Journal of the
Royal Statistical Society

SERIES C
Applied
Statistics

Appl. Statist. (2012)
61, *Part* 2, *pp.* 219–235

Phase II trial design with Bayesian adaptive randomization and predictive probability

Guosheng Yin,

University of Hong Kong, People's Republic of China

and Nan Chen and J. Jack Lee

University of Texas M. D. Anderson Cancer Center, Houston, USA

[Received April 2010. Final revision June 2011]

Summary. We propose a randomized phase II clinical trial design based on Bayesian adaptive randomization and predictive probability monitoring. Adaptive randomization assigns more patients to a more efficacious treatment arm by comparing the posterior probabilities of efficacy between different arms. We continuously monitor the trial using the predictive probability. The trial is terminated early when it is shown that one treatment is overwhelmingly superior to others or that all the treatments are equivalent. We develop two methods to compute the predictive probability by considering the uncertainty of the sample size of the future data. We illustrate the proposed Bayesian adaptive randomization and predictive probability design using a phase II lung cancer clinical trial, and we conduct extensive simulation studies to examine the operating characteristics of the design. By coupling adaptive randomization and predictive probability approaches, the trial can treat more patients with a more efficacious treatment and allow for early stopping whenever sufficient information is obtained to conclude treatment superiority or equivalence. The design proposed also controls both the type I and the type II errors and offers an alternative Bayesian approach to the frequentist group sequential design.

Keywords: Adaptive randomization; Bayesian inference; Clinical trial ethics; Group sequential method; Posterior predictive distribution; Randomized trial; Type I error; Type II error

1. Introduction

In a conventional phase II trial, an experimental therapy is examined for any antidisease activity in a single-arm setting first. If the new drug shows promising efficacy, it can be evaluated further in a randomized phase II trial or brought forward into a phase III study for confirmatory testing. The end point in an early phase II clinical trial is typically a short-term measure of the treatment efficacy. For example, if a patient receiving treatment achieves complete or partial response within a predefined period of evaluation, the clinical response status Y, a binary outcome, is defined as 1; otherwise it takes the value 0.

Typically, single-arm phase II trials are conducted with a comparison with a historical or a standard response rate. Two-stage or multistage designs are often implemented to increase the efficiency of the trial by allowing for early termination of the trial if the treatment is deemed inefficacious or efficacious after partial data have been observed. Gehan (1961), Simon (1989), Fleming (1982) and Chang *et al.* (1987) proposed phase II designs based on the multiple-testing procedure and group sequential theory. In the Bayesian framework, Thall and Simon (1994)

Address for correspondence: J. Jack Lee, Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, PO Box 301402, Houston, TX 77230-1402, USA.
E-mail: jjlee@mdanderson.org

0035–9254/12/61219

provided some practical guidelines on how to implement a phase II trial. The trial is monitored continuously so that the Bayesian posterior probability is updated after observing every new outcome. Decisions are made adaptively throughout the conduct of the trial until the maximum sample size has been reached. At any time during the conduct of the trial, on the basis of the cumulated data, one can stop the trial and claim that the experimental drug is promising, or not promising or continue the trial because of a lack of convincing evidence to inform a decision. Lee and Liu (2008) developed a continuous Bayesian monitoring scheme based on the predictive probability (PP) for single-arm phase II trials. The PP is obtained by calculating the probability of rejecting the null hypothesis should the trial be conducted to the maximum planned sample size given the interim observed data and assuming that the current trend continues. In the PP framework, one can evaluate the chance that the trial will show a conclusive result at the end of the study, given the current information. Then, the decision to continue or to stop the trial can be made according to the strength of the PP. Comparing with the inference making based on the posterior probability, the PP approach resembles more closely the clinical decision-making process by projecting into the future on the basis of the interim data. Moreover, the PP approach has a higher early stopping probability under the null hypothesis, and the rejection region has a smoother transition compared with the posterior probability approach.

Often, a successful single-arm phase II trial does not necessarily translate to a success of definitive efficacy testing in a phase III trial. One main reason for this is the inherent nature of a single-arm phase II trial, in which the efficacy of a new treatment is compared with historical data or with the standard response rate. Such a comparison is less objective and can often be biased owing to substantial differences in patient populations, study conduct, end point evaluation and medical facilities between the current study and the historical data. Therefore, randomized phase II trials have been proposed to bridge the gap between a successful single-arm phase II trial and a full scale phase III evaluation. As in phase III trials, randomized phase II trials compare the experimental drug with a standard drug in a randomized setting but with a less stringent definition of efficacy and a larger type I error rate. The use of a randomized phase II trial design has become more popular in drug development because it allows for greater objectivity in the assessment of the efficacy of a new treatment. However, such a phase II study should not be considered a poor man's phase III trial and used as a substitute for a more rigorous evaluation of efficacy (Lee and Feng, 2005; Ratain and Sargent, 2009).

In clinical trials, patients are often randomized to different treatments to balance patients' characteristics and to eliminate selection bias and potential confounding factors. This is usually achieved through fixed randomization, which assigns patients to each treatment with a prespecified probability of randomization. However, it may not be ethically desirable to use a fixed probability of randomization such as equal randomization (ER) throughout the trial. This is because interim results based on cumulating data in an on-going trial may indicate that one treatment is likely to be superior to the other; therefore, the clinician's preference would be to provide the superior treatment to more patients. To address the ethical consideration, outcome-based or response adaptive randomization (AR) has been proposed. Response AR assigns a new patient to a more efficacious arm with a higher probability based on the cumulated response data. This enhances the individual ethics design in which more patients participating in the trial are assigned to the superior treatment as the trial proceeds (Flehinger *et al.*, 1972; Louis, 1975, 1977; Berry and Eick, 1995; Karrison *et al.*, 2003; Hu and Rosenberger, 2006; Thall and Wathen, 2007; Zhang and Rosenberger, 2007; Cheng and Berry, 2007; Lee *et al.*, 2010).

One such trial which was recently considered at the University of Texas M. D. Anderson Cancer Center is a neoadjuvant lung cancer trial. Neoadjuvant chemotherapy or new targeted agents are given to lung cancer patients before surgery with the intent of shrinking the tumour such that better disease control and a smaller surgical field can be achieved. Eligible patients are to be randomized to carboplatin plus paclitaxel (the standard chemotherapy) or an AKT inhibitor plus an MEK inhibitor (new targeted agents). Patients will be treated for 4 weeks before surgery. The primary end point of the trial is the 4-week clinical response status. We contemplated several design options including an ER design without early stopping, a group sequential design with ER using the Hwang–Shih–DeCani α-spending function (Hwang *et al.*, 1990) and futility stopping (DeMets and Ware, 1982), and a Bayesian AR design with PP monitoring.

Motivated by this lung cancer trial, we propose a randomized phase II design with Bayesian adaptive randomization and predictive probability (BARPP) monitoring. Owing to AR, the future sample size in each arm becomes unknown; however, such information is essential for computing the PP. We develop two approaches to approximate the PP, which is used for adaptive decision making in the trial conduct. We characterize the design to achieve the usual frequentist properties, such as controlling the type I and type II errors. At any given time, if there is a high probability that one treatment is better than the other, we would stop the trial and declare superiority; if there is a high probability that the treatments are similar in terms of efficacy, we would stop the trial and declare equivalence; otherwise, we would continue the trial. Through the use of AR, more patients are treated with the better treatment. Our method combines the advantages of Bayesian AR with PP to develop a flexible and ethical trial design.

The rest of this paper is organized as follows. In Section 2, we introduce the notation and propose the randomized phase II design using the BARPP monitoring. In Section 3, we demonstrate how to calibrate the design parameters and present simulation studies to examine the design properties under different practical scenarios. We give concluding remarks in Section 4.

The programs that were used to analyse the data can be obtained from

http://www.blackwellpublishing.com/rss

2. Bayesian trial design

2.1. Predictive probability

Suppose that we compare K treatments in a K-arm randomized phase II trial. Let p_k be the response rate of treatment k, and assign p_k a prior distribution of beta(α_k, β_k), for $k=1,\ldots,K$. If, among n_k subjects treated in arm k, we observe x_k responses, then

$$X_k \sim \text{binomial}(n_k, p_k),$$

and the posterior distribution of p_k is

$$p_k|(X_k=x_k) \sim \text{beta}(\alpha_k+x_k, \beta_k+n_k-x_k).$$

If the maximum sample size in arm k is N_k, then the number of responses in the future N_k-n_k patients, Y_k, follows a beta–binomial distribution:

$$Y_k|(X_k=x_k) \sim \text{beta–binomial}(N_k-n_k, \alpha_k+x_k, \beta_k+n_k-x_k).$$

When $Y_k=y_k$, the posterior distribution of the response rate given the current and future data is

$$p_k|(X_k=x_k, Y_k=y_k) \sim \text{beta}(\alpha_k+x_k+y_k, \beta_k+N_k-x_k-y_k).$$

For ease of exposition, we consider two treatments to illustrate the design, i.e. $K=2$. We specify a clinically meaningful treatment difference δ, and a threshold probability θ_{T}. If

$$P(|p_2-p_1|>\delta|X_1=x_1, X_2=x_2, Y_1=y_1, Y_2=y_2) \geqslant \theta_{\mathrm{T}},$$

we claim non-equivalence of the two treatments, i.e. one treatment is superior to the other. However, Y_1 and Y_2 are the future data, which have still not been observed at the current decision-making stage. We can average out the randomness in Y_1 and Y_2 by computing the PP as follows:

$$\begin{aligned}\mathrm{PP} &= E_{Y_1,Y_2}[I\{P(|p_2-p_1|>\delta|X_1=x_1, X_2=x_2, Y_1, Y_2) \geqslant \theta_{\mathrm{T}}\}] \\ &= \sum_{y_1=0}^{N_1-n_1} \sum_{y_2=0}^{N_2-n_2} P(Y_1=y_1|X_1=x_1)\, P(Y_2=y_2|X_2=x_2) \\ &\quad \times I\{P(|p_2-p_1|>\delta|X_1=x_1, X_2=x_2, Y_1=y_1, Y_2=y_2) \geqslant \theta_{\mathrm{T}}\} \qquad (1)\end{aligned}$$

where $I\{\cdot\}$ is the indicator function. PP denotes the PP to claim superiority at the end of the trial.

Following the work of Lee and Liu (2008), we need to specify the lower and upper cut-off probabilities for adaptive decision making in the trial conduct. The decision rules based on the PP are as follows.

(a) Equivalence stopping: if PP $<\theta_{\mathrm{L}}$, then we stop the trial and accept the null hypothesis to claim treatment equivalence.

(b) Superiority stopping: if PP $>\theta_{\mathrm{U}}$, then we stop the trial and reject the null hypothesis to claim a superior treatment arm.

We can maintain the frequentist type I and type II error rates by calibrating the design parameters $(N, \delta, \theta_{\mathrm{T}}, \theta_{\mathrm{L}}, \theta_{\mathrm{U}})$, where N is the maximum sample size of the trial, $N=N_1+N_2$.

A trial design based on the PP allows for continuous monitoring. If the two treatments have similar efficacy effects, or if one treatment is overwhelmingly better than the other, the trial can be stopped early when sufficient evidence has accumulated. This would result in a smaller expected sample size, and hence a more efficient trial. At the end of the trial, we either declare that one treatment is better than the other, or the equivalence of two treatments.

2.2. Response adaptive randomization

Response AR enhances the individual ethics in clinical trials by assigning more patients to the putatively better treatments on the basis of the interim data. For the stability of parameter estimation and randomization at the beginning of the trial, there is typically a prelude of ER before AR takes effect. First, ER is applied to a fixed number of subjects and, subsequently, the remaining subjects are adaptively randomized to a superior arm with a higher probability. Following the work of Thall and Wathen (2007), we denote the randomization probability as

$$\pi = \frac{P(p_2>p_1|X_1=x_1, X_2=x_2)^{\tau}}{P(p_2>p_1|X_1=x_1, X_2=x_2)^{\tau} + \{1-P(p_2>p_1|X_1=x_1, X_2=x_2)\}^{\tau}}. \qquad (2)$$

We assign the next cohort of patients to arm 2 with probability π, and to arm 1 with probability $1-\pi$. We use the tuning parameter τ to control the AR rate; if $\tau=0$, then $\pi=0.5$, leading to ER. A larger value of τ would lead to a higher imbalance in allocation of patients between the two arms and vice versa. Such Bayesian AR takes into consideration both the estimated efficacy rates and their variability. In contrast, using only the point estimates, $\pi=\hat{p}_2/(\hat{p}_1+\hat{p}_2)$, as the assigning probability to arm 2 does not account for the variability.

The PP in equation (1) can be easily calculated if the total sample sizes in arms 1 and 2, N_1 and N_2, are known and fixed. However, N_1 and N_2 can only be known *a priori* in the fixed randomization procedure. In the case of response AR, the probability of assignment for each incoming subject changes throughout the trial. Therefore, N_1 and N_2 in equation (1) are not fixed any more, which poses a new challenge in computing the PP.

In what follows, we propose two different ways to compute the PP. The first method is more rigorous but more computationally intensive, and the second applies an approximation but is relatively fast. In our numerical studies, we have found that these two approaches produce very similar results and thus lead to very close design operating characteristics.

2.2.1. *Method 1 of computing the predictive probability*

Once AR is in effect, the total numbers of subjects in arm 1 and arm 2, N_1 and N_2, become random, whereas the number of remaining subjects in the trial, m, is fixed, if the trial is not allowed for early termination. Let Z be the number of subjects who would be assigned to arm 2; then $Z \sim \text{binomial}(m, \pi)$, i.e.

$$P_Z(z|X_1 = x_1, X_2 = x_2) = \binom{m}{z} \pi^z (1-\pi)^{m-z}. \tag{3}$$

To obtain the PP, we first average over Y_1 and Y_2 conditioning on $Z = z$, and then average over Z according to the binomial distribution in equation (3). Following this route,

$$\begin{aligned}\text{PP} = \sum_{z=0}^{m} \sum_{y_1=0}^{m-z} \sum_{y_2=0}^{z} & P_Z(z|X_1 = x_1, X_2 = x_2) \\ & \times P(Y_1 = y_1|X_1 = x_1, Z = z)\, P(Y_2 = y_2|X_2 = x_2, Z = z) \\ & \times I\{P(|p_2 - p_1| > \delta | X_1 = x_1, Y_1 = y_1, X_2 = x_2, Y_2 = y_2) \geqslant \theta_{\mathrm{T}}\},\end{aligned}$$

which can be quite computationally intensive owing to the additional summation that marginalizes over Z. This method enumerates all the possibilities of the future sample sizes; we refer to it as method 1.

2.2.2. *Method 2 of computing the predictive probability*

The first method involves three embedded summations and is computationally expensive. The second approach is to approximate $N_k - n_k$ by the expected number of subjects assigned to arm k for $k = 1, 2$, i.e. $N_1 - n_1 = m(1-\pi)$ and $N_2 - n_2 = m\pi$. This is a direct approximation based on the currently observed data which does not impose any further computational difficulties.

Although the total sample size of the trial is fixed, the remaining sample size m is not fixed if the trial is allowed for early termination. Early termination of a trial is an extra feature of a study design. As will be seen in Section 3, the design parameters are calibrated in a two-stage sequential procedure: we first choose δ and θ_{T} without early stopping and then select the early stopping parameters θ_{U} and θ_{L}. The design parameters are calibrated in such a sequential order to avoid the intertwining effects of early stopping.

2.3. *Multiple-treatment arms*

When we consider multiple treatments with $K > 2$ in a randomized trial, we assume that there is one standard treatment and $K - 1$ experimental treatments. Let p_1 denote the response rate of the standard arm and $p_{\max}$ denote the treatment with the highest efficacy among $(p_2, \ldots, p_K)$.

Then, the PP of selecting the best arm at the end of the trial is

$$\begin{aligned}\mathrm{PP}=&\sum_{y_1=0}^{N_1-n_1}\ldots\sum_{y_K=0}^{N_K-n_K}P(Y_1=y_1|X_1=x_1)\ldots P(Y_K=y_K|X_K=x_K)\\&\times I\{P(|p_{\max}-p_1|>\delta|X_1=x_1,Y_1=y_1;\ldots;X_K=x_K,Y_K=y_K)\geqslant\theta_{\mathrm{T}}\},\end{aligned}\qquad(4)$$

where $(X_1,\ldots,X_K)$ are the currently observed data and $(Y_1,\ldots,Y_K)$ are the future data in the K arms.

The Bayesian AR procedure needs to accommodate comparisons between these K arms. There are many ways to construct the randomization probabilities. For example, we first obtain the average of the posterior samples of the response rates,

$$\bar{p}=\frac{1}{K}\sum_{k=1}^{K}p_k,$$

and then compute the posterior probability of

$$\lambda_k=P(p_k>\bar{p}|X_1=x_1,\ldots,X_K=x_K).$$

We would assign the next cohort of patients to arm k with probability $\pi_k=\lambda_k^{\tau}/\Sigma_{j=1}^{K}\lambda_j^{\tau}$. This leads to a multinomial distribution with the remaining number of subjects $m=N_1+\ldots+N_K-n_1-\ldots-n_K$. We can also replace $\bar{p}$ with p_1, or define π_k as the probability that arm k has the largest response rate among all treatments.

Let Z_k be the number of subjects that would be assigned to arm k; then $(Z_1,\ldots,Z_K)\sim$ multinomial$(m;\pi_1,\ldots,\pi_K)$, i.e.

$$P_{Z_1,\ldots,Z_K}(z_1,\ldots,z_K|X_1=x_1,\ldots,X_K=x_K)=\frac{m!}{z_1!\ldots z_K!}\pi_1^{z_1}\ldots\pi_K^{z_K},$$

with $\Sigma_{k=1}^{K}z_k=m$ and $\Sigma_{k=1}^{K}\pi_k=1$. To obtain the PP, we first average over $(Y_1,\ldots,Y_K)$ conditioning on $(Z_1=z_1,\ldots,Z_K=z_K)$, and then average over all the Z_ks according to the multinomial distribution:

$$\begin{aligned}\mathrm{PP}=&\sum_{z_1=0}^{m}\ldots\sum_{z_K=0}^{m}\sum_{y_1=0}^{z_1}\ldots\sum_{y_K=0}^{z_K}\frac{m!}{z_1!\ldots z_K!}\pi_1^{z_1}\ldots\pi_K^{z_K}\\&\times P(Y_1=y_1|X_1=x_1,Z_1=z_1)\ldots P(Y_K=y_K|X_K=x_K,Z_K=z_k)\\&\times I\{P(|p_{\max}-p_1|>\delta|X_1=x_1,Y_1=y_1;\ldots;X_K=x_K,Y_K=y_K)\geqslant\theta_{\mathrm{T}}\},\end{aligned}$$

subject to $\Sigma_{k=1}^{K}z_k=m$. The computation increases multiplicatively with respect to the number of treatment arms. However, we can easily generalize method 2 of computing the PP by using the multinomial distribution and the expected number of subjects assigned to arm k, $N_k-n_k=m\pi_k$.

3. Simulation studies

3.1. Parameter calibration

In practice, we need to calibrate the five design parameters $(N,\delta,\theta_{\mathrm{T}},\theta_{\mathrm{L}},\theta_{\mathrm{U}})$ on the basis of the desired type I error rate and power in the trial. We first specify N, and then take a two-stage procedure to calibrate the main design parameters $(\delta,\theta_{\mathrm{T}})$, and the early termination parameters $(\theta_{\mathrm{L}},\theta_{\mathrm{U}})$ for equivalence or superiority.

In the first stage, we set $\theta_{\mathrm{L}}=0$ and $\theta_{\mathrm{U}}=1$, so that the trial would not be terminated early, to determine the threshold values of δ and θ_{T}. We performed a series of simulation studies with different values of δ and θ_{T} and compared the corresponding type I error rates and powers.

Recall the neoadjuvant lung cancer trial that was mentioned in Section 1; in this phase II trial, we chose N to control both the type I error rate (10% or less) and the power (at least 80%). One of the two treatments (say, arm 1) under investigation was the standard chemotherapy with a known efficacy rate: $p_1 = 0.2$. We assumed that the new treatment would double the response rate, i.e. $p_2 = 0.4$.

The total sample size was set as $N = 160$, although the actual sample size could be much less owing to early termination of the trial. The first 40 patients ($n_1 = n_2 = 20$) were equally randomized to the two arms and thereafter patients were adaptively randomized on the basis of the posterior probabilities of comparing the response rates of the two treatments after observing every single outcome. The tuning parameter τ was taken as 0.5 (Thall and Wathen, 2007) and the randomization rates were restricted between 0.1 and 0.9 to prevent having very unbalanced randomization rates. To allow the likelihood to dominate the posterior distribution, we took a relatively non-informative prior distribution of beta$(2, 2)$ for both p_1 and p_2. We varied δ from 0.02 up to 0.09, and θ_{T} from 0.70 up to 0.90. We carried out 10000 simulated clinical trials. For each of the paired values of (δ, θ_T), we obtained the type I error rate and power as listed in Table 1.

Considering the null cases in the left-hand panel of Table 1, all entries of the type I error rates below the boundary line of the staircase curve are 10% or less, for which the paired values of $(\delta, \theta_{\mathrm{T}})$ satisfy our requirement. Simultaneously, under the alternative cases, we need to find the paired values of $(\delta, \theta_{\mathrm{T}})$ that lead to a power of 80% or higher. These correspond to the power values above the staircase curve in the right-hand panel of Table 1. The overlapping tinted area meets both the type I error and the power constraints. With a clinically meaningful range of equivalence of $\delta = 0.05$, we chose $\theta_{\mathrm{T}} = 0.85$ for further study. It is worth noting that a higher power value corresponds to a higher type I error rate. The null cases cover $p_1 = p_2 = p$ for p between 0.2 and 0.4, and we chose $p = 0.4$ to report as it corresponds to the case with the largest type I error rate.

In the second stage, fixing $\delta = 0.05$ and $\theta_{\mathrm{T}} = 0.85$, we followed a similar procedure to calibrate $(\theta_{\mathrm{L}}, \theta_{\mathrm{U}})$, which determine the early termination of a trial due to equivalence or superiority respectively. Although the design allows monitoring after every outcome becomes available, from the computational and practical point of view, we opted to monitor the trial for early termination with a cohort size of 10. We explored method 1 by enumerating all the possibilities

Table 1. Type I error rates and power values under the null hypothesis of $p_1 = p_2 = 0.4$ and alternative hypothesis of $p_1 = 0.2$ and $p_2 = 0.4$ by varying the design parameters δ and θ_{T}†

δ	*Null cases–results for the following θ_{T}:*					*Alternative cases–results for the following θ_{T}:*				
	0.70	*0.75*	*0.80*	*0.85*	*0.90*	*0.70*	*0.75*	*0.80*	*0.85*	*0.90*
0.02	1.000	1.000	0.797	0.427	0.230	1.000	1.000	0.991	0.967	0.919
0.03	0.844	0.517	0.362	0.228	0.124	0.991	0.978	0.955	0.918	0.859
0.04	0.429	0.317	0.224	0.142	0.080	0.967	0.949	0.920	0.864	0.788
0.05	0.291	0.218	0.157	*0.097*	0.053	0.945	0.912	0.875	*0.822*	0.735
0.06	0.214	0.158	0.112	0.073	0.038	0.917	0.878	0.833	0.765	0.669
0.07	0.157	0.119	0.080	0.056	0.028	0.878	0.835	0.785	0.716	0.617
0.08	0.111	0.086	0.060	0.034	0.016	0.845	0.797	0.741	0.667	0.564
0.09	0.093	0.064	0.043	0.026	0.013	0.800	0.753	0.680	0.612	0.501

†The step curves indicate the 10% type I error and 80% power boundaries. The tinted areas are the overlapping parameters that satisfy the design constraints. The values chosen are in italics.

Table 2. Type I error rates and power values by varying the design parameters θ_L and θ_U using method 1 and method 2 (fixing $\delta = 0.05$ and $\theta_T = 0.85$)†

θ_U	*Null cases–results for the following θ_L:*					*Alternative cases–results for the following θ_L:*				
	0.00	*0.05*	*0.10*	*0.15*	*0.20*	*0.00*	*0.05*	*0.10*	*0.15*	*0.20*
Method 1										
0.95	0.126	0.116	0.112	0.106	0.095	0.839	0.823	0.792	0.783	0.747
0.96	0.120	0.120	0.105	0.107	0.096	0.828	0.815	0.799	0.772	0.739
0.97	0.115	0.106	0.097	0.092	0.092	0.827	0.813	0.791	0.765	0.734
0.98	0.107	0.105	0.092	0.091	0.082	0.827	0.814	0.791	0.764	0.734
0.99	0.105	*0.099*	0.095	0.080	0.079	0.820	*0.803*	0.796	0.766	0.723
1.00	0.101	0.097	0.088	0.083	0.074	0.822	0.806	0.789	0.755	0.733
Method 2										
0.95	0.125	0.116	0.113	0.104	0.099	0.840	0.822	0.793	0.782	0.747
0.96	0.121	0.115	0.106	0.102	0.088	0.829	0.816	0.799	0.772	0.734
0.97	0.112	0.110	0.099	0.096	0.082	0.829	0.813	0.789	0.765	0.735
0.98	0.112	0.101	0.096	0.092	0.085	0.826	0.815	0.796	0.765	0.731
0.99	0.102	*0.096*	0.087	0.076	0.071	0.819	*0.802*	0.796	0.767	0.722
1.00	0.101	0.093	0.088	0.080	0.070	0.820	0.806	0.790	0.760	0.731

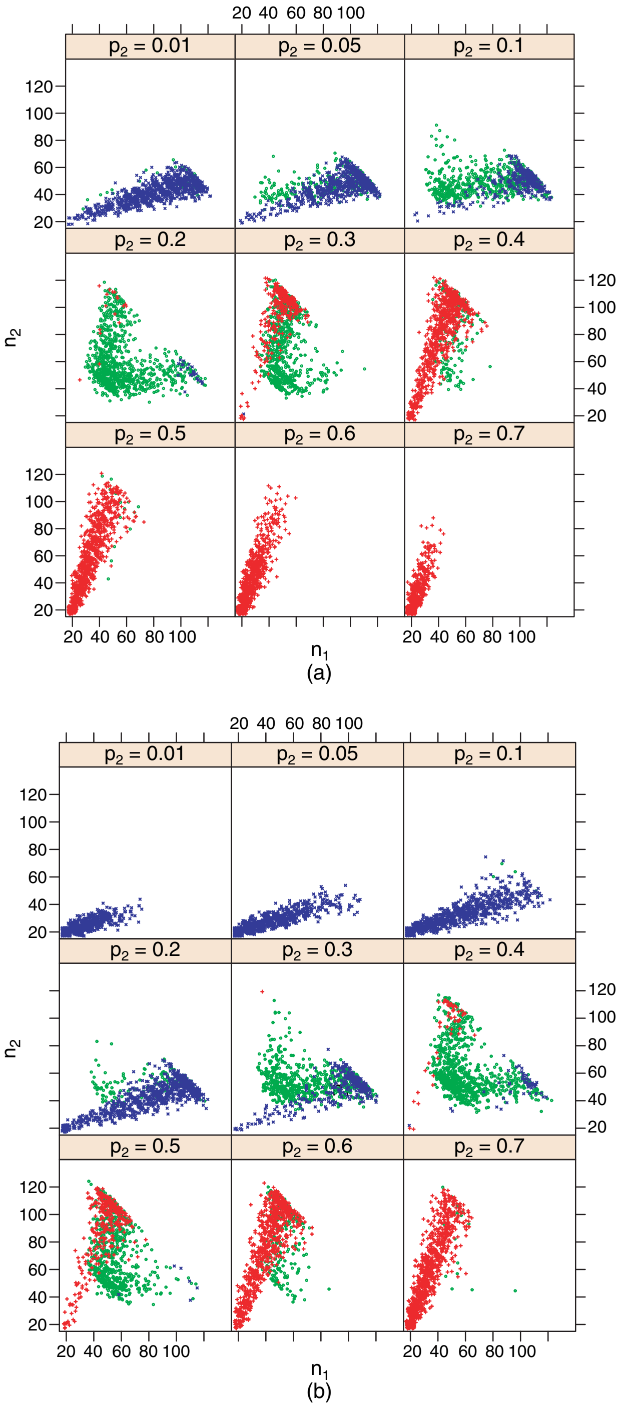
†The step curves indicate the 10% type I error and 80% power boundaries. The tinted areas are the overlapping parameters that satisfy the design requirements, and the chosen values are in italics.

of the future sample sizes and method 2 by using the expected future sample sizes to compute the PPs. In Table 2, we can see that the type I error rates and powers obtained from methods 1 and 2 are very close, which implies that using the expected number of future subjects in method 2 gives a very good approximation to the results from all possible future sample sizes. Our goal is still to maintain a type I error rate of 10% or lower and to achieve a power of 80% or higher when the trial is allowed to terminate early. There are multiple pairs of (θ_L, θ_U) that satisfy our design requirements, as indicated by the values in the tinted areas of Table 2, from which we selected $\theta_L = 0.05$ and $\theta_U = 0.99$.

3.2. Selected scenarios

To examine the performance of the proposed design with the BARPP, we carried out a series of simulation studies under various scenarios. We varied the true response rate p_1 from 0.1 to 0.4 and, for each fixed value of p_1, we set p_2 at a value from 0.01 to 0.8. In all the simulations, we fixed the design parameters as $N = 160, \delta = 0.05, \theta_T = 0.85, \theta_L = 0.05$ and $\theta_U = 0.99$ on the basis of the two-stage parameter calibration procedure that was described in the previous section. We replicated 10000 clinical trials for each configuration.

Fig. 1 illustrates the decision and sample size distributions with various values of p_1 and p_2. The colour and the co-ordinates of each point indicate the final decision and the number of patients assigned to each arm respectively. For a better view, the points are slightly jittered to break the ties and only 1000 trials are presented. When $p_1 = p_2$, the green points (shown in circles with a decision of $p_1 = p_2$) take a dominant role, indicating that the two treatments are equivalent; the red points (shown in plus symbols with a decision of $p_1 < p_2$) and the blue points (shown in crosses with a decision of $p_1 > p_2$) take roughly symmetric positions at the two corners. The small numbers of red and blue points depict that the stochastic nature of the



(a)

(b)

Fig. 1. Sample size and decision distributions for various values of p_1 and p_2, with the BARPP designs (the value of p_2 varies from 0.01 to 0.7 whereas the value of p_1 is fixed at (a) 0.2 and (b) 0.4; for each p_1- and p_2-combination, 1000 trials were simulated; each point on the plot corresponds to one trial; the *x*-co-ordinate and *y*-co-ordinate of each point indicate the number of patients in arm 1 and arm 2 respectively; the colour of each point indicates the decision made at the end of each trial): ×, $p_1 > p_2$; ○, $p_1 = p_2$; +, $p_1 < p_2$

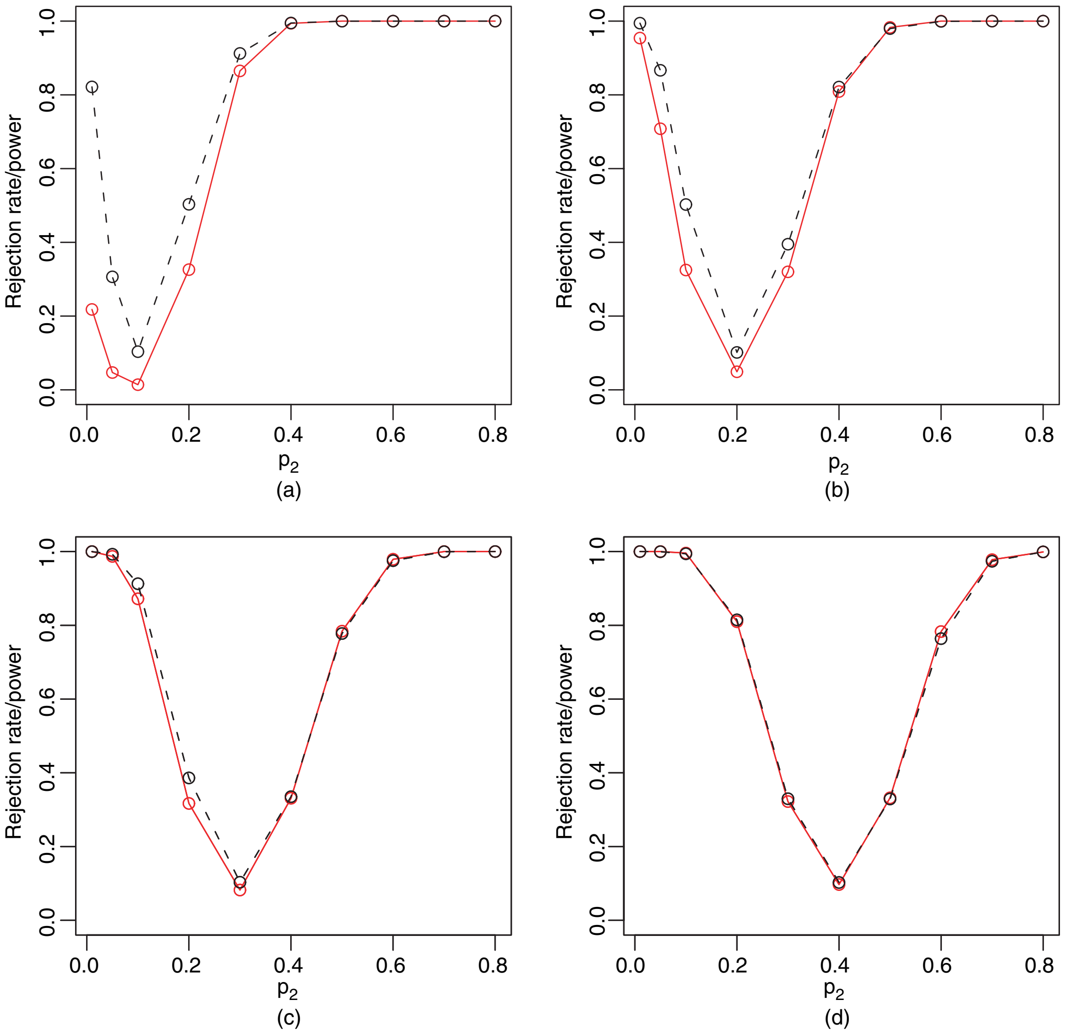
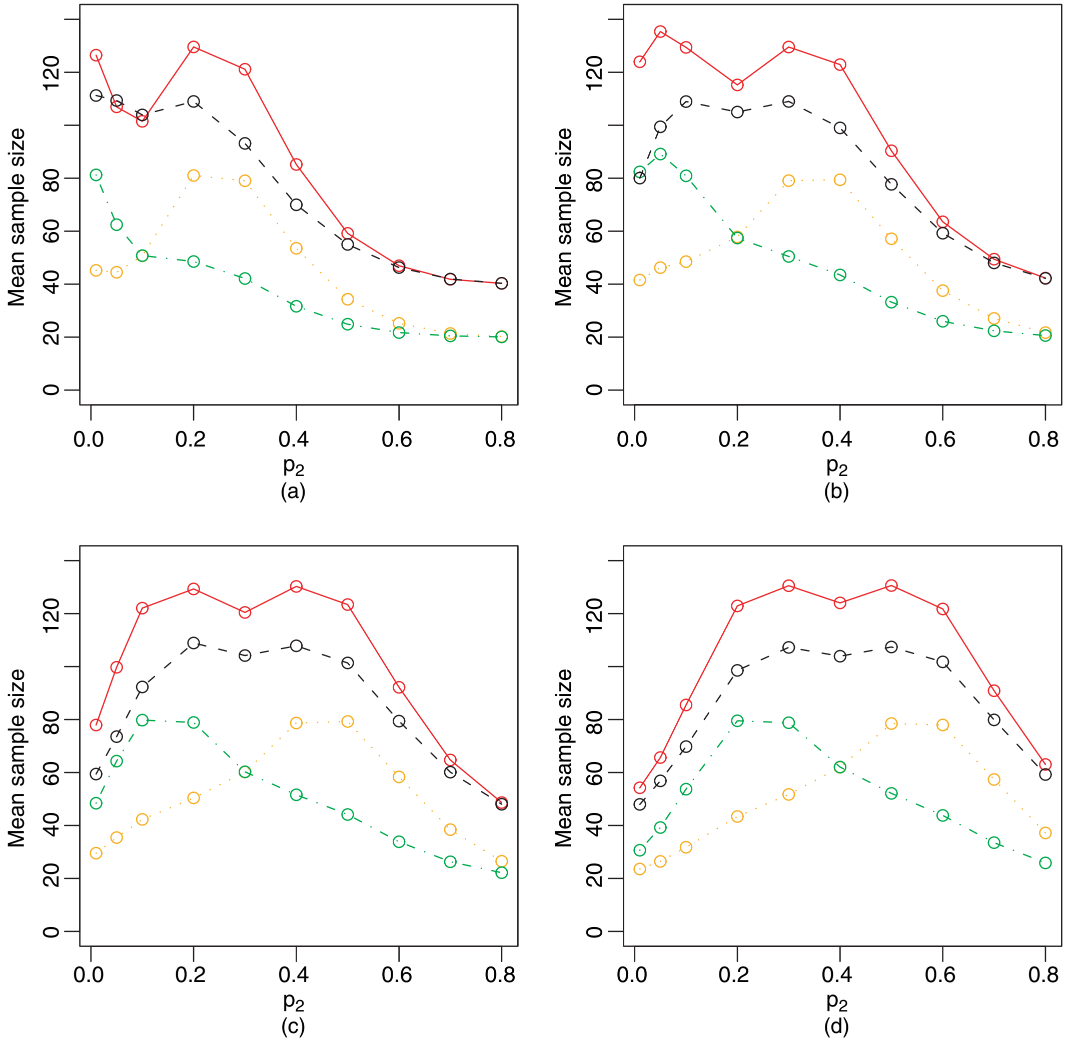


Fig. 2. Rejection rates of H_0 and power values by using the BARPP (———) and GS (-------) methods at various values of p_2, while the response rate of arm 1 is fixed at (a) $p_1 = 0.1$, (b) $p_1 = 0.2$, (c) $p_1 = 0.3$ and (d) $p_1 = 0.4$

responses may result in an imbalance of sample allocation between the two arms, and also lead to incorrect final conclusions. When the difference between p_1 and p_2 is large (e.g. $p_1 = 0.2$ and $p_2 = 0.7$, or $p_1 = 0.4$ and $p_2 = 0.01$), AR assigns most of the patients to the superior arm and almost all the simulated trials were terminated early. When the difference between p_1 and p_2 is small (e.g. $p_1 = 0.2$ and $p_2 = 0.3$, or $p_1 = 0.4$ and $p_2 = 0.3$), the treatments were claimed to be either equivalent or different and many trials used a large number of patients.

We illustrate the percentages of rejecting the null hypothesis under various scenarios in Fig. 2. The value of p_1 is fixed and the value of p_2 varies from 0.01 to 0.8. The curves that were obtained from methods 1 and 2 are indistinguishable; hence, only one curve for the BARPP design is shown. The minimum percentage of rejecting the null case is always located at $p_1 = p_2$ for each scenario, which corresponds to the type I error rate. Our method yielded a minimum rejection rate of 0.014, 0.049, 0.082 and 0.097 at the null cases with $p_1 = 0.1, 0.2, 0.3, 0.4$ respectively. The



(a)

(b)

(c)

(d)

Fig. 3. Mean sample size on arm 1 (·-·-·-·-·) and arm 2 (· · · · · ·), and the mean total sample size of the BARPP (———) and GS (-------) methods at various values of p_2, while the response rate of arm 1 is fixed at (a) $p_1 = 0.1$, (b) $p_1 = 0.2$, (c) $p_1 = 0.3$ and (d) $p_1 = 0.4$

power curves typically have a ‘V’ shape because the power increases as p_2 moves away from p_1 to either the left-hand or the right-hand side.

To compare our design with the frequentist approach, in Fig. 2 we also present the corresponding power values calculated from the group sequential (GS) design by using the R package `gsDesign` (`http://gsdesign.r-forge.r-project.org/`). Given a significance level of 0.1 and a power of 80% under the alternative case with $p_1 = 0.2$ and $p_2 = 0.4$, the upper and lower boundary values at each group sequential test were calculated with the Hwang–Shih–DeCani spending function (Hwang *et al.*, 1990), for which the upper design parameter $\lambda = -4$ yielded the O’Brien–Fleming type of boundary (O’Brien and Fleming, 1979) for efficacy stopping and the lower design parameter $\lambda = -2$ was taken for futility stopping. Both futility (or equivalence) and efficacy stopping were considered in the GS design to make it comparable with the BARPP method. The number of patients in each group under the GS design was also

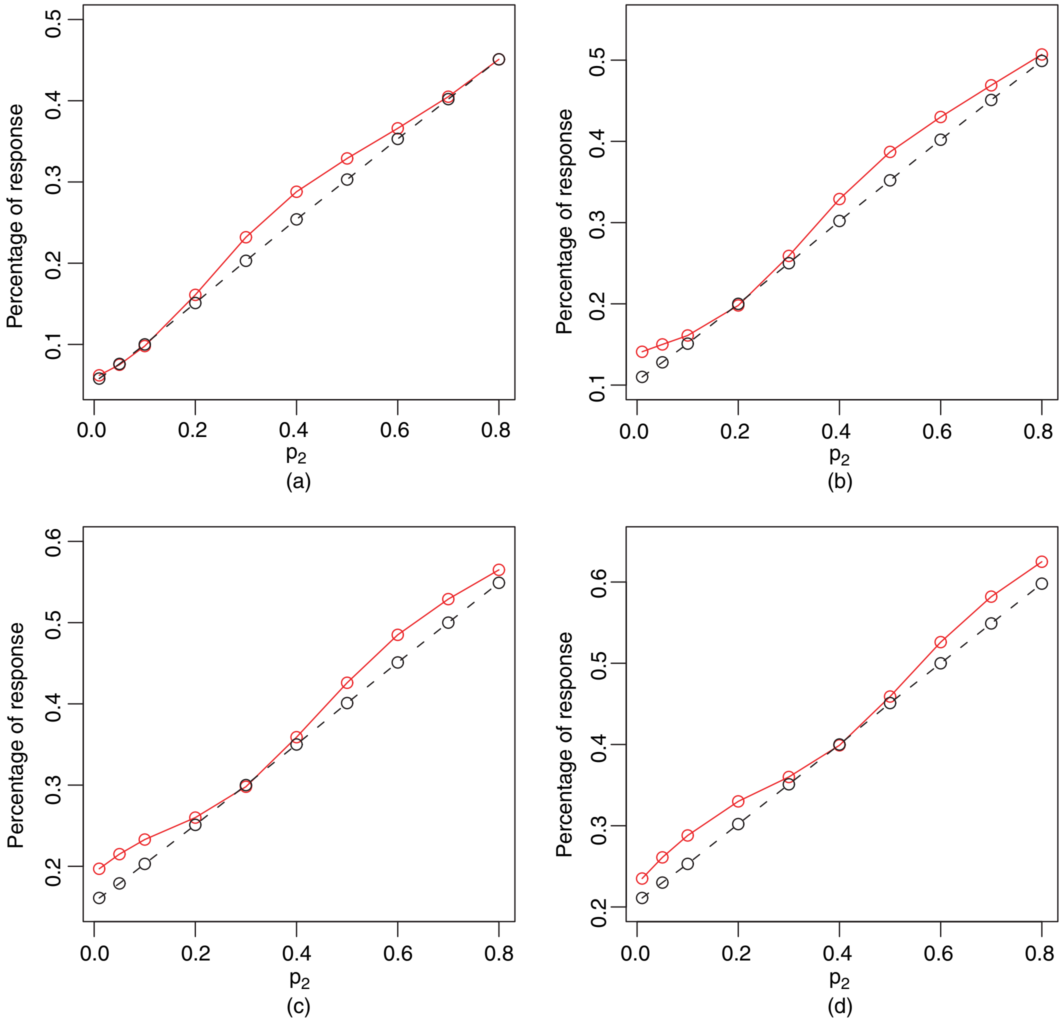


Fig. 4. Percentages of patients' responses by using the BARPP (———) and GS (-------) methods at various values of p_2, while the response rate of arm 1 is fixed at (a) $p_1 = 0.1$, (b) $p_1 = 0.2$, (c) $p_1 = 0.3$ and (d) $p_1 = 0.4$

set as 10 with five patients in each arm. Equal randomization is applied throughout with the maximum number of patients at 140. No early termination was allowed for the first 40 patients and thereafter the GS boundaries were applied. On the basis of 10000 simulations, the GS method also produced a V-shaped power curve similar to that using the BARPP. In scenarios with $p_1 = 0.3$ or $p_1 = 0.4$, the curves of the BARPP and GS designs are almost identical. However, for scenarios with $p_1 = 0.1$ or $p_1 = 0.2$, the power values by using the GS design are higher than those by using the BARPP design. This is because the BARPP design takes a more conservative approach to controlling type I errors across different null response rates and thus the BARPP has lower type I error rates.

Fig. 3 illustrates the numbers of patients who were allocated to arm 1 and arm 2, and the total sample sizes under various scenarios. It can be seen that more patients were randomized to a more efficacious treatment arm by using the BARPP method. When $p_1 = p_2$, patients were essentially equally randomized to the two arms by using AR. When the difference between the

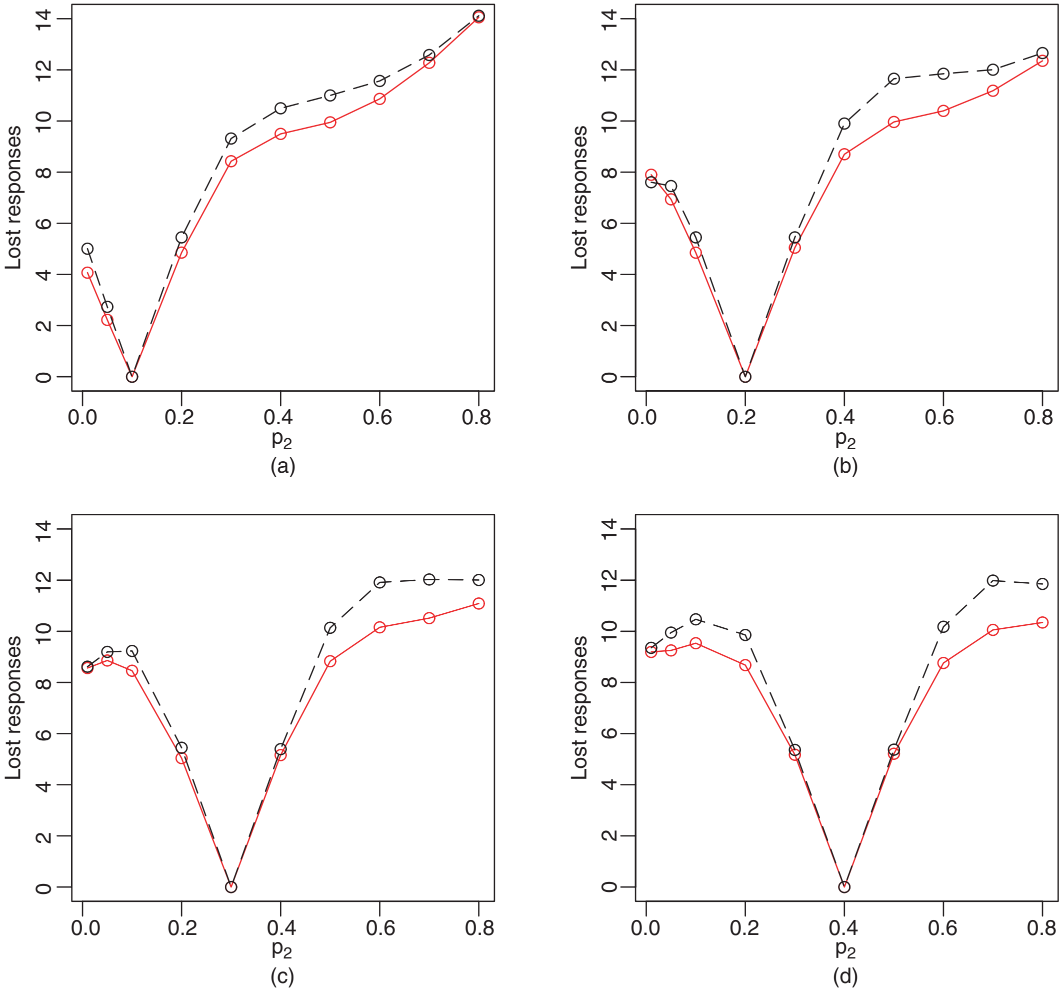


Fig. 5. Numbers of lost responses by using the BARPP (———) and GS (-----) methods at various values of p_2, while the response rate of arm 1 is fixed at (a) $p_1 = 0.1$, (b) $p_1 = 0.2$, (c) $p_1 = 0.3$ and (d) $p_1 = 0.4$

two response rates was substantially large, early stopping took place very quickly in the AR stage, which led to small sample sizes in both arms. When p_2 increases while fixing p_1 at a certain value, the number of patients who were assigned to arm 2 increases and, as a result, the overall percentage of patient responses increases. The total sample size of the BARPP method is slightly larger than that of the GS design, which is mainly caused by AR in the BARPP method. Owing to the provision allowing for early stopping, both the BARPP and the GS design are more efficient and more ethical than the fixed sample size design. Allowing for early stopping is an important design consideration for randomized phase II trials (Lee and Feng, 2005).

Fig. 4 shows a comparison of the percentages of patient responses between the BARPP and the GS methods. It can be seen that the overall response rate of the BARPP method is higher than that of the GS method when the values of p_1 and p_2 are different. When $p_1 = p_2$, the percentages of response are the same between the two methods because patients are also equally randomized using the BARPP method. When the value of $|p_1 - p_2|$ lies around 0.3, we observe the biggest difference in the overall response rate between the two methods. For $p_1 = 0.1$ and

$p_2 = 0.3$, the overall response rates of the BARPP and the GS methods were 0.233 and 0.203 respectively, and, for $p_1 = 0.2$ and $p_2 = 0.4$, the corresponding response rates were 0.33 and 0.301. Despite the substantial difference in sample size between the two arms (for example, the averaged sample sizes of treatments 1 and 2 are 43 and 79 respectively, for the latter case; Fig. 1), AR achieves only a modest 10% gain in the overall response rate compared with ER.

When the difference between p_1 and p_2 is larger than 0.3, early termination occurs very quickly after ER of the first 40 patients, and thus the number of patients who were assigned in the AR stage becomes very small. This would in turn lead to a small difference in the percentage of response between the BARPP and the GS methods. For example, in Fig. 3(a), when $p_1 = 0.1$ and $p_2 = 0.7$ or $p_2 = 0.8$, i.e. treatment 2 is overwhelmingly superior to treatment 1, the trial is stopped soon after the initial ER stage to claim superiority of treatment 2, and the total sample size is very small (41.1 and 40.1 for the cases of $p_2 = 0.7$ and $p_2 = 0.8$ respectively). Comparing Figs 2 and 3, it is interesting that the power still increases even when the sample size decreases. Because of trial early termination based on the PP, the sample size can be substantially reduced if a decision can be made in the middle of the trial.

As suggested by the Associate Editor, we can measure the number of lost responses due to treating patients with the worse treatment, i.e. the number of patients who were assigned to the worse treatment arm multiplied by $|p_2 - p_1|$. In Fig. 5, we can see that the lost responses in the BARPP design are lower than that in the GS design, mainly because of AR. Moreover, we also explored the BARPP design without equivalence stopping and the findings are quite similar, except that the trials may run until reaching the maximum sample size when p_1 and p_2 are close to each other. The added feature of AR in the BARPP design assigns more patients to the better treatment arm, leading to more imbalance between the two arms. The imbalance in allocation of patients may result in a loss of statistical power. Hence, the sample size that is required for the BARPP is typically larger than that for the GS design. In addition, we also observed more variability in the sample size of the BARPP design. Overall, the BARPP design performed very well in terms of frequentist properties, such as maintaining the type I error rate and achieving the power desired.

4. Discussion

To make the best use of resources and to select promising candidate treatments for a phase III trial carefully, there is an increasing need for randomized phase II trial designs. Using PPs to guide the phase II trial design is appealing to clinical investigators. It is desirable to terminate a trial if the cumulative evidence is sufficiently strong to draw a definitive conclusion in the middle of the trial conduct. Adding AR further enhances the individual ethics of the clinical trial by allocating more patients to more effective treatments, and it results in an increase in the overall trial response. Designs that evaluate short-term responses, such as binary outcomes, are ideal for the application of Bayesian AR, which can be implemented in an almost realtime fashion. We have proposed two different approaches to solving the issue of random future sample sizes in computing the PP, both of which lead to essentially identical trial operating characteristics. However, the computation time for method 2 is only about 4% of that required for method 1.

Several design parameters can be calibrated to meet the goals for various designs. For example, we chose to randomize equally 25% of the patients at the beginning of the trial to learn about the treatment efficacy before randomizing patients adaptively. We also constrained the randomization probability to be within [0.1, 0.9]. In addition, we chose the randomization tuning parameter $\tau = 0.5$ to avoid extreme imbalance in randomizing patients. All those choices limited the utility of AR, which could be applied more aggressively. Furthermore, we only performed simulation studies based on two-arm trials. As was reported recently, only limited advantages of AR are observed in two-arm trials (Korn and Freidlin, 2011), and the advantages of AR can be more pronounced in multiarm trials (Berry, 2011). The trade-off between ER and AR is that ER is favoured for group ethics in terms of achieving higher statistical power whereas AR is favoured for individual ethics such that patients can be treated better during the trial. In addition, there is a price to be paid for AR: as a result of imbalance of the sample size between the two groups, the average sample size of AR is larger than that of the GS design with ER. Although the BARPP method could lead to a larger trial, the treatment effect of the better arm can be estimated more precisely as a result of more patients being treated in the better arm. Treating more patients with more efficacious treatments can also lead to other tangential benefits (or harms) that are not captured by the response rate alone. With more patients treated in the more efficacious arm, more tissue specimens can be acquired to facilitate the analysis of biomarkers. However, AR designs also require additional infrastructure for implementing the trials.

The size of the cohort for evaluating the stopping rules can also be changed depending on how frequently the trial is monitored. The ability to choose the prior distribution is a unique strength of Bayesian methods. Additional information about the efficacy of treatment external to the trial, if available, can be naturally incorporated in the prior distribution. We chose a relatively non-informative prior to put more emphasis on the observed data for decision making. As is true in every design, the design parameters should be chosen to reflect the available information that is relevant to the trial. Apart from AR, our design has similar operating characteristics to those of the frequentist GS design. Our goal is not to 'beat' the frequentist design based on the frequentist operating characteristics, but to propose a comparable Bayesian solution to the problem. In the meantime, extensive simulation studies have been conducted to evaluate the operating characteristics such as the percentage of correct decisions, the maximum sample size, the proportion of patients who are randomized to the more effective treatments and the overall response rate, to ensure that desirable properties can be achieved. From a practical point of view, the response AR is more applicable to trials with short-term end points. Its applicability also depends on the relative time of the duration of accrual and the time required to measure the response. Sufficient learning from the observed patients is required for the success of response AR, regardless of the Bayesian or frequentist designs. Well-defined eligibility criteria should be implemented to ensure a comparable population of patients throughout the trial. If a drift in patients' characteristics occurs, it could lead to biased information on the treatment effect. However, a randomized study is still preferred to a non-randomized study. The bias should be prevented in the first place by enrolling patients with homogeneous characteristics. Covariate-adjusted analysis can be performed to attenuate the bias if it occurs. In addition, selection bias and reporting bias need to be examined carefully (Bauer *et al.*, 2010). It is also known that response AR can result in an overestimated treatment effect (Hu and Rosenberger, 2006). Up to 15% bias is observed in our randomization studies (the data are not shown). Hence, the observed effect size from an AR trial should be somewhat discounted when planning for future trials.

In summary, we proposed a Bayesian design as an extension to the frequentist design by coupling the Bayesian AR with PP approaches. We can attain the following advantages.

(a) After the initial ER phase, via AR more patients are preferentially allocated to the more effective treatment on the basis of the interim data.

(b) The trial is monitored frequently to examine the strength of the cumulative information for interim decision making:

    (i) if one treatment is superior to the other, stop the trial and declare superiority;

    (ii) if the two treatments have similar efficacy, stop the trial and declare equivalence;

(iii) otherwise, continue the trial until the maximum sample size has been reached.

Under the Bayesian framework, the inference is consistent with the likelihood principle. The decision making is based on the prior and the strength of the observed data. Because the inference is not constrained in a fixed study design, it is more flexible in terms of the frequency and time for the interim analysis. Valid inference still can be drawn even when the study condition deviates from what was originally planned. However, some disadvantages of the proposed design are noted as well.

(a) The design is calibrated to control both type I and type II errors, which requires extensive computation in the planning stage.
(b) In terms of the overall response rate, the gain of using AR is only moderate compared with ER, particulary when the early stopping rule is implemented.
(c) A relatively larger sample size is required to achieve the power desired, because the allocation of patients becomes unbalanced by using AR. As a result, the variance of the sample size is larger than that of the ER design.

Regardless of the frequentist or Bayesian, ER or AR, group sequential or PP approaches, the goal of designing efficient and ethical trials to draw accurate inference is the same. There is no single best design universally. Our paper expands some of the previous work (Grossman *et al.*, 1994; Rosner and Berry, 1995; Emerson *et al.*, 2007) and offers an appealing alternative for designing randomized phase II trials.

Acknowledgements

We thank the Associate Editor, two referees and the Joint Editor for many insightful suggestions which strengthened the work immensely. We also thank Valen Johnson, Gary Rosner and Diane Liu for helpful discussions. This research was supported in part by a grant from the Research Grants Council of Hong Kong, the US National Institutes of Health, grants CA16672 and CA97007, and M. D. Anderson University Cancer Foundation grant 80094548.

References

Bauer, P., Koenig, F., Brannatha, W. and Poscha, M. (2010) Selection and bias—two hostile brothers. *Statist. Med.*, **29**, 1–13.

Berry, D. A. (2011) Adaptive clinical trials: the promise and the caution. *J. Clin. Oncol.*, **29**, 606–609.

Berry, D. A. and Eick, S. G. (1995) Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Statist. Med.*, **14**, 231–246.

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

Cheng, Y. and Berry, D. A. (2007) Optimal adaptive randomized designs for clinical trials. *Biometrika*, **94**, 673–689.

DeMets, L. D. and Ware, H. J. (1982) Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika*, **69**, 661–663.

Emerson, S. S., Kittelson, J. M. and Gillen, D. L. (2007) Bayesian evaluation of group sequential clinical trial designs. *Statist. Med.*, **26**, 1431–1449.

Flehinger, B. J., Louis, T. A., Robbins, H. and Singer, B. (1972) Reducing the number of inferior treatments in clinical trials. *Proc. Natn. Acad. Sci. USA*, **69**, 2993–2994.

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

Gehan, E. A. (1961) The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *J. Chron. Dis.*, **13**, 346–353.

Grossman, J., Parmar, M. K. B., Spiegelhalter, D. J. and Freedman, L. S. (1994) A unified method for monitoring and analysing controlled trials. *Statist. Med.*, **13**, 1815–1826.

Hu, F. and Rosenberger, W. F. (2006) *The Theory of Response-adaptive Randomization in Clinical Trials*. Hoboken: Wiley.

Hwang, I., Shih, W. J. and DeCani, J. S. (1990) Group sequential designs using a family of type I error probability spending functions. *Statist. Med.*, **9**, 1439–1445.

Karrison, T., Huo, D. and Chappell, R. (2003) Group sequential, response-adaptive designs for randomized clinical trials. *Contr. Clin. Trials*, **24**, 506–522.

Korn, E. L. and Freidlin, B. (2011) Outcome-adaptive randomization: is it useful? *J. Clin. Oncol.*, **29**, 771–776.

Lee, J. J. and Feng, L. (2005) Randomized phase II designs in cancer clinical trials: current status and future directions. *J. Clin. Oncol.*, **23**, 4450–4457.

Lee, J. J., Gu, X. and Liu, S. (2010) Bayesian adaptive randomization designs for targeted agent development. *Clin. Trials*, **7**, 584–596.

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

Louis, T. A. (1975) Optimal allocation in sequential tests comparing the means of two Gaussian populations. *Biometrika*, **62**, 359–369.

Louis, T. A. (1977) Sequential allocation in clinical trials comparing two exponential survival curves. *Biometrics*, **33**, 627–634.

O'Brien, P. C. and Fleming, T. R. (1979) A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549–556.

Ratain, M. J. and Sargent, D. J. (2009) Optimising the design of phase II oncology trials: the importance of randomization. *Eur. J. Cancer*, **45**, 275–280.

Rosner, G. L. and Berry, D. A. (1995) A Bayesian group sequential design for a multiple arm randomized clinical trial. *Statist. Med.*, **14**, 381–394.

Simon, R. (1989) Optimal 2-stage designs for phase-II clinical trials. *Contr. Clin. Trials*, **10**, 1–10.

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

Thall, P. F. and Wathen, J. K. (2007) Practical Bayesian adaptive randomisation in clinical trials. *Eur. J. Cancer*, **43**, 859–866.

Zhang, L. and Rosenberger, W. F. (2007) Response-adaptive randomization for survival trials: the parametric approach. *Appl. Statist.*, **56**, 153–165.