Published in final edited form as:

Stat Biosci. 2018 August ; 10(2): 420–438. doi:10.1007/s12561-017-9199-7.

Bayesian Adaptive Randomization and Trial Monitoring with Predictive Probability for Time-to-event Endpoint

Guosheng Yin,

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

Nan Chen, and

Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, U.S.A., nchen2@mdanderson.org

J. Jack Lee

Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, U.S.A., jjlee@mdanderson.org

Abstract

There has been much development in Bayesian adaptive designs in clinical trials. In the Bayesian paradigm, the posterior predictive distribution characterizes the future possible outcomes given the currently observed data. Based on the interim time-to-event data, we develop a new phase II trial design by combining the strength of both Bayesian adaptive randomization and the predictive probability. By comparing the mean survival times between patients assigned to two treatment arms, more patients are assigned to the better treatment on the basis of adaptive randomization. We continuously monitor the trial using the predictive probability for early termination in the case of superiority or futility. We conduct extensive simulation studies to examine the operating characteristics of four designs: the proposed predictive probability adaptive randomization design, the predictive probability equal randomization design, the posterior probability adaptive randomization design, and the group sequential design. Adaptive randomization designs using predictive probability and posterior probability yield a longer overall median survival time than the group sequential design, but at the cost of a slightly larger sample size. The average sample size using the predictive probability method is generally smaller than that of the posterior probability design.

Keywords

Bayesian adaptive design; Clinical trial; Power; Randomization; Survival data; Type I error; Type II error

1 Introduction

In oncology, phase II clinical trials are designed to provide an initial assessment of a drug's therapeutic effects. In order to facilitate timely decision making on whether to proceed to a phase III trial for confirmative evaluation, the primary endpoint in a phase II trial is typically short-term and dichotomous, characterizing the patient's clinical response to treatment.

Phase II trials can be either single-arm, comparing the response rate for a new treatment with the reference response rate estimated from historical data, or multi-arm, with patients randomized among different treatments. Gehan (1961) proposed a two-stage design for phase II oncology trials to evaluate whether a treatment has a certain minimum response rate. Simon (1989) developed an optimal and a minimax two-stage design to test whether a new treatment reaches a target response rate beyond that of the standard treatment. All these designs assume a frequentist hypothesis testing framework and control both the type I and type II error rates. Frequentist designs, however, have the limitation of fixed stage sizes. Such trials need to be paused to evaluate the outcome of patients in the current stage before moving to the next stage. The trial conduct, hence, is rigid and can be difficult to follow. Fleming (1982) and Chang et al. (1987) studied multiple testing and group sequential methods for phase II trial designs. In the Bayesian paradigm, Thall and Simon (1994) provided practical guidelines regarding how to implement a phase II trial design by continuously monitoring every new outcome based on computing the posterior probability of the response rate. Bayesian trial designs are more flexible, which allows for more frequent monitoring throughout the trial conduct. The trial may be terminated at any time if the accumulating data provide sufficient evidence to indicate that the experimental drug is promising or not promising. If there is a lack of sufficient evidence to support either decision, the trial continues to enroll the next patient.

An alternative Bayesian approach is to apply the Bayesian posterior predictive probability, which is a very useful way to monitor trial conduct before enrolling all of the planned subjects (Lee and Liu, 2008). Using the predictive probability renders great flexibility in trial monitoring. We can adaptively make decisions to terminate a trial for superiority or futility before reaching the maximum sample size. The essence of predictive monitoring is to evaluate whether a trial will reach a definitive conclusion should the trial continue to the end, given the current information. In contrast, trial monitoring using the posterior probability bases the decision upon the data that have accumulated thus far without projecting any future outcomes. One distinctive feature of predictive monitoring is that trial decisions depend on the future sample size (Yin, 2012). If there is still a large number of patients to enroll in the future, the decision on trial termination can be harder to reach in contrast to the situation in which there are only a few more subjects to enroll because the outcomes of more future patients carry more variability that can overturn the previous result. The predictive probability synthesizes the evidence between the current data and the projection of future data to make an informed decision at an interim point during the trial. At any time during the course of a trial, based on the accumulated data, one can stop the trial and claim that the experimental drug is promising or not promising, or continue the trial due to a lack of convincing evidence to render a decision. Lee and Liu (2008) proposed a continuous monitoring design based on predictive probability for a binary endpoint by calculating the probability of rejecting the null hypothesis should the trial be conducted to the maximum sample size given the observed interim data in a single-arm phase II trial. Yin, Chen, and Lee (2012) extended predictive monitoring to randomized trials while still using a binary endpoint. By using a binary endpoint, the computation of the predictive probability is straightforward due to the conjugacy of the beta prior distribution with the binomial likelihood function. A Bayesian predictive probability method has also been proposed for

two-stage designs to increase the success of the phase III trial (Cellamare et al. 2015). The successful application of model-based predictive probability approaches in the I-SPY 2 trial has been reported recently (Park et al. 2016; Rugo et al. 2016).

In conventional phase II oncology clinical trials, the short-term binary outcome is often the tumor response. For example, the most commonly used chemotherapies are cytotoxic agents, which may disrupt tumor growth, cause cancer cell death, and result in tumor shrinkage. However, many drugs that have shown promising therapeutic effects in phase II trials eventually fail in phase III trials. This undesirably high failure rate shows the limitation of using the short-term endpoint of the binary tumor response to predict the long-term survival endpoint. Furthermore, many phase II trials are single-arm studies that lack a concurrent control group. It is well known that single-arm trials can yield biased and overly optimistic results. On the other hand, randomized phase II trials can provide a more objective comparison between new and standard treatments (Lee and Feng, 2005; Ratain and Sargent, 2009).

As the development of personalized medicine advances rapidly, more targeted agents are emerging and being prepared for testing. Targeted agents are often cytostatic and exhibit therapeutic effects through tumor control in contrast to tumor shrinkage. Hence, the use of tumor response as the primary endpoint may not be ideal. Instead, progression-free survival may be a more appropriate way to characterize the therapeutic effects of cytostatic agents (Korn et al., 2001). Furthermore, the recent development of cancer immunotherapies shows that effective treatments may not cause quick tumor shrinkage but rather yield tumor control over a longer term (Hoos et al., 2010; Hoos et al., 2016). A recent review of the use of adaptive designs in oncology can be found in Zang and Lee (2014).

For time-to-event endpoints, Louis (1977) studied sequential allocation in clinical trials by comparing survival curves under the exponential model assumption. Zhang and Rosenberger (2007) developed optimal response-adaptive randomization for clinical trials with survival endpoints using exponential and Weibull models. To address the emerging challenges of phase II trial design described above, we propose a two-arm Bayesian adaptive randomization design with interim predictive probability monitoring for a time-to-event endpoint.

As the time-to-event endpoint requires a longer follow-up, it becomes more computationally challenging to continuously update the predictive probability. Moreover, due to adaptive randomization, the future sample size in each arm is unknown, even though such information is essential for computing the predictive probability. As a remedy, we propose to approximate the future sample size when we compute the predictive probability. To determine the operating characteristics, we examine the frequentist properties of the proposed design with the goal of controlling the type I error rate and achieving a sufficiently high power. We choose the maximum sample size and design parameters to achieve desirable frequentist properties. At any time during the trial, if there is a high predictive probability that one treatment is better than the other, we stop the trial and declare superiority; if there is a high predictive probability that the two treatments have similar efficacy, we stop the trial and declare equivalence; otherwise, we continue the trial until we

have exhausted the maximum sample size. Through the use of adaptive randomization, more patients are treated with the better treatment. In comparison with other available designs such as group sequential (GS) methods, our proposed design using the time-to-event predictive probability (TPPr) appears to be more efficient and flexible.

The rest of this article is organized as follows. In Section 2, we introduce the notation and the survival model for the time-to-event endpoint and propose the statistical design by incorporating Bayesian adaptive randomization and predictive probability monitoring. For a comparison and better understanding of the TPPr, we also develop the posterior monitoring for randomized phase II trials with a time-to-event endpoint (TPoP). In Section 3, we illustrate how to calibrate the design parameters and present simulation studies to examine the design's operating characteristics under different practical scenarios. We give concluding remarks in Section 4.

2 Statistical Model for Time-to-Event Endpoint

2.1 Exponential Model

For oncology trials, we often encounter time-to-event outcomes, for example, progression-free survival and overall survival times. Let T denote the underlying failure time, which may be right-censored, and let C denote the censoring time. We denote the observed time as $X = \min(T, C)$ with a censoring indicator $\delta = \mathbb{I}(T \leq C)$, i.e., if $\delta = 1$, then $X = T$, which is the failure time; and if $\delta = 0$, then $X = C$, which is the censoring time.

In a two-arm trial, we assume that the survival times for both the experimental and standard treatment arms follow exponential distributions, but with means of μ_1 and μ_2 , respectively. The probability density function of an exponential distribution is given by

$$f(t) = \frac{1}{\mu} \exp(-t/\mu),$$

and the corresponding survival function is

$$S(t) = \exp(-t/\mu).$$

The null hypothesis we consider is equivalence of the two treatments in terms of the efficacy and the alternative hypothesis is non-equivalence of the two treatments. Let δ be a threshold for a clinically meaningful treatment difference, and let θ_T be a cutoff probability for the strength of evidence for decision making. If the hazard ratio between treatments 1 and 2 given the observed data D satisfies

$$\Pr(|\log(\mu_2/\mu_1)| > \delta | D) \geq \theta_T, \quad (1)$$

we claim non-equivalence of the two treatments, i.e., one treatment is superior to the other.

During the trial conduct, the decision is adaptively made as to whether the trial should be stopped early for superiority or equivalence of the two treatments, and thus the follow-up time for each subject is censored at the decision-making time if the event of interest has not occurred yet. Therefore, we in fact observe (X_i, Δ_i) at the interim monitoring time, which yields the likelihood function

$$L(\mu|D) = \prod_{i=1}^n f(x_i)^{\Delta_i} S(x_i)^{1-\Delta_i} = \mu^{-\sum_{i=1}^n \Delta_i} \exp\left(-\frac{1}{\mu} \sum_{i=1}^n x_i\right),$$

where $D = \{(x_i, \Delta_i), i = 1, \dots, n\}$ denotes the observed data, and n is the sample size that has accumulated thus far.

2.2 Posterior Probability

In the Bayesian framework, we specify a prior distribution for μ ; for example, we may choose an inverse-gamma (IG) prior distribution for μ ,

$$\mu \sim \text{IG}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mu^{-\alpha-1} \exp(-\beta/\mu),$$

where $\alpha > 0$ and $\beta > 0$. Since the IG distribution is conjugate with the exponential likelihood function, the posterior distribution of μ also follows an IG distribution,

$$p(\mu|D) \propto \mu^{-\sum_{i=1}^n \Delta_i - \alpha - 1} \exp\left(-\frac{\sum_{i=1}^n x_i + \beta}{\mu}\right),$$

that is, $\mu|D \sim \text{IG}(\sum_{i=1}^n \Delta_i + \alpha, \sum_{i=1}^n x_i + \beta)$. As one step further, the posterior predictive distribution of the survival time T is given by

$$p(t|D) = \int p(t|\mu)p(\mu|D)d\mu,$$

where $p(t|\mu)$ is the density function of T . However, there is typically no closed form for $p(t|D)$.

The trial conduct can be based on the posterior probability of comparing the mean survival times of the two treatments. Under the exponential survival model, the mean survival time is the reciprocal of the hazard. More specifically, we can compute the posterior probability (PoP) of one treatment being better than the other based on the logarithm of the hazard ratio between treatments 1 and 2, as shown by equation (1). If $\text{PoP} > \theta_T$, we claim non-equivalence of the two treatments, i.e., one treatment is superior to the other. We further need to specify the lower and upper cutoff probabilities η_L and η_U for adaptive decision making during the trial conduct. The decision rules are given as follows:

- i. Superiority stopping: If $\text{PoP} > \eta_U$, we stop the trial to claim a superior treatment.

- ii. Equivalence stopping: If $\text{PoP} < \eta_L$, we stop the trial to claim the equivalence of the two treatments.

The design parameters (η_U, η_L) can be calibrated via simulation studies to achieve desirable trial operating characteristics.

2.3 Predictive Probability

On the other hand, we may base our adaptive decisions during the trial conduct upon the posterior predictive probability of whether the expected trial outcome is positive or negative at the completion of the trial. Let m be the remaining number of patients to be recruited in the trial, i.e., the sample size of the future data. Let $\{\tilde{T}_1, \dots, \tilde{T}_m\}$ denote the survival times of the future m patients, which may be administratively censored by the study ending time τ . To facilitate decision making, we specify a threshold for a clinically meaningful treatment difference, δ , and a threshold for the probability, θ_T , corresponding to the strength of evidence for making a decision. If the hazard ratio between treatments 1 and 2 satisfies

$$\Pr(|\log(\mu_2/\mu_1)| > \delta | D, \tilde{T}_1, \dots, \tilde{T}_m) \geq \theta_T,$$

we claim non-equivalence of the two treatments, i.e., one treatment is superior to the other. However, $\{\tilde{T}_1, \dots, \tilde{T}_m\}$ are the future data and are still not observed at the decision-making time. As a result, we can average out the randomness in all the $\tilde{T}$ s by computing the predictive probability (PrP) as

$$\begin{aligned} \text{PrP} = E_{\tilde{T}_1, \dots, \tilde{T}_m} \left[I\{\Pr(|\log(\mu_2/\mu_1)| > \delta | D, \tilde{T}_1, \dots, \tilde{T}_m) \geq \theta_T\} \right] &= \int \dots \int I\{\Pr(|\log(\mu_2/\mu_1)| \\ &> \delta | D, \tilde{t}_1, \dots, \tilde{t}_m) \geq \theta_T\} \times p(\tilde{t}_1 | D) \dots p(\tilde{t}_m | D) d\tilde{t}_1 \dots d\tilde{t}_m \end{aligned} \quad (2)$$

where $p(\tilde{t}_1 | D), \dots, p(\tilde{t}_m | D)$ are the posterior predictive distributions of $\tilde{T}_1, \dots, \tilde{T}_m$, respectively.

We only observe the current data D , but have not observed the future data $\tilde{T}$, and also there is no closed form for $p(\tilde{t} | D)$. Thus, the computation of the PrP proceeds as follows:

1. Sample from the posterior distribution of μ ,

$$\mu | D \sim \text{IG} \left(\sum_{i=1}^n \Delta_i + \alpha, \sum_{i=1}^n x_i + \beta \right).$$

2. Sample from the posterior predictive distribution,

$$\tilde{t} | \mu \sim \text{Exp}(\mu).$$

We can iterate the above two steps for m times to obtain $\{\tilde{t}_1, \dots, \tilde{t}_m\}$. Due to possible administrative censoring by the study ending time τ , we in fact observe $\tilde{D} = \{(\tilde{x}_i, \tilde{\Delta}_i), i = 1, \dots, m\}$, where $\tilde{x}_i = \min(\tilde{t}_i, \tau)$ and $\tilde{\Delta}_i = I(\tilde{t}_i > \tau)$.

Combining the observed data D and the future data $\tilde{D}$, we obtain the posterior distribution of μ given both D and $\tilde{D}$,

$$p(\mu | D, \tilde{D}) \propto \mu^{-\sum_{i=1}^n \Delta_i - \alpha - 1} \exp\left(-\frac{\sum_{i=1}^n x_i + \beta}{\mu}\right) \mu^{-\sum_{i=1}^m \tilde{\Delta}_i} \exp\left(-\frac{\sum_{i=1}^m \tilde{x}_i}{\mu}\right).$$

Therefore, $\mu | (D, \tilde{D})$ also follows an IG distribution,

$$\mu | (D, \tilde{D}) \sim \text{IG}\left(\sum_{i=1}^n \Delta_i + \sum_{i=1}^m \tilde{\Delta}_i + \alpha, \sum_{i=1}^n x_i + \sum_{i=1}^m \tilde{x}_i + \beta\right). \quad (3)$$

For each treatment arm, we can carry out the aforementioned computational procedure, that is to sample μ_k ($k = 1, 2$) as in (3), and then compute the predictive probability in (2). We further need to specify the lower and upper cutoff probabilities θ_L and θ_U for adaptive decision making in the trial conduct. The decision rules are given as follows:

- i. Superiority stopping: If $\text{PrP} > \theta_U$, we stop the trial to claim a superior treatment.
- ii. Equivalence stopping: If $\text{PrP} < \theta_L$, we stop the trial to claim the equivalence of the two treatments.

The design parameters $(\theta_T, \theta_U, \theta_L)$ as well as the maximum sample size can be calibrated from simulation studies to achieve desirable trial operating characteristics, e.g., to maintain the type I error rate and achieve a desirable level of power. In general, θ_T and θ_U are large values close to 1, and θ_L is a small value close to 0.

2.4 Bayesian Adaptive Randomization

Historically, equal randomization (ER) is the standard process for assigning patients to treatment arms in clinical trials due to its ease of implementation and the state of equipoise (i.e., the genuine uncertainty about which treatment is more beneficial). Response- or outcome-based adaptive randomization (AR) utilizes the responses of the patients currently treated in the trial to determine the treatment allocation for future participants (Berry and Eick, 1995; Karrison, Huo, and Chappell, 2003; and Cheng and Berry, 2007). The purpose of AR is to increase the number of patients assigned to the superior treatment arm, so that more patients in the trial would benefit from the better treatment based upon the evidence that has accumulated from the data. In general, ER emphasizes “collective ethics” and maximizes the statistical power given the sample size. On the other hand, AR weighs in more on “individual ethics” with more patients treated with the more effective treatment but resulting imbalance of sample size and decreasing of statistical power. Although the utilization of AR versus ER is debatable in practice, there is a tradeoff between the two randomization schemes and each trial should be considered individually for the suitability of

AR. The literature on comparing AR versus ER is vast and is still expanding (Hu and Rosenberger, 2006; Korn and Freidlin, 2011; Lee et al., 2012; Hey and Kimmelman, 2015 and its related commentaries). Especially, in several recent biomarker-based adaptive trials, such as BATTLE (Zhou et al., 2008; Kim et al., 2011), BATTLE 2 (Gu et al., 2013; Papadimitrakopoulou et al., 2016), and I-SPY 2 trials, AR was implemented to improve trial participants' response.

We base the assignment probability upon the posterior probability of comparing the mean survival time. More specifically, we calculate

$$\pi = \Pr(\mu_1 > \mu_2 | D),$$

and assign patients to arm 1 with probability π and to arm 2 with probability $1 - \pi$. To avoid extreme cases, we can implement a stage of ER before AR takes place to avoid extreme allocation based on little data at the beginning of the trial. In addition, we confine the randomization probability between 0.1 and 0.9. Furthermore, we can take a more general approach to AR by incorporating a power transformation,

$$\pi = \frac{\Pr(\mu_1 > \mu_2 | D)^\gamma}{\Pr(\mu_1 > \mu_2 | D)^\gamma + \{1 - \Pr(\mu_1 > \mu_2 | D)\}^\gamma}.$$

Such a transformation has been explored in Thall and Wathen (2007). The parameter γ controls the extent of imbalance during AR.

3 Numerical Study

3.1 Parameter Calibration and Setup

Before implementing our time-to-event PrP (TPrP) design, we need to calibrate the design parameters, including $(\theta_T, \theta_U, \theta_L)$ and δ . As θ_U and θ_L are for early termination, we take a two-stage procedure to calibrate these design parameters. In stage 1, we set $\theta_U = 1$ and $\theta_L = 0$ such that the trial will not be terminated early, and we focus on choosing an appropriate sample size, θ_T and δ . In stage 2, we further select the values of θ_U and θ_L while fixing the values of θ_T and δ .

For our simulation study, we considered the null hypothesis $H_0 : \mu_1 = \mu_2 = 1.0$ and the alternative hypothesis $H_1 : \mu_1 = 1.0$ and $\mu_2 = 1.5$. For the adaptive randomization (AR), the maximum sample size in the TPrP design was set as 200, with the first 60 patients equally randomized between the two arms. The accrual rate was 5 patients per month and the maximum trial period was 5 years. No early stopping was allowed if no event was observed or the number of patients was less than 30 in any treatment arm. In the TPrP AR design, the prior distribution of μ for each arm was IG(0.01, 0.01). We took 5,000 posterior samples of μ for the PoP calculation. We performed simulations to calibrate the cutoff probability values to satisfy both the type I error rate (below 10%) and power (above 80%) constraints. For

each configuration, we carried out 10,000 simulated trials to summarize the operating characteristics.

The first stage of parameter calibration was to determine the cutoff values for the main design parameters θ_T at different values of δ . In this step, we did not allow for early stopping in the design and varied the value of $\exp(\delta)$ from 1.05 to 1.25, and θ_T from 0.6 to 0.9. The resulting type I error rate and power values are presented in Table 1. When $\delta = \log(1.15)$, we selected the cutoff value of $\theta_T = 0.8$ to yield 10% for the type I error rate and 80% for the power. In the second stage of parameter calibration, we fixed δ and θ_T at the chosen values in stage 1, and varied the values of θ_U and θ_L : $\theta_U = 0.993, \dots, 0.999$ and $\theta_L = 0.02, \dots, 0.06$. The type I error rates and power values are given in Table 2. To control both the type I and type II error rates, we selected $\theta_U = 0.997$ and $\theta_L = 0.04$ for the simulation studies presented next. Similar calibration of design parameters can be applied to the time-to-event PoP (TPoP) design. In the TPoP adaptive randomization (TPoP AR) method, we selected the same maximum number of patients, 200, $\delta = \log(1.15)$ and the cutoff value of $\theta_T = 0.8$ as those in the predictive probability method. As shown in Table 3, the cutoff values of $\eta_U = 0.997$ and $\eta_L = 0.25$ were calibrated through a similar way to that used in the predictive probability method. We set the AR parameter $\gamma = 1.0$ for both the TPoP AR and TPrP AR methods. We used the same strategy to calibrate the cutoff values for the TPrP equal randomization (TPrP ER) method. In particular, we selected the maximum number of patients as 180, $\delta = \log(1.15)$, and the cutoff values of $\theta_T = 0.8$, $\theta_U = 0.988$ and $\theta_L = 0.04$.

3.2 Example

To illustrate the usage of the time-to-event predictive probability adaptive randomization (TPrP AR) design, we simulate a trial by assuming the true mean survival times of two arms to be $\mu_1 = 1.0$ and $\mu_2 = 1.5$ respectively, while keeping the other parameters the same as those in Section 3.1. The progress of the trial is shown in Figure 1. After the ER stage (at time = 1.03 years), the numbers of patients assigned to the control arm and the new treatment arm are both 30. The numbers of events on the control arm and the new treatment arm are 12 and 6, respectively. Because the number of events on the control arm is more than that on the new treatment arm, the predictive probability value is 0.888 and no stopping occurs for futility or efficacy. At time = 1.81 years, the numbers of patients respectively assigned to the control arm and the new treatment arm are 46 and 72, and the corresponding numbers of events are 29 and 26. More patients are randomized to the new treatment arm than to the control arm due to the use of AR in the design, and the predictive probability value is 0.917 at this moment. At time = 2.61 years, the predictive probability value is 0.9973 (exceeding the threshold value of 0.997), which activates early stopping for efficacy and concludes that the new treatment is superior. The total number of patients treated is 156 at the end of the trial. Because the new treatment is superior to the control treatment in this trial, fewer patients are randomized to the control and more patients are randomized to the superior treatment (52 versus 104), which allows more patients to receive the better treatment. The numbers of events on the control arm and the new treatment arm are 41 and 48, respectively.

We compare the average probability of a patient being randomized to the better treatment between the TPrP AR method and the ER of the GS method in Figure 2. In the initial stage of the trial, little information is known about the outcomes from two treatments, thus we apply ER for the first 60 patients (enrolled in one year) in both the TPrP AR method and the GS method. After the first 60 patients are enrolled, we accumulate more information about treatment efficacy from the patients' outcomes and, based on that information, we then use AR to assign the remaining patients in the TPrP AR method while ER continues to be used in the GS method. The average probability of a patient being randomized to the better treatment increases from 50% during the ER stage to 76.8% at the beginning of the AR stage. With the enrollment of more patients, the average probability of a patient being randomized to the better treatment gradually increases and its value reaches 87.8% after patient enrollment is finished. From the trace plots (grey lines) of some simulated TPrP AR trials, we can observe that in most of the trials, the probability of a patient being randomized to the better treatment arm increases very quickly after the ER stage and thus more patients are treated with the better treatment as more outcome information is collected.

3.3 Simulation Results

To systematically evaluate the performance of the proposed TPrP AR design, we made a thorough comparison with the TPrP ER design, the TPoP AR design and the GS method (Jennison and Turnbull, 2000). We considered two-arm trial designs using a time-to-event endpoint, e.g., progression-free survival. In the group sequential design, we took up to 18 interim stages and used the two-sided log-rank test for each analysis. Under the same null and alternative hypotheses, the total number of patients under the GS design and the TPrP ER design was set as 180 to yield comparable results.

We simulated four sets of scenarios by fixing the mean survival time for patients assigned to arm 1 at $\mu_1 = 0.5, 1, 1.5$, and 2, respectively. For each set of simulations, we varied μ_2 , the mean survival time for patients assigned to arm 2, from 0.5 to 2.0. We present the rejection rates, i.e., power, under different scenarios for the TPrP AR, TPrP ER, TPoP AR, and GS designs in Figure 3. The rejection rates of the four methods are very close to each other for almost all scenarios to ensure that all designs have comparable type I error rates and power. The largest difference in the rejection rates between these methods is only 0.02; see panel (c) of Figure 3 for the scenarios with $\mu_1 = 1.5$ and $\mu_2 = 1.25$. The lowest point of all the plots is located at the null case, i.e., when $\mu_1 = \mu_2$, which is in fact the type I error rate. All the methods maintained the type I error rate around 10%, and power increased as the difference between μ_1 and μ_2 became larger.

We compared the expected sample sizes of the GS, TPoP AR, TPrP AR and TPrP ER methods for different scenarios, as shown in Figure 4. Notably, the expected sample sizes of the TPrP ER design are less than those from the GS design, which demonstrates the superior early stopping property of the TPrP ER design compared with the GS design. In particular, when the difference between μ_1 and μ_2 is in the range of 0.5–1.0, the expected sample sizes of the TPrP ER are about 6%–14% less than those of the GS method. The largest difference between the expected sample sizes of the two methods occurs at $\mu_1 = 1.0$ and $\mu_2 = 2.0$ and,

under that scenario the expected sample size of TPrP ER (119.3) is 13.9% less than that of the GS method (135.9).

For the adaptive randomization design, as shown in panel (a) of Figure 4, the total sample sizes of the TPrP AR method are smaller than those of the GS method when the difference between μ_1 and μ_2 is larger than 0.75, but are larger than those of the GS method when that difference is smaller than 0.5. Similar trends can be observed in other panels of Figure 4 as well. At the null case with $\mu_1 = \mu_2 = 0.5$, the number of patients in each arm using the TPrP AR method is the same, despite using AR. As the difference between μ_1 and μ_2 increases, more patients are assigned to arm 2. Nevertheless, if that difference becomes sufficiently large, the trials would be stopped earlier, which in turn mitigates the advantages of AR.

In general, due to the use of AR, the sample sizes of the TPrP AR method and the TPoP AR method are larger than those of the GS method so as to compensate for the power loss caused by imbalanced patient allocation. When the difference between μ_1 and μ_2 is larger than 1.0, the total sample sizes of the TPrP AR are close to those of the GS method. When the difference between μ_1 and μ_2 is small (less than 1.0), the total sample sizes of the TPrP AR are about 7–9% more than those of the GS method. The largest difference between the total sample sizes of the two methods occurs at $\mu_1 = 1.0$ and $\mu_2 = 0.75$, where the total sample size of TPrP AR (181.4) is 10% more than that of the GS method (164.9).

In Figure 5, we present the average median survival times for patients under the TPrP AR, TPoP AR, TPrP ER and GS methods, respectively. Generally speaking, the average median survival times using the TPrP AR method and TPoP AR method are very close and both are longer than that using the GS method due to the use of AR. When the values of μ_1 and μ_2 are identical, the average median survival times from the TPrP AR method and the TPoP AR method are very close to that from the GS method, as expected. When the difference between μ_1 and μ_2 becomes larger (e.g., $|\mu_1 - \mu_2| = 0.25$), the advantage of the TPrP AR method and the TPoP AR method on the average median survival time becomes more evident. For example, when $\mu_1 = 2.0$ and $\mu_2 = 1.0$, the median survival time is 1.18, 1.19, and 1.12 years using the TPrP AR, TPoP AR, and GS methods, respectively. In this case, AR results in approximately 6% gain in the median survival time for the whole group. However, when the difference between μ_1 and μ_2 is larger than 1.0, the difference in the median survival times by using AR versus ER becomes smaller, which is caused by early stopping.

When μ_1 and μ_2 are identical or the difference between μ_1 and μ_2 is small, the average median survival times of the TPrP ER and GS methods are very close to each other. When the difference between μ_1 and μ_2 is around 1.0, the average median survival times of the GS method are about 3%–5% longer than those from the TPrP ER method. The difference may be caused by the different expected sample sizes and early stopping in the two methods.

At the late stage of the trial, more patients are assigned to the better treatment in the AR method and, for illustration, we present the one year survival rates of the last 50 patients under the TPrP AR and TPrP ER methods in Figure 6. The one year survival rates of the last 50 patients under the TPrP AR method are significantly higher than those of the TPrP ER method. For instance, when $\mu_1 = 1.0$ and $\mu_2 = 2.0$, the one year survival rate of the last 50

patients under the TPrP AR method is 17% higher than that of the TPrP ER method. When the difference between μ_1 and μ_2 is around 1.0, the average one year survival rates of the last 50 patients under the TPrP AR method are 10%–17% higher than those of the TPrP ER method.

To examine the robustness of the TPrP AR method with respect to other distributions, we simulate the time-to-event data from Weibull and gamma distributions, respectively. In particular, the shape parameters of Weibull and gamma distributions are set as 1.5 and 0.7, and the scale parameters are calculated to make the mean of the survival times to be 1.0 and 1.5, which are consistent with the mean values of the null and alternative scenarios as described previously. As shown in Table 4, when the shape parameter is 0.7, the type I error rate is inflated to certain extent under the TPrP AR design when the scale parameter is smaller than 1 for the Weibull model and greater than 1 for the gamma model, while the GS method maintains the type I error rate well. Due to the tradeoffs between the type I error rate and power, the power of the TPrP AR design appears to be higher than the GS method under the alternative in those scenarios. On the other hand, when the shape parameter is 1.5, TPrP AR method resulted in a lower than specified 10% type I error. To remedy such situations, the Cox proportional hazards model can be used in the Bayesian framework and thus the TPrP AR design would rely upon the proportional hazards assumption instead of the exponential distribution. Nevertheless, the computation using the Bayesian semiparametric Cox model would be much more intensive. One possibility of circumventing prior specification on the baseline hazard is to use the partial likelihood function instead of the full likelihood, yet the predictive distribution of survival times may still require the baseline modeling structure.

4 Discussion

As targeted therapy development grows in the field of oncology, more cytostatic agents become available for testing. These agents typically do not cause tumor shrinkage, and thus the typical endpoint of tumor response in the conventional phase II trials may not be suitable. In addition, newly developed immunotherapies may not result in early tumor shrinkage. Thus, progression-free survival rather than tumor response has been recommended as a more appropriate endpoint when evaluating such agents. For ease of computation, the exponential model and the corresponding hazard ratio are explored in modeling progression-free survival times. Other semiparametric survival models, such as the Cox proportional hazards model, can be used as well, which however would involve much more extensive elaboration and computation. In addition, alternative metrics other than the hazard ratio, such as the (log) difference or ratio of median survival, residual mean/median survival (Ma and Yin, 2010), or the area under the survival curve (Yuan and Yin, 2009; and Uno et al., 2014), can be explored to quantify the superiority of one treatment over the other.

Moreover, to better select promising candidate treatments for phase III trials, randomized phase II trials are becoming commonplace. Using predictive probabilities to guide phase II trial designs is appealing, as we can terminate a trial earlier if the predicted trial outcome is convincing. This may save sample size substantially. Adaptive randomization tends to allocate more patients to the more effective treatment, and thus results in more favorable

responses in the trial. However, early stopping generally reduces the advantage of adaptive randomization when sufficient evidence accumulates. Extensive simulation studies have shown that the TPrP AR method requires a slightly larger sample size because adaptive randomization may lead to an unbalanced number of patients in each arm. The tradeoff is that the TPrP AR method assigns more patients to the better treatment arm and thus leads to a higher overall treatment benefit.

Acknowledgments

We thank the three referees for their constructive and insightful comments that led to significant improvements in the article. This research was supported in part by a grant (17326316) from the Research Grants Council of Hong Kong, the U.S. National Institutes of Health (grants CA16672 and CA97007), and MD Anderson UCF (grant 80094548).

References

- Berry DA, Eick SG. 1995; Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine*. 14(3):231–246. [PubMed: 7724909]
- Cellamare M, Sambucini V. 2015; A randomized two-stage design for phase II clinical trials based on a Bayesian predictive approach. *Statistics in Medicine*. 34(6):1059–1078. [PubMed: 25545805]
- Chang MN, Therneau TM, Wieand HS, Cha SS. 1987; Designs for group sequential phase II clinical trials. *Biometrics*. 43:865–874. [PubMed: 3427171]
- Cheng Y, Berry DA. 2007; Optimal adaptive randomized designs for clinical trials. *Biometrika*. 94:673–689.
- Fleehinger BJ, Louis TA, Robbins H, Singer B. 1972; Reducing the number of inferior treatments in clinical trials. *Proc. Nat. Acad. of Sci. US*. 69:2993–94.
- Fleming TR. 1982; One-sample multiple testing procedure for phase II clinical trials. *Biometrics*. 38:143–151. [PubMed: 7082756]
- Gehan EA. 1961; The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Disease*. 13:346–353.
- Gu X, Yin G, Lee JJ. 2013; Bayesian two-step Lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemporary clinical trials*. 36(2):642–650. [PubMed: 24075829]
- Hey SP, Kimmelman J. 2015; Are outcome-adaptive allocation trials ethical? *Clinical Trials*. 12(2): 102–106. [PubMed: 25649106]
- Hoos A, Eggermont AM, Janetzki S, Hodi FS, Ibrahim R, Anderson A, Humphrey R, Blumenstein B, Old L, Wolchok J. 2010; Improved endpoints for cancer immunotherapy trials. *Journal of the National Cancer Institute*. 102(18):1388–1397. [PubMed: 20826737]
- Hoos A. 2016; Development of immuno-oncology drugs [mdash] from CTLA4 to PD1 to the next generations. *Nature Reviews Drug Discovery*. 15(4):235–247.
- Hu, F, Rosenberger, WF. *The Theory of Response-adaptive Randomization in Clinical Trials*. Wiley; Hoboken, New Jersey: 2006.
- Jennison, C, Turnbull, BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC; 2000.
- Karrison T, Huo D, Chappell R. 2003; Group sequential, response-adaptive designs for randomized clinical trials. *Controlled Clinical Trials*. 24:506–522. [PubMed: 14500050]
- Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein GR, Tsao A, Stewart DJ, Hicks ME, Erasmus J, Gupta S, Alden CM. 2011; The BATTLE trial: personalizing therapy for lung cancer. *Cancer discovery*. 1(1):44–53. [PubMed: 22586319]
- Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. 2001; Clinical trial designs for cytostatic agents: are new approaches needed? *Journal of Clinical Oncology*. 19:265–272. [PubMed: 11134222]

- Korn EL, Freidlin B. 2011; Outcome-adaptive randomization: Is it useful? *Journal of Clinical Oncology*. 29:771–776. [PubMed: 21172882]
- Lee JJ, Feng L. 2005; Randomized phase II designs in cancer clinical trials: current status and future directions. *Journal of Clinical Oncology*. 23:4450–4457. [PubMed: 15994154]
- Lee JJ, Liu DD. 2008; A predictive probability design for phase II cancer clinical trials. *Clinical Trials*. 5:93–106. [PubMed: 18375647]
- Lee JJ, Chen N, Yin G. 2012; Worth adapting? Revisiting the usefulness of outcome-adaptive randomization. *Clinical Cancer Research*. 18:4498–4507. [PubMed: 22753588]
- Louis TA. 1977; Sequential allocation in clinical trials comparing two exponential survival curves. *Biometrics*. 33:627–634. [PubMed: 338043]
- Ma Y, Yin G. 2010; Semiparametric median residual life model and inference. *Canadian Journal of Statistics*. 38:665–679.
- O'Brien PC, Fleming TR. 1979; A multiple testing procedure for clinical trials. *Biometrics*. 35:549–556. [PubMed: 497341]
- Papadimitrakopoulou V, Lee JJ, Wistuba II, Tsao AS, Fossella FV, Kalhor N, Gupta S, Byers LA, Izzo JG, Gettinger SN, Goldberg SB. 2016; The BATTLE-2 study: A biomarker-integrated targeted therapy study in previously treated patients with advanced nonsmall-cell lung cancer. *Journal of Clinical Oncology*. 34(30):3638–3647. [PubMed: 27480147]
- Park JW, Liu MC, Yee D, Yau C, van't Veer LJ, Symmans WF, Paoloni M, Perlmutter J, Hylton NM, Hogarth M, DeMichele A. 2016; Adaptive randomization of neratinib in early breast cancer. *New England Journal of Medicine*. 375(1):11–22. [PubMed: 27406346]
- Ratain MJ, Sargent DJ. 2009; Optimising the design of phase II oncology trials: The importance of randomization. *European Journal of Cancer*. 45:275–280. [PubMed: 19059773]
- Rugo HS, Olopade OI, DeMichele A, Yau C, van't Veer LJ, Buxton MB, Hogarth M, Hylton NM, Paoloni M, Perlmutter J, Symmans WF. 2016; Adaptive randomization of veliparib carboplatin treatment in breast cancer. *New England Journal of Medicine*. 375(1):23–34. [PubMed: 27406347]
- Simon R. 1989; Optimal 2-stage designs for phase-II clinical trials. *Controlled Clinical Trials*. 10:1–10. [PubMed: 2702835]
- Thall PF, Simon R. 1994; Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics*. 50:337–349. [PubMed: 7980801]
- Thall PF, Wathen JK. 2007; Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*. 43:859–866. [PubMed: 17306975]
- Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, Schrag D, Takeuchi M, Uyama Y, Zhao L, Skali H, Solomon S, Jacobus S, Hughes M, Packer M, Wei L-J. 2014; Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*. 32:2380–2385. [PubMed: 24982461]
- Yin, G. *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*. John Wiley & Sons; Hoboken, New Jersey: 2012.
- Yin G, Chen N, Lee JJ. 2012; Phase II trial design with Bayesian adaptive randomization and predictive probability. *Journal of Royal Statistical Society C – Applied Statistics*. 61:21–235.
- Yuan Y, Yin G. 2009; Bayesian dose finding by jointly modeling toxicity and efficacy as time-to-event outcomes. *Journal of Royal Statistical Society C – Applied Statistics*. 58:719–736.
- Zang Y, Lee JJ. 2014; Adaptive clinical trial designs in oncology. *Chinese Clinical Oncology*. 3(4)
- Zhang L, Rosenberger WF. 2007; Response-adaptive randomization for survival trials: the parametric approach. *Applied Statistics*. 56:153–165.
- Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. 2008; Bayesian adaptive design for targeted therapy development in lung cancer: a step toward personalized medicine. *Clinical Trials*. 5(3):181–193. [PubMed: 18559407]

Fig. 1.

A simulated trial based on the time-to-event predictive probability adaptive randomization (TPrP AR) design with $\mu_1 = 1.0$ and $\mu_2 = 1.5$: (a) numbers of patients and events on the control and new treatment arms respectively, (b) predictive probability values at different times in the trial, where the red (solid) and blue (dotted) lines correspond to the predictive probability values and threshold values.

Fig. 2.

Comparison of the randomization probabilities between the time-to-event predictive probability adaptive randomization (TPrP AR) method and equal randomization (ER) in the group sequential (GS) method, with $\mu_1 = 1.0$ and $\mu_2 = 1.5$. We also present trace plots of some simulated TPrP AR (grey) and GS (light blue) trials.

Fig. 3. Rejection rates and power values using the time-to-event predictive probability methods (TPrP AR and TPrP ER), posterior probability adaptive randomization method (TPoP AR) and group sequential (GS) method at different values of μ_2 , while the value of μ_1 is fixed at 0.5, 1.0, 1.5, and 2.0, respectively.

Fig. 4.

Comparison of the mean sample size on each arm and the mean total sample size (arm 1 + arm 2) for the time-to-event predictive probability methods (TPrP AR and TPrP ER), posterior probability adaptive randomization method (TPoP AR) and group sequential (GS) method.

Fig. 5.

Comparison of the average median survival time after trials are finished (5 years) using the time-to-event predictive probability methods (TPrP AR and TPrP ER), posterior probability adaptive randomization method (TPoP AR) and group sequential (GS) method.

Fig. 6.

Comparison of the average one year survival rates using the time-to-event predictive probability methods (TPrP AR and TPrP ER). AR50 and ER50 correspond to the average 1 year survival rates of the last 50 patients under the TPrP AR method and the TPrP ER method, respectively. AR and ER correspond to the average 1 year survival rates of overall patients under the TPrP AR method and the TPrP ER method, respectively. ER and ER50 are overlapped each other due to the very tiny difference between them.

Table 1

Stage 1 parameter calibration for the TPrP AR design with no early stopping: Type I error rates and power values under the null hypothesis with $\mu_1 = \mu_2 = 1.0$ and the alternative hypothesis with $\mu_1 = 1.0$ and $\mu_2 = 1.5$ by varying the design parameter θ_T at different values of δ . The solid line indicates the 10% type I error rate and 80% power boundaries. The shaded areas are the overlapping parameters that satisfy the design constraints, and the chosen values are in boldface.

θ_T	Null cases e^δ					Alternative cases e^δ				
	1.05	1.10	1.15	1.20	1.25	1.05	1.10	1.15	1.20	1.25
0.600	0.575	0.385	0.263	0.163	0.101	0.983	0.967	0.940	0.889	0.833
0.650	0.490	0.327	0.218	0.125	0.079	0.977	0.954	0.915	0.862	0.788
0.700	0.414	0.256	0.164	0.100	0.061	0.964	0.939	0.894	0.833	0.752
0.750	0.337	0.204	0.123	0.072	0.044	0.957	0.919	0.862	0.793	0.699
0.800	0.265	0.163	0.094	0.054	0.029	0.935	0.885	0.813	0.732	0.634
0.850	0.195	0.111	0.061	0.031	0.018	0.904	0.835	0.756	0.655	0.555
0.900	0.118	0.064	0.036	0.018	0.010	0.852	0.766	0.659	0.554	0.452

Table 2

Stage 2 parameter calibration for the TPrP AR design: Type I error rates and power values under the null hypothesis with $\mu_1 = \mu_2 = 1.0$ and the alternative hypothesis with $\mu_1 = 1.0$ and $\mu_2 = 1.5$ by varying the design parameters θ_U and θ_L (fixing $e^\delta = 1.15$ and $\theta_T = 0.8$). The solid line indicates the 10% type I error rate and 80% power boundaries. The shaded areas are the overlapping parameters that satisfy the design constraints, and the chosen values are in boldface.

θ_U	Null cases		θ_L			Alternative cases		θ_L		
	0.02	0.03	0.04	0.05	0.06	0.02	0.03	0.04	0.05	0.06
0.993	0.108	0.107	0.105	0.104	0.102	0.813	0.810	0.805	0.800	0.795
0.994	0.106	0.105	0.103	0.102	0.100	0.812	0.809	0.805	0.799	0.794
0.995	0.103	0.102	0.101	0.099	0.098	0.812	0.809	0.804	0.798	0.793
0.996	0.102	0.101	0.099	0.098	0.096	0.810	0.807	0.803	0.797	0.792
0.997	0.100	0.099	0.097	0.096	0.094	0.809	0.806	0.801	0.795	0.791
0.998	0.098	0.097	0.095	0.094	0.092	0.808	0.805	0.801	0.795	0.790
0.999	0.096	0.095	0.093	0.091	0.090	0.807	0.804	0.799	0.794	0.789

Table 3

Parameter calibration for the TPoP AR design: Type I error rates and power values under the null hypothesis with $\mu_1 = \mu_2 = 1.0$ and the alternative hypothesis with $\mu_1 = 1.0$ and $\mu_2 = 1.5$ by varying the design parameters η_U and η_L (fixing $e^\delta = 1.15$ and $\theta_T = 0.8$). The solid line indicates the 10% type I error rate and 80% power boundaries. The shaded areas are the overlapping parameters that satisfy the design constraints, and the chosen values are in boldface.

η_U	Null cases					Alternative cases				
	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
0.993	0.115	0.113	0.112	0.102	0.088	0.824	0.824	0.809	0.786	0.692
0.994	0.112	0.113	0.108	0.102	0.087	0.816	0.816	0.808	0.779	0.689
0.995	0.112	0.107	0.105	0.098	0.080	0.816	0.815	0.808	0.779	0.688
0.996	0.105	0.106	0.101	0.098	0.080	0.814	0.814	0.808	0.778	0.686
0.997	0.104	0.103	0.098	0.090	0.080	0.814	0.814	0.807	0.776	0.686
0.998	0.102	0.102	0.098	0.084	0.076	0.812	0.812	0.807	0.772	0.686
0.999	0.102	0.098	0.098	0.084	0.074	0.810	0.807	0.807	0.771	0.684

Table 4

Examination of the robustness of the TPrP AR (using the exponential distribution) versus the GS method (using the log-rank test) when data are simulated from Weibull or gamma distributions with the shape parameter k , scale parameter λ , mean μ , and variance σ^2 .

Distribution	Arm 1				Arm 2				Rejection rate	
	k_1	λ_1	μ_1	σ_1^2	k_2	λ_2	μ_2	σ_2^2	TPrP AR	GS
Weibull	1.5	1.108	1.0	0.461	1.5	1.108	1.0	0.461	0.036	0.105
Weibull	1.5	1.108	1.0	0.461	1.5	1.662	1.5	1.037	0.950	0.985
Weibull	0.7	0.79	1.0	2.139	0.7	0.79	1.0	2.139	0.181	0.106
Weibull	0.7	0.79	1.0	2.139	0.7	1.185	1.5	4.812	0.613	0.562
Gamma	1.5	0.667	1.0	0.667	1.5	0.667	1.0	0.667	0.060	0.105
Gamma	1.5	0.667	1.0	0.667	1.5	1.0	1.5	1.5	0.897	0.992
Gamma	0.7	1.428	1.0	1.428	0.7	1.428	1.0	1.428	0.137	0.100
Gamma	0.7	1.428	1.0	1.428	0.7	2.143	1.5	3.214	0.707	0.519