $S(t_0)$, and c_3 of 0.03 provides similar PET as the MiniMax design. In Figure 2, c_3 is fixed to be 0.03. The highest c_2 (e.g., $c_2 = 10$) has the lowest probability of rejecting the H_0, resulting in the lowest power at $S(t_0) = 0.25$ as well as the lowest type I error at $S(t_0) = 0.1$. We found that the choice of $c_2 = 3$ and $c_3 = 0.03$ gives type I error of 0.03 and power of 0.9. To obtain approximately type I error of 0.05 and power of approximately 0.85, to match the MiniMax design, we reduced the sample size by 32% to $n = 32$ given $c_2 = 3$ and $c_3 = 0.03$. With this sample size, we investigated the operating characteristics of trials with different design parameters such as different hypotheses (p_1 and p_2), different patient accrual rates, different timing of the interim look, a stopping rule that allows stopping for both efficacy and futility, and different prespecified values for types I and II errors. Without loss of generality, we used the exponential method for data that are exponentially distributed.

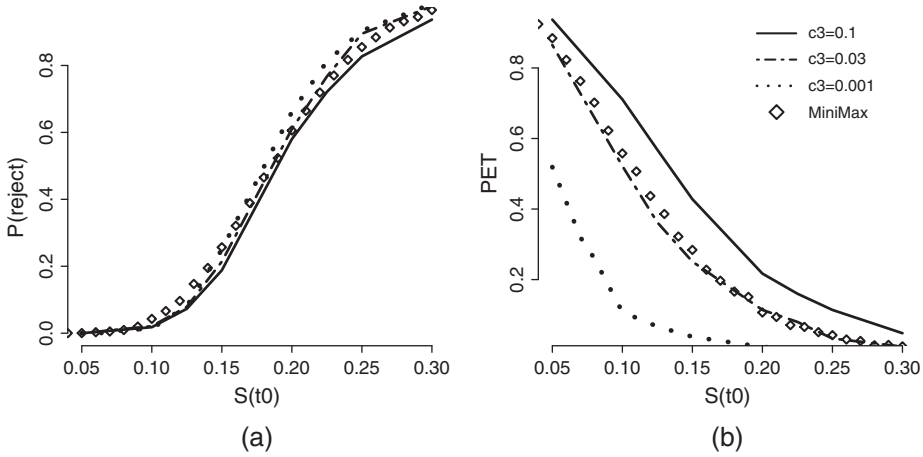

Figure 1. (a) Probability of rejecting H_0 for various $S(t_0)$, given three choices of c_3 with c_2 fixed to be 5, and (b) the corresponding probability of early stopping when p_1 is true for various $S(t_0)$, given three choices of c_3 with c_2 fixed to be 5 (on the basis of 1000 trial simulations and 1000 ($B = 1000$) simulations in calculating the expected Bayes risk of continuation).

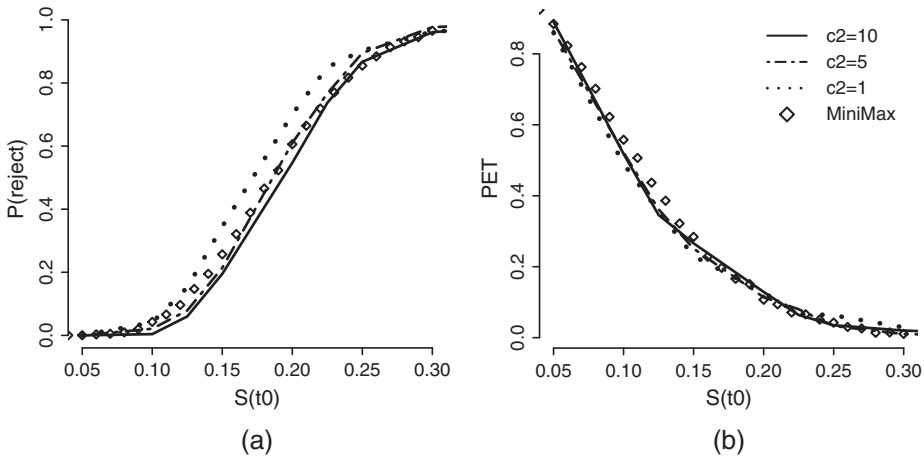

Figure 2. (a) Probability of rejecting H_0 for various $S(t_0)$, given three choices of c_2 with c_3 fixed to be 0.03, and (b) the corresponding probability of early stopping when p_1 is true for various $S(t_0)$, given three choices of c_2 with c_3 fixed to be 0.03 (on the basis of 1000 trial simulations and 1000 ($B = 1000$) simulations in calculating the expected Bayes risk of continuation).

3.2. Trial properties with different design parameters

3.2.1. Trials with different hypotheses. We base the results shown in Table II on 5000 simulations of the same trial and 1000 ($B = 1000$) simulations in calculating the Bayes risk of continuation. With $c_2 = 3$ and $c_3 = 0.03$, all the studied cases have type I error of approximately 0.05 and power of approximately 0.85 using about 30% less total sample size than the MiniMax design. The Bayes designs significantly shorten the average trial duration with much smaller expected sample sizes, compared with other designs. For example, in the first scenario, it shortens the total study length by 4 months if $S(t_0)$ is 0.1 and by 7 months if $S(t_0)$ is 0.25, compared with the MiniMax design. The savings in sample size are even more under the alternative hypothesis because of the dramatically reduced total sample size and seamless enrollment. The PET is low in the Bayes design under the null hypothesis. This, nonetheless, does not result in longer trial duration and larger expected sample size because of the dramatically reduced total sample size. And increasing n_1 would increase the PET as will be presented in the next section. The outperformance of the Bayes designs over the Simon's designs is mainly due to making use of more information with the time-to-event data rather than the dichotomized data in Simon's designs.

Table II. Operating characteristics of the exponential method for trials with different hypotheses given type I error of approximately 0.05 and power of approximately 0.85, with $c_2 = 3$ and $c_3 = 0.03$, based on exponentially distributed survival data.

		Under H_0			Under H_1		
Design	$n(n_1)$ [a]	PET	EN	ETSL[b]	PET	EN	ETSL
$0.10 vs 0.25$							
MiniMax[c]	47(24)	0.57	34	17	0.04	46	23
Optimal[d]	50(20)	0.68	30	15	0.09	47	24
H-ETSL[e]	52(33)	0.62	40	16	0.09	50	22
Bayes	32(16)	0.34	27	13	0.04	31	16
$0.2 vs 0.4$							
MiniMax	37(17)	0.55	26	15	0.05	36	21
Optimal	51(17)	0.76	25	13	0.12	47	25
H-ETSL	49(31)	0.67	37	14	0.08	48	21
Bayes	26(13)	0.25	23	12	0.03	26	14
$0.3 vs 0.45$							
MiniMax	76(35)	0.51	55	24	0.04	75	33
Optimal	86(29)	0.64	49	22	0.09	81	36
H-ETSL	98(48)	0.68	64	23	0.08	94	37
Bayes	54(27)	0.38	44	18	0.04	53	23
$0.5 vs 0.7$							
MiniMax	44(22)	0.60	31	16	0.04	43	24
Optimal	52(17)	0.67	28	15	0.09	49	26
H-ETSL	66(37)	0.69	46	17	0.06	64	27
Bayes	30(15)	0.23	27	13	0.03	30	16

PET, probability of early stopping; EN, expected sample size; ETSL, expected total study length.
[a]n in the Bayes designs is about 70% of the sample size calculated from the corresponding MiniMax design, and n_1 is half of the total sample size.
[b]The unit of ETSL is months.
[c]MiniMax design is targeted at minimizing the maximum sample size under the null hypothesis.
[d]Optimal design is targeted at minimizing the expected sample size under the null hypothesis.
[e]H-ETSL is targeted at minimizing the ETSL. The sample size is calculated using the OptimDes function in the *OptimPhase2* package in R while fixing the power 0.85 and type I error 0.05. The operating characteristics are obtained with the SimDes function in the *OptimPhase2* package.

On the contrary, the H-ETSL design, although it successfully avoids trial suspension, will not necessarily reduce the trial duration.

The optimal designs give smaller expected sample size and shorter averaged trial duration than the MiniMax designs under the null hypotheses (the opposite is true under the alternatives). We will use the optimal design for the comparisons for the rest of the article.

3.2.2. Trials with the interim look taken at a different time point. As illustrated in Table III, larger n_1 gives higher PET under the null hypothesis. Increasing n_1 from 16 to 24 further reduces the trial duration by 1 month. This fact suggests that interim analysis may be performed after we accumulate more data information (larger n_1 or longer follow-up), which will provide us with more evidence to terminate the trial early if the null hypothesis is true, thereby increasing the PET and shortening the trial duration especially when the patient actual is fast. In contrast with the Bayes designs, smaller n_1 in Simon's designs gives larger PET when the optimal design is compared with the MiniMax design (Table II). This reflects a fundamental problem in frequentist designs based on hypothesis testing, in which the hypotheses are set up to collect evidence to prove the treatment is promising (i.e., reject the H_0), which leads to the tendency to declare the treatment bad with smaller sample sizes.

Table III. Operating characteristics of the exponential method for hypothesis of 0.1 vs 0.25, with the interim look taken at different time points given type I error of approximately 0.05 and power of approximately 0.85, with $c_2 = 3$ and $c_3 = 0.03$, based on exponentially distributed survival data.

	Under H_0			Under H_1		
$n(n_1)$	PET	EN	ETSL	PET	EN	ETSL
32(8)	0.12	29	15	0.03	31	16
32(16)	0.34	27	13	0.04	31	16
32(24)	0.61	27	12	0.06	31	16

PET, probability of early stopping; EN, expected sample size; ETSL, expected total study length.

3.2.3. Trials with stopping for both efficacy and futility. Stopping for both futility and efficacy is naturally embedded within the decision-theoretic framework. As demonstrated in Table IV, type I error and power are increased slightly by allowing stopping for efficacy as well as for futility. Increasing c_2 (such as $c_2 = 4$) would lower the probability of rejection (that is, the type I error and power) by approximately 2% .

3.2.4. Trials with different interpatient arrival times. As presented in Figure 3, the Bayes design has the smallest averaged trial duration for a wide range of accrual rates under both the null and alternative hypotheses. The H-ETSL design only shows some benefit over the optimal design under the alternative hypothesis when the accrual rate is very fast.

3.2.5. Trials with different type I error and power. Different clinical trials may have different requirements for the type I error and power. In cancer trials, other choices may be as follows: (1) type I error of 5% with power of 80% or (2) type I error of 10% with power of 90 %. With 70 % of the total sample size calculated from the MiniMax design, we found, as shown in Table V, that $c_2 = 4$ and $c_3 = 0.03$ give type I error of approximately 0.05 and power of approximately 0.80, and $c_2 = 1$ and $c_3 = 0.02$ give type I error of approximately 0.1 and power of approximately 0.90.

3.3. Simulation using the Weibull method

Thus far, we have assumed that survival data is exponentially distributed. If the event rates are not constant, however, a more complex distribution is required. In this section, we applied the Weibull method for data generated from Weibull distributions with various shape parameters (denoted by α) given $S(t_0) = 0.1$ or 0.25. We base the results in Table VI on 500 simulated trials and 250 ($B = 250$) simulations to calculate the expected Bayes risk of continuation. With $c_2 = 3$ and $c_3 = 0.03$, the type I error is approximately 0.05 and power is approximately 0.85, regardless of different shape parameters.

Table IV. Operating characteristics of the exponential method with stopping for both futility and efficacy, for different hypotheses with $c_2 = 3$ and $c_3 = 0.03$, based on exponentially distributed survival data.

	Under H_0				Under H_1			
$n(n_1)$	Type I error	PET	EN	ETSL	Power	PET	EN	ETSL
$0.10 vs 0.25$								
32(16)	0.07	0.36	26	13	0.86	0.31	27	13
$0.2 vs 0.4$								
26(13)	0.08	0.29	22	12	0.85	0.23	23	12
$0.3 vs 0.45$								
54(27)	0.07	0.39	43	18	0.85	0.31	46	19
$0.5 vs .0.7$								
30(15)	0.08	0.27	26	13	0.88	0.24	26	13

PET, probability of early stopping; EN, expected sample size; ETSL, expected total study length.

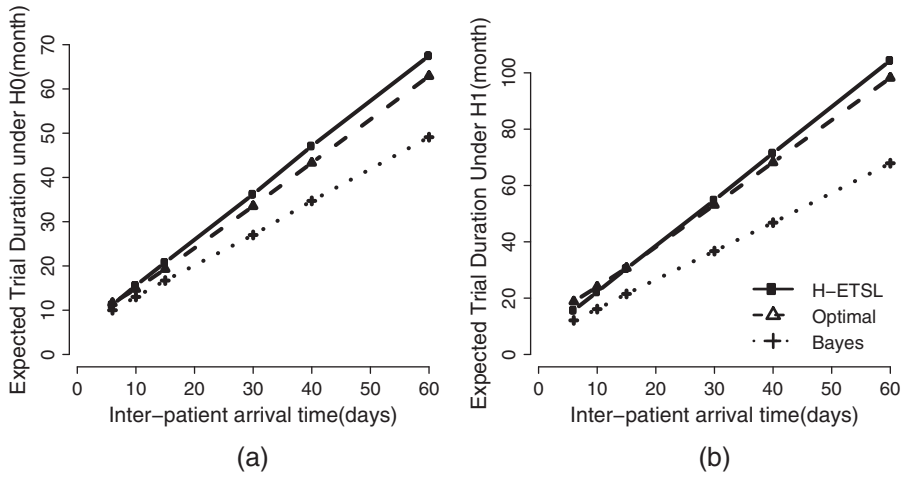

Figure 3. (a) Expected trial duration when $S(t_0) = 0.1$ for various interpatient arrival times; and (b) expected trial duration when $S(t_0) = 0.25$ for various interpatient arrival times.

Table V. Operating characteristics of the exponential method, for different hypotheses with different type I error and power, based on exponentially distributed survival data.

		Under H_0				Under H_1			
Design	$n(n_1)$	Type I error	PET	EN	ETSL	Power	PET	EN	ETSL
$c_2 = 4$ and $c_3 = 0.03$									
$0.1 vs 0.25$									
Optimal	43(18)	0.05	0.73	25	13	0.80	0.14	40	22
Bayes	28(14)	0.05	0.29	24	12	0.80	0.04	27	15
$0.2 vs 0.4$									
Optimal	43(13)	0.05	0.75	20	12	0.80	0.16	38	22
Bayes	24(12)	0.05	0.24	21	12	0.79	0.03	24	14
$0.3 vs 0.45$									
Optimal	82(30)	0.05	0.58	52	23	0.80	0.07	79	35
Bayes	46(23)	0.05	0.33	39	17	0.82	0.05	45	21
$0.5 vs 0.7$									
Optimal	43(15)	0.05	0.70	24	13	0.80	0.13	39	23
Bayes	26(13)	0.04	0.19	24	13	0.81	0.02	26	14
$c_2 = 1$ and $c_3 = 0.02$									
$0.1 vs 0.25$									
Optimal	50(21)	0.10	0.67	31	15	0.90	0.07	48	24
Bayes	28(14)	0.09	0.24	25	13	0.87	0.04	27	15
$0.2 vs 0.4$									
Optimal	37(17)	0.10	0.56	26	15	0.90	0.05	36	21
Bayes	26(13)	0.09	0.19	23	13	0.90	0.03	26	14
$0.3 vs 0.45$									
Optimal	81(27)	0.10	0.59	51	23	0.90	0.07	78	34
Bayes	50(25)	0.11	0.28	43	19	0.90	0.04	49	22
$0.5 vs 0.7$									
Optimal	45(21)	0.10	0.66	29	15	0.90	0.07	43	24
Bayes	28(14)	0.10	0.15	26	14	0.90	0.02	28	15

PET, probability of early stopping; EN, expected sample size; ETSL, expected total study length.

Table VI. Operating characteristics of the Weibull method for hypothesis of 0.1 vs 0.25, given type I error of approximately 0.05 and power of approximately 0.85 except that BayesII gives power approximately 0.87 and type I error of approximately 0.07, with $c_2 = 3$ and $c_3 = 0.03$, based on Weibull distributed survival data.

		Under H_0			Under H_1		
Design	$n(n_1)$	PET	EN	ETSL	PET	EN	ETSL
	$\alpha = 0.5$						
Optimal	50(20)	0.68	30	14	0.09	47	24
H-ETSL	53(31)	0.66	39	15	0.09	51	22
Bayes	32(16)	0.36	26	13	0.04	31	16
Bayes	32(24)	0.58	27	12	0.03	32	16
BayesII	32(16)	0.38	26	12	0.34	27	13
	$\alpha = 1$						
Optimal	50(20)	0.68	30	15	0.09	47	24
H-ETSL	52(33)	0.62	41	16	0.09	50	22
Bayes	32(16)	0.23	28	14	0.03	32	16
Bayes	32(24)	0.43	29	13	0.03	32	16
BayesII	32(16)	0.22	29	14	0.25	28	14
	$\alpha = 1.5$						
Optimal	50(20)	0.68	30	15	0.09	47	24
H-ETSL	52(34)	0.62	41	16	0.09	50	22
Bayes	32(16)	0.15	30	15	0.03	32	16
Bayes	32(24)	0.38	29	13	0.02	32	16
BayesII	32(16)	0.19	29	15	0.18	29	15
	$\alpha = 2$						
Optimal	50(20)	0.67	30	16	0.09	47	24
H-ETSL	52(34)	0.61	41	16	0.09	50	22
Bayes	32(16)	0.14	30	15	0.03	32	16
Bayes	32(24)	0.33	29	14	0.03	32	16
BayesII	32(16)	0.19	29	14	0.19	29	14

PET, probability of early stopping; EN, expected sample size; ETSL, expected total study length.

Again, the Weibull method significantly shortens the trial duration with only 70% of the sample size from the MiniMax design. The PET decreases as the median failure time increases (as measured by α) because more events (evidence) are observed with smaller median failure times, which leads to shorter trial duration and smaller expected sample size.

Table VI also suggests that waiting longer (larger n_1) to make the interim decision could result in higher PET and shorter trial duration under the null hypothesis (compare $n_1 = 16$ versus $n_1 = 24$).

3.3.1. Robustness with the Weibull method. We examined the robustness of the Weibull method for data generated from log-normal and Gamma distributions. As we can see from Table VII, the Weibull method is very robust in the cases studied.

3.4. Simulation using the grouped-data method

If time-to-event data can be assumed to follow a parametric distribution, the Weibull method could be performed because of its robustness presented in the last section. In this section, we evaluate a design for trials for which the majority of the events are observed at the scheduled tumor assessment time points. The grouped-data method was applied to survival data generated from various distributions such as Weibull, log-normal, and Gamma distributions. The time points of the grid were chosen to be $0 < 2 < 4 < 6$ under the assumption that the appointment is scheduled every 2 months. Prior parameters of $\tau_0 = 0.1$ and $\eta = 0.014$ ($\kappa_0 \equiv 1$) imply that a priori, $P(S(t_0) < 0.1) = 0.16$ and $P(S(t_0) > 0.25) = 0.8$.

Table VII. Operating characteristics of the Weibull method for hypothesis of 0.1 vs 0.25, with $c_2 = 3$ and $c_3 = 0.03$, based on Weibull, Gamma, and log-normal distributed survival data.

		Under H_0				Under H_1			
Design	$n(n_1)$	Type I error	PET	EN	ETSL	Power	PET	EN	ETSL
$\mathcal{LN}(\mu,2)$									
Optimal	50(20)	0.05	0.68	30	14	0.85	0.09	47	23
Bayes	32(24)	0.05	0.73	26	10	0.80	0.10	31	16
$\mathcal{LN}(\mu,1)$									
Optimal	50(20)	0.05	0.67	30	15	0.85	0.02	47	24
Bayes	32(24)	0.05	0.64	27	11	0.86	0.06	32	16
$\mathcal{G}(1,b)$									
Optimal	50(20)	0.05	0.68	30	15	0.85	0.09	47	24
Bayes	32(24)	0.05	0.41	29	13	0.85	0.04	32	16
$\mathcal{G}(2,b)$									
Optimal	50(20)	0.05	0.68	30	15	0.85	0.09	47	24
Bayes	32(24)	0.05	0.45	28	13	0.82	0.05	32	16

PET, probability of early stopping; EN, expected sample size; ETSL, expected total study length.

But this prior distribution of $S(t_0)$ is flat, so that the influence of the prior on the posterior distribution is small. We increased n_1 from 24 to 30 to gain more information at the end of stage I.

With the results shown in Table VIII with 500 repeated trials and 250 ($B = 250$) simulations for the expected Bayes risk of continuation, the grouped-data method shortens the trial duration only under the alternative hypotheses because of the seamless recruitment. With the same total sample size as in the MiniMax design, the significantly lower PET in the Bayes design (even based on six more patients than in the MiniMax design) results in longer trial duration under the null hypothesis. With slower accrual or longer follow-up per patient, the trial could have higher PET. We also found that allowing stopping for both efficacy and futility increased the type I error to about 10%–15% (results not shown). Therefore, we would suggest that a trial only be allowed to stop for futility when the grouped-data method is applied.

4. Application of the methods to a sarcoma clinical trial

The primary objective of this trial is to evaluate the 6-month PFS rate in patients with advanced sarcoma treated with oral cyclophosphamide and sirolimus (OCR). The trial was originally designed using the Simon MiniMax design. Patients who are progression free and alive at 6 months following the initiation of study treatment will be considered a success. A 25% or greater 6-month PFS rate from OCR treatment would be considered as showing potential activity of the combination, thus worthy of further study. A 10% or less 6-month PFS rate from OCR is considered uninteresting for additional study. With a type I error rate of 5%, the study of 47 evaluable patients will provide 85% power to detect an effective treatment. Tumor imaging will be performed every 2 months until tumor progression or until the patient completes 12 cycles of treatment. Each cycle is 28 days long. According to the MiniMax design, 24 patients will be enrolled in stage I. If three or more patients are alive and free from sarcoma progression 6 months after enrollment, an additional 23 patients will be enrolled in the second stage of the study. If nine or more of the 47 patients are alive and free of progression 6 months after enrollment, the proposed treatment will be declared as having anti-sarcoma activity worthy of further study. The trial is now closed for accrual, and the trial was paused for about 2 months until three patients reached 6 months follow-up and were free of progression and thereafter continued to stage II. The trial declared the OCR treatment as promising after nine patients were progression free after 6 months in stage II.

If we could design the trial using the Bayes design, we would choose $c_2 = 3$ and $c_3 = 0.03$ for type I error of 0.05 and power of 0.85 as specified in the MiniMax design. Then, we assume that patients

Table VIII. Operating characteristics of the grouped-data method for hypothesis of 0.1 vs 0.25, given type I error of approximately 0.05 and power of approximately 0.85, based on Weibull, Gamma, and log-normal distributed survival data.

		Under H_0			Under H_1		
Design	$n(n_1)$	PET	EN	ETSL	PET	EN	ETSL
	$\mathcal{W}(0.5,\gamma)$						
Optimal	50(20)	0.68	30	14	0.09	47	24
Bayes	47(30)	0.48	39	16	0.03	47	21
	$\mathcal{W}(1.5,\gamma)$						
Optimal	50(20)	0.68	30	15	0.09	47	24
Bayes	47(30)	0.25	43	19	0.02	47	21
	$\mathcal{LN}(\mu,2)$						
Optimal	50(20)	0.68	30	14	0.09	47	23
Bayes	47(30)	0.53	38	16	0.03	46	21
	$\mathcal{LN}(\mu,1)$						
Optimal	50(20)	0.67	30	15	0.09	47	24
Bayes	47(30)	0.32	42	18	0.02	47	21
	$\mathcal{G}(1,b)$						
Optimal	50(20)	0.68	30	15	0.09	47	24
Bayes	47(30)	0.33	41	18	0.03	47	21
	$\mathcal{G}(2,b)$						
Optimal	50(20)	0.68	30	15	0.09	47	24
Bayes	47(30)	0.24	43	19	0.02	47	22

PET, probability of early stopping; EN, expected sample size; ETSL, expected total study length.

Table IX. Oral cyclophosphamide and sirolimus trial data.

	Time to progression/death (days)
Stage I (D_1)	$103, 112, 212^+, 210^+, 112, 111, 56, 59, 111, 31, 142^+, 110, 61, 133^+, 133^+, 93^+, 51, 87^+, 56, 74^+, 67^+, 53^+, 31^+, 11$
Stage II (D_2)	$103, 112, 225, 360^+, 112, 111, 56, 59, 111, 31, 162, 110, 61, 223, 360^+, 225, 51, 253, 56, 107, 219, 140, 55, 11, 112, 113, 53, 112, 56, 184^+, 180^+, 56, 57, 54, 61, 28, 56, 56, 49, 100^+, 95, 85^+, 52, 62, 65^+, 56, 42^+$

in stage II are enrolled 2 months earlier than their true enrollment dates such that there is no trial suspension at the end of stage I. The end of stage II is defined when nine patients reach the time t_0 and are free of progression under this hypothetical enrollment schedule. In contrast to the Simon's designs, which require waiting for one more event from the nine patients that are event free and have not been followed-up for 6 months, the Bayes design will make the interim decision on the basis of the data D_1 in Table IX.

In this trial, although the design should lead to interval censored data, the variation of patient scheduling, death without confirmed progression, and clinical progression that triggers an earlier detection of the progression, in our view, justify using a parametric distribution for the event time data. Because the mean estimate of α is about 1.6 based on the data D_1, the Weibull method was used. We then tried the Weibull method using 68% of the sample size as in the MiniMax design (n=32) with n_1=24. In the first

Table X. Expected losses and decisions at two stages.

		Stage I				Stage II	
$n(n_1)$	Action	ELoss	ρ_0	ρ_c	Decision	ELoss	Decision
Weibull							
47(24)	Accept	0.64	0.08	0.04	Continue	0.38	Reject
	Reject	0.08				0.02	
32(24)	Accept	0.64	0.08	0.05	Continue	0.75	Reject
	Reject	0.08				0.01	
Grouped data							
47(30)	Accept	0.65	0.02	0.03	Continue	0.40	Reject
	Reject	0.02				0.02	

design, the posterior expected loss of a rejection and an acceptance decision are 0.08 and 0.64, respectively (see Table X). The smaller of these two is 0.08, which is the Bayes risk of immediate stopping. Because it is larger than the expected Bayes risk of continuation ($\rho_c = 0.04$), the trial is continued to the second stage. At the end of the trial, the posterior expected loss of a rejection decision is smaller (0.02 vs 0.38). Therefore, we reject the H_0. In the first two designs, ρ_0 is the same because D_1 is the same in both designs, whereas ρ_c is smaller in the first design because more information (less risk) is expected if we would enroll more patients in the second stage given the same c_3.

We note that using the grouped-data method with the same time points of grids and priors as specified in the simulation studies, we would also reject H_0.

All the three designs give the same conclusion as the MiniMax design that the OCR is effective, but the Bayes designs have shorter trial duration.

5. Discussion

There are three reasons for the use of decision-based Bayesian approach in designing clinical trials with interim looks. First, it is natural to make 'stop' or 'continue' decisions on the basis of the risk of 'stop' and 'continue' if the trial will continue to the final stage. Once we decide to stop, the terminal decision is based on the observed data alone (with noninformative priors). Second, the terminal decision rule (reject or accept) is constructed by weighting the evidence of two hypotheses. Third, calibrating the two cost parameters to obtain desired type I error and power makes the decision-based approaches feasible in clinical trial designs, and the robustness of the properties of the designs to cost parameters, irrespective of patient accrual rates, hypotheses, timing of the interim look, and three proposed methods, makes the design very appealing.

Compared with the commonly used Simon's designs, the decision-based Bayesian designs have the following advantages: (1) no need to pause the trial at the end of stage I to wait for all the patients to reach the landmark point; (2) significantly shorter average trial duration and much smaller sample size; (3) a flexible design that could easily add early stopping for efficacy as well as for futility; and (4) reasonable design properties (such as PET), which are based on weight of evidence in the data rather than awkward frequentist hypothesis tests.

In addition, the proposed methods can be easily adapted to randomized trials with survival endpoints. In such trials, the 'difference' in the survival rate between treatment and control groups will be monitored. This 'difference' could be represented as a parameter in the accelerated failure time model or Cox regression model. The terminal decision rule and stopping rule will be constructed using the posterior and predictive distribution of this parameter.

In this paper, we did not include patients that are lost to follow-up or dropout before reaching the landmark point. This information can be easily incorporated into the estimation methods for the Bayesian model [25]. In contrast, designs that use binary endpoints do not have a natural way to incorporate this partial information.

We used quite noninformative prior to reflect limited experience with the experimental regimen under considerations. In certain special situations in which more experience with experimental treatment is available, other priors could be used.

Assuming exponential distributions, trials can be designed with many stages because of the simple sufficient statistics. The grid method using backward induction makes the computation time increase

linearly with the number of stages. However, the Weibull and grouped-data methods are practically limited to two stages mainly because there are not a small number of sufficient statistics [30, 31]. For these two methods, Bayes risk was calculated in a forward fashion, for which the computation time increase exponentially with the number of stages, which makes adding one more interim look computationally infeasible although theoretically possible.

In summary, Bayesian decision-based approaches are feasible in designing clinical trials with appealing properties. As more and more experience accumulates with the application of this approach in real trials, it may be possible to design the trials by directly specifying the costs rather than treating them as tuning parameters.

Acknowledgements

Many thanks to Frank E. Harrell for his encouragement to initiate this research. The authors gratefully acknowledge the constructive comments of an associate editor and referees.

References

1. Chan JK, Ueda SM, Sugiyama VE, *et al.* 578 Analysis of phase II studies on targeted agents and subsequent phase III trials: what are the 579 predictors for success? *J Clin Oncol* 2008; **26**:1511–1518.
2. Korn EL, Arbuck SG, Pluda JM, *et al.* Clinical trial designs for cytostatic agents: are new approaches needed? *J Clin Oncol* 2001; **19**:265–72.
3. Stone A, Wheeler C, Barge A. Improving the design of phase II trials of cytostatic anticancer agents. *Contemporary Clinical Trials* 2007; **28**:138–145.
4. Panageas KS, Ben-Porat L, Dickler MN, *et al.* When you look matters: the effect of assessment schedule on progression-free survival. *J Natl Cancer Inst* 2007; **99**:428–32.
5. Gehan EA. The determination of the number of patients required in a preliminary and a follow up 555 trial of a new chemotherapeutic agent. *J Chronic Dis* 1961; **13**:346–353.
6. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; **38**(558):143–151.
7. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; **10**:1–10.
8. Herndon JE. A design alternative for two-stage, phase II, multicenter cancer clinical. *Contemporary Clinical Trials* 1998; **19**(5):440–450.
9. Case LD, Morgan TM. Design of Phase II cancer trials evaluating survival probabilities. *BMC Medical Research Methodology* 2003; **3**:6. DOI: 10.1186/1471-2288-3-6.
10. Lin DY, Shen L, Ying Z, Breslow NE. Group sequential designs for monitoring survival probabilities. *Biometrics* 1996; **52**:1033–1041. DOI: 10.1200/JCO.2009.22.4329.
11. Huang B, Talukder E, Thomas N. Optimal two-stage phase II designs with long-term endpoints. *American Statistical Association, Statsitics in Biopharmaceutical Research* 2010; **2**:51–60. DOI: 10.1198/sbr.2010.09001.
12. Follman DA, Albert PS. Bayesian monitoring of event rates with censored data. *Biometrics* 1999; **55**(2):603–607.
13. Rosner GL. Bayesian monitoring of clinical trials with failure-time endpoint. *Biometrics* 2005; **61**:239–245.
14. Cheung YK, Thall PF. Monitoring the rates of composite events with censored data in phase II clinical trials. *Biometrics* 2002; **58**:89–97.
15. Thall PF, Wooten LH, Tannir NM. Monitoring event times in early phase clinical trials: some practical issues. *Clin Trials* 2005; **2**(6):467–78.
16. Thall PF, Simon R, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* 1995; **14**:357–379.
17. Lachin J. A review of methods for futility stopping based on conditional power. *Statistics in Medicine* 2005; **24**: 2747–2764. DOI: 10.1002/sim.2151.
18. Pepe MS, Anderson GL. Two-stage experimental designs: early stopping with a negative result. *Applied Statistics* 1992; **41**:181–190.
19. Berry DA. Bayesian clinical trials. *Nature Reviews Drug Discovery* 2006; **5**:27–36. DOI: 10.1038/nrd1927.
20. Staquet MJ, Sylvester RJ. A decision theory approach to phase II clinical trials. *Biomedicine* 1977; **26**(4):262–266.
21. Sylvester RJ, Staquet MJ. Design of phase II clinical trials in cancer using decision theory. *Cancer Treatment Report* 1980; **64**(2-3):519–524.
22. Sylvester RJ. A Bayesian approach to the design of phase II clinical trials. *Biometrics* 1988; **44**(3):823–836.
23. Brunier HC, Whitehead J. Sample sizes for phase II clinical trials derived from Bayesian decision theory. *Statistics in Medicine* 1994; **13**(23-24):2493–2502.
24. Stallard N. Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics* 1998; **54**(1):279–294.
25. Zhao L, Woodworth G. Bayesian decision sequential analysis with survival endpoint in phase II clinical trials. *Statistics in Medicine* 2009; **28**:1339–1352. DOI: 10.1002/sim.3544.
26. Berry SM, Carlin BP, Lee J, Muller P. *Bayesian Adaptive Methods for Clinical Trials*, Chapman & Hall/ CRC Biostatistics Series. Boca Raton: CRC Press, 2011.
27. Albert J. *Bayesian Computation with R*. Springer: New York, 2009.

28. Kalbfleisch J. Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society* 1978; **40**:214–221.
29. Ibrahim JG, Chen MH, Sinha D. *Bayesian Survival Analysis*. Springer: New York, 2001. 50–53.
30. Muller P, Berry DA, Grieve AP, *et al.* Simulation-based sequential Bayesian design. *Journal of Statistical Planning and Inference* 2007; **137**(10):3140–3150. DOI: 10.1016/j.jspi.2006.05.021.
31. Brockwell AE, Kadane JB. Sequential analysis by griding sufficient statistics. *J. Comput. Graph. Statist* 2003; **12**:566–584.