BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

# Comparison of Bayesian and frequentist group-sequential clinical trial designs

Nigel Stallard[1*], Susan Todd[2], Elizabeth G. Ryan[3] and Simon Gates[3]

## Abstract

**Background:** There is a growing interest in the use of Bayesian adaptive designs in late-phase clinical trials. This includes the use of stopping rules based on Bayesian analyses in which the frequentist type I error rate is controlled as in frequentist group-sequential designs.

**Methods:** This paper presents a practical comparison of Bayesian and frequentist group-sequential tests. Focussing on the setting in which data can be summarised by normally distributed test statistics, we evaluate and compare boundary values and operating characteristics.

**Results:** Although Bayesian and frequentist group-sequential approaches are based on fundamentally different paradigms, in a single arm trial or two-arm comparative trial with a prior distribution specified for the treatment difference, Bayesian and frequentist group-sequential tests can have identical stopping rules if particular critical values with which the posterior probability is compared or particular spending function values are chosen. If the Bayesian critical values at different looks are restricted to be equal, O'Brien and Fleming's design corresponds to a Bayesian design with an exceptionally informative negative prior, Pocock's design to a Bayesian design with a non-informative prior and frequentist designs with a linear alpha spending function are very similar to Bayesian designs with slightly informative priors.
This contrasts with the setting of a comparative trial with independent prior distributions specified for treatment effects in different groups. In this case Bayesian and frequentist group-sequential tests cannot have the same stopping rule as the Bayesian stopping rule depends on the observed means in the two groups and not just on their difference. In this setting the Bayesian test can only be guaranteed to control the type I error for a specified range of values of the control group treatment effect.

**Conclusions:** Comparison of frequentist and Bayesian designs can encourage careful thought about design parameters and help to ensure appropriate design choices are made.

**Keywords:** Adaptive design, Interim analysis, Type I error rate, Sequential analysis, Sequential design

## Background

An increasing desire for efficiency in clinical trials has led to growing interest in adaptive designs. Frequentist group-sequential designs enable interim analyses to be performed during the conduct of a clinical trial without inflation of the overall type I error rate [1]. With an increased application of Bayesian methods in clinical trials, a number of researchers have proposed Bayesian group sequential methods [2, 3].

Not all proponents of Bayesian sequential designs consider exact control of the type I error rate essential [4]. Some, however, have suggested that the stopping rules for Bayesian group sequential designs should also be chosen in such a way that the frequentist type I error rate is controlled [2, 5, 6], particularly in the setting of phase III or late phase II clinical trials, when it is often considered desirable to control the risk of a false positive result, that is an erroneous conclusion that a new treatment is efficacious.

There are a number of published examples of trials using a Bayesian stopping rule chosen to control the type I error

*Correspondence: n.stallard@warwick.ac.uk
[1]Statistics and Epidemiology, Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, UK
Full list of author information is available at the end of the article

rate. Hueber et al. [7] (see also [8] for additional statistical details) describe a Bayesian group-sequential trial comparing secukinumab with placebo for the treatment of Crohn's disease. The outcome is the change in Crohn's Disease Activity Index (CDAI), which was taken to be normally distributed. Prior distributions were specified separately for the placebo and secukinumab effects, with the former being informative and the latter non-informative. Analyses were planned after 30 and 60 patients, when the trial could be stopped if both (i) the posterior probability that secukinumab was superior to the placebo exceeded 95%, and (ii) there was at least a 50% posterior probability that the change in CDAI due to secukinumab was superior to that for placebo by at least fifty. The type I error rate for this design was calculated using the R package `gsbDesign`[9] and shown to be 1.2% if the change in CDAI due to placebo was as anticipated.

A Bayesian group-sequential trial with a binary primary outcome is described by Wilber et al. [10]. This randomised trial compared antiarrhythmic drug therapy with catheter ablation for the treatment of paroxysmal atrial fibrillation. The primary outcome was the observation of protocol-defined treatment failure. Analyses were planned after 150, 175, 200 and 230 patients, with a stopping rule based on the posterior probability of superiority of the experimental treatment over the control exceeding 98%, giving a type I error rate of 0.025.

The increasing use of Bayesian sequential designs that control the frequentist type I error rate has led to a growing body of work comparing Bayesian and frequentist group sequential trial methods [3, 5, 8, 11–14]. This paper adds to this work. In contrast to some authors who draw comparisons between underlying Bayesian and frequentist paradigms, our focus is a practical one, in which we compare Bayesian and frequentist group sequential tests in terms of their boundary values and operating characteristics. We consider specifically the setting of normally distributed data or test statistics. This facilitates comparison between Bayesian and frequentist group sequential methods as the latter have been largely developed in this setting.

We consider separately Bayesian designs in which a single treatment effect is considered, either in a single-arm trial or with a prior specified directly for the difference between experimental and control treatments, and in which treatment effects have independent prior distributions. In the one-parameter setting frequentist and Bayesian group-sequential designs can be identical if sufficient flexibility in choice of design parameters is allowed [12], and we show that frequentist and Bayesian group-sequential designs may be very similar for common choices of stopping rules. In the two-parameter setting we show that the frequentist and Bayesian designs cannot correspond, and show that in this case the Bayesian group-sequential designs can only control the type I error rate for specified values of the control group treatment effect.

## Methods

### Notation and problem formulation

#### *Single arm trials with normally distributed data*

Suppose we conduct a group-sequential single-arm clinical trial of some experimental treatment with up to $K$ analyses of a single sample of normally distributed data with a cumulative total of $n_k$ observations at look $k, k = 1, \ldots, K$.

At each look the data observed up to that point will be analysed and a decision made whether or not to continue to the next look. We will only consider stopping the trial for a positive result, that is for efficacy. Additional stopping for futility is considered in the "Discussion" section.

Denoting by $Y_i$ the observed value for patient $i$, we will assume this is normally distributed with mean $\theta$ and known variance $\sigma^2$. We wish to draw inference on $\theta$ and will assume that parameterisation is such that $\theta = 0$ corresponds to the experimental treatment being of equal efficacy to some specified reference value or standard treatment effect, with positive values of $\theta$ (and hence of $Y_i$) indicative of superiority of the experimental treatment.

Let $\bar{Y}_k = \sum_{i=1}^{n_k} Y_i/n_k$ denote the mean value from the cumulative sample at look $k$. This is the sufficient statistic for $\theta$ at look $k$. It is helpful to write the distribution in terms of the inverse of the variance, known as the information, and set $I_k = n_k/\sigma^2$. We then have $\bar{Y}_1, \ldots, \bar{Y}_K$ multivariate normal with

$$\begin{pmatrix} \bar{Y}_1 \\ \vdots \\ \bar{Y}_K \end{pmatrix} \sim N\left( \begin{pmatrix} \theta \\ \vdots \\ \theta \end{pmatrix}, \begin{pmatrix} I_1^{-1} & I_2^{-1} & \cdots & I_K^{-1} \\ I_2^{-1} & I_2^{-1} & \cdots & I_K^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ I_K^{-1} & I_K^{-1} & \cdots & I_K^{-1} \end{pmatrix} \right) \quad (1)$$

with a similar multivariate normal distribution for the standardised test statistics, $\bar{Y}_1\sqrt{I_1}, \ldots, \bar{Y}_K\sqrt{I_K}$.

In a frequentist setting, we will test the null hypothesis, $H_0 : \theta \leq 0$ against the one-sided alternative, $\theta > 0$, concluding that the experimental treatment is superior to the standard if this null hypothesis is rejected. The test will be based on the observed values of $\bar{Y}_1, \ldots, \bar{Y}_K$, stopping and rejecting the null hypothesis at look $k$ if $\bar{Y}_k$ is sufficiently large as described in more detail below.

In a Bayesian setting, inference will be based on the posterior distribution for $\theta$ given the observed data. Basing the likelihood on (1), a normal prior for $\theta$ is conjugate. Given prior distribution $\theta \sim N\left(\theta_0, I_0^{-1}\right)$ the posterior distribution for $\theta$ following observation of $\bar{Y}_k = \bar{y}_k$ at look $k$ is given by

$$\theta \mid \bar{y}_k \sim N\left(\frac{\theta_0 I_0 + \bar{y}_k I_k}{I_0 + I_k}, \frac{1}{I_0 + I_k}\right) \quad (2)$$

(see [15] Section 5.2). If this posterior distribution is sufficiently indicative of a positive treatment effect the trial will be stopped with the conclusion that the experimental treatment is superior to the standard or reference value. More details are given below.

The value of $I_0$ gives a measure of the prior information. In particular, letting $I_0$ approach 0 gives a flat improper normal prior.

#### *Single arm trials with non-normal data*

For non-normal data, tests can be based on the assumed distributional form parameterised in terms of the treatment effect, which will again be denoted by $\theta$. An analytic form of the posterior distribution may be available if a conjugate prior distribution is used.

Alternatively, in many cases, if $n_1, \ldots, n_K$ are sufficiently large, we can obtain an estimate $\hat{\theta}_k$ for the treatment effect based on the data at look $k$ with $\hat{\theta}_1, \ldots, \hat{\theta}_K$ approximately following the multivariate normal distribution (1) for some $I_1, \ldots, I_K$. It is common to use this approximate distributional form in a frequentist group-sequential test [16], enabling use of these estimates in place of the single sample means and applying methods based on the normal distribution (1) even without normally distributed data, or with normal data when the variance cannot be assumed to be known.

An illustration in the setting of a single sample of binomial data is given below.

#### *Comparative trials*

Suppose now we have two groups; group 0, the control group and group 1, the experimental treatment group. Let $Y_{ji}$ denote the response from patient $i$ in group $j$, assumed to be normally distributed with known variance, with $Y_{ji} \sim N\left(\mu_j, \sigma_j^2\right), j = 0, 1$. We wish to draw inference on the treatment difference given by $\theta = \mu_1 - \mu_0$. We will again assume larger values of $Y_{ji}$ are preferable so that larger values of $\theta$ correspond to the superiority of the experimental treatment to the control treatment.

At analysis $k$, suppose that we have a total of $n_{jk}$ observations from group $j$, and let $\bar{Y}_{jk} = \sum_{i=1}^{n_{jk}} Y_{ji}/n_{jk}, j = 0, 1, k = 1, \ldots, K$. Writing $I_{jk} = n_{jk}/\sigma_j^2$, we have $\bar{Y}_{j1}, \ldots, \bar{Y}_{jK}$ multivariate normal with $\bar{Y}_{jk} \sim N\left(\mu_j, I_{jk}^{-1}\right)$ and $\text{cov}(\bar{Y}_{jk}, \bar{Y}_{jk'}) = I_{jk'}^{-1}$ if $k < k'$.

A sufficient statistic for $\theta$ at look $k$ is $D_k = \bar{Y}_{1k} - \bar{Y}_{0k}$, with $D_1, \ldots, D_K$ following the multivariate normal distribution as in (1) with $I_k = \left(\sigma_1^2/n_{1k} + \sigma_0^2/n_{0k}\right)^{-1}$.

In a frequentist setting, we will test $H_0 : \theta \leq 0$ against $\theta > 0$ based on the observed values of $D_1, \ldots, D_K$, stopping and rejecting the null hypothesis at look $k$, concluding that the experimental treatment is superior to the control, if $D_k$ is sufficiently large, as described in more detail below.

In a Bayesian setting, we may specify the prior distribution for the treatment effect in two ways. The first is to specify a prior distribution for the treatment difference, $\theta$, directly. Suppose again that $\theta$ has a normal prior distribution with $\theta \sim N\left(\theta_0, I_0^{-1}\right)$. At look $k$ the posterior distribution for $\theta$ given observed value $D_k = d_k$ is given by

$$\theta \mid d_k \sim N\left(\frac{\theta_0 I_0 + d_k I_k}{I_0 + I_k}, \frac{1}{I_0 + I_k}\right). \quad (3)$$

The alternative is to specify independent prior distributions for $\mu_0$ and $\mu_1$, update these separately to obtain posterior distributions for $\mu_0$ and $\mu_1$ and then use these to obtain a posterior distribution for $\theta$. This approach is considered in detail below in the section entitled "Comparison of frequentist and Bayesian group-sequential approaches - two parameter case".

For non-normal data, or when the variance cannot be assumed known, we often again have estimates of the treatment effect, $\hat{\theta}_k$, approximately normally distributed, so that the distributional form (1) can be used. As in the two-sample case with normally distributed data, in the Bayesian setting we can either specify a prior for $\theta$ directly or specify independent prior distributions for treatment effects in the two groups.

### Bayesian group-sequential approach

In a Bayesian sequential trial, inference at look $k$ will be based on the posterior distribution for $\theta$ given in the single group case by (2), in the two sample case when a prior distribution is specified for $\theta$ directly by (3) and in the two sample case when prior distributions are given for $\mu_0$ and $\mu_1$ by the expression (10) given below.

A common approach is to stop the trial, concluding that the experimental treatment is superior to the control if the posterior probability that $\theta$ exceeds 0 given the observed data is sufficiently large. In detail, critical values, $p_k, k = 1, \ldots, K$, will be specified and the trial will stop as soon as

$$Pr(\theta > 0 \mid \text{data at look } k) \geq p_k. \quad (4)$$

Considering stopping to conclude the experimental treatment is superior to the control to be equivalent to rejection of $H_0$, the frequentist type I error rate of this Bayesian sequential procedure can be calculated by noting that $Pr(\theta > 0 \mid \text{data at look } k)$ is a random variable since it depends on the observed data. Control of the type I error rate is thus achieved if

$$Pr(Pr(\theta > 0 \mid \text{data at look } k) \geq p_k \text{ some } k \leq K; \theta = 0) = \alpha. \quad (5)$$

It has been suggested that $p_1, \ldots, p_K$ should be chosen to satisfy this condition [2].

A number of alternatives to the stopping criterion (4) above have also been proposed. For example, the trial might be stopped to declare the experimental treatment superior at look $k$ if the posterior probability that $\theta$ exceeds some specified positive target value, or the predictive probability that the experimental treatment would be found superior if the trial continued to the final analysis, is sufficiently large [8, 17, 18].

Although, in general, different values for $p_1, \ldots, p_K$ could be specified, often a common value $p_1 = \cdots = p_K$ is used [2], with this value chosen to satisfy (5). We will consider both the general and this specific case in the examples below.

In many settings the probability on the left hand side of (5) can most easily be calculated via simulation methods [2]. In the case of single- or two-sample normally distributed data considered here, since, for a specified prior distribution, the posterior probability (4) depends on $\bar{Y}_k$, it can be calculated analytically from the joint distribution (1), for example in R using the `gsbDesign` [9] or code available from the first author.

### Frequentist group-sequential approaches

In a frequentist setting, the null hypothesis, $H_0 : \theta \leq 0$, will be rejected, and the trial stopped at look $k$ if $\bar{Y}_k\sqrt{I_k} \geq u_k$ for some $u_k$ in the single-sample case or if $D_k\sqrt{I_k} \geq u_k$ in the two sample case. As the forms of the joint distributions for $\bar{Y}_1, \ldots, \bar{Y}_K$ and $D_1, \ldots, D_K$ are identical, we will here consider only the single-sample case.

To control the type I error rate at some specified level $\alpha$, it is required to choose $u_1, \ldots, u_K$ with $Pr(\bar{Y}_k\sqrt{I_k} \geq u_k, \text{some} k \leq K; \theta) \leq \alpha$ for all $\theta \leq 0$. The form (1) means that this is satisfied if

$$Pr(\bar{Y}_k\sqrt{I_k} \geq u_k \text{ some } k \leq K; \theta = 0) = \alpha. \quad (6)$$

As the requirement (6) is insufficient to specify $u_1, \ldots, u_K$, a number of approaches have been proposed as described in the next two subsections.

#### *Pocock's test and O'Brien and Fleming's test*

Pocock [19] and O'Brien and Fleming [20] propose methods with equally-spaced looks, that is, using the notation introduced above, with $I_k = kI_K/K, k = 1, \ldots, K$. O'Brien and Fleming suggest stopping if $\bar{Y}_k I_k$ exceeds some fixed value, that is taking $u_k = c/\sqrt{I_k}$. Pocock suggests stopping if the standardised difference $\bar{Y}_k I_k^{1/2}$ exceeds a fixed value, that is taking $u_k = c$. In each case, the constant value for $c$ is found so as to satisfy (6). These values are tabulated for certain $K$ and $\alpha$ [19, 20], or can be obtained from a numerical search, noting that the probability in (6) can be expressed in terms of the multivariate normal distribution function which may be evaluated numerically, for example in R using function `pmvnorm` in the `mvtnorm` package [21].

#### *Spending function approaches*

Slud and Wei [22] suggest introducing greater flexibility to sequential designs that satisfy (6) by specifying the type I error rate "spent" at each look. In detail, they specify $\alpha_1 \leq \cdots \leq \alpha_K = \alpha$, then obtain $u_k, k = 1, \ldots, K$, such that the probability under the null hypothesis of stopping at or before look $k$, say at some look $k'$ with $k' \leq k$, is equal to $\alpha_k$, that is

$$Pr(\bar{Y}_{k'}\sqrt{I_{k'}} \geq u_{k'} \text{ some } k' \leq k; \theta = 0) = \alpha_k. \quad (7)$$

This approach was extended by Lan and DeMets [23], who proposed that $\alpha_1, \ldots, \alpha_K$ be given by a function $\alpha^*(t)$ of the information time, with $t$ at look $k$ equal to $I_k/I_K$ so that $\alpha_k = \alpha^*(I_k/I_K), k = 1, \ldots, K$. For general choice of non-decreasing $\alpha^*$ with $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$, the approaches of Slud and Wei and Lan and DeMets are equivalent provided $I_1, \ldots, I_K$ are specified in advance. By defining the functional form of $\alpha^*$, the Lan and DeMets approach enables calculation of $u_1, \ldots, u_K$ to satisfy (6) when $I_1, \ldots, I_K$ are not given in advance, providing they are independent of $\bar{Y}_1, \ldots, \bar{Y}_K$.

Lan and DeMets give forms for the spending function $\alpha^*(t)$ corresponding approximately to the Pocock test, with $\alpha^*(t) = \alpha \log(1+(e-1)t)$, and the O'Brien and Fleming test, with $\alpha^*(t) = 2(1 - \Phi(z_\alpha/\sqrt{t}))$, where $\Phi$ denotes the distribution function for a standard normal and $z_\alpha$ denotes $\Phi^{-1}(1-\alpha)$, the upper $100\alpha$ percentile of the standard normal distribution. Exact spending functions for these tests for a given number of looks can be obtained numerically from the joint distribution (1) [24]. Alternative spending function forms have been suggested [1, 25], including as a special case the linear spending function $\alpha^*(t) = \alpha t$.

The stopping boundary values $u_1, \ldots, u_K$ may be computed recursively[1]; at look $k$, supposing $u_1, \ldots, u_{k-1}$ and $I_1, \ldots, I_k$ are known, we can use the joint distribution of $\bar{Y}_1, \ldots, \bar{Y}_k$ for $\theta = 0$ from (1) along with a numerical search to find $u_k$ to satisfy (7). These calculations can be performed in R using the `gsBound` in the `gsDesign` package [26] or code available from the first author.

### Examples

To compare the Bayesian and frequentist group-sequential methods, we illustrate the two approaches using three simplified examples. These are described below.

#### *Example 1: Single-arm trial with normally distributed data*

Consider a single-arm trial with the outcome for patient $i$ equal to $Y_i$ with $Y_i \sim N(\theta, \sigma^2)$ for some known $\sigma$. Suppose that $\theta = 0$ corresponds to a null value and $\theta = 1$ to a worthwhile treatment effect. We will assume that the trial is conducted in up to five stages, that is $K = 5$, with these of equal size so that the number of patients included in the first $k$ stages is $n_k = nk/K$. We will further assume that $n_K = 10\sigma^2$. With this sample size a fixed sample size trial with a hypothesis test conducted at a two-sided 5% level would have power of approximately 90%. This gives $I_1, \ldots, I_5 = 2, \ldots, 10$.

We will consider a range of prior distributions for $\theta$. We will take $I_0$ equal to 0 (non-informative), 0.5 and 1 (that is with weight equivalent to one twentieth and one tenth of the total information available from the trial) as well as a very informative prior distribution with $I_0 = 20$, and will take $\theta_0$ equal to $-0.25$, 0, 0.25 and 0.5, recalling that 0 and 1 correspond to null and worthwhile treatment effects. Density functions for the range of prior distributions considered are shown in Fig. 1. The prior mean, $\theta_0$, increases across the columns moving from left to right and the prior information, $I_0$, decreases as we move down the rows. The vertical lines correspond to the null and worthwhile treatment effects of 0 and 1. Only one plot is given in the lowest row as when $I_0 = 0$ the prior distribution does not depend on $\theta_0$.

#### *Example 2: Single-arm trial with binary data*

Consider, as a second example, a single-arm trial with a binary outcome corresponding to success or failure for each patient. Suppose that the trial has up to four looks with 25, 50, 75 and 100 patients and assume that we wish to determine whether the true success rate, which will be denoted by $\pi$, exceeds a control rate, $\pi_0$, assumed to be 0.5, using a non-informative prior distribution for $\pi$.

#### *Example 3: Two-arm trial with normally distributed data*

The third example is a two-arm trial with up to five equally-sized stages with the outcome for patient $i$ in group $j$ ($j = 0, 1$) equal to $Y_{ij}$ with $Y_{ij} \sim N\left(\mu_j, \sigma_j^2\right)$ for some known $\sigma_j$, where we assume $\sigma_1 = \sigma_0$.

Denoting the treatment difference $\mu_1 - \mu_0$ by $\theta$, we will, as in Example 1 above, assume that $\theta = 1$ represents a worthwhile treatment effect. Assuming at stage $k$ we have included a total of $n_k$ patients in each of the two trial arms, we will set $I_k = n_k/2\sigma^2$ and, again as in Example 1, take $I_1, \ldots, I_5 = 2, \ldots, 10$.

Suppose that $\mu_1$ and $\mu_0$ have independent normal prior distributions with $\mu_j \sim N\left(\mu_{j0}, I_{j0}^{-1}\right)$, with a moderately informative prior distribution for $\mu_0$ with $\mu_{00} = 0$ and $I_{00} = 0.5$, and a noninformative prior distribution for $\mu_1$ with $I_{10} = 0$. The treatment difference $\theta$ thus has a non-informative prior distribution with $I_0 = 0$.

## Results

### Comparison of frequentist and Bayesian group-sequential approaches - single parameter case

In this section we consider the setting in which we either have a single sample or are comparing two groups but specify a prior distribution for the treatment effect, $\theta$, directly rather than giving separate prior distributions for $\mu_1$ and $\mu_0$. As noted above, in this case the two-sample setting is essentially identical to the single-sample settings, so that we will consider only the latter specifically.

Suppose that the maximum number of looks, $K$, the information at these looks, $I_1, \ldots, I_K$ and, for the Bayesian design, the prior distribution parameters, $\theta_0$ and $I_0$ are specified.

The posterior distribution for $\theta$ at look $k$ in this case is given by (2) so that the posterior probability that $\theta$ exceeds 0 is given by

$$Pr(\theta > 0 \mid \bar{y}_k, I_k) = 1 - \Phi\left(\frac{-\bar{y}_k I_k - \theta_0 I_0}{\sqrt{I_0 + I_k}}\right). \tag{8}$$

Given some choice of $p_1, \ldots, p_K$, for the Bayesian design using stopping criterion (4) expression (8) means that the trial will be stopped at look $k$ if $\bar{Y}_k\sqrt{I_k} \geq u_k^B$ where

$$u_k^B = \frac{-\theta_0 I_0 - \sqrt{I_0 + I_k}\,\Phi^{-1}(1 - p_k)}{\sqrt{I_k}} \tag{9}$$

so that the Bayesian trial, like the frequentist one, will stop whenever $\bar{Y}_k$, or equivalently the standardised $\bar{Y}_k\sqrt{I_k}$, is sufficiently large.

#### *Sequential tests with general $\alpha_1, \ldots, \alpha_K$ or $p_1, \ldots, p_K$*

With $u_k^B$ as given by (9), let $\alpha_k^B = Pr(\bar{Y}_{k'}\sqrt{I_k} \geq u_{k'}^B \text{ some} k' \leq k; \theta = 0)$. This may be calculated from the multivariate normal distribution of $\bar{Y}_1\sqrt{I_1}, \ldots, \bar{Y}_k\sqrt{I_k}$ following from (1). Setting $k = K$ enables analytic calculation of the frequentist type I error rate for the Bayesian test.

Setting $\alpha_k = \alpha_k^B$ and constructing a frequentist design using these $\alpha_1, \ldots, \alpha_K$ values will give a frequentist group-sequential boundary identical to the Bayesian one.

Similarly, given frequentist group sequential spending function values $\alpha_1, \ldots, \alpha_K$, we can obtain $u_1, \ldots, u_K$ to satisfy (7). A Bayesian design with $p_k = 1 - \Phi\left((-u_k\sqrt{I_k} - \theta_0 I_0)/\sqrt{I_0 + I_k}\right)$, $k = 1, \ldots, K$, will then be identical to this frequentist one.

Thus, as noted by Emerson et al. [12], if we allow full flexibility over the choice of $p_1, \ldots, p_K$ for the Bayesian group-sequential design and $\alpha_1, \ldots, \alpha_K$ for the frequentist design, subject respectively to the constraint on overall type I error rate (5) or (6), the classes of frequentist group sequential and Bayesian designs are identical.

Similarly, if Bayesian sequential boundaries are constructed using the posterior probability that $\theta$ exceeds a

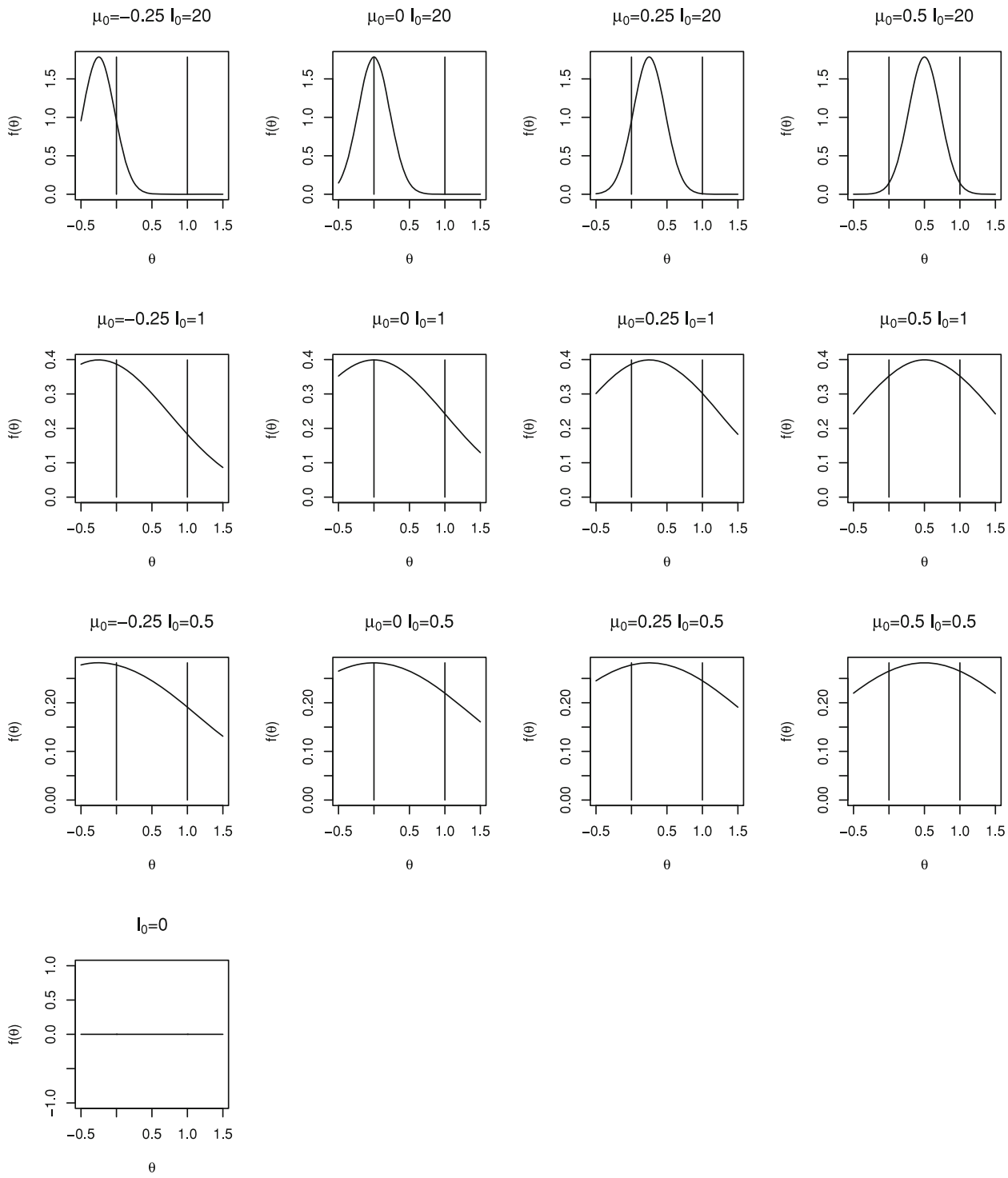


**Fig. 1** Densities for range of prior distributions for Bayesian sequential designs for Example 1

positive target value or the posterior predictive probability of a final positive result, the fact that both of these are monotonically increasing in $\bar{Y}_k$ means that the stopping boundaries are again of the form $\bar{Y}_k\sqrt{I_k} \geq u_k^B$ for some $u_1^B, \ldots, u_K^B$, so that these still correspond to a frequentist boundary for appropriate choice of $\alpha_1, \ldots, \alpha_K$ and vice versa [12]. The same result holds for sequential tests based on Bayes factors provided these are constructed so as to

be monotonically increasing in $\bar{Y}_k$, as is the case, for example, when a point null at $\theta = 0$ is compared to a 'one-sided' prior with support for positive $\theta$ only.

#### *Specific group-sequential tests: Single-arm trial with normally distributed data*

Although in principle, $p_1, \ldots, p_K$ and $\alpha_1, \ldots, \alpha_K$ may be chosen arbitrarily, in practice, constraints may be put on the values used. In this case frequentist and Bayesian group sequential tests may not correspond. In this section we construct frequentist group-sequential designs with a linear alpha spending function and with alpha spending functions corresponding to the Pocock design and the O'Brien and Fleming design, comparing these with Bayesian tests with stopping criteria given by (4) with $p_1 = \cdots = p_K$.

Consider Example 1 above with the range of prior distributions illustrated in Fig. 1. In each case we used stopping criterion (4) and took $p_1 = \cdots = p_K$, finding the common value to give overall type I error rate of $\alpha = 0.025$.

Figure 2 shows critical values, $u_1^B, \ldots, u_5^B$, (plotted as circles) for the Bayesian tests with different prior distributions. Each plot corresponds to a different prior distribution, the layout of plots in the figure matching those in Fig. 1. Note that a different scale is used for the plots in the uppermost row. Using a similar format, Fig. 3 shows the cumulative type I error spent by each look for the tests shown in Fig. 2. Critical values and cumulative type I error spent are also given in Table 1.

It can be seen that more informative or more negative priors lead to a smaller chance of stopping at earlier interim analyses; this makes sense as more information is required to overcome the prior and obtain a posterior probability $pr(\theta > 0 \mid \bar{y}_k) \geq p_k$. Other than for the most informative priors considered, it appears that the choice of $\theta_0$ has relatively little impact; in these cases the value of $I_0$ is small relative to $I_K$ so that the prior distribution makes relatively little contribution to the posterior distribution and hence to the stopping decision.

Figures 2 and 3 and Table 1 also show stopping boundaries and type I error spending functions for O'Brien and Fleming's test, Pocock's test and the frequentist test with a linear spending function, that is with $\alpha^*(t) = \alpha t$, for five equally-spaced analyses. Boundary values and type I error spent at each look for the different tests (omitting those with $I_0 = 20$ and $\theta_0 > -0.025$) are also given in Table 1, together with the value of $p_1 = \cdots = p_K$ required to give overall type I error rate of 0.025 for the Bayesian designs.

It can be seen that stopping boundaries and type I error spent for the O'Brien and Fleming test are nearly identical to those for the Bayesian test with prior distribution with $\theta_0 = -0.25$ and $I_0 = 20$. In this case the form of the stopping boundary, with stopping very unlikely at interim analyses but relatively likely at the final analysis, is only achieved if very strong negative prior opinion is held. This prior distribution was included specifically because of this similarity; it is hard to imagine anyone conducting a trial if they had such a strongly negative prior opinion of the effect of the treatment under investigation.

The similarity between Pocock's test and the Bayesian test with a non-informative prior distribution for $\theta$ can also be noted. For a non-informative prior, that is with $I_0 = 0$, (9) gives $u_k^B = -\Phi^{-1}(1 - p_k)$ so that taking $p_1 = \cdots = p_K$ corresponds to taking $u_1^B = \cdots = u_K^B$. Thus in this case the Bayesian test with $p_k$ chosen to control the overall error rate is identical to Pocock's test when looks are equally spaced in terms of information.

For moderately informative prior distributions, that is for $I_0$ equal to 0.5 or 1, the Bayesian test appears to be similar to the frequentist test with $\alpha^*(t) = \alpha t$ for the reasonably wide range of $\theta_0$ values considered.

#### *Specific group-sequential tests: Single-arm trial with binary data*

Consider next Example 2 above. In this case a Bayesian sequential test can be based on the exact binomial distribution of the data. In detail, denoting by $X_k$ the number of successes observed from the $n_k$ patients observed up to look $k$, $k = 1, \ldots, 4$, we can take $X_k \sim Bin(n_k, \pi)$. A beta prior distribution is conjugate and a non-informative prior is $\pi \sim Beta(1, 1)$, or equivalently $\pi \sim U[0, 1]$. The posterior distribution at look $k$ after observing $X_k = x_k$ is then $\pi \mid x_k, n_k \sim Beta(x_k + 1, n_k - x_k + 1)$.

To be consistent with the notation above, where $\theta$ denotes the treatment effect with $\theta = 0$ corresponding to the null hypothesis, we can take $\theta = \pi - \pi_0$. The trial will stop to claim that $\theta > 0$, or equivalently, $\pi > \pi_0$, if the posterior probability $Pr(\pi > \pi_0 \mid x_k, n_k) \geq p_k$ for some $p_k$.

Taking $p_1 = \cdots = p_k$, for a given value of $p_1$, critical values in terms of the required number of successes at each look can be found by calculating this posterior probability for a range of possible $x_k$ values. These in turn can be used to calculate the resulting frequentist type I error rate under the null hypothesis $H_0 : \theta = 0$ or equivalently in this case, $\pi = \pi_0 = 0.5$, either by simulation or calculation and summation of the appropriate binomial probabilities. A numerical search can then be used to find the value of $p_1$ at which the type I error rate is controlled at a specified level.

For a four-look test with a non-informative $Beta(1, 1)$ prior distribution for $\pi$, the type I error rate is controlled at level 0.05 for $p_1 = \cdots = p_4 = 0.977$. The critical values for the test in terms of the total number of successes observed at looks 1 to 4 are then respectively 18, 33, 47 and 61.

A frequentist group-sequential analysis can be based on the normal approximation (1) for $\hat{\theta} = X_k/n_k -$

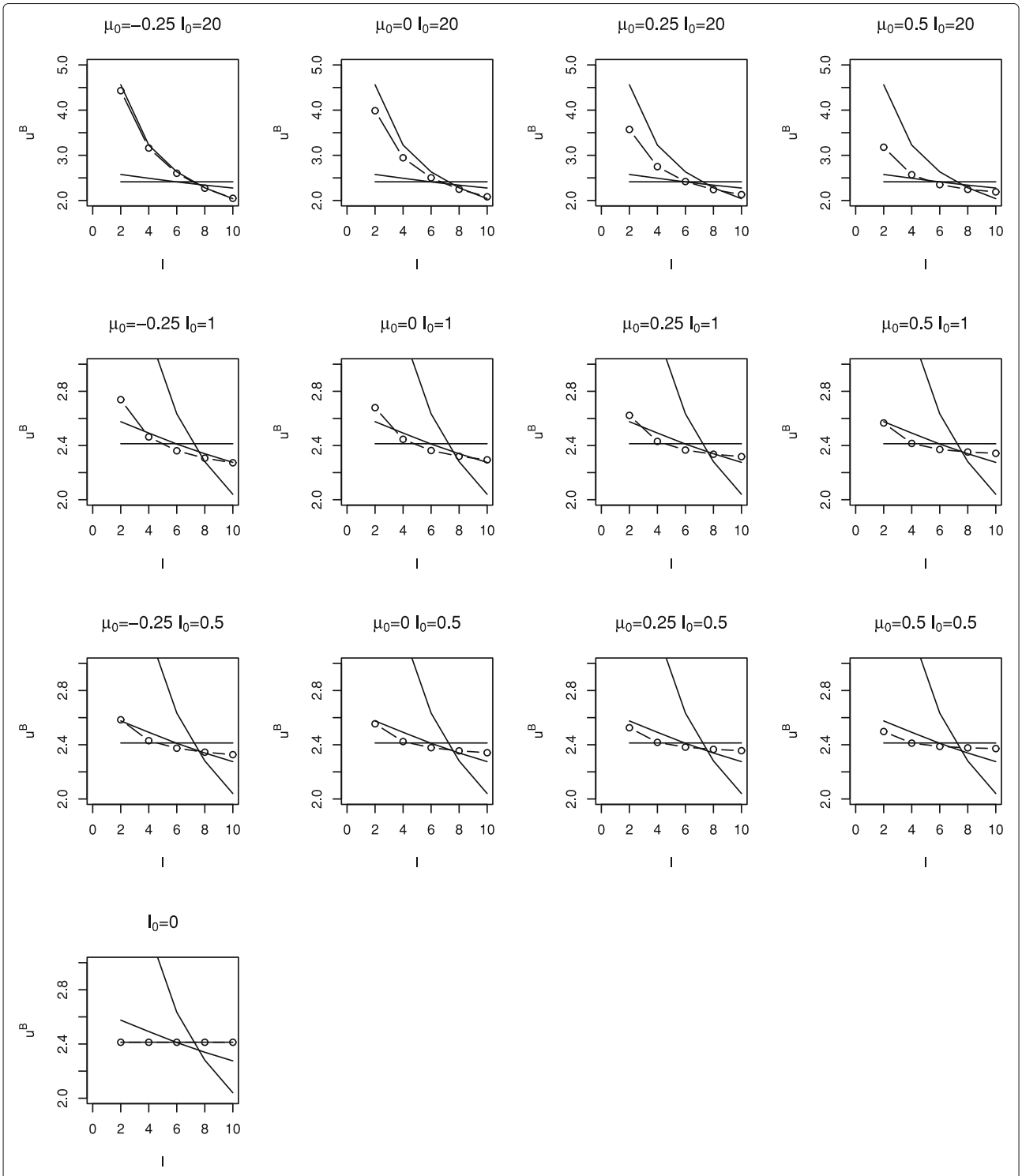


**Fig. 2** Stopping boundaries for Bayesian sequential tests with 5 looks using prior distributions from Figure 1 (∘). Solid lines give boundaries for O'Brien and Fleming test (steep sloping lines), Pocock test (horizontal lines) and for frequentist test with $\alpha^*(t) = \alpha t$ (shallow sloping lines)

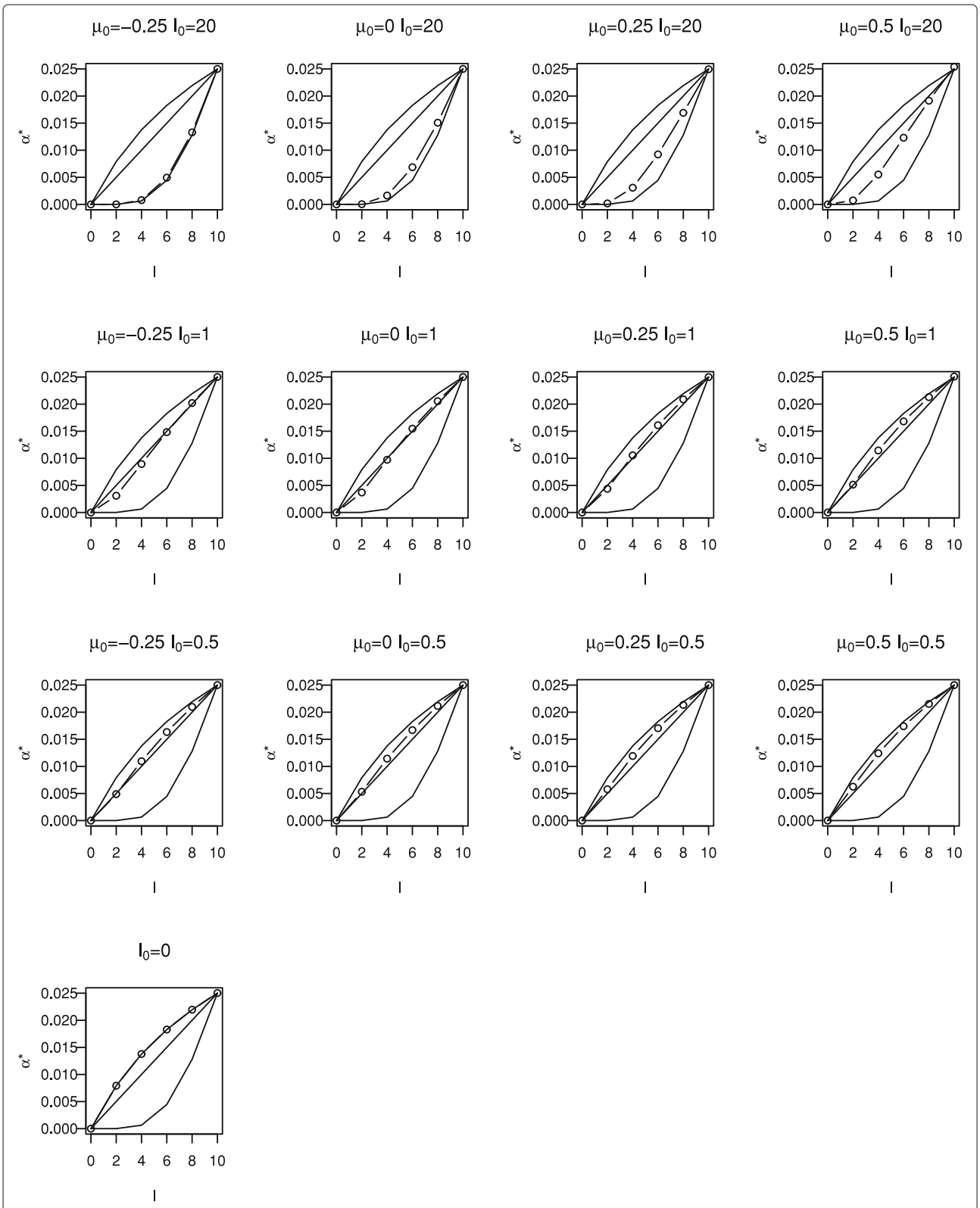


**Fig. 3** Cumulative type I error spent for Bayesian sequential tests shown in Fig. 2 (o). Solid lines give boundaries for O'Brien and Fleming test (lower line), Pocock test (upper line) and for frequentist test with $\alpha^*(t) = \alpha t$ (middle line)

**Table 1** Boundary values and type I error rate spent for Bayesian and frequentist five-look group sequential tests

| Bayesian tests | | | | |
|---|---|---|---|---|
| $I_0$ | $\theta_0$ | $p_1 = \cdots = p_5$ | $u_1^B, \ldots, u_5^B$ | $\alpha_1^B, \ldots, \alpha_5^B$ |
| 20.0 | -0.25 | 0.6063 | 4.43, 3.16, 2.60, 2.27, 2.05 | 0.0000, 0.0008, 0.0049, 0.0133, 0.0250 |
| 1.0 | -0.25 | 0.9818 | 2.74, 2.46, 2.36, 2.31, 2.27 | 0.0031, 0.0089, 0.0148, 0.0202, 0.0250 |
| 1.0 | 0.00 | 0.9856 | 2.68, 2.45, 2.36, 2.32, 2.29 | 0.0037, 0.0097, 0.0155, 0.0205, 0.0250 |
| 1.0 | 0.25 | 0.9889 | 2.62, 2.43, 2.37, 2.34, 2.32 | 0.0044, 0.0105, 0.0161, 0.0209, 0.0250 |
| 1.0 | 0.50 | 0.9914 | 2.57, 2.42, 2.37, 2.35, 2.34 | 0.0051, 0.0114, 0.0168, 0.0213, 0.0251 |
| 0.5 | -0.25 | 0.9872 | 2.58, 2.43, 2.37, 2.35, 2.33 | 0.0049, 0.0110, 0.0163, 0.0210, 0.0250 |
| 0.5 | 0.00 | 0.9888 | 2.55, 2.42, 2.38, 2.36, 2.34 | 0.0053, 0.0114, 0.0167, 0.0212, 0.0250 |
| 0.5 | 0.25 | 0.9903 | 2.53, 2.42, 2.38, 2.37, 2.36 | 0.0058, 0.0119, 0.0171, 0.0213, 0.0250 |
| 0.5 | 0.50 | 0.9916 | 2.50, 2.41, 2.39, 2.38, 2.37 | 0.0063, 0.0124, 0.0174, 0.0215, 0.0250 |
| 0.0 | 0.00 | 0.9921 | 2.41, 2.41, 2.41, 2.41, 2.41 | 0.0079, 0.0138, 0.0183, 0.0220, 0.0250 |
| Frequentist tests | | | | |
| | | | $u_1, \ldots, u_5$ | $\alpha_1, \ldots, \alpha_5$ |
| O'Brien & Fleming | | | 4.56, 3.23, 2.63, 2.28, 2.04 | 0.0000, 0.0006, 0.0045, 0.0128, 0.0250 |
| Pocock | | | 2.41, 2.41, 2.41, 2.41, 2.41 | 0.0079, 0.0138, 0.0183, 0.0219, 0.0250 |
| $\alpha^*(t) = \alpha t$ | | | 2.58, 2.49, 2.41, 2.34, 2.28 | 0.0050, 0.0100, 0.0150, 0.0200, 0.0250 |

$\pi_0$ and $I_k^{-1} = \pi_0(1-\pi_0)/n_k$. A four-look frequentist group-sequential Pocock test constructed based on this approximation would stop for $\hat{\theta}_k\sqrt{I_k} \geq u_k$ with $u_k = 2.067$, that is for $X_k \geq 0.5n_k + 2.067\sqrt{n_k}/2$, giving stopping boundaries in terms of $X_k$ for $n_k = 25, 50, 75$ and 100 of 17.7, 32.3, 46.5 and 60.3. Rounding these up to integers gives stopping boundary values identical to those for the Bayesian test with a non-informative prior distribution.

#### *Specific group-sequential tests: Two-arm trial with normally distributed data*

We next consider Example 3 above, using only the prior information given by the prior distribution for the treatment difference $\theta$, that is the non-informative prior distribution with $I_0 = 0$.

The distribution of the observed difference between the treatment means at looks 1 to $K, D_1, \ldots, D_K$ follows a multivariate normal distribution of the same form as that of the mean values $\bar{Y}_1, \ldots, \bar{Y}_K$ in the single-group case, with $I_k$ now taken to be $n_k/2\sigma^2$. Setting $p_1 = \cdots = p_K$ and taking this value so as to control the overall type I error rate to be 0.025, thus gives critical values, $u_k$, now for $D_k\sqrt{I_k}$, equal to 2.41 at all looks, exactly as in single-arm case with a non-informative prior distribution for $\theta$.

### Comparison of frequentist and Bayesian group-sequential approaches - two parameter case

Consider now the setting in which we are comparing two groups of normally distributed data and, in the Bayesian setting, specify separate independent normal prior distributions for $\mu_1$ and $\mu_0$.

Suppose that the prior distributions are given by $\mu_j \sim N\left(\mu_{j0}, I_{j0}^{-1}\right), j = 0, 1$. Given observation of $\bar{Y}_{jk} = \bar{y}_{jk}$, the posterior distribution for $\mu_j$ is given by

$$\mu_j \mid \bar{y}_{jk} \sim N\left(\frac{\mu_{j0}I_{j0} + \bar{y}_{jk}I_{jk}}{I_{j0} + I_{jk}}, \frac{1}{I_{j0} + I_{jk}}\right).$$

As $\mu_0$ and $\mu_1$ have independent prior distributions, their posterior distributions are also independent, so that the posterior distribution for $\theta$ is given by

$$\theta \mid \bar{y}_{1k}, \bar{y}_{0k} \sim N\left(\frac{\mu_{10}I_{10} + \bar{y}_{1k}I_{1k}}{I_{10} + I_{1k}} - \frac{\mu_{00}I_{00} + \bar{y}_{0k}I_{0k}}{I_{00} + I_{0k}}, \frac{1}{I_{10} + I_{1k}} + \frac{1}{I_{00} + I_{0k}}\right). \quad (10)$$

Note that although in this case the prior distribution for $\theta$ is again normal, with $\theta \sim N(\theta_0, I_0)$ with $\theta_0 = \mu_{10} - \mu_{00}$ and $I_0^{-1} = I_{10}^{-1} + I_{00}^{-1}$, the posterior distribution given by (10) is not generally the same as (3) that was obtained when the prior distribution for $\theta$ was considered directly.

It is shown in Appendix A that the posterior variance of $\theta$ when separate prior distributions are given for $\mu_1$ and $\mu_0$ given by (10) is always smaller than that given by (3) when only the prior distribution for $\theta$ is used. With independent prior distributions for $\mu_1$ and $\mu_0$, the posterior distribution depends on $\bar{y}_{1k}$ and $\bar{y}_{0k}$, and not just on the difference $d_k = \bar{y}_{1k} - \bar{y}_{0k}$. Assuming $\mu_1$ and $\mu_0$ are independent means that $\theta$ is not independent of $\mu_1 + \mu_0$. Thus although $D_k$ is sufficient for $\theta$, we can also learn about $\theta$ by learning about $\mu_1 + \mu_0$, for which $D_k$ is not sufficient. We therefore gain information by knowing $\bar{y}_{1k} + \bar{y}_{0k}$ as well as $\bar{y}_{1k} - \bar{y}_{0k}$, that is by having information on both $\bar{y}_{1k}$ and $\bar{y}_{0k}$, leading to a smaller posterior variance.

Suppose that, as in the single parameter case, we stop the trial as soon as we have $Pr(\theta > 0 \mid \text{data at look } k) \geq p_k$, and that we wish to choose $p_1, \ldots, p_K$ so as to control the type I error rate to be at most $\alpha$, that is to satisfy (5).

It is shown in Appendix B that, irrespective of the values of $p_1, \ldots, p_k$, the stopping regions for frequentist and Bayesian group-sequential tests cannot coincide other than in the special case with $I_{1k}/(I_{10} + I_{1k}) = I_{0k}/(I_{00} + I_{0k}), k = 1, \ldots, K$, when the posterior distribution for $\theta$ is exactly the same as that obtained directly from a single prior distribution for $\theta$ without considering prior distributions for the means of the two groups separately,

With independent prior distributions for $\mu_1$ and $\mu_0$ the posterior distribution of $\theta$ depends on $\bar{y}_{1k}$ and $\bar{y}_{0k}$. The probability in (5) thus depends on $\mu_0$ and $\mu_1$ and the requirement that this is controlled at level $\alpha$ when $\theta = 0$ requires that it is controlled when $\mu_1 = \mu_0$ for all values of $\mu_0$. Appendix B shows that beacuse the mean of the posterior distribution for $\theta$ when $\mu_1 = \mu_0$ depends on $\mu_0$, this is impossible.

For the two-arm Bayesian group-sequential trial with five looks in Example 3 above, controlling the one-sided type I error rate to be 0.025 when $\mu_1 = \mu_0 = 0$ requires $p_1 = \cdots = p_5 = 0.9884$.

Figure 4 shows the one-sided type I error rate for this design for a range of $\mu_0$ values with, in each case, $\mu_1 = \mu_0$ so that $\theta = 0$. It can be seen that in this case although the type I error rate is controlled for $\mu_0 = 0$, the type I error rate increases above the desired level for $\mu_0 > 0$. The figure also shows the prior distribution for $\mu_0$, showing that error rate inflation would occur for plausible values of $\mu_0$.

## Discussion

Our comparison has been restricted on the whole to group-sequential tests based on normally distributed test statistics. Although some exact or non-normal frequentist group-sequential test methods have been proposed [27–29] the assumption of normality is common in this setting. In Bayesian group-sequential tests it is more common to use non-normal distributions, with simulation methods being used if necessary to calculate operating characteristics. The decision to focus on normally distributed test statistics was made so as to put Bayesian and frequentist designs in a similar setting, facilitate comparison and identify relationships, such as that between the Pocock test and the Bayesian test with a non-informative prior distribution, which might otherwise not be apparent. As can be seen from the binary data example above, where the Pocock test and the exact Bayesian test give identical stopping rules, in practice asymptotic normality can be a reasonable assumption.

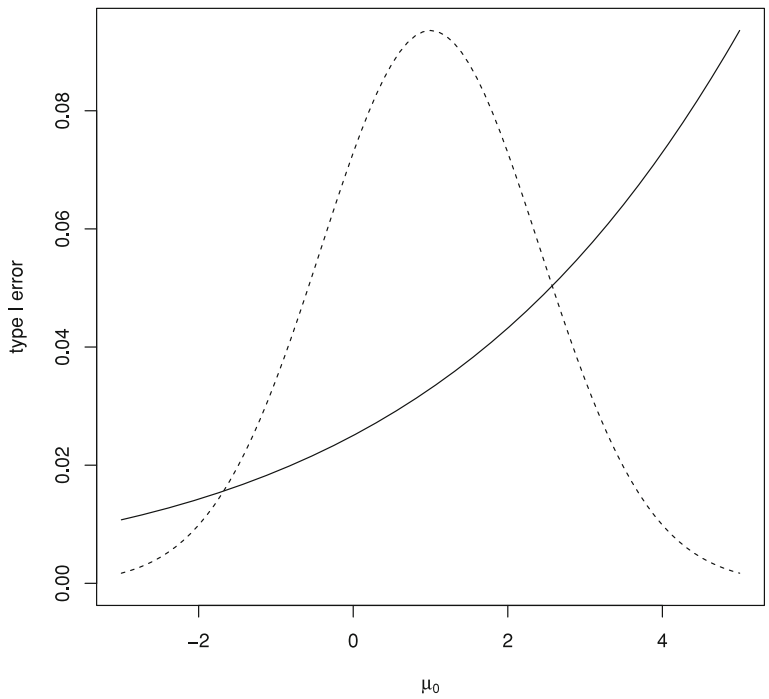


**Fig. 4** Type I error rate for Bayesian test with $K = 5$ and $p_1 = \cdots = p_5 = 0.9884$ for range of true $\mu_0$ values along with density (not to scale) for the prior distribution for $\mu_0$

We have considered stopping for a positive result only. In practice, with both frequentist and Bayesian group-sequential designs, it is often desirable to allow stopping when a lack of efficacy is clear, that is for futility. Futility stopping rules can be divided into those that are binding, when the rule is specified in advance and must be adhered to in order to maintain the required properties of the design, and those that are non-binding, where a more flexible approach can be taken. As stopping for futility cannot lead to a positive claim of efficacy, it can only decrease the type I error rate. Thus with a non-binding futility stopping rule, it is desirable to control the type I error rate even if no futility stopping occurs, that is in the case when the trial is only stopped for a positive result as considered above. The use of a binding futility stopping rule will change the operating characteristics of the group-sequential tests.

We have focussed on comparison of Bayesian and frequentist group-sequential designs for single-arm and comparative studies. These are just one type of adaptive design, which can include many other features including adaptive exploration of a dose-response relationship, adaptive randomisation, dropping of arms in multi-arm trials, incorporation of multiple endpoints and sample size reestimation. Frequentist methods that guarantee control of error rates are available for some of these problems such as sample size re-estimation [30] but in some other cases construction of decision rules for frequentist methods can be challenging. Bayesian methods can be accompanied by simulations to verify operating characteristics under a likely range of scenarios for a wide variety of adaptations for which rigorous proof of error rate control is not available.

## Conclusions

Although Bayesian and frequentist group-sequential approaches are based on fundamentally different paradigms, in practice, when used for the analysis of a clinical trial, both provide an indication of the efficacy of an experimental treatment. This means that a comparison of Bayesian and frequentist test can be helpful to understand the frequentist operating characteristics for Bayesian tests and the Bayesian model and prior distributional assumptions that could lead to a particular frequentist test. This has been our aim in this paper.

Focussing on a setting in which test statistics can be assumed to be normally distributed, we have shown that in comparative trials with independent prior distributions specified for treatment effects in different groups, stopping rules from Bayesian and frequentist group-sequential designs cannot generally correspond. In this case the Bayesian group-sequential design can then only control the type I error rate for specified values of the control group treatment effect. Conversely, in single-arm trials, or when a prior distribution is specified for the treatment difference, stopping rules for Bayesian and frequentist group-sequential tests can be identical if full flexibility for both classes of designs is allowed, or can closely correspond for common choices of design parameters.

O'Brien and Fleming's design was found to correspond closely to a Bayesian design with an exceptionally informative negative prior, this prior leading to the very small probability of early stopping for this design. The fact that such a prior is unlikely to represent prior belief suggests that the use of this design might not be appropriate without very careful thought.

In a similar way, noting that the Bayesian design with a non-informative prior and $p_1 = \cdots = p_K$ corresponds to a Pocock design suggests that this might also not be generally appropriate given the criticism that this design gives too high a probability of early stopping [31]. This illustrates the importance of appropriate choice of a prior distribution, rather than the general use of a non-informative prior. Evaluation of the frequentist properties can be useful in understanding the influence of the prior distribution in a Bayesian group-sequential design in which the overall type I error rate is controlled.

Bayesian adaptive methods are often more bespoke than frequentist approaches, with simulations used to evaluate their performance not only for a range of treatment effect scenarios but also allowing for anticipated data patterns arising from, for example, delayed responses, multiple endpoints including early outcomes, or different recruitment and drop-out rates. This can require more design work than the use of a more standard frequentist method but can be advantageous in that design choices and their consequences are considered carefully. It is recommended that if frequentist methods are used, equal care should be taken over design choices and their properties explored, using simulations if necessary.

## Appendix A: Comparison of posterior variances for comparative trials with single or independent prior distributions

Suppose we are in the two-group setting and have independent prior distributions with $\mu_j \sim N\left(\mu_{j0}, I_{j0}^{-1}\right), j = 0,1$ and that we have observation of $\bar{Y}_{jk}$ with $\bar{Y}_{jk} \sim N(\theta, I_k^{-1}), j = 0,1, k = 1,\ldots,K$, so that the posterior distribution for $\theta$ is given by (10).

Considering only the single parameter $\theta$, the posterior distribution is given by (10) with $\theta_0 = \mu_{10} - \mu_{00}$, $I_0^{-1} = I_{00}^{-1} + I_{10}^{-1}$ and $I_k = \left(I_{0k}^{-1} + I_{1k}^{-1}\right)^{-1}$.

Let $I_{[1]k}$ and $I_{[2]k}$ denote the inverses of the posterior variance for $\theta$ in the one-parameter and two-parameter cases respectively. We will show that $I_{[1]k} \leq I_{[2]k}$.

We will denote by $r_0$ the ratio $I_{10}/I_{00}$, so that $I_{10} = r_0 I_{00}$, by $r_k$ the ratio $I_{1k}/I_{0k}$, and by $\Lambda_k$ the ratio $r_k/r_0$ so that

$r_k = \Lambda_k r_0$ and $I_{1k} = \Lambda_k r_0 I_{0k}$. Without loss of generality, we will take $I_{0k} = 1$ so that $I_{1k} = \Lambda_k r_0$. We then have $I_{[1]k} = I_{00}/\left(1 + r_0^{-1}\right) + 1/\left(1 + (\Lambda_k r_0)^{-1}\right)$ and $I_{[2]k} = 1/\left((I_{00}+1)^{-1} + (r_0 I_{00} + r_0\Lambda_k)^{-1}\right)$.

Letting $R_k$ denote the ratio $I_{[1]k}/I_{[2]k}$ and differentiating this with respect to $\Lambda_k$ yields

$$\frac{dR_k}{d\Lambda_k} = \frac{r_0 I_{00}(a\Lambda_k^2 + b\Lambda_k + c)}{(I_{00}+1)(1+r_0)(I_{00}+\Lambda_k)^2(\Lambda_k r_0 + 1)^2}.$$

with $a = -(r_0 I_{00} + 2r_0 + 1)$, $b = 2(r_0 - I_{00})$ and $c = I_{00}(r_0 + 2) + 1$. Note that the derivative is defined for all $\Lambda_k \geq 0$ as $I_{00}$ and $r_0$ are both positive. Setting the numerator to zero and solving the quadratic, we find that $R_k$ has stationary points at $\Lambda_k = 1$ and $-(r_0 I_{00} + 2I_{00} + 1)/(r_0 I_{00} + 2r_0 + 1)$. The second of these is negative as $I_{00}$ and $r_0$ are positive, so that the only stationary point with $\Lambda_k \geq 0$ is at $\Lambda_k = 1$ when $R_k = 1$

The second derivative of $R_k$ with respect to $\Lambda_k$ at $\Lambda_k = 1$ is equal to $-2r_0 I_{00}(I_{00}+1)^{-2}(r_0+1)^{-2}$, and so is negative, confirming that the turning point is a maximum so that $R_k \leq 1$, and hence $I_{[1]k} \leq I_{[2]k}$, as stated.

## Appendix B: Type I error rate for Bayesian comparative trial with independent prior distributions

The requirement (5) that the error rate is controlled at level $\alpha$ in the two-paramter case can be stated as

$$Pr(Pr(\theta > 0 \mid \bar{Y}_{1k}, \bar{Y}_{0k}) \geq p_k \text{ some } k \leq K; \mu_1 = \mu_0) \leq \alpha \text{ for all } \mu_0. \tag{11}$$

We can rewrite the posterior distribuion (10) as $\theta \mid \bar{y}_{1k}, \bar{y}_{0k} \sim N\left(M_k, I_{[2]k}^{-1}\right)$ with $I_{[2]k}^{-1} = (I_{10} + I_{1k})^{-1} + (I_{00} + I_{0k})^{-1}$ and

$$M_k = \frac{\mu_{10} I_{10} + \bar{y}_{1k} I_{1k}}{I_{10} + I_{1k}} - \frac{\mu_{00} I_{00} + \bar{y}_{0k} I_{0k}}{I_{00} + I_{0k}}. \tag{12}$$

The posterior probability $pr(\theta > 0 \mid \bar{y}_{1k}, \bar{y}_{0k})$ is thus equal to $1 - \Phi\left(-M_k I_{[2]k}^{1/2}\right)$. This exceeds $p_k$ whenever $M_k \geq -\Phi^{-1}(1-p_k) I_{[2]k}^{-1/2}$.

Hence in this case the stopping decision for the Bayesian sequential test depends on $\bar{Y}_{1k}$ and $\bar{Y}_{0k}$ via $M_k$ and the frequentist operating characteristics for the Bayesian sequential test can be obtained from the joint distribution of $M_1, \ldots, M_K$.

It follows from (12) and (1) that $M_1, \ldots, M_K$ are multivariate normal with

$$E(M_k) = \frac{\mu_{10} I_{10} + \mu_1 I_{1k}}{I_{10} + I_{1k}} - \frac{\mu_{00} I_{00} + \mu_0 I_{0k}}{I_{00} + I_{0k}}.$$

When $\mu_1 = \mu_0$, we have

$$E(M_k) = \frac{\mu_{10} I_{10}}{I_{10} + I_{1k}} - \frac{\mu_{00} I_{00}}{I_{00} + I_{0k}} + \mu_0 \left(\frac{I_{1k}}{I_{10} + I_{1k}} - \frac{I_{0k}}{I_{00} + I_{0k}}\right).$$

If $\frac{I_{1k}}{I_{10}+I_{1k}} - \frac{I_{0k}}{I_{00}+I_{0k}} > 0$, we have $E(M_k) \to \infty$ as $\mu_0 \to \infty$, and if $\frac{I_{1k}}{I_{10}+I_{1k}} - \frac{I_{0k}}{I_{00}+I_{0k}} < 0$, we have $E(M_k) \to \infty$ as $\mu_0 \to -\infty$. In neither of these cases, then, is it possible to satisfy (11) for all values of $\mu_0$ other than in the trival case with $p_1 = 1$, when stopping is impossible.

### Abbreviations

CDAI: Crohn's disease activity index

### Acknowledgements

The authors are grateful to an Editor and two referees for their comments on an earlier draft of the paper.

### Authors' contributions

NS conceived and undertook the research with feedback from all other authors. NS, ST, ER and SG discussed and commented on the manuscript. NS, ST, ER and SG read and approved the final manuscript.

### Funding

NS, SG and EGR were supported by a Medical Research Council (MRC) Methodology Research Grant (Grant number: MR/N02828/1) during the conduct of this research. The funder had no role in the design of the study, collection, analysis and interpretation of data, or in writing the manuscript.

### Availability of data and materials

Not applicable: no data or materials were used in this research.

### Ethics approval and consent to participate

Not applicable: no patient data were used in this research.

### Consent for publication

Not applicable: no patient data were used in this research.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1]Statistics and Epidemiology, Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, UK. [2]Department of Mathematics and Statistics, University of Reading, Reading, UK. [3]Cancer Research UK Clinical Trials Unit, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK.

Received: 20 September 2019 Accepted: 23 December 2019
Published online: 07 January 2020

### References

1. Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. Boca Raton: Chapman & Hall; 2000.
2. Berry SM, Carlin BP, Lee JJ, Müller P. Bayesian Adaptive Methods for Clinical Trials. Boca Raton: CRC Press; 2011.
3. Zhu H, Yu Q. A Bayesian sequential design using alpha spending function to control type I error. Stat Methods Med Res. 2017;26:2184–69.
4. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials. J R Stat Soc Ser A. 1994;157:357–416.
5. Ryan EG, Bruce J, Metcalfe AJ, Stallard N, Lamb SE, Viele K, Young D, Gates S. Using Bayesian adaptive designs to improve phase III trials: a respiratory care example. BMC Med Res Methodol. 2019;19:99.
6. Food and Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry. 2019. https://www.fda.gov/media/78495/download. Accessed: 3 Jan 2020.
7. Hueber W, Sands BE, Lewitzky S, Vandemeulebroecke M, Reinisch W, Higgins PDR, Wehkamp J, Feagan BG, Yao MD, Karczewski M, Karczewski J, Pezous N, Bek S, Bruin G, Mellgard B, Berger C, Londei M, Bertolino AP, Tougas G, Travis SPL. Secukinumab, a human anti-IL-17A monoclonal antibody, for moderate to severe crohn's disease: unexpected results of a randomised, double-blind placebo-controlled trial. Gut. 2012;61:1693–700.
8. Gsponer T, Gerber F, Bornkamp G, Ohlssen D, Vandemeulebroecke M, Schmidli H. A practical guide to Bayesian group sequential designs. Pharm Stat. 2014;13:71–80.

9. Gerber F, Gsponer T. Package 'gsbDesign'. 2016. http://CRAN.R-project.org/web/packages/gsbDesign/gsbDesign.pdf. Accessed: 3 Jan 2020.
10. Wilber DJ, Pappone C, Neuzil P, Paola AD, Marchlinski F, Natale A, Macle L, Daoud EG, Calkins H, Hall B, Reddy V, Augello G, Reynolds MR, Vinekar C, Liu CY, Berry SM, Berry DA. Comparison of antiarrhythmic drug therapy and radiofrequency catheter ablation in patients with paroxysmal atrial fibrillation. J Am Med Assoc. 2010;303:333–40.
11. Emerson SS, Kittelson JM, Gillen DL. Bayesian evaluation of group sequential clinical trial designs. Stat Med. 2007;26:1431–49.
12. Emerson SS, Kittelson JM, Gillen DL. Frequentist evaluation of group sequential clinical trial designs. Stat Med. 2007;26:5047–80.
13. Campbell G. Similarities and differences of Bayesian designs and adaptive designs for medical devices: a regulatory view. Stat Biopharm Res. 2013;5:356–68.
14. Shi H, Yin G. Control of type I error rates in Bayesian sequential designs. Bayesian Anal. 2018. https://doi.org/10.1214/18-ba1109.
15. Bernado JM, Smith AFM. Bayesian Theory. Chichester: Wiley; 2000.
16. Jennison C, Turnbull BW. Group sequential analysis incorporting covariate information. J Am Stat Assoc. 1997;92:1330–41.
17. Saville BR, Connor JT, Ayers GD, Alvarez J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. Clin Trials. 2014;11:485–93.
18. Mujagic E, Zwimpfer T, Marti WR, Zwahlen M, Hoffmann H, Kindler C, Fux C, Misteli H, Iselin L, Lugli AK, Nebiker CA, von Holzen U, Vinzens F, von Strauss M, Reck S, Kraljević M, Widmer AF, Oertli D, Rosenthal R, Weber WP. Evaluating the optimal timing of surgical antimicrobial prophylaxis: study protocol for a randomized controlled trial. Trials. 2014;15:188.
19. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977;64:191–9.
20. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979;35:549–56.
21. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Bornkamp B, Mæchler M, Hothorn T. Package 'mvtnorm'. 2018. http://CRAN.R-project.org/web/packages/mvtnorm/mvtnorm.pdf. Accessed: 3 Jan 2020.
22. Slud EV, Wei LJ. Two-sample repeated significance tests based on the modified Wilcoxon statistics. J Am Stat Assoc. 1982;77:862–8.
23. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70:659–63.
24. Proschan M, Lan KKG, Wittes JT. Statistical Monitoring of Clinical Trials: A Unified Approach. New York: Springer; 2006.
25. Kim K, DeMets DL. Design and analysis of group sequential tests based on the type I error spending rate function. Biometrika. 1987;74:149–54.
26. Anderson K. Package 'gsDesign'. 2016. http://CRAN.R-project.org/web/packages/gsDesign/gsDesign.pdf. Accessed: 3 Jan 2020.
27. Jennison C, Turnbull BW. Exact calculations for sequential $t$, $\chi^2$ and $F$ tests. Biometrika. 1991;78:133–41.
28. Stallard N, Todd S. Exact sequential tests for single samples of discrete responses using spending functions. Stat Med. 2000;19:3051–64.
29. Stallard N, Rosenberger WF. Exact group-sequential designs for clinical trials with randomized play-the-winner allocation. Stat Med. 2002;21:467–80.
30. Cui L, Hung HMJ, Wang S-J. Modification of sample size in group sequential clinical trials. Biometrics. 1999;55:853–7.
31. Pocock S, White I. Trials stopped early: too good to be true? Lancet. 1999;353:943–4.


## Publisher's Note