Bayesian Analysis (2019) **14**, Number 2, pp. 399–425

Control of Type I Error Rates in Bayesian Sequential Designs

Haolun Shi[*] and Guosheng Yin[†]

Abstract. Bayesian approaches to phase II clinical trial designs are usually based on the posterior distribution of the parameter of interest and calibration of certain threshold for decision making. If the posterior probability is computed and assessed in a sequential manner, the design may involve the problem of multiplicity, which, however, is often a neglected aspect in Bayesian trial designs. To effectively maintain the overall type I error rate, we propose solutions to the problem of multiplicity for Bayesian sequential designs and, in particular, the determination of the cutoff boundaries for the posterior probabilities. We present both theoretical and numerical methods for finding the optimal posterior probability boundaries with α-spending functions that mimic those of the frequentist group sequential designs. The theoretical approach is based on the asymptotic properties of the posterior probability, which establishes a connection between the Bayesian trial design and the frequentist group sequential method. The numerical approach uses a sandwich-type searching algorithm, which immensely reduces the computational burden. We apply least-square fitting to find the α-spending function closest to the target. We discuss the application of our method to single-arm and double-arm cases with binary and normal endpoints, respectively, and provide a real trial example for each case.

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

Keywords: Bayesian design, group sequential method, multiple testing, phase II clinical trial, posterior probability, type I error rate.

1 Introduction

Along with the frequentist method, one of the popular paradigms in clinical trial designs is the Bayesian approach, where samples are treated as fixed and the parameter of interest is assigned a prior probability distribution to represent the uncertainty about its value; see, e.g., Berry (2006, 2011), Berger and Berry (1988), Efron (1986, 2005) and Yin (2012). The posterior distribution of the parameter is continuously updated with regard to the accrued samples. Bayesian approaches allow incorporating useful information into the prior distribution and are usually more efficient provided that the prior distribution is sensible. Inferences are made based on the posterior distribution of the parameter of interest, which can be updated as the trial accumulates more data. Along this direction, Thall and Simon (1994) proposed a Bayesian single-arm phase II clinical trial design that continually evaluates the posterior probability that the experimental drug is superior to

[*] Department of Statistics and Actuarial Science, The University of Hong Kong, 91 Pokfulam Road, Hong Kong, shl2003@connect.hku.hk

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

https://doi.org/10.1214/18-BA1109

the standard of care, where the response rate of the new treatment is compared with a fixed cutoff boundary at each interim analysis during the trial. Because the comparison is made multiple times during the study, the design involves the problem of multiple testing, and a failure to make proper adjustment for multiplicity is known to induce potential inflation in the type I error rate. As an illustration of multiple testing problems in the Bayesian setting, consider a random sample $\{y_1, \ldots, y_n\}$ from $N(\theta, 1)$, and we are interested in $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$. From a frequentist viewpoint, the test statistic at the kth interim analysis is $\bar{y}_k = \sum_{i=1}^{n_k} y_i/n_k$, where n_k is the cumulative sample size up to stage k. The decision rule for the O'Brien–Fleming type (O'Brien and Fleming, 1979) group sequential design is $\sqrt{n_k}\bar{y}_k > C_{\mathrm{OF}}(K, \alpha)\sqrt{K/k}$, $k = 1, \ldots, K$, where K is the total number of analyses planned for the trial and $C_{\mathrm{OF}}(K, \alpha)$ is the critical constant for the design. By contrast, assuming a flat prior distribution for θ, $f(\theta) \propto 1$, under a Bayesian approach, the posterior at interim analysis k is $\theta|\bar{y}_k \sim N(\bar{y}_k, 1/n_k)$. If we employ the decision rule that the posterior probability of H_1 should be greater than $1-\alpha$, our decision boundary would be $\sqrt{n_k}\bar{y}_k > \Phi^{-1}(1-\alpha)$, and thus there is no penalty for multiple testing in the Bayesian setting. To effectively control the overall type I error in Bayesian sequential designs, we study the problem of multiplicity, specifically, how the cutoff boundaries for the posterior probabilities should be determined.

The issue of multiplicity either involves testing multiple hypotheses simultaneously, or testing a single hypothesis repeatly over time. For the former, extensive research has been conducted under the Bayesian paradigm, e.g., see Gopalan and Berry (1998), Berry and Hochberg (1999), Scott and Berger (2006), Labbe and Thompson (2007), Guindani et al. (2009), and Guo and Heitjan (2010). For the latter, various Bayesian clinical trial designs involving sequential testing of a single hypothesis have been proposed, e.g., see Thall and Simon (1994), Lee and Liu (2008), Thall et al. (1995), Heitjan (1997), Rosner and Berry (1995), and Gsponer et al. (2014). However, a comprehensive and unified approach to controlling the overall type I error rate and accounting for the multiplicity is rarely discussed. With a focus on the binary endpoint, Zhu and Yu (2017) adopted a numerical search method for calibrating the operating characteristics of a Bayesian sequential design in terms of the α-spending function. Murray et al. (2016) developed a computational algorithm for calibrating the spending of the type I error rate for utility-based sequential trials with multinomial endpoints. Murray et al. (2017) adopted a simulation-based approach to calibrating the empirical α-spending function for a Bayesian design with two co-primary semicompeting time-to-event endpoints, which incorporates three interim analyses.

In the frequentist group sequential design, multiplicity is explicitly considered to control the overall type I error rate, e.g., see Pocock (1977), O'Brien and Fleming (1979), Wang and Tsiatis (1987), Eales and Jennison (1992), and Barber and Jennison (2002). Limited research has been conducted on the problem of multiplicity adjustment for Bayesian sequential designs. Bayesian multiple testing procedure should be ideally conducted in a fully decision-theoretic framework, where a loss function and related parametric assumptions are explicitly specified; e.g., see Lewis and Berry (1994), Christen et al. (2004), Müller et al. (2007), and Ventz and Trippa (2015). However, in clinical trials, regulatory bodies (e.g., the Food and Drug Administration) often require explicit evidence that the frequentist error rates are well maintained. As a result, it is a common practice to evaluate the frequentist properties of a Bayesian design based on simulations, which require simulating a large number of repetitions of the trial conduct and different trial designs may require different simulation setups. Our goal is to provide a more unified framework to directly control the type I error rate for Bayesian sequential designs. We propose both theoretical and numerical approaches to maintaining the overall type I error rate for Bayesian designs that involve multiple comparisons using posterior probabilities, such that the designs' operating characteristics mimic those of the commonly used group sequential methods. By carefully calibrating design parameters, Bayesian methods can effectively maintain the frequentist type I and type II error rates at the nominal levels. Although such a calibrated Bayesian design has similar operating characteristics to its frequentist counterpart, Bayesian designs bring more flexility to the trial conduct, e.g., adaptive randomization based on the posterior distribution, or prediction of trial success using posterior predictive distributions. Moreover, when historical data are available, Bayesian approaches allow the incorporation of historical information via informative priors, which would lead to savings in the sample size.

The rest of this article is organized as follows. In Section 2, we describe a motivating example where a Bayesian design may inflate the type I error rate if no adjustment is made to account for multiplicity. In Sections 3, we develop the Bayesian sequential designs using posterior probabilities and describe the methods for maintaining the frequentist error rates for single- and double-arm designs. Section 4 extends our methods to trials with normal endpoints, and Section 5 compares the operating characteristics of our methods with those of a Bayesian continuous monitoring scheme. Section 6 presents examples of design applications. Section 7 concludes the article with some remarks.

2 Motivating Example

Thall and Simon (1994) proposed a Bayesian single-arm design for phase II trials. The design continually evaluates the efficacy of the experimental treatment by monitoring the binary outcomes and makes adaptive decisions throughout the trial. Let p_E denote the response rate of the experimental drug, and let p_S denote that of the standard drug. We are interested in testing the hypotheses

$$H_0\colon p_E \leq p_S = p_{\text{null}} \quad \text{versus} \quad H_1\colon p_E > p_S = p_{\text{null}}.$$

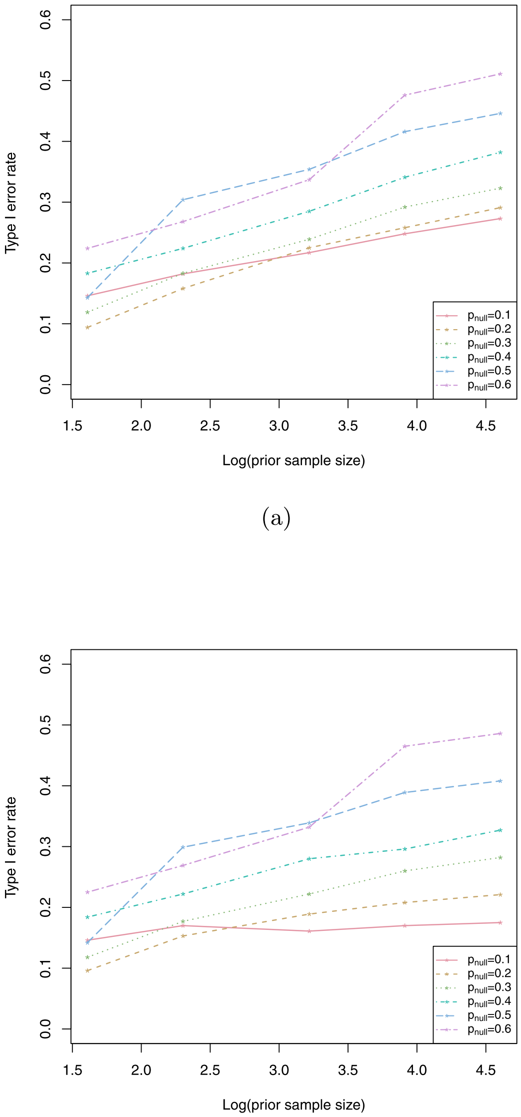
We assume beta prior distributions for these two response rates, $p_E \sim \text{Beta}(a_E, b_E)$ and $p_S \sim \text{Beta}(a_S, b_S)$, where the prior mean of p_S is set to equal to p_{null}. Typically, historical information for the standard treatment is often available and we may set p_{null} to be the estimate from the historical data. We inflate the prior variance of p_S to account for the uncertainty due to between-trial effects, i.e., the differences between the historical trials and the current trial. The beta prior for p_E is usually much more diffuse, with a large variance reflecting the fact that little information is known about the experimental drug. For example, we may assume a vague prior distribution for the experimental drug, $p_E \sim \text{Beta}(0.2, 0.8)$, which is often considered to be equivalent to the information of only one subject. For the standard drug, suppose that we have observed 200 responses among 1000 subjects in historical studies, we may set p_{null} to be the historical sample proportion while discounting the historical information by enlarging the prior variance; for example, we may assume $p_S \sim \text{Beta}(20, 80)$, which contains the amount of information equivalent to 100 patients (Morita et al., 2008).

Suppose the trial has accrued n subjects to receive the experimental treatment, and we observe y responses among them. Let D denote the observed data (n, y). Due to the conjugate nature of the beta prior distribution when combined with a binomial likelihood function, the posterior distribution of p_E is still beta, $p_E|D \sim \text{Beta}(a_E + y, b_E + n - y)$. Let $f(p; a, b)$ and $F(p; a, b)$ denote the probability density function and the cumulative distribution function of a $\text{Beta}(a, b)$ distribution, respectively. We compute the posterior probability that the experimental response rate is larger than the standard rate in the form of

$$\Pr(p_E > p_S|D) = \int_0^1 \{1 - F(p; a_E + y, b_E + n - y)\} f(p; a_S, b_S) dp. \qquad (1)$$

Let θ_U and θ_L denote the upper and lower boundaries for the posterior probability. At each step of the trial, we compute the value of $\Pr(p_E > p_S|D)$ and claim the experimental drug promising if it is larger than θ_U, unpromising if it is smaller than θ_L, or proceed to enroll the next subject if it lies between the two. At the end of the trial when the prespecified maximum sample size $N_{\max}$ is exhausted, if $\Pr(p_E > p_S|D) > \theta_U$, the drug is concluded to be promising, otherwise unpromising.

The design by Thall and Simon (1994) in fact suffers from the issue of multiplicity, as the same hypothesis $p_E > p_S$ is tested repeatedly and the drug can be declared as promising over any interim result that produces a posterior probability larger than θ_U. We define the type I error rate as the probability of rejecting the null hypothesis when $p_E = p_{\text{null}}$, and assess the degree of its inflation under $N_{\max} = 60$ and $\theta_U = 0.9$, and assess the type I error rates under different prior sample sizes of p_S. The prior distribution of p_E has a mean equal to that of p_S and a prior sample size of 1. We simulate one million trials by generating random samples from the $\text{Bernoulli}(p_E)$ distribution, and calculate the empirical type I error rate as the proportion of times the trial results lead to positive conclusions. Figure 1 shows the type I error rates under different prior sample sizes of p_S, for Bayesian sequential designs without futility stopping and with futility stopping θ_L, respectively. A more informative prior of p_S would induce a larger type I error rate. The intuition behind such a pattern is that with a larger prior sample size, the prior distribution of p_S is more centered at p_{null}, which makes it easier for the trial results to reach a high posterior probability of $p_E > p_S$. Moreover, a larger value of p_{null} appears to be associated with a higher type I error rate. A possible reason for this phenomenon could be that for smaller values of p_{null}, the posterior probability decision boundaries tend to be more conservative and more difficult to reach at $p_E = p_{\text{null}}$ when the cumulative sample size is relatively low. As an example, consider the case when the cumulative sample size is 10, and under a prior sample size of 1000, the probability of reaching the posterior probability boundary under $p_{\text{null}} = 0.1$ is 0.07, whereas that under $p_{\text{null}} = 0.6$ is as high as 0.17. The type I error rate inflation is slightly ameliorated with a futility stopping scheme by setting $\theta_L = 0.1$. In all cases, the type I error rate exceeds the nominal level of $1 - \theta_U = 0.1$, and in some extreme cases the type I error rate can be inflated up to 0.5. Therefore, it is recommended that



(a)

(b)

Figure 1: Type I error rates under different prior sample sizes for p_S (a) without futility stopping and (b) with futility stopping $\theta_L = 0.1$ under the Bayesian single-arm continuous monitoring design by Thall and Simon (1994).

for the Bayesian sequential design where the same hypothesis is tested multiple times, the decision boundaries should be carefully adjusted and calibrated to prevent inflation of the overall type I error rate, particularly when the design involves a strong degree of information borrowing from historical trials, i.e., an informative prior distribution or a large prior sample size. Similar findings on the type I error rate inflation are also noted in Jennison and Turnbull (2000), Chapter 18.

3 Bayesian Sequential Design with Binary Endpoints

3.1 Single-Arm Design

We propose a Bayesian sequential design based on posterior probabilties for single-arm phase II trials with binary outcomes. Our goal is to not only maintain the overall type I error rate, but also calibrate the decision boundaries such that the design's operating characteristics mimic those of the commonly used group sequential designs. Let K be the total number of analyses to be conducted throughout the trial and let m be the number of samples in each group, i.e., we conduct one analysis every time m additional subjects are enrolled. Let c_k be the efficacy decision boundary at stage k, $k = 1, \ldots, K$. At the kth interim stage, the posterior probability of the experimental response rate being larger than the standard rate is

$$P(H_1|D_k) = P(p_E > p_S|D_k),$$

where D_k is the cumulative data up to stage k. If $P(H_1|D_k) > c_k$, we stop the trial and declare treatment efficacy, otherwise we enroll the next group of m patients and conduct another analysis at stage $k+1$, or if $k = K$, i.e., we reach the end of the trial, we declare treatment futility. We define the type I error rate to be the probability of declaring treatment efficacy when $p_E = p_{\text{null}}$. The amount of the type I error rate spent at stage k, denoted as α_k, is defined to be the probability of reaching the efficacy boundary of stage k. Let $b(y; n, p)$ denote the binomial probability mass function and let $I(\cdot)$ denote the indicator function, and then

$$\alpha_k = \sum_{y_1=0}^{m} \sum_{y_2=0}^{m} \cdots \sum_{y_k=0}^{m} \left\{ I(P(H_1|D_k) > c_k) \prod_{j=1}^{k-1} I(P(H_1|D_j) \leq c_j) \prod_{i=1}^{k} b(y_i; m, p_{\text{null}}) \right\}. \tag{2}$$

The overall type I error rate is thus $\sum_{k=1}^{K} \alpha_k$, which can be maintained at the nominal level α if we set $\sum_{k=1}^{K} \alpha_k \leq \alpha$.

3.2 Posterior Probability Boundaries: Numerical Method

Our goal is to search for the optimal set $\{c_k : k = 1, \ldots, K\}$ that yields the closest fit to a prespecified target α-spending function while controlling the overall type I error rate. It is evident that the search space is of high dimension and a full enumeration method would be computationally intensive. To overcome this issue, we shrink the search space to the region where the optimal solution most probably lies so that the numerical approach to the problem becomes feasible. We first establish that $P(H_1|D_k)$

is an increasing function of y_k, the cumulative number of responses at stage k, because in the integration of (1), it is true that $F(p; a_E{+}y{+}1, b_E{+}n{-}y{-}1) < F(p; a_E{+}y, b_E{+}n{-}y)$ by the two lemmas in Thall and Simon (1994). Based on such a monotonic relationship, we can avoid the computationally intensive integration when calculating $P(H_1|D_k)$ and directly calibrate u_k, which is the boundary for the cumulative number of responses at stage k, i.e., if $y_k > u_k$, we declare the drug promising. In other words, we translate the information in the probability domain c_k to the number of responses u_k. As the spending of the type I error rate α_k is a function of all the upper boundaries up to stage k, we denote it as a function $\alpha_k(\mathbf{U}_k)$, where $\mathbf{U}_k = (u_1, \ldots, u_k)^T$ is a vector of design parameters. Let $\alpha(k)$ denote the target amount of type I error rate to be spent at stage k and α the overall type I error rate. More specifically, we propose the sandwich-type searching algorithm which can immensely reduce the computational burden, and the detail is described as follows.

1. At step $k = 1$, we iterate j from 0 to m.

   (i) For each j, we compute the amount of type I error rate spent at stage 1 when $u_1 = j$, denoted as $\alpha_1(\mathbf{U}_{1j})$, where $\mathbf{U}_{1j}$ is a scalar equal to j, corresponding to an upper boundary value of j at the end of the first stage.

   (ii) We find the pair $(u^*_{1j}, u^*_{1j} + 1)$, such that $\alpha_1(\mathbf{U}^*_{1j}) < \alpha(1) < \alpha_1(\mathbf{U}^\dagger_{1j})$, where $\mathbf{U}^*_{1j}$ and $\mathbf{U}^\dagger_{1j}$ are two scalars equal to u^*_{1j} and $u^*_{1j} + 1$, respectively. Let $\mathcal{A}_1$ be the set consisting of $\mathbf{U}^*_{1j}$ and $\mathbf{U}^\dagger_{1j}$.

2. At step k, $1 < k < K$, we iterate through each vector in $\mathcal{A}_{k-1}$.

   (i) Let $\mathbf{U}_{k-1}$ denote the design vector consisting of the values of upper boundaries up to the $(k-1)$th interim analysis, and denote its last element as $n_{\min}$. Fixing $\mathbf{U}_{k-1}$, we iterate j from $n_{\min}$ to km.

   (ii) We find the pair $(u^*_{kj}, u^*_{kj} + 1)$, such that $\alpha_k(\mathbf{U}^*_{kj}) < \alpha(k) < \alpha_k(\mathbf{U}^\dagger_{kj})$, where $\mathbf{U}^*_{kj}$ and $\mathbf{U}^\dagger_{kj}$ can be obtained by appending u^*_{kj} and $u^*_{kj} + 1$ to the current design vector $\mathbf{U}_{k-1}$, respectively. The vectors $\mathbf{U}^*_{kj}$ and $\mathbf{U}^\dagger_{kj}$ represent the two sets of upper boundaries up to stage k whose amounts of cumulative error rate spending are closest to the target; the error rates under the design vector $\mathbf{U}^*_{kj}$ are under-spent while those under $\mathbf{U}^\dagger_{kj}$ are over-spent. Let $\mathcal{A}_k$ be the set consisting of vectors $\mathbf{U}^*_{kj}$ and $\mathbf{U}^\dagger_{kj}$.

3. At step $k = K$, we iterate through each vector in $\mathcal{A}_{K-1}$.

   (i) For each vector in $\mathcal{A}_{K-1}$, denoted as $\mathbf{U}_{K-1}$, we calculate the type I error rate spent up to the $(K-1)$th stage as $\alpha_* = \sum_{k=1}^{K-1} \alpha_k(\mathbf{U}_k)$, where $\mathbf{U}_k$ contains the first k elements of the vector $\mathbf{U}_{K-1}$.

   (ii) We find the decision boundary u_K such that $\alpha_k(\mathbf{U}_K) < \alpha - \alpha_*$, where $\mathbf{U}_K$ is obtained by appending u_K to $\mathbf{U}_{K-1}$.

4. Among all the obtained vectors $\mathbf{U}_K$ in the last step, we choose the one that yields the smallest L_2-distance to the target α-spending function, i.e., minimizing

$$\sum_{k=1}^{K}\{\alpha_k(\mathbf{U}_k) - \alpha(k)\}^2,$$

where $\mathbf{U}_k$ consists of the first k elements in $\mathbf{U}_K$. Based on the increasing relationship between $P(H_1|D_k)$ and y_k, we can then find the corresponding c_k such that $y_k > u_k$ is equivalent to $P(H_1|D_k) > c_k$.

Steps 1 to 3 identify the sets of upper boundaries under which the amounts of cumulative type I error rate spending are closest to the target, and step 4 selects the best set of boundaries with the smallest L_2-distance to the target function. In the first step, only one pair of design vectors are identified. In each subsequent step k, further pairs are identified and appended to the design vector in the set $\mathcal{A}_{k-1}$ from the previous step. The total number of design vectors assessed at step k is 2^k. We can also minimize the maximum difference between $\alpha_k(\mathbf{U}_k)$ and $\alpha(k)$ in the last step, which would give similar results.

As the proposed numerical algorithm seeks to minimize the squared distance between the empirical spending function and the target, it is robust, accurate and flexible, which can accommodate any types of α-spending functions in the group sequential methods, including the commonly used Pocock, O'Brien–Fleming, and Wang–Tsiatis types (Jennison and Turnbull, 2000). The specification of the target function depends on the preferences on how the spendings of the type I error rate should be distributed over the interim analyses.

3.3 Posterior Probability Boundaries: Theoretical Method

There exists an asymptotic connection between the Bayesian approach based on posterior probabilities and the frequentist method using p-values. Dudley and Haughton (2002) studied the asymptotic normality of the posterior probability of half-spaces. In particular, let Θ be an open subset of a Euclidean space $\mathbf{R}^d$. A half-space $\mathcal{H}$ is a set satisfying a linear inequality,

$$\mathcal{H} = \{\boldsymbol{\theta} : \mathbf{a}^T\boldsymbol{\theta} \geq b\},$$

where $\boldsymbol{\theta} \in \Theta$, $\mathbf{a} \in \mathbf{R}^d$ and b is a scalar, and let $\partial\mathcal{H}$ represent the boundary hyperplane of $\mathcal{H}$,

$$\partial\mathcal{H} = \{\boldsymbol{\theta} : \mathbf{a}^T\boldsymbol{\theta} = b\}.$$

Examples of half-spaces under the context of clinical trials are $\{p_E : p_E > p_{\text{null}}\}$ for single-arm trials, or $\{(p_E, p_S) : p_E > p_S\}$ for double-arm trials.

Let y_i denote the observed data whose probability density function is $f(y_i, \boldsymbol{\theta})$, for $i = 1, \ldots, n$. The likelihood ratio statistic for testing the null hypothesis $H_0 : \boldsymbol{\theta} \in \partial\mathcal{H}$ is

$$\Delta_m = 2\log L(\hat{\boldsymbol{\theta}}) - 2\log L(\tilde{\boldsymbol{\theta}}),$$

where $\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta})$, and $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ are the maximum likelihood estimates for $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\theta} \in \partial\mathcal{H}$, respectively. Let S_n denote the signed root likelihood ratio statistics, i.e., if $\hat{\boldsymbol{\theta}} \notin \mathcal{H}$, $S_n = -\sqrt{\Delta_n}$; otherwise, $S_n = \sqrt{\Delta_n}$. Let $\pi_n(\mathcal{H})$ denote the posterior probability of the half space given the data with sample size n.

Theorem 1. *Under the regularity conditions in Dudley and Haughton (2002), we have*

(i) If $\mathcal{H}_n$ is a sequence of the same half-space, indexed by the cumulative sample sizes, then as $n \to \infty$, $\pi_n(\mathcal{H}_n)/\Phi(S_n) \to 1$ almost surely, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable.

(ii) For cumulative sample sizes $n_1, \ldots, n_K$, the joint statistics $\{\Phi^{-1}(\pi_{n_1}(\mathcal{H}_{n_1})), \ldots, \Phi^{-1}(\pi_{n_K}(\mathcal{H}_{n_K}))\}$ converges in distribution to $(S_{n_1}, \ldots, S_{n_K})$, which follows a multivariate normal distribution asymptotically.

The proof of the first part of the theorem can be found in Dudley and Haughton (2002), and the second part follows from the continuous mapping theorem and the argument on the joint canonical distribution of $(S_{n_1}, \ldots, S_{n_K})$ in Jennison and Turnbull (1997), Scharfstein et al. (1997), and Jennison and Turnbull (2000), Chapter 11.2.

Based on the theoretical results, we propose a method to find the set of c_k by connecting the Bayesian and the frequentist group sequential designs. Specifically, let $\{z_k; k = 1, \ldots, K\}$ denote a series of critical constants obtained from the frequentist group sequential method and, without loss of generality, we assume that all the z_k's are positive. We set c_k to be equal to $\Phi(z_k)$, because the decision rules using the posterior probabilities of $\mathcal{H}_1, \ldots, \mathcal{H}_k$ are asymptotically equivalent to $S_1, \ldots, S_k$ being greater than $z_1, \ldots, z_k$, respectively, which leads to the correct type I error rate spending of $\alpha_1, \ldots, \alpha_k$ based on the canonical distribution in the group sequential design.

3.4 Double-Arm Design

Design Specification

In addition to the single-arm Bayesian sequential design, we also study the properties of a double-arm Bayesian sequential design that uses the posterior probability at the interim analyses. Consider a double-arm clinical trial with dichotomous outcomes, let p_E denote the response rate of the experimental drug, and let p_S denote that of the standard drug. Let K denote the total number of analyses and let m denote the sample size per arm in each group. If we consider the one-sided hypothesis test, we are interested in examining whether the experimental drug is superior to the standard of care,

$$H_0\colon p_E \le p_S \quad \text{versus} \quad H_1\colon p_E > p_S.$$

Under the Bayesian framework, we assume beta prior distributions for p_E and p_S, i.e., $p_E \sim \text{Beta}(a_E, b_E)$ and $p_S \sim \text{Beta}(a_S, b_S)$. At the kth interim analysis, the cumulative number of patients accrued in each arm is km. If the numbers of responses in the experimental and standard arms are y_E and y_S respectively, the binomial likelihood functions can be formulated as

$$P(y_g|p_g) = \binom{km}{y_g} p_g^{y_g}(1-p_g)^{km-y_g}, \quad g = E, S.$$

The posterior distributions of p_E and p_S are given by

$$\begin{aligned} p_E|y_E &\sim \text{Beta}(a_E + y_E, b_E + km - y_E), \\ p_S|y_S &\sim \text{Beta}(a_S + y_S, b_S + km - y_S), \end{aligned}$$

whose density functions are denoted by $f(p_E|y_E)$ and $f(p_S|y_S)$, respectively. Let c_k be a prespecified cutoff probability boundary at stage k. Based on the posterior probability, we can construct a Bayesian sequential testing procedure, so that the experimental treatment is declared as promising if

$$\Pr(p_E > p_S|y_E, y_S) \geq c_k,$$

where

$$\Pr(p_E > p_S|y_E, y_S) = \int_0^1 \int_{p_S}^1 f(p_E|y_E) f(p_S|y_S) dp_E dp_S.$$

Otherwise, we fail to declare treatment efficacy.

To control the overall type I error rate, we may adopt the theoretical method that connects the Bayesian sequential design with the frequentist group sequential method by setting $c_k = \Phi(z_k)$, where $\{z_k; k = 1, \ldots, K\}$ is a series of critical constants obtained from the frequentist group sequential method.

Extension to Biomarker Design

Wason et al. (2015) proposed a Bayesian adaptive design for analyzing the relationships between biomarkers and the experimental treatment effects. Consider a biomarker trial where there are L biomarkers and a total of J experimental treatment arms and one control arm for the standard drug. When a patient is enrolled into the study, a test is conducted to obtain his/her biomarker profile. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iL})$ denote the biomarker profile of the ith patient, where $X_{il} = 1$ if the expression of biomarker l is positive for the ith patient, and $X_{il} = 0$ otherwise. Patients are equally randomized to all treatment arms, and we denote $\mathbf{T}_i = (T_{i1}, \ldots, T_{iJ})$ as the treatment assignment vector, where $T_{ij} = 1$ if the ith patient is allocated to the jth experimental treatment, and when all entries of $\mathbf{T}_i$ are zero, the patient is assigned to the control arm. Let Y_i denote the binary endpoint for the ith patient and let p_i denote the probability of treatment success. The design utilizes a Bayesian logistic regression model to characterize the treatment effect of the drug, the biomarker and their interaction, which is represented as

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(p_i), \\ \log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \sum_{j=1}^{J} \beta_j T_{ij} + \sum_{l=1}^{L} \gamma_l X_{il} + \sum_{l=1}^{L} \sum_{j=1}^{J} \delta_{jl} X_{il} T_{ij}, \end{aligned}$$

where β_0 is the intercept, β_j represents the effect of the jth experimental treatment, γ_l represents the effect of the lth biomarker, and δ_{jl} represents the effect of the treatment and biomarker interaction. Noninformative normal prior distributions are specified for all regression coefficients.

A total of $J(L+1)$ hypotheses are tested at the interim or the final analysis. In particular, the set of alternative hypotheses are $\{H_1^{(j,l)}|j=1,\ldots,J;l=0,\ldots,L\}$. When $l>0$, the hypothesis $H_1^{(j,l)}$ represents the case where the jth experimental treatment is superior to the standard treatment in patients with positive biomarker l.

$$H_1^{(j,l)}:\beta_j+\delta_{jl}>0,\qquad l>0.$$

When $l=0$, the hypothesis $H_1^{(j,l)}$ represents the case where the experimental treatment is superior for patients who have no positive biomarker profiles,

$$H_1^{(j,0)}:\beta_j>0.$$

At the kth interim or final analysis, where $k=1,\ldots,K$, we compute the posterior probabilities $\Pr(H_1^{(j,l)}|D_k)$, and if it is larger than c_k, the superiority of the experimental treatment can be declared for the corresponding subgroup of patients. The theoretical method for controlling the overall type I error rate can be adopted for such a Bayesian sequential design, i.e., we may set $c_k=\Phi(z_k)$.

Numerical Evaluation

We apply the theoretical posterior probability boundaries for controlling the type I error rate in a double-arm sequential biomarker design. We assume that there are $J=2$ experimental treatments and $L=2$ biomarkers, and the trial involves $K=4$ analyses with sample size 480. Patients are equally randomized to the two treatment arms and the control arm. Wason et al. (2015) recommended controlling the familywise error rate (FWER) in a range of 0.4 to 0.5. As a total of 6 hypotheses are to be tested, we adopt Bonferroni's method for controlling the FWER at 0.48, i.e., we set the significance level for each hypothesis test to be $0.48/6=0.08$.

To examine the effectiveness of controlling the type I error rate, we consider a null case where β_j and δ_{jl} are all zero for $j=1,2$ and $l=1,2$, and $\beta_0=\gamma_1=\gamma_2=0.1$. Based on 1000 trial replications, we compute the empirical type I error rates spent at the interim analyses using the theoretical posterior probability boundaries with the Pocock type and O'Brien–Fleming type α-spending functions, which are exhibited in Table 1. Due to symmetry, we only need to show results for the two alternative hypotheses $H_1^{(1,1)}:\beta_1+\delta_{11}>0$ and $H_1^{(1,0)}:\beta_1>0$. Because the endpoint is binary, slight deviation between the total empirical type I error rate and the target level 0.08 is observed. Nevertheless, the theoretical method controls the type I error rate spending in accordance with the specified α-spending function.

α-spending function	α_1	α_2	α_3	α_4	$\sum_{k=1}^{K} \alpha_k$
$H_1^{(1,0)} : \beta_1 > 0$					
Pocock	0.040	0.022	0.021	0.009	0.092
O'Brien–Fleming	0.004	0.015	0.035	0.034	0.088
$H_1^{(1,1)} : \beta_1 + \delta_{11} > 0$					
Pocock	0.035	0.023	0.017	0.016	0.090
O'Brien–Fleming	0.002	0.013	0.039	0.032	0.086

Table 1: Type I error rate spendings in Bayesian sequential biomarker designs with binary endpoints using the theoretical posterior probability boundaries with the Pocock type and O'Brien–Fleming type α–spending functions.

4 Bayesian Sequential Design with Normal Endpoints

4.1 Design Specification

Consider a single-arm trial with a continuous endpoint from the normal distribution $N(\mu, \sigma^2)$. Let y_i denote the observed outcome for the ith subject in the experimental arm and let n denote the number of observations. We assign a prior distribution $N(\mu_0, \sigma_0^2)$ to the mean μ, and for simplicity we assume the variance σ^2 to be known. The likelihood can be expressed as $\prod_{i=1}^{n} \phi(y_i; \mu, \sigma^2)$, where $\phi(\cdot; \mu, \sigma^2)$ denotes the normal density function with mean μ and variance σ^2. Based on the conjugacy of a normal prior distribution under a normal likelihood, the posterior distribution of μ follows $N(\mu_*, \sigma_*^2)$, where

$$\mu_* = \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2} \right) \sigma_*^2,$$
$$\sigma_*^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1}.$$

We formulate the null and alternative hypotheses as

$$H_0 : \mu \leq \delta \quad \text{versus} \quad H_1 : \mu > \delta,$$

where δ is the minimum value of μ that warrants further investigation. We reject the null hypothesis if $\Pr(\mu > \delta | D) > c$, $D = \{y_1, \ldots, y_n\}$, which is equivalent to $(\mu_* - \delta)/\sigma_* > \Phi^{-1}(c)$, and it can be further expressed as

$$\bar{y} > Q(c; \mu_0, \sigma_0, \sigma) = \frac{\sigma^2}{n} \left\{ \frac{\Phi^{-1}(c)}{\sigma_*} + \frac{\delta}{\sigma_*^2} - \frac{\mu_0}{\sigma_0^2} \right\},$$

where $\bar{y} = \sum_{i=1}^{n} y_i / n$ and $Q(\cdot)$ is a nondecreasing function of c.

To control the overall type I error rate for a series of sequential tests, we can equate $Q(c; \mu_0, \sigma_0, \sigma)$ to the corresponding critical constant in the group sequential design.

Under the group sequential methodology, we reject the null at an interim analysis if $\sqrt{n}(\bar{y}-\delta)/\sigma > z$, or $\bar{y} > z\sigma/\sqrt{n}+\delta$, where z is a known critical constant from the interim boundaries in the group sequential test with the same specification of the overall type I error rate, power and α-spending function. The value of c can be solved by setting

$$Q(c;\mu_0,\sigma_0,\sigma) = z\sigma/\sqrt{n}+\delta.$$

In the case with a two-arm trial, we are interested in comparing the means of the endpoints between the experimental and control arms, denoted by μ_E and μ_S respectively. Under a normal likelihood function with normal prior distributions on the means, the posterior distributions of μ_E and μ_S are both normal. The posterior distribution of $\mu_E - \mu_S$ is also normal and, as a result, the decision boundary c can be derived along similar lines.

4.2 Commensurate Prior

One of the advantages of Bayesian trial designs is the ability to incorporate useful historical information in the prior distribution, which, if adopted correctly, leads to higher power and saving in sample size. Hobbs et al. (2011) proposed several classes of commensurate prior distributions for normal endpoints. Commensurate priors can adaptively adjust the amount of information borrowing from the historical data according to the degree of commensurability between the data in the historical trials and the current one.

We consider a class of commensurate prior distributions proposed by Hobbs et al. (2011) called the location commensurate prior. Let μ_S and μ_H be the mean parameter for the current and historical data respectively, and let D_H denote the historical data. The location commensurate prior is a hierarchical construct where we first specify a prior distribution $p(\tau)$ for the commensurability parameter $\tau > 0$, which serves as the primary mechanism for adjusting the influence of prior information relative to its commensurability with the data in the current trial. Conditional on the commensurability parameter τ, we center the prior of μ_S at the historical mean μ_H, i.e., a normal distribution with mean μ_H and precision τ (i.e., variance $1/\tau$), and multiply it with the historical likelihood, which results in a prior of the form,

$$p(\mu_S|D_H,\mu_H,\tau) \propto L(\mu_H|D_H)p(\mu_S|\mu_H,\tau)p_0(\mu_S).$$

As $\tau \to 0$, $p(\mu_S|D_H,\mu_H,\tau)$ approaches $p_0(\mu_S)$, such that the historical data are completely ignored due to noncommensurability; and as $\tau \to \infty$, $p(\mu_S|D_H,\mu_H,\tau)$ approaches $L(\mu_S|D_H)p_0(\mu_S)$, leading to full exchangeability between the historical and current data, and thus the current and historical data are equally weighed and can be simply merged.

Assume that the historical data follow a normal distribution, $N(\mu_H,\sigma_H^2)$, with sample size n_H, and the current data in the standard arm follow $N(\mu_S,\sigma_S^2)$. Let $\bar{y}_H$ denote the historical sample mean, and let $\hat{\sigma}_H^2$ denote the maximum likelihood estimator of σ_H^2. We specify $p(\tau)$ to be a Gamma$(\nu\tilde{\tau},\nu)$ distribution with mean $\tilde{\tau}$ and variance $\tilde{\tau}/\nu$,

$p(\mu_S|\mu_H, \tau)$ a normal distribution with mean μ_H and precision τ, and $p_0(\mu_S) \propto 1$. The location commensurate prior for μ_S under such hierarchical models can be derived as

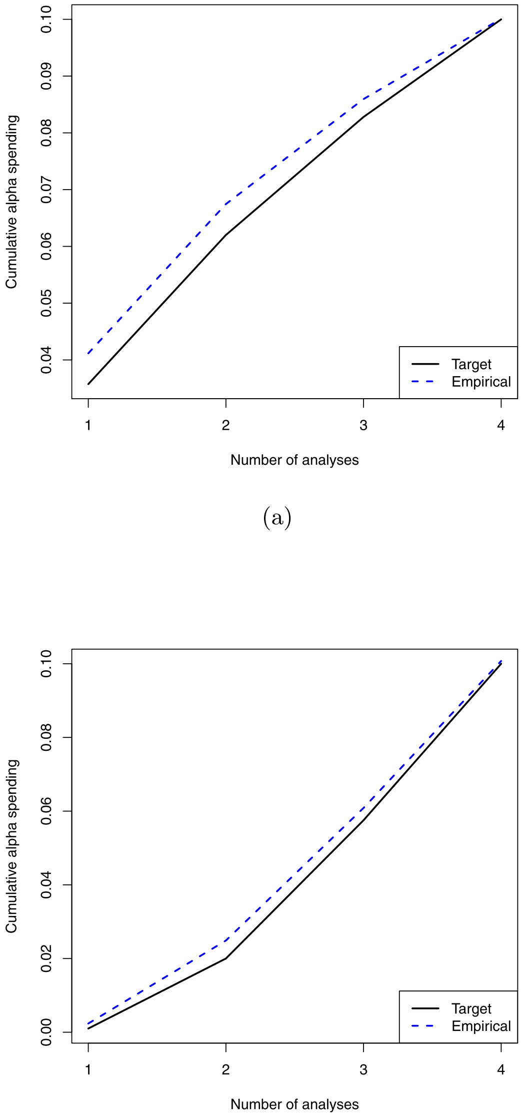
$$p(\mu_S, \sigma_S^2, \tau|D_H) \propto \phi\left(\mu_S\middle|\bar{y}_H, \frac{1}{\tau} + \frac{\hat{\sigma}_H^2}{n_H}\right) \times \frac{1}{\sigma_S^2} \times p(\tau).$$

We apply the theoretical approach to determining posterior probability boundaries in a Bayesian double-arm sequential trial that utilizes the commensurate prior in the standard arm. For the experimental arm, we assume a vague prior $N(\mu_0, \sigma_0)$ for μ_E. At the kth interim analysis, if the posterior probability $\Pr(\mu_E > \mu_S|D, D_H) > c_k$, where D denotes the data in the current trial, we terminate the trial and declare treatment superiority; otherwise, we continue to recruit the next group of patients.

4.3 Numerical Evaluation

We conduct two simulation studies to study the trial performance with informative priors: one for a single-arm design and the other for a double-arm design. In the single-arm study, we compute the empirical type I error rate based on normal endpoints with mean μ and variance 1. We are interested in testing whether μ is greater than $\delta = 0$. The sample size is 200 and a total of $K = 4$ analyses are considered. Our desired type I error rate is $\alpha = 0.1$. We set the prior distribution of μ to be $N(0, 100)$. We simulate one million trials by generating random samples from the standard normal distribution $N(0, 1)$ and the proportion of times the null hypothesis is rejected is defined as the empirical type I error rate. Figure 2 shows the target α-spending functions versus the empirical ones for the Pocock type and O'Brien–Fleming type boundaries, respectively. Clearly, the proposed method maintains the type I error rate under the nominal level and the empirical pattern of the type I error rate spent at each stage is close to that of the target α-spending function.

For the double-arm trial, we apply the theoretical method for calculating the posterior probability boundaries by setting $c_k = \Phi(z_k)$ where z_k's are the critical constants from the frequentist group sequential designs. The commensurate prior is adopted for the standard arm to facilitate information borrowing from the historical data with sample size $n_H = 200$. As the variances in the current and historical trials are not of direct interest, for simplicity we assume that variances are known to be 1 in both trials. Four interim analyses are involved and the sample size in each arm is 200. As we are interested in the influence of the commensurability parameter τ and the historical mean parameter μ_H on the current trial's operating characteristics, we consider various values of μ_H and commensurate prior distributions for τ. In particular, we consider cases where $\mu_H = \mu_S$ and $\mu_H = \mu_S \pm 0.05$; and we specify $p(\tau)$ to be a Gamma$(\nu\tilde{\tau}, \nu)$ distribution and consider the cases with $\tilde{\tau} = \nu = 1$ and $\tilde{\tau} = \nu = 1000$, i.e., respective prior means of 1 and 1000 and prior variances of 1, which correspond to weak and strong degrees of information borrowing. Based on 1000 trial replications, we compute the empirical type I error rates spent under the null where $\mu_S = \mu_E = 0.5$, and the power under the alternative where $\mu_S = 0.5$ and $\mu_E = 0.8$.



(a)

(b)

Figure 2: The target α-spending functions versus the empirical counterparts of (a) the Pocock type and (b) the O'Brien–Fleming type boundaries for a single-arm study with normal endpoints.

Table 2 shows the type I error rate spending and power under various values of historical means, commensurate priors and α–spending functions. Compared with those with a weak degree of information borrowing ($\tilde{\tau} = \nu = 1$), the cases with highly infor-

μ_H	α-spending function	α_1	α_2	α_3	α_4	$\sum_{k=1}^{K} \alpha_k$	Power
$\tilde{\tau}$=1, ν=1, $c_k = \Phi(z_k)$							
0.45	Pocock	0.039	0.032	0.031	0.025	0.127	0.765
	O'Brien–Fleming	0.001	0.024	0.043	0.057	0.125	0.750
0.5	Pocock	0.035	0.035	0.022	0.028	0.120	0.767
	O'Brien–Fleming	0.000	0.026	0.041	0.064	0.131	0.758
0.55	Pocock	0.043	0.026	0.039	0.018	0.126	0.752
	O'Brien–Fleming	0.003	0.022	0.060	0.050	0.135	0.745
$\tilde{\tau}$=1, ν=1, $c_k = \Phi(z_k) + \zeta_k$							
0.45	Pocock	0.033	0.025	0.028	0.020	0.106	0.733
	O'Brien–Fleming	0.001	0.018	0.035	0.041	0.095	0.699
0.5	Pocock	0.027	0.033	0.020	0.020	0.100	0.767
	O'Brien–Fleming	0.000	0.020	0.028	0.052	0.100	0.758
0.55	Pocock	0.037	0.025	0.032	0.014	0.108	0.752
	O'Brien–Fleming	0.000	0.018	0.042	0.031	0.091	0.745
$\tilde{\tau}$=1000, ν=1000, $c_k = \Phi(z_k)$							
0.45	Pocock	0.056	0.045	0.036	0.033	0.170	0.888
	O'Brien–Fleming	0.004	0.030	0.064	0.076	0.174	0.878
0.5	Pocock	0.028	0.022	0.022	0.024	0.096	0.833
	O'Brien–Fleming	0.000	0.011	0.032	0.055	0.098	0.809
0.55	Pocock	0.016	0.012	0.010	0.005	0.043	0.760
	O'Brien–Fleming	0.000	0.008	0.019	0.022	0.049	0.762

Table 2: Type I error rate spending and power in Bayesian sequential designs with normal endpoints using the commensurate priors and the theoretical posterior probability boundaries with the Pocock type and O'Brien–Fleming type α-spending functions.

mative priors ($\tilde{\tau} = \nu = 1000$) have higher power values, but may suffer from inflation or over-stringentness in the type I error rate when μ_H deviates from μ_S. When the historical mean is over/under-estimated, the design would have higher/lower values of power. Clearly, under the Pocock type and O'Brien–Fleming type α–spending functions, the pattern of type I error rate spending matches the desired target. It is worth emphasizing that under the theoretical posterior probability boundaries, the overall type I error rate might not be controlled exactly at the target level, particularly when a complex and informative prior distribution is adopted. To achieve an exact control of the type I error rate, we may compute the empirical type I error rate by simulating a large number of trials, and adjust the posterior probability boundaries to be $c_k = \Phi(z_k) + \zeta_k$, where the value of ζ_k can be easily calibrated via grid search or bisectional search, such that the empirical type I error rate can be maintained at the desired level. For example, we may set $\zeta_k = \{1 - \Phi(z_k)\} \times u$, where $0 \leq u \leq 1$, and perform numerical calibration on the value of u. The middle part of Table 2 shows the spendings of the type I error rate where u is calibrated to be 0.141 and 0.237 under the null case ($\mu_H = 0.5$) with the Pocock type and the O'Brien–Fleming type boundaries, respectively.

5 Design Comparison

Wathen and Thall (2008) proposed a Bayesian doubly optimal group sequential design, abbreviated as "BDOGS", which optimizes the expected utility function under the frequentist constraints. As a comparison, we consider the BDOGS design for the binary endpoint, which takes an interim look once every time a new outcome is observed. At each interim analysis, the posterior probability is updated and compared with a boundary function $P_U(n) = a_U + b_U(n/N_{\max})^{c_U}$, where n is the cumulative sample size and $N_{\max}$ is the maximum sample size, to decide whether the trial should be stopped for efficacy. The design calibration involves finding the optimal values for the parameters (a_U, b_U, c_U) such that the expected utility is optimized under the constraints on the type I error rate and power. The expected utility under the BDOGS design is specified to be the average of the expected sample sizes under the null and the alternative hypotheses. Calculations of the expected utility, the type I error rate and power are conducted via a forward simulation approach (Carlin et al., 1998), and a simple grid search method is used for finding the optimal parameters (a_U, b_U, c_U).

The proposed Bayesian group sequential designs incorporating the Pocock type and O'Brien–Fleming type α-spending functions are compared with the BDOGS design, under the constraints of the type I error rate being at most 0.1 and power at least 0.8. Considering a binary endpoint, the single-arm design aims to test the hypotheses H_0: $p_E \leq 0.2$ versus H_1: $p_E > 0.4$. For the Bayesian group sequential designs with $K = 4$ analyses, the minimally required group sizes are 11 for the Pocock type design and 9 for the O'Brien–Fleming type design. For the BDOGS design, we specify $N_{\max} = 45$ and the optimal parameters are found to be $(a_U, b_U, c_U) = (0.985, -0.015, 0.600)$.

Figure 1 in the Supplementary Material (Shi and Yin, 2018) shows the stopping boundaries for the three designs under comparison, and Figure 2 in the Supplementary Material exhibits the distributions of the spendings of the type I error rates. Except for the first interim analysis under the O'Brien–Fleming type design, the boundary values of the Bayesian group sequential designs are smaller than those of the BDOGS design, as the latter requires much more interim looks. In terms of the type I error rate spendings, the Bayesian group sequential design allows specifying the pattern of the distribution of the spendings, whereas the BDOGS design is less flexible and the majority of the spendings are distributed at the first few analyses. The expected sample sizes are similar across the three designs, which are 30.7, 32.7 and 30.1 for the BDOGS, Pocock and O'Brien–Fleming types of designs, respectively.

6 Clinical Trial Application

6.1 Acute Myeloid Leukemia Trial

Thall and Simon (1994) described a single-arm clinical trial using fludarabine + ara-C + granulocyte colony stimulating factor (G-CSF) in the treatment of acute myeloid leukemia. The study aimed at assessing whether the addition of G-CSF to the standard therapy (fludarabine + ara-C) can improve the clinical outcomes of the patients. The complete remission of the disease is defined as the binary endpoint of the study. We

		Bayesian Posterior Probability			Frequentist Z-test		
α-spending function	k	Cutoff c_k	α_k	$\sum_{k=1}^{K} \alpha_k$	Z_k	α_k	$\sum_{k=1}^{K} \alpha_k$
Pocock	1	(0.923,0.963)	0.0432		1.7299	0.0432	
	2	(0.940,0.965)	0.0227		1.7299	0.0227	
	3	(0.957,0.973)	0.0111		1.7299	0.0217	
	4	(0.933,0.954)	0.0213	0.0983	1.7299	0.0170	0.1046
O'Brien–Fleming	1	(0.993,0.998)	0.0029		2.8141	0.0029	
	2	(0.965,0.981)	0.0198		1.9898	0.0198	
	3	(0.934,0.957)	0.0318		1.6247	0.0318	
	4	(0.905,0.933)	0.0355	0.0900	1.4070	0.0355	0.0900

Table 3: Bayesian single-arm sequential designs with binary endpoints using the numerical calibration method versus frequentist group sequential designs using Z-test with the Pocock type and O'Brien–Fleming type α-spending functions.

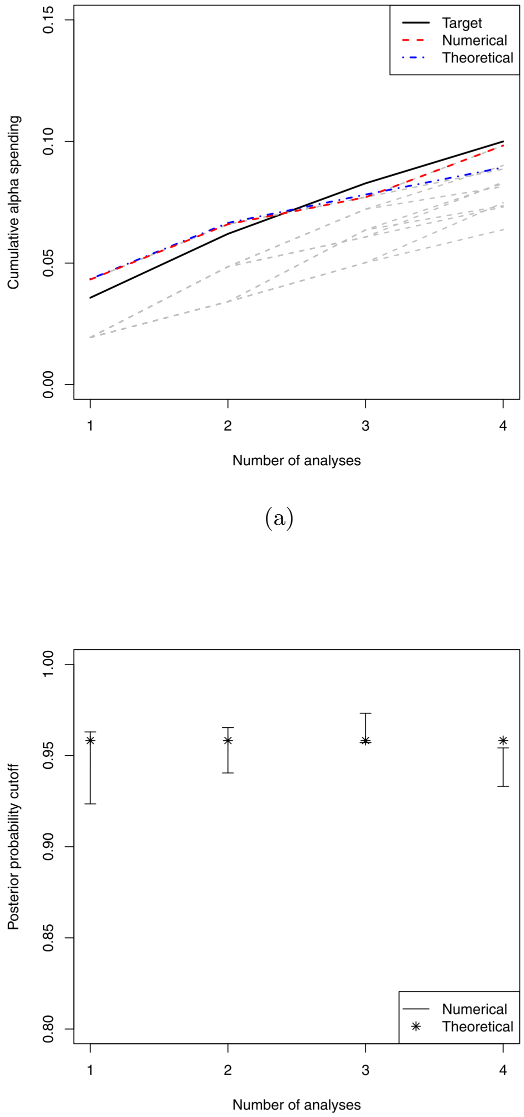
provide an illustrative trial example where the proposed design is adopted. The trial has a maximum sample size of $N_{\max} = 160$, and there are a total of $K = 4$ analyses to be carried out in the trial. We take the type I error rate $\alpha = 0.1$, $p_{\text{null}} = 0.2$ and noninformative prior $p_E \sim \text{Beta}(0.2, 0.8)$. For the Pocock design and the O'Brien–Fleming design, the α-spending functions are respectively given by

$$\alpha(t) = \alpha\log\{1 + (e-1)t\} \qquad \text{Pocock type,}$$

$$\alpha(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t}) \qquad \text{O'Brien–Fleming type,}$$

where $t \in [0, 1]$ denotes the information fraction, taking values of 0.25, 0.5, 0.75, and 1 in our case, and $z_{\alpha/2}$ denotes the $100(1-\alpha/2)$th percentile of the standard normal distribution. The target type I error rate to be spent at the kth analysis is thus $\alpha(k/4) - \alpha\{(k-1)/4\}$. We provide Bayesian sequential designs whose empirical type I error spending functions are respectively calibrated towards those of the two classical group sequential designs.

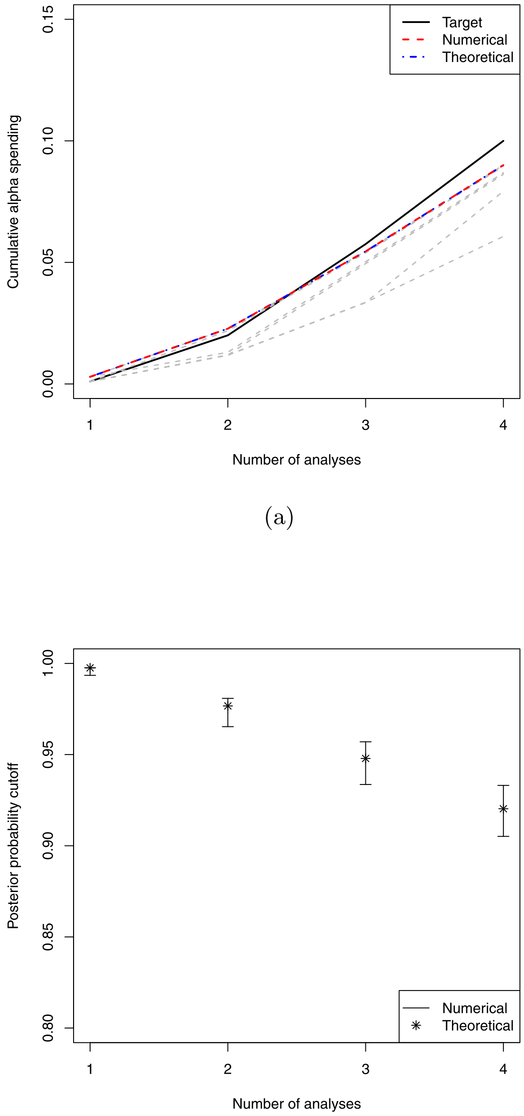
Table 3 shows the values of the posterior probability cutoff and the α-spending function at each interim analysis for the Bayesian Pocock type and O'Brien–Fleming type sequential designs, respectively. It is worth emphasizing that because $P(H_1|D_k)$ is discrete and takes a finite number of values, the upper cutoff c_k can take any value within a certain interval to satisfy the type I error constraint. For example, in the Pocock type sequential design, the first cutoff c_1 can be any value within the interval (0.923,0.963). Figures 3 and 4 show the posterior probability cutoff intervals and the empirical type I error spending functions versus the target for the Pocock and O'Brien–Fleming designs, respectively. Because the endpoint is binary, exact calibration to the target function is not possible. Therefore, the empirical spending function under the proposed methods would slightly deviate from the target one. For the Pocock type design, the numerical method (dashed) and the theoretical method (dot–dashed) yield similar solutions, although the former produces a closer fit to the target α-spending function. Similar to the constant critical values in the frequentist Pocock design, the posterior probability



(a)

(b)

Figure 3: Bayesian single-arm sequential trial design with binary endpoints under the Pocock type α-spending function with (a) the target α-spending function versus numerical and theoretical approaches to finding the cutoff boundaries, where the grey dashed lines represent all the feasible designs obtained from the numerical searching algorithm and the bold dashed line is the one closest to the target; and (b) the posterior probability cutoff value (theoretical) and interval (numerical) at each interim analysis.



(a)

(b)

Figure 4: Bayesian single-arm sequential trial design with binary endpoints under the O'Brien–Fleming type α-spending function with (a) the target α-spending function versus numerical and theoretical approaches to finding cutoff boundaries, where the grey dashed lines represent all the feasible designs obtained from the numerical searching algorithm and the bold dashed line is the one closest to the target; and (b) the posterior probability cutoff value (theoretical) and interval (numerical) at each interim analysis.

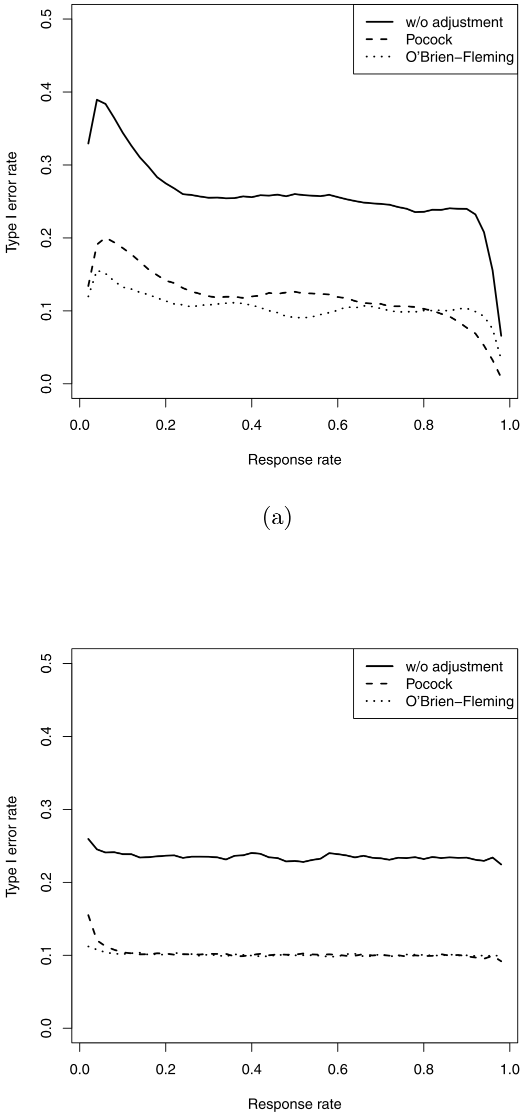
cutoffs are also constant with a value of 0.958 under the theoretical calibration, while the critical intervals under the numerical approach are also close to each other with substantial overlappings. For the O'Brien–Fleming type design, the theoretical method produces the posterior probability cutoffs of 0.998, 0977, 0.948, and 0.920, while the numerical method leads to cutoff intervals that tend to decline throughout the trial. The O'Brien–Fleming sequential design imposes more stringent posterior probability cutoffs at the early stages of the trial, and then gradually relaxes the cutoffs as the trial progresses.

Jennison and Turnbull (1989) provided a formulation of the repeated confidence intervals across interim analyses under a group sequential design. As a counterpart in the Bayesian paradigm, a similar notion of the repeated credible interval can be naturally developed. In particular, we obtain the posterior distribution of the parameter of interest, and adopt the highest posterior density interval repeatedly based on the type I error rate spent for each interim analysis. Figure 3 in the Supplementary Material shows the repeated credible intervals when the number of responses attains the efficacy boundary, i.e., the minimum value of y_k such that $P(H_1|D_k) > c_k$ is satisfied, which can be back-solved based on the monotonic relationship between y_k and the posterior probability $P(H_1|D_k)$. As more data are accumulated in each interim analysis, the width of the repeated credible interval decreases.

6.2 Soft Tissue Sarcoma Trial

Maki et al. (2007) described a phase II randomized study comparing the efficacy of gemcitabine alone and the gemcitabine–docetaxel combination in the treatment of metastatic soft tissue sarcoma. Based on the binary outcomes of tumor response, the study aimed at determining whether the addition of docetaxel could improve the efficacy of gemcitabine. For illustrative purpose, we applied the proposed design to the trial and examined the empirical type I error rates under various types of sequential boundaries. We experimented the total sample size of 50 and 500 respectively, and for both cases $K = 5$ interim analyses were considered. We took non-informative prior distributions, $p_S \sim \text{Beta}(0.2, 0.8)$ and $p_E \sim \text{Beta}(0.2, 0.8)$, and the type I error rate was controlled at $\alpha = 0.1$. The type I error rates were computed as the probabilities of trial success with $p_E = p_S$ at different values of p_E (or p_S).

Figure 5 shows the type I error rates under different types of boundaries. Due to the finite sample size, the joint distribution of the test statistics at the interim analyses may deviate from the multivariate normal canonical distribution under the frequentist group sequential framework. As a result, the type I error rates can be different from the nominal level. As expected, when no adjustment is made to account for the multiplicity, i.e., the posterior probability boundaries are all set equal to $1 - \alpha = 0.9$ throughout the trial, the type I error rate is inflated up to the level of 25%. Both the Pocock type and O'Brien–Fleming type boundaries work well for the large-sample cases, but suffer from slight inflation of the type I error rate when the response rate is very low for the small-sample cases.



(a)

(b)

Figure 5: Type I error rates under different types of boundaries in Bayesian double-arm sequential designs with binary endpoints and $K = 5$, (a) sample size of 50 subjects per arm and (b) sample size of 500 subjects per arm. The solid line represents the type I error rate with a fixed posterior probability cutoff of 0.9 throughout the trial, and the dashed and dotted lines correspond to those with the posterior probability cutoffs calibrated using the Pocock type and O'Brien–Fleming type α-spending functions.

7 Discussion

Controlling the type I error rate for clinical trial designs that involve multiple interim assessments on the posterior probabilities is often a neglected aspect in Bayesian sequential designs. Although Bayesian methods can serve as a useful and flexible alternative to conventional frequentist designs, it is crucial to understand its frequentist properties. As shown in the motivating example, a failure to account for the multiplicity in a Bayesian trial may lead to a severe inflation of the type I error rate. The proposed method connects the aspect of multiple testing in Bayesian designs with that of the frequentist group sequential method. Although the theoretical method is primarily applied to the binary case with an assumed beta prior distribution on the response rate, it can be used under a more general family of prior and posterior models, as long as the regularity conditions for the asymptotic properties of the posterior probability are satisfied.

We consider both the binary and normal endpoints and single- and double-arm trials. We develop a numerical approach as well as establishing a connection based on the asymptotic properties of the posterior probability between the Bayesian sequential design and the frequentist counterpart. The numerical method involves the calculation of the exact type I error rate. When the number of analyses and the group size are large, the summation of the product of binomial probabilities could be computationally intensive. To overcome this issue, simulation-based computation or normal approximation to the binomial distribution might be preferred.

For the numerical approach, the sandwich-type algorithm can be generalized to more complex model settings. The error rate formulation would be similar to that in (2) except for the binomial distribution, which can be approximated using a simulation-based approach. A more general formulation of error rates can be implemented by first setting up a null parametric model representing H_0, and then simulating a large number of trials, and the proportion of trials that reach the decision boundary at each interim step can be used as an approximation to the type I error rate. Based on the prespecified error rates, suitable design parameters can be calibrated either by a grid-based search or a bisection approach in order to yield the desirable pattern of the type I error rate spendings. Examples of such calibration methods can be found in Murray et al. (2016) and Murray et al. (2017).

For the theoretical approach, we assume a non-informative prior distribution when assessing the design's operating characteristics, and show that the type I error rates can be well maintained under such a prior assumption. It should be emphasized that the theorem in Dudley and Haughton (2002) only holds asymptotically, and simulation studies on finite-sample performance of the design might be necessary for assessing the adequacy of the theoretical boundaries. In the settings where the prior distribution is highly informative or the sample size is relatively small, the empirical performance of the decision boundaries under the theoretical approach might not be satisfactory. It is then advised to adopt the numerical approach, either by explicitly formulating the error rates as discussed in this paper, or by averaging the number of error cases with computer simulation.

Supplementary Material

Supplementary Material of the Control of type I error rates in Bayesian sequential designs (DOI: 10.1214/18-BA1109SUPP; .pdf).

References

Barber, S. and Jennison, C. (2002). "Optimal asymmetric one-sided group sequential tests." *Biometrika*, 89(1): 49–60. MR1888345. doi: https://doi.org/10.1093/biomet/89.1.49. 400

Berger, J. O. and Berry, D. A. (1988). "Statistical analysis and the illusion of objectivity." *American Scientist*, 76(1): 159–165. 399

Berry, D. A. (2006). "Bayesian clinical trials." *Nature Reviews Drug Discovery*, 5(1): 218–226. 399

Berry, D. A. (2011). "Adaptive clinical trials in oncology." *Nature Reviews Drug Discovery*, 9(1): 199–207. 399

Berry, D. A. and Hochberg, Y. (1999). "Bayesian perspectives on multiple comparisons." *Journal of Statistical Planning and Inference*, 82(1): 215–227. MR1736444. doi: https://doi.org/10.1016/S0378-3758(99)00044-0. 400

Carlin, B. P., Kadane, J. B., and Gelfand, A. E. (1998). "Approaches for optimal sequential decision analysis in clinical trials." *Biometrics*, 54(3): 964–975. 415

Christen, J. A., Müller, P., Wathen, J. K., and Wolf, J. (2004). "Bayesian randomized clinical trials: A decision-theoretic sequential design." *Canadian Journal of Statistics*, 32(4): 387–402. MR2125852. doi: https://doi.org/10.2307/3316023. 400

Dudley, R. M. and Haughton, D. (2002). "Asymptotic normality with small relative errors of posterior probabilities of half-spaces." *The Annals of Statistics*, 30(5): 1311–1344. MR1936321. doi: https://doi.org/10.1214/aos/1035844978. 406, 407, 421

Eales, J. D. and Jennison, C. (1992). "An improved method for deriving optimal one-sided group sequential tests." *Biometrika*, 79(1): 13–24. MR1158514. doi: https://doi.org/10.1093/biomet/79.1.13. 400

Efron, B. (1986). "Why isn't everyone a Bayesian." *The American Statistician*, 40(1): 1–5. MR0828575. doi: https://doi.org/10.2307/2683105. 399

Efron, B. (2005). "Bayesians, frequentists, and scientists." *Journal of the American Statistical Association*, 100(1): 1–5. MR2166064. doi: https://doi.org/10.1198/016214505000000033. 399

Gopalan, R. and Berry, D. A. (1998). "Bayesian multiple comparisons using Dirichlet process priors." *Journal of the American Statistical Association*, 93(443): 1130–1139. MR1649207. doi: https://doi.org/10.2307/2669856. 400

Gsponer, T., Gerber, F., Bornkamp, B., Ohlssen, D., Vandemeulebroecke, M., and Schmidli, H. (2014). "A practical guide to Bayesian group sequential designs." *Pharmaceutical Statisics*, 13(1): 71–80. 400

Guindani, M., Müller, P., and Zhang, S. (2009). "A Bayesian discovery procedure." *Journal of the Royal Statistical Society. Series B*, 71(5): 905–925. MR2750250. doi: https://doi.org/10.1111/j.1467-9868.2009.00714.x. 400

Guo, M. and Heitjan, D. F. (2010). "Multiplicity-calibrated Bayesian hypothesis tests." *Biostatistics*, 11(3): 473–483. 400

Heitjan, D. F. (1997). "Bayesian interim analysis of phase II cancer clinical trials." *Statistics in Medicine*, 16(1): 1791–1802. 400

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., and Sargent, D. J. (2011). "Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials." *Biometrics*, 67(1): 1047–1056. MR2829239. doi: https://doi.org/10.1111/j.1541-0420.2011.01564.x. 411

Jennison, C. and Turnbull, B. W. (1989). "Interim analyses: the repeated confidence interval approach." *Journal of the Royal Statistical Society. Series B*, 51(1): 305–361. MR1017201. 419

Jennison, C. and Turnbull, B. W. (1997). "Group sequential analysis incorporating covariate information." *Journal of the American Statistical Association*, 92(440): 1330–1341. MR1615245. doi: https://doi.org/10.2307/2965403. 407

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: Chapman & Hall/CRC. MR1710781. 404, 406, 407

Labbe, A. and Thompson, M. E. (2007). "Multiple testing using the posterior probabilities of directional alternatives, with application to genomic studies." *The Canadian Journal of Statistics*, 35(1): 51–68. MR2345374. doi: https://doi.org/10.1002/cjs.5550350107. 400

Lee, J. J. and Liu, D. D. (2008). "A predictive probability design for phase II cancer clinical trials." *Clinical Trials*, 5(2): 93–106. 400

Lewis, R. J. and Berry, D. A. (1994). "Group sequential clinical trials, A classical evaluation of Bayesian decision-theoretic designs." *Journal of the American Statistical Association*, 89(1): 1528–1534. 400

Maki, R. G., Wathen, J. K., Patel, S. R., Priebat, D. A., Okuno, S. H., Samuels, B., Fanucchi, M., Harmon, D. C., Schuetze, S. M., Reinke, D., Thall, P. F., Benjamin, R. S., Baker, L. H., and Hensley, M. L. (2007). "Randomized phase II study of gemcitabine and docetaxel compared with gemcitabine alone in patients with metastatic soft tissue sarcomas: results of sarcoma alliance for research through collaboration study 002." *Journal of Clinical Oncology*, 25(1): 2755–2763. 419

Morita, S., Müller, P., and Thall, P. F. (2008). "Determining the effective sample size of a parametric prior." *Biometrics*, 64(2): 595–602. MR2432433. doi: https://doi.org/10.1111/j.1541-0420.2007.00888.x. 402

Müller, P., Berry, D. A., Grieve, A. P., Smith, M., and Krams, M. (2007). "Simulation-based sequential Bayesian design." *Journal of Statistical Planning and Inference*, 137(1): 3140–3150. MR2364157. doi: https://doi.org/10.1016/j.jspi.2006.05.021. 400

Murray, T. A., Thall, P. F., and Yuan, Y. (2016). "Utility-based designs for randomized comparative trials with categorical outcomes." *Statistics in Medicine*, 35(24): 4285–4305. MR3554963. doi: https://doi.org/10.1002/sim.6989. 400, 421

Murray, T. A., Thall, P. F., Yuan, Y., McAvoy, S., and Gomez, D. R. (2017). "Robust treatment comparison based on utilities of semi-competing risks in non-small-cell lung cancer." *Journal of the American Statistical Association*, 112(517): 11–23. MR3646549. doi: https://doi.org/10.1080/01621459.2016.1176926. 400, 421

O'Brien, P. C. and Fleming, T. R. (1979). "A multiple testing procedure for clinical trials." *Biometrics*, 35(3): 549–556. 400

Pocock, S. J. (1977). "Group sequential methods in the design and analysis of clinical trials." *Biometrika*, 64(2): 191–199. 400

Rosner, G. L. and Berry, D. A. (1995). "A Bayesian group sequential design for a multiple arm randomized clinical trial." *Statistics in Medicine*, 14(14): 381–394. 400

Scharfstein, D. O., Tsiatis, A. A., and Robins, J. M. (1997). "Semiparametric efficiency and its implication on the design and analysis of group sequential studies." 92(440): 1342–1350. 407

Scott, J. G. and Berger, J. O. (2006). "An exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136(7): 2144–2162. MR2235051. doi: https://doi.org/10.1016/j.jspi.2005.08.031. 400

Shi, H. and Yin, G. (2018). "Supplementary Material of the Control of type I error rates in Bayesian sequential designs" *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1109SUPP. 415

Thall, P. F. and Simon, R. (1994). "Practical Bayesian guidelines for phase IIB clinical trials." *Biometrics*, 50(2): 337–349. MR1294683. doi: https://doi.org/10.2307/2533377. 399, 400, 401, 402, 403, 405, 415

Thall, P. F., Simon, R., and Estey, E. H. (1995). "Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes." *Statistics in Medicine*, 14(1): 357–379. 400

Ventz, S. and Trippa, L. (2015). "Bayesian designs and the control of frequentist characteristics: A practical solution." *Biometrics*, 71(1): 218–226. MR3335366. doi: https://doi.org/10.1111/biom.12226. 400

Wang, S. K. and Tsiatis, A. A. (1987). "Approximately optimal one-parameter boundaries for group sequential trials." *Biometrics*, 43(1): 193–199. MR0882780. doi: https://doi.org/10.2307/2531959. 400

Wason, J. M. S., Abraham, S. E., Baird, R. D., Gournaris, I., Vallier, A., Brenton, J. D., Earl, H. M., and Mander, A. P. (2015). "A Bayesian adaptive design for biomarker trials with linked treatments." *British Journal of Cancer*, 113(5): 699–705. 408, 409

Wathen, J. K. and Thall, P. F. (2008). "Bayesian adaptive model selection for optimizing group sequential clinical trials." *Statistics in Medicine*, 27(27): 5586–5604. MR2573771. doi: https://doi.org/10.1002/sim.3381. 415

Yin, G. (2012). *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*. Hoboken: John Willey & Sons, Inc. 399

Zhu, H. and Yu, Q. (2017). "A Bayesian sequential design using alpha spending function to control type I error." *Statistical Methods in Medical Research*, 26(5): 2184–2196. MR3712227. doi: https://doi.org/10.1177/0962280215595058. 400

Acknowledgments

We thank the associate editor, the referees, and the editor for their many constructive and insightful comments that have led to significant improvements in the article. The research was supported in part by a grant (grant number 17326316) from the Research Grants Council of Hong Kong.