University of Texas, MD Anderson Cancer Center

UT MD Anderson Cancer Center Department of Biostatistics
Working Paper Series

Year 2008 *Paper* 47

Bayesian Design of Single-Arm Phase II Clinical Trials with Continuous Monitoring

Valen E. Johnson* John D. Cook†

*University of Texas M.D. Anderson Cancer Center, vejohnson@mdanderson.org

†University of Texas M. D. Anderson Cancer Center, jdcook@mdanderson.org

http://biostats.bepress.com/mdandersonbiostat/paper47

Bayesian Design of Single-Arm Phase II Clinical Trials with Continuous Monitoring

Valen E. Johnson and John D. Cook

Abstract

Many "Bayesian" clinical trial designs use posterior credible intervals as tools to define stopping boundaries for inferiority, futility, or superiority. However, the thresholds on posterior credible intervals that trigger termination of a trial are determined by frequentist operating characteristics. This practice can result in substantial overlap between the credible intervals associated with, say, stopping a trial for superiority and stopping a trial for inferiority, which severely limits the interpretation of posterior probability statements. In this article, we use formal Bayesian hypothesis tests to design single-arm phase II clinical trials. By using non-local prior densities to define null and alternative models, we obtain exponential convergence of Bayes factors under both null and alternative models. When compared to other commonly used Bayesian and frequentist designs, we show that our method provides better operating characteristics, uses fewer patients per correct decision, and provides more directly interpretable results. We also demonstrate that designs based on Bayesian hypothesis tests eliminates a potential source of bias often associated with Bayesian trial designs.

Bayesian Design of Single-Arm Phase II Clinical Trials with Continuous Monitoring

Valen E. Johnson*

John D. Cook†

July 27, 2008

Abstract

Many "Bayesian" clinical trial designs use posterior credible intervals as tools to define stopping boundaries for inferiority, futility, or superiority. However, the thresholds on posterior credible intervals that trigger termination of a trial are determined by frequentist operating characteristics. This practice can result in substantial overlap between the credible intervals associated with, say, stopping a trial for superiority and stopping a trial for inferiority, which severely limits the interpretation of posterior probability statements. In this article, we use formal Bayesian hypothesis tests to design single-arm phase II clinical trials. By using non-local prior densities to define null and alternative models, we obtain exponential convergence of Bayes factors under both null and alternative models. When compared to other commonly used Bayesian and frequentist designs, we show that our method provides better operating characteristics,

*University of Texas M. D. Anderson Cancer Center, vejohnson@mdanderson.org

†University of Texas M. D. Anderson Cancer Center, cook@mdanderson.org

uses fewer patients per correct decision, and provides more directly interpretable results. We also demonstrate that designs based on Bayesian hypothesis tests eliminates a potential source of bias often associated with Bayesian trial designs.

Keywords: Simon two-stage design, Thall-Simon design, efficacy trial, Bayes factors, Gibbs inequality, non-local prior density

1 Introduction

Bayesian methodology has recently played an increasingly prominent role in the conduct of clinical trials, particularly in early phase clinical trials. Most of this development has relied on Bayesian inferential techniques, usually premised on the assumption of weakly informative or non-informative prior densities. Using such priors, decisions to terminate clinical trials for futility, inferiority, or superiority are made by examining the content of posterior credible intervals on response rates or survival times (e.g., Thall & Simon, 1994; George et al. 1994; Heitjan, 1997; Thall et al. 1995; Fayers et al., 1997; Simon, 1999; Tan & Machin, 2002; Thall et al., 2005). In practice, the thresholds above which the content of a posterior credible interval triggers termination of a trial are determined by frequentist operating characteristics, and the intervals defining superiority and inferiority often overlap. This makes posterior probability statements derived from such trials difficult to interpret.

Paradoxically, although clinical trials represent statistical tests of the relative efficacy of experimental therapies or diagnostic methods, Bayesian testing methodology is seldom used in their design. We assert that this paradox stems from two sources: a misperception surrounding the use of informative prior distributions to define alternative hypotheses, and the resulting substitution of vague prior densities to achieve this purpose.

Bayesian hypothesis tests require the specification of full probability models for data

and parameters under two or more competing models. Each probability model includes a parametric sampling density for data and a prior density on model parameters. In an effort to avoid "biasing" the result of a test, there is a temptation to specify vague or objective prior densities on model parameters under the alternative hypothesis (e.g., Berger & Pericchi, 1996, 1998; Bertolino et al. 2000; O'Hagan, 1995, 1997; Moreno et al. 1998). As we demonstrate below, however, mis-specification of the prior density under the alternative model can only decrease the expected weight of evidence in favor of the alternative model. Thus, there is no danger that proponents of an experimental treatment can bias the results of a Bayesian test in favor of the alternative model by specifying an overly optimistic alternative model.

The use of vague prior specifications to define alternative models leads to the specification of what might be called local alternative hypotheses. Informally speaking, local alternative probability models are models that assign positive probability to regions of the parameter space that are consistent with the null hypothesis. Under regularity conditions stated below, the use of local alternative hypotheses leads to convergence rates of only $O_p(n^{-1/2})$ or slower in favor of true null hypotheses, but convergence at exponential rates in favor of true alternative models. Thus, the use of local alternative hypotheses in clinical trials for efficacy can make it essentially impossible to stop a trial in favor of a null hypothesis of no beneficial treatment effect. It is probably for this reason that most Bayesian clinical trial designs employ stopping rules based on posterior credible intervals rather than posterior model probabilities.

The remainder of this article is organized as follows. In Section 2, we demonstrate that mis-specification of the alternative model in a single-arm phase II trial increases the expected weight of evidence in favor of the null model of no (additional) treatment benefit. In Section 3, we propose a new class of prior densities for the definition of alternative models in single-arm phase II clinical trials with interim monitoring. This class of prior densities

provides exponential convergence of Bayes factors in favor of both true null and true alternative models. Examples of clinical trials designed using tests based on these priors are presented in Section 4. Trials with both binary and time-to-event (TTE) patient outcomes are considered, and comparisons are made to common Bayesian designs based on posterior credible intervals. In Section 5 we compare our design to perhaps the most commonly used frequentist design for trials with interim monitoring, the Simon two-stage design (Simon, 1989). We conclude with discussion and comments in Section 6.

2 An inequality for expected weight of evidence

Let $x_1, \ldots, x_n = \{\mathbf{x}_n\}$ denote independent and identically distributed random variables representing n patient outcomes, and suppose that x_1 has density function $f(\cdot \,|\, \boldsymbol{\theta})$ with respect to a q-dimensional parameter $\boldsymbol{\theta} \in \Theta \subset \mathcal{R}^q$. Define the null and alternative hypotheses according to

$$H_0 : \boldsymbol{\theta} \sim \pi_0(\boldsymbol{\theta}), \qquad\qquad H_1 : \boldsymbol{\theta} \sim \pi_1(\boldsymbol{\theta}), \tag{1}$$

respectively. The function π_1 is assumed to be a continuous probability density function defined with respect to Lebesgue measure, and π_0 is assumed to be either a point mass concentrated at, say, a known response rate θ_0, or is also a continuous probability density function defined with respect to Lebesgue measure. Let $m_i(\mathbf{x}_n)$, $i = 0, 1$, denote the marginal density of the data $\mathbf{x}_n$ under each hypothesis. That is,

$$m_i(\mathbf{x}_n) = \int_\Theta \left[\prod_{j=1}^{n} f(x_j \,|\, \boldsymbol{\theta}) \right] \pi_i(\boldsymbol{\theta}) \; d\boldsymbol{\theta}. \tag{2}$$

If α denotes the prior odds in favor of the alternative hypothesis, then the posterior odds between H_1 and H_0 given $\mathbf{x}_n$ can be expressed

$$\frac{\mathbf{Pr}(H_1 \,|\, \mathbf{x}_n)}{\mathbf{Pr}(H_0 \,|\, \mathbf{x}_n)} = \frac{m_1(\mathbf{x}_n)}{m_0(\mathbf{x}_n)} \times \frac{\alpha}{1-\alpha}. \tag{3}$$

The first term on the right-hand side of (3) is the ratio of the marginal densities and is the *Bayes factor* (BF) between the hypotheses. The logarithm of the BF is called the *weight of evidence*.

The expected weight of evidence in a hypothesis test has the following property. Suppose in the hypothesis test specified in (1) that the data-generating value of $\boldsymbol{\theta}$ is actually drawn from density π_t, and let $m_t(\mathbf{x}_n)$ denote the corresponding marginal density of $\mathbf{x}_n$. If the sample space does not depend on $\boldsymbol{\theta}$, then Gibbs' inequality implies that

$$\begin{aligned}\int_{\mathcal{X}} m_t(\mathbf{x}_n) \log\left[\frac{m_t(\mathbf{x}_n)}{m_0(\mathbf{x}_n)}\right] d\mathbf{x}_n &- \int_{\mathcal{X}} m_t(\mathbf{x}_n) \log\left[\frac{m_1(\mathbf{x}_n)}{m_0(\mathbf{x}_n)}\right] d\mathbf{x}_n \\ &= \int_{\mathcal{X}} m_t(\mathbf{x}_n) \log\left[\frac{m_t(\mathbf{x}_n)}{m_1(\mathbf{x}_n)}\right] d\mathbf{x}_n \\ &\geq 0.\end{aligned}$$

That is,

$$\int_{\mathcal{X}} m_t(\mathbf{x}_n) \log\left[\frac{m_t(\mathbf{x}_n)}{m_0(\mathbf{x}_n)}\right] d\mathbf{x}_n > \int_{\mathcal{X}} m_t(\mathbf{x}_n) \log\left[\frac{m_1(\mathbf{x}_n)}{m_0(\mathbf{x}_n)}\right] d\mathbf{x}_n. \tag{4}$$

Equality holds only if $m_t(\mathbf{x}_n) = m_1(\mathbf{x}_n)$ almost everywhere. If the marginal density under the null hypothesis is known—as it is assumed to be in a single-arm phase II clinical trial—then mis-specification of the prior density assumed for the experimental treatment under the alternative model decreases the expected weight of evidence in favor of the alternative model. From a regulatory perspective, this means that investigators cannot manipulate trial outcomes by specifying an overly optimistic prior in favor of the experimental treatment. This situation contrasts sharply with inferential approaches toward summarizing trial evidence—in which overly optimistic priors can introduce potentially serious biases in favor of the experimental treatment—and provides a strong argument in favor of designing Bayesian trials using testing methodology rather than inferential techniques and posterior credible intervals.

3 Non-local alternative models

We now define classes of alternative models that provide exponential convergence of the Bayes factor in favor of both true null and true alternative hypotheses.

Consider a test of the null hypothesis

$$H_0 : \theta \sim \pi_0(\theta), \tag{5}$$

where $\pi_0(\theta) > 0$ for all $\theta \in \Theta_0 \subset \Theta \subset \mathcal{R}$ and $\pi_0(\theta) = 0$ for all $\theta \in \Theta - \Theta_0$, versus the alternative hypothesis

$$H_1 : \theta \sim \pi_1(\theta), \tag{6}$$

where $\pi_1(\theta) > 0$ for all $\theta \in \Theta - \Theta_0$, and for some $\epsilon, \zeta > 0$

$$\pi_1(\theta) > \epsilon \quad \text{for all} \quad \{\theta \in \Theta : \inf_{\theta_0 \in \Theta_0} |\theta - \theta_0| < \zeta\}. \tag{7}$$

Condition (7) is the defining property of a *local alternative hypothesis* (or *local alternative prior density*).

Similarly, if for every $\epsilon > 0$ there exits $\zeta > 0$ such that

$$\pi_2(\theta) < \epsilon \quad \text{for all} \quad \{\theta \in \Theta : \inf_{\theta_0 \in \Theta_0} |\theta - \theta_0| < \zeta\}, \tag{8}$$

then we define π_2 to be a *non-local alternative prior density.*

Most objective Bayesian testing methods result in local alternative priors, at least when the null hypothesis is true. For example, fractional Bayes factors (O'Hagan, 1995, 1997) and intrinsic Bayes factors (Berger and Pericchi, 1996, 1998) both produce local priors when tested against true null hypotheses. Similarly, intrinsic priors (Bertolino et al. 2000) are often centered on point null hypothesis values, and thus are also local.

Johnson and Rossell (2008) studied the large sample properties of BFs under local alternative models. Under regularity conditions that apply in most clinical trials, they demonstrated that the convergence rate of the BF in favor of a local alternative hypothesis against

a true point null hypothesis is only $O_p(n^{-1/2})$, and that the posterior odds in favor of local alternative hypothesis against true composite hypotheses is $O_p(1)$. In the latter case, this means that the posterior odds are not consistent as the sample size n becomes large. Against a false, non-local null hypothesis, however, convergence of the posterior odds in favor of the alternative hypothesis occurs at exponential rate.

To counter the slow convergence rates in favor of true null hypothesis that obtain under local alternative hypotheses, Johnson and Rossell (2008) proposed a class of inverse moment (iMOM) prior densities. For the case of a point null hypotheses $H_0 : \theta = \theta_0$, members of this alternative class of prior densities can be expressed as

$$\pi_I(\theta; \theta_0, k, \nu, \tau) = \frac{k\tau^\nu}{\Gamma(\nu/2k)} \left[(\theta - \theta_0)^2\right]^{-\frac{\nu+1}{2}} \exp\left\{ - \left[\left(\frac{\theta - \theta_0}{\tau} \right)^2 \right]^{-k} \right\} \tag{9}$$

for $k, \nu, \tau > 0$ and $\theta \in \mathcal{R}$. When the null hypothesis is true, the Bayes factor in favor of an alternative hypothesis defined using an iMOM prior density, say $BF_n(1|0)$, satisfies

$$p \lim_{n\to\infty} n^{-k/(k+1)} \log BF_n(1|0) = c, \qquad c < 0. \tag{10}$$

Thus, convergence of the posterior odds to the true model occurs at exponential rate under both true null and true alternative hypotheses when both models are assigned non-zero prior probability.

The density function (9) has modes at

$$\hat{\theta} = \theta_0 \pm \tau \left[\frac{2k}{\nu + 1} \right]^{1/2k}. \tag{11}$$

For values of $\nu = 2k$, the distribution function corresponding to density (9) is available in closed form, which makes normalization of the density on restricted intervals straightforward. This property is convenient when the range of θ values is restricted to, say, a subset of the unit interval. Convenient default values for these parameters are $k = 1$ and $\nu = 2$, for which

the distribution function is

$$\Pi_I(\theta;\theta_0,k,\nu,\tau) = \begin{cases} \frac{1}{2} - \frac{1}{2}\exp\left[-\tau(\theta-\theta_0)^{-2}\right], & \theta < \theta_0 \\ \frac{1}{2}, & \theta = \theta_0 \\ \frac{1}{2} + \frac{1}{2}\exp\left[-\tau(\theta-\theta_0)^{-2}\right], & \theta > \theta_0 \end{cases} \tag{12}$$

The tails of the distribution are similar to the tails of a Student t distribution on 3 degrees of freedom. A plot of this density for $\tau = 0.05$, $\theta_0 = 0.2$, $k = 1$ and $\nu = 2$ appears in Figure 1.

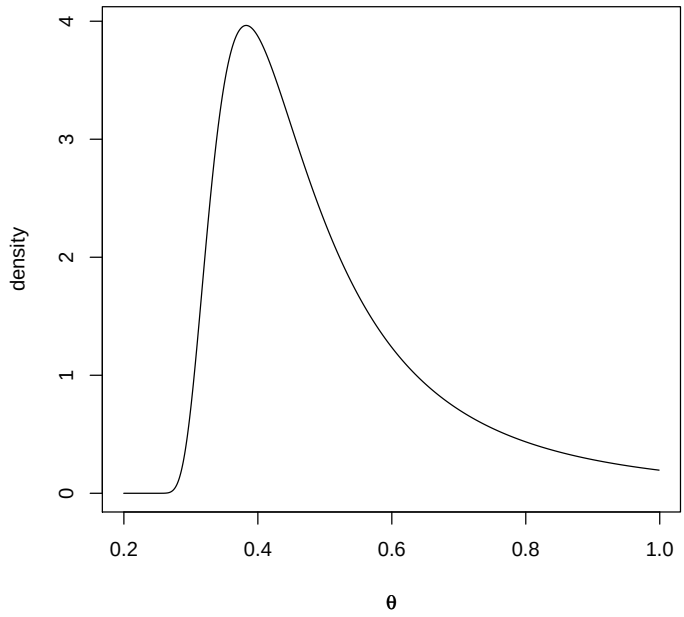

Figure 1: Illustration of iMOM density restricted to the interval (0,0.2). Parameters of this density are $\theta_0 = 0.2$, $\tau = 0.05$, $k = 1$, and $\nu = 2$.

4 Examples

We begin by examining the performance of common credible-interval based designs to formal test-based designs in single-arm, phase II clinical trials with either binary or time-to-event outcomes and continuous monitoring. Single-arm trials refer to trials in which all patients

receive a single, experimental therapy. Patient outcomes are then compared to responses expected under standard therapy. Continuous monitoring implies that a decision to terminate a trial is made after each patient outcome is observed. In general, trials may be stopped for *superiority* (the experimental treatment is judged to be better), *inferiority* (the standard treatment is judged to be at least as good), or *futility* (it is unlikely the trial will produce a conclusive result, given observed patient outcomes). Phase II trials represent the most common design to which Bayesian methods are applied.

4.1 Phase II trials with binary outcomes

Suppose that an experimental treatment for a specific type of cancer is to be tested against a standard treatment. Suppose further that historical data suggest that the response rate achieved by the standard treatment follows a Beta(20, 80) distribution, and that the trial sponsor expects to achieve approximately a 30% response rate with the experimental drug. A maximum of 50 patients are available for enrollment, and all patients are assigned to treatment with the experimental treatment. Patient outcomes are assumed to follow independent Bernoulli distributions with success probability θ. For simplicity, we assume that patient outcomes are known immediately following treatment.

To compare the operating characteristics of trial designs based on hypothesis tests to trials based on designs based on posterior credible intervals, we consider the following hypothetical trial designs.

Design 1. This design is based on a hypothesis test using the iMOM prior. Define null and alternative hypotheses according to

$$H_0 : \theta = 0.2 \quad \text{versus} \quad H_1 : \pi_1(\theta) \propto \pi_I(\theta; 0.2, 1, 2, 0.015)\, I_{(0.2,1)}(\theta).$$

The alternative prior density is truncated to the interval $(0.2, 1)$ and has a mode at

0.3. Equal prior odds are assigned to null and alternative hypotheses, and the null hypothesis has been represented by a point mass on 0.2. This trial is terminated after the nth patient for superiority if $\Pr(H_1 \mid \mathbf{x}_n) > 0.9$, and is terminated for inferiority if $\Pr(H_0 \mid \mathbf{x}_n) > 0.9$.

Design 2. This trial is a Thall and Simon (1994) design, an example of one of the most commonly used single-arm Bayesian phase II designs for binary outcomes. Letting θ_S denote response rate under standard therapy and assuming that $\theta_S \sim$ Beta(200, 800), this trial is stopped for superiority if $\Pr(\theta > \theta_S | \mathbf{x}_n) > 0.976$, and is stopped for inferiority if $\Pr(\theta < 0.1 + \theta_S) > 0.99$. Following guidelines suggested by Thall and Simon, the prior density assumed for θ in this design is assumed to be a Beta(0.6,1.4) distribution, which corresponds to the "equivalent" of two observations having success rate 0.3. We note that the "superiority interval" and "inferiority interval" overlap on the interval $(0.2, 0.3)$. This means that it is possible to terminate the trial for superiority (inferiority) even when the probability assigned to the inferiority (superiority) interval substantially exceeds 0.5.

Design 3. This design is based on Heitjan (1997), another common Bayesian design that permits continuous monitoring. An optimist's prior for θ is assumed to be a Beta(1.3, 1.7) density, while a pessimist's prior for θ is assumed to be a Beta(11, 41) distribution. The trial is stopped for superiority if the pessimist's posterior probability that θ exceeds 0.3 is 0.0023, and is stopped for inferiority if the optimist's posterior probability that θ is less than 0.2 exceeds 0.999.

Finally, we add a fourth trial as a straw man to illustrate the futility of designing trials using vaguely specified local alternative models.

Design 4. This trial treats the pessimist and optimist priors specified in Design 3 as the

null and alternative model for θ. Like Design 1, this trial stops when the probability of either hypothesis exceeds 0.9.

The first three trials were designed to have similar type I errors when $\theta = 0.2$ and power when $\theta = 0.3$. Each trial is considered inconclusive if no decision has been reached by enrollment of patient 51. The decision rule for each trial is evaluated after the outcome of each patient becomes available.

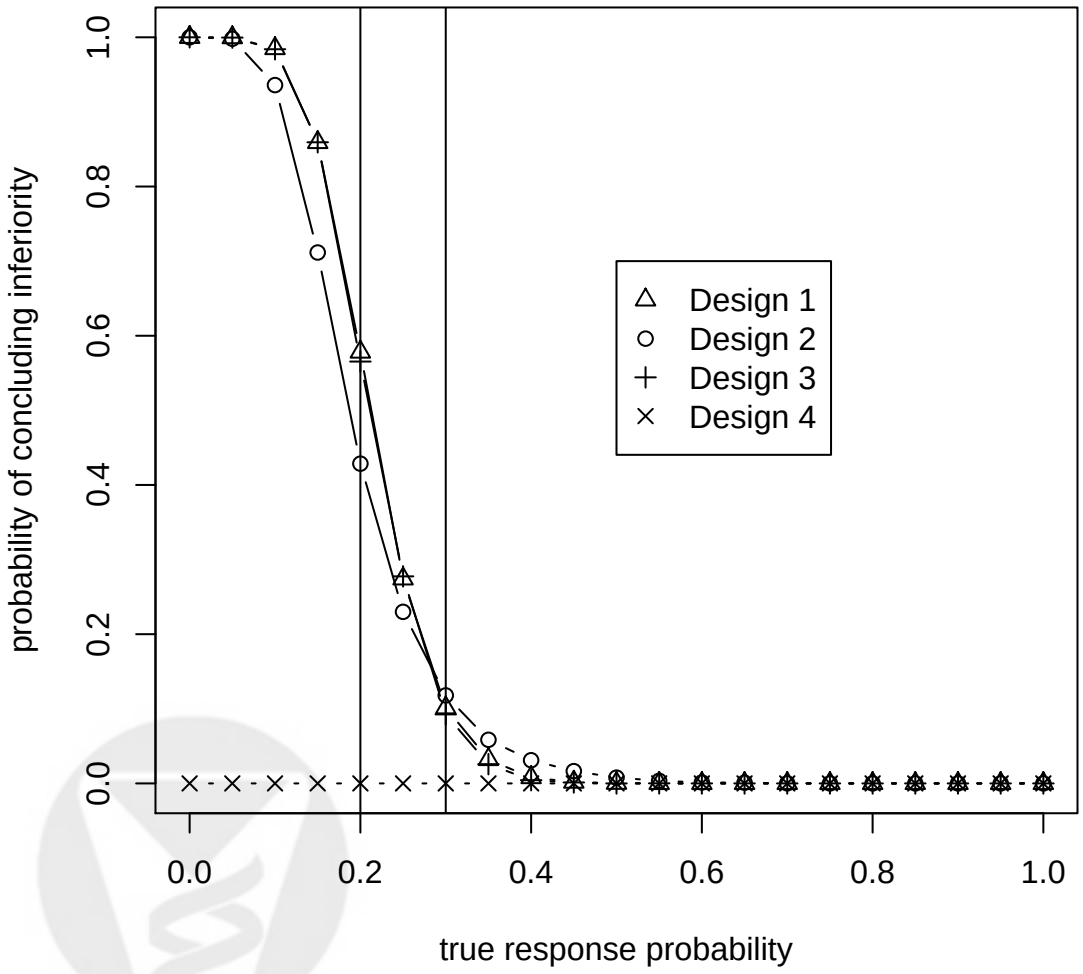

Figure 2: Probabilities of concluding that a trial agent is inferior to a standard treatment, when the standard treatment has success probability 0.2.

The operating characteristics of the four trial designs are depicted in Figures 2–4. Figure 2

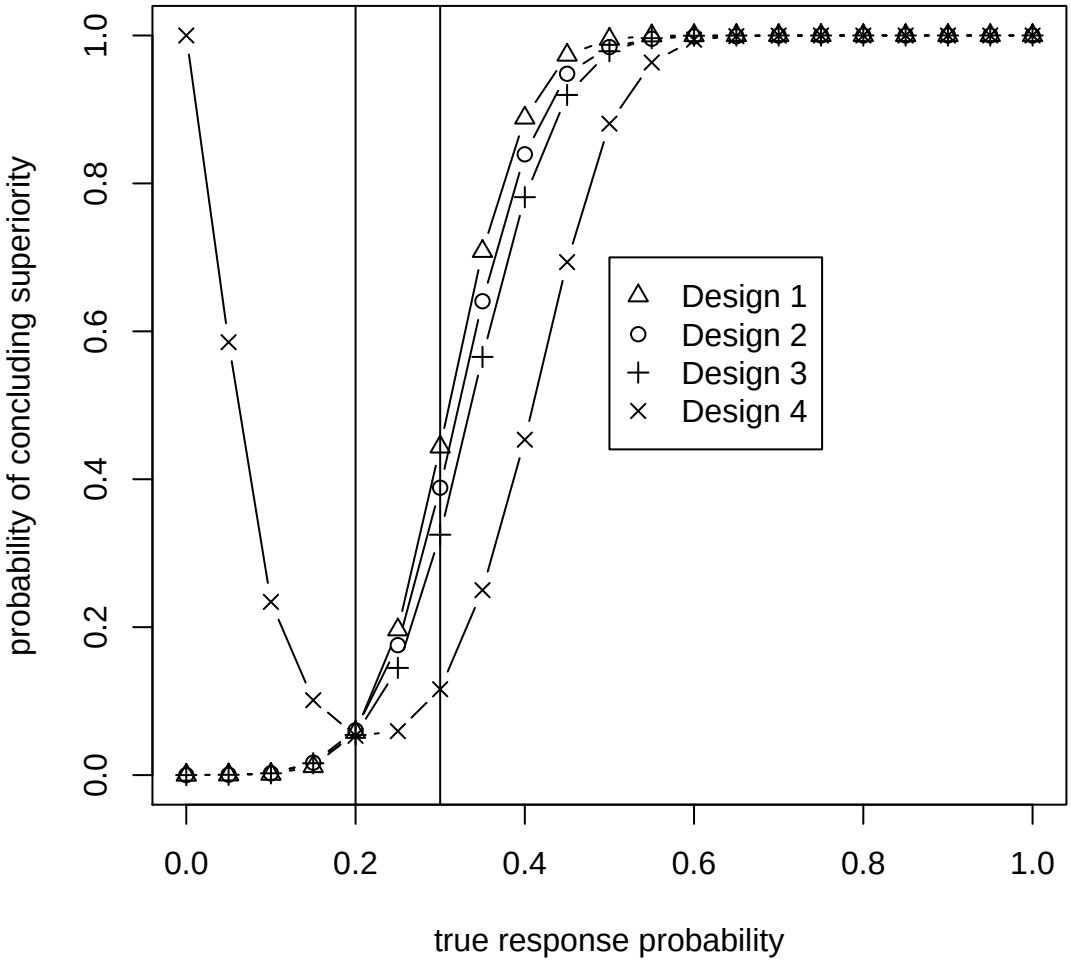

Figure 3: Probabilities of concluding that a trial agent is superior to a standard treatment, when the standard treatment has success probability 0.2.

shows that the probability of stopping for inferiority is highest under Design 1 for all values of $\theta < 0.3$, although Design 3 does nearly as well. At the null value of 0.2, the probabilities that Designs 1 and 3 stop for inferiority are 35% and 31% higher than Design 2. The frequentist probabilities for stopping for superiority are displayed in Figure 3. Design 1 exhibits higher probabilities of stopping for superiority for all values of $\theta > 0.2$. At the targeted rate of 0.3, it is 14% more likely to stop for superiority than Design 2, and 37% more likely to stop for superiority than Design 3.

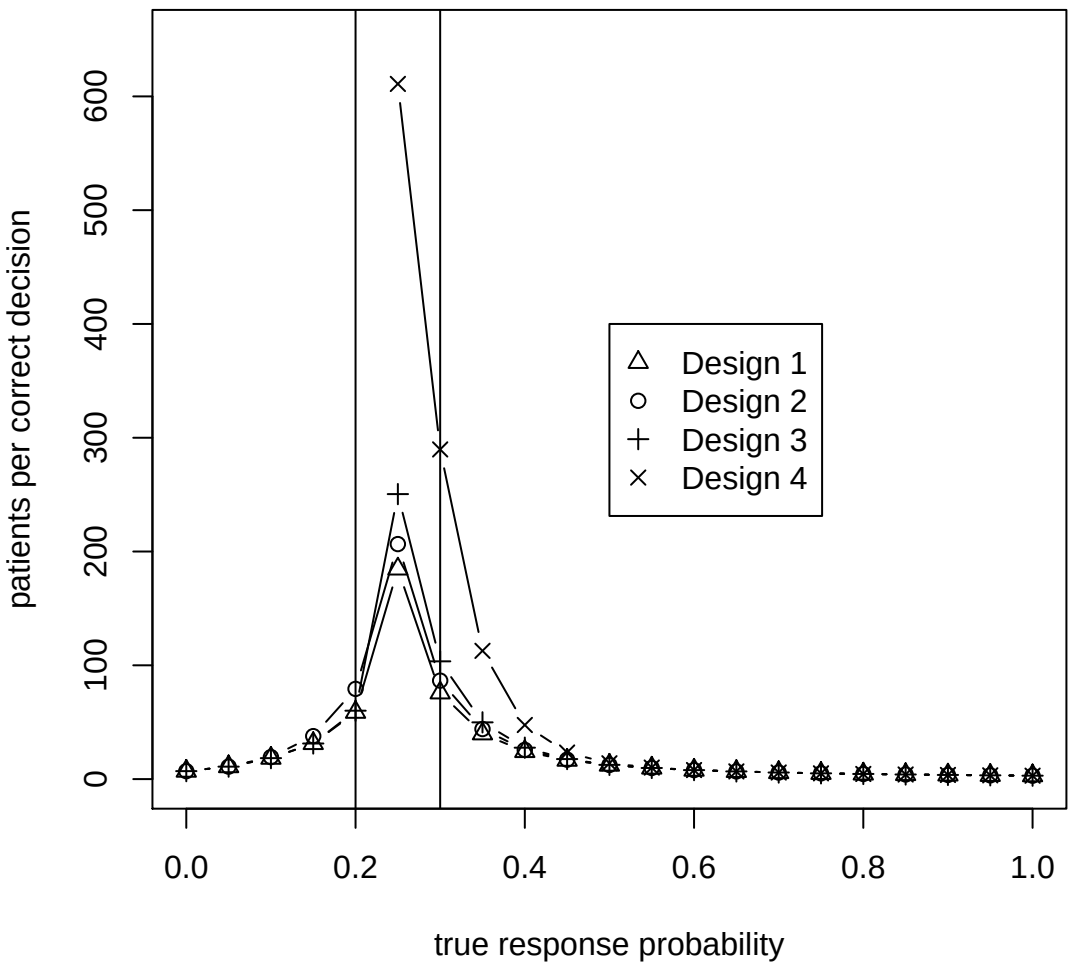

Figure 4: Average number of patients accrued per trial, divided by the probability that the trial resulted in a correct decision.

Results are mixed when the trial agent has a true response probability of 0.25. At this value, Designs 1 and 3 are more likely to stop for inferiority than Design 2, while Design 1 is also most likely to stop for superiority. Because the targeted success rate for the trial agent was 0.3 in all three designs, it is not clear whether a value of 0.25 should be considered a success.

We next compared the average number of patients used in each trial design. However, comparisons between the expected numbers of patients used in credible-interval based de-

signs and test-based designs are complicated by the fact that credible-interval based designs terminate more frequently for an incorrect decision. To make such comparisons fair, we calculated the ratio of the average number of patients treated to the probability of correctly stopping the trial. In making this comparison, we assumed that it was correct to reject the experimental treatment when $\theta < 0.3$. Figure 4 displays these ratios. Design 1—the test-based design—is most efficient for all values of θ according to this metric.

Finally, we note that Design 4 performs abysmally according to all criteria. Because of the local alternative hypothesis, Design 4 never results in early termination of a trial for inferiority. It is also much slower to conclude superiority for values of $\theta > 0.2$. Because the alternative hypothesis assigns prior mass to values of $\theta < 0.2$, it is also likely to conclude superiority for small values of θ. In those regions where Design 4 is able to appropriately terminate a trial, the average number of patients required to make a correct decision is larger than each of the other designs. Although an extreme example, this design illustrates the pitfalls that arise if local prior densities are naïvely specified to define alternative hypotheses.

4.2 Phase II trials with TTE outcomes

We next consider a hypothetical single-arm trial of a experimental agent designed to prolong progression free survival (PFS) in a given population of cancer patients. We assume that the mean PFS under standard therapy is 6 months, and that the experimental agent is hypothesized to extend this period to approximately 8 months. We suppose that PFS follows an exponential distribution, and we let θ and θ_S denote the mean survival time in months for the experimental and standard treatments, respectively. We consider two designs: one based on hypothesis tests using an iMOM alternative prior, and the second based on posterior credible intervals.

Design 5. Under the test-based design, the null and alternative hypotheses are defined

according to

$$H_0 : \theta = 6 \quad H_1 : \pi_1(\theta) \propto \pi_I(\theta; 6, 1, 2, 6)\, I_{(6,\infty)}(\theta) \tag{13}$$

The alternative prior density is an iMOM density truncated to the interval $(6, \infty)$ with a mode at 8. Equal prior odds are assigned to the null and alternative hypotheses and the null hypothesis is represented by a point mass at 6. This trial is terminated after the n'th patient for superiority if $\Pr(H_1 \,|\, \mathbf{x}_n) > 0.9$, and is terminated for inferiority if $\Pr(H_0 \,|\, \mathbf{x}_n) > 0.9$.

Design 6. This design is based on methodology described in Thall et al. (2005), and, like its binary counterpart, represents one of the most commonly used single-arm Bayesian phase II designs for TTE outcomes. In this trial, θ_S is assumed to be drawn from a inverse gamma distribution with parameters 20 and 1200 (i.e., $IG(200, 1200)$), while θ is assigned a $IG(3, 12)$ prior distribution. Note that the prior means of both parameters is approximately 6. This trial is stopped for superiority if

$$\Pr(\theta_S < \theta_E \,|\, x_n) \geq 0.94,$$

and for inferiority if

$$\Pr(\theta_S + 2 < \theta_E \,|\, x_n) \leq 0.07.$$

As in the previous example, enrollment for both trials is capped at 50, and trials not stopped by patient 51 are considered inconclusive. Both trials have been specified so that the probability of stopping for inferiority when $\theta = 8$ is 0.2, and for stopping for superiority when $\theta = 6$ is 0.05. Stopping criteria are evaluated after the outcome of each patient becomes known. As in Design 2, the superiority and inferiority intervals overlap in Design 6.

Figures 5 and 6 depict the probability that Designs 5 and 6 result in a conclusive termination of the trial when at most 50 patients are enrolled. As in the case of binary endpoints,

the test-based design (Design 5) stops significantly more often for inferiority when the true value of $\theta_E < 8$, and stops for superiority more often when $\theta_E > 6$ than does Design 6. For example, when the experimental and standard treatments both yield six-month mean survival time, Design 5 correctly concludes inferiority in 84% of trials, while Design 6 makes this determination in only 67% of trials. Design 5 thus provides a 25% increase in the probability of a correct decision when the null hypothesis of no additional treatment benefit is true.

Design 5, like Design 1 of the previous section, also has the advantage of providing a clear interpretation of trial results. Namely, the posterior probability of each hypothesis has an exact frequentist interpretation at the end of every trial. That is, among trials for which the (final) posterior probability of the null hypothesis is p, in repeated sampling of the full probability model there is exactly probability p that trial data was drawn from the null model.

5 Comparison to Simon two-stage design

We conclude by comparing Bayesian test-based designs to an optimal frequentist two-stage design (Simon, 1989). Assuming a null response rate of $p_0 = 0.20$ and a targeted response rate of $p_1 = 0.40$, the parameters of both designs were chosen so that the Type I and II errors were 5% and 20%, respectively. Under the optimality criterion described by Simon, the two-stage design requires a maximum of 43 patients, 13 in the first stage and 30 in the second.

The null hypothesis under the Bayesian design was assumed to be a point mass at $\theta = 0.2$, and the alternative hypothesis was represented by an iMOM density with $\tau = 0.26$, $k = 1$ and $\nu = 1$. The Bayesian design utilized continuous monitoring and rejected the experimental

treatment if the posterior probability of the alternative model dropped below 0.037. These parameter values were selected so as to match the type I and II errors of the Simon design. Enrollment under the Bayesian design was also restricted to be 43 patients.

As in the previous section, we compared the operating characteristics of these trials by varying the true probability of response between 0 to 1 in increments of 0.05. For each response probability, we simulated 10,000 trials. Denoting the Simon design by Design 7 and the Bayesian design by Design 8, Figure 7 shows that the probabilities of rejecting the experimental treatments are similar under both designs. However, because the Bayesian design performs continuous monitoring, it uses fewer patients to reject treatments that fall below the targeted value of 0.4. This fact is illustrated in Figure 8, where two stopping criteria for the Simon design have been implemented. In the "naïve Simon" design (Design 7a), a trial could be stopped only after 13 or 43 patients. In a simple modification of this design, a trial was stopped as soon as it became clear that there would not be enough successes to either continue the trial or to accept the experimental treatment at the end of the study (Design 7b).

Figure 8 (as well as Figures 10 and 12 below) demonstrate a critical feature of the test-based design: these deisgns enable earlier termination of trials of ineffective treatments. In practice, this means that significantly fewer patients are deprived of the clinical benefit of existing therapies.

Figures 9-10 provide similar comparisons to the Simon two-stage designs corresponding $p_0 = 0.4$ and $p_1 = 0.6$; Figures 11-12 provide similar comparisons to the Simon two-stage designs corresponding $p_0 = 0.3$ and $p_1 = 0.4$. These figures exhibit similar properties to those described above for the case $p_0 = 0.2$ and $p_1 = 0.4$.

6 Discussion

A primary obstacle to the use of Bayesian clinical trial designs has been the hesitancy of practitioners to specify prior distributions on model parameters. When inferences regarding the outcome of a trial are based on posterior credible intervals, these concerns stem from the fact that the prior density retains a direct influence on the posterior credible regions reported, regardless of the outcome of the trial.

The situation is fundamentally different when viewed within the context of a Bayesian hypothesis test. In this setting, the alternative hypothesis, by definition, represent the investigator's belief regarding the distribution of the trial agent under an assumption that the trial agent is effective. Such an assumption does not represent a bias; it simply reflects the fact that—under the alternative hypothesis—the new treatment is assumed to be better than the standard. Mis-specification of the alternative hypothesis has the effect of decreasing the expected weight of evidence in favor of the experimental treatment, which means that proponents of an experimental treatment cannot, on average, bias a test in favor of the experimental treatment. The definition of BFs using the non-local prior distributions proposed in this paper thus facilitates the conduct of clinical trials using Bayesian methods by eliminating one of the major obstacles associated with their use.

For trials designed with fixed endpoints, it follows from the Neyman-Pearson lemma that Bayesian testing procedures provide optimal tests of a specified size (recall that the marginal density of the data is explicitly defined under both null and alternative models within the Bayesian paradigm). Although the Neyman-Pearson result does not extend to trials with continuous monitoring, results presented in this article suggest that Bayesian test-based designs provide better operating characteristics than trials designed using either posterior credible intervals or Simon two-stage designs. Test-based designs also provide a cleaner interpretation of trial outcomes. As mentioned earlier, among trials for which the

posterior probability of the null hypothesis is p, in repeated sampling of the full probability model there is exactly probability p that trial data were drawn from the null model.

Finally, we note that software to design phase II single-arm trials using iMOM specification of alternative hypotheses is available at http://biostatistics.mdanderson.org/SoftwareDownload/. The program *nonlocal1* provides stopping boundaries based on iMOM parameters as user input, while *nonlocal2* provides stopping boundaries that provide specified type I and II errors against specified point null and alternative hypotheses.

References

[1] Berger, J. O. and Pericchi, L. R. (1996), "The intrinsic Bayes factor for model selection and prediction," *Journal of the American Statistical Association*, **91**, 109-122.

[2] Berger, J. O. and Pericchi, L. R. (1998), "Accurate and stable Bayesian model selection: The median intrinsic Bayes factor," *Sankhyá, Series B,* **60**, 1-18.

[3] Bertolino, F., Moreno, E. Racugno, W. (2000), "Bayesian model selection approach to analysis of variance under heteroscedasticity," *Journal of the Royal Statistical Society, Series D: The Statistician*, **49**, 503-517.

[4] Fayers, P. M., Ashby, D. and Parmar, M. K. (1997), Bayesian data monitoring in clinical trials. Statistics in Medicine, 16, 1413-1430

[5] George, S. L., Li, C., Berry, D. A. and Green, M. R. (1994), Stopping a clinical trial early: Frequentist and Bayesian approaches applied to a CALGB trial in non-small-cell lung cancer. Statistics in Medicine, 13, 1313-1327.

[6] Heitjan D.F. (1997), Bayesian interim analysis of phase II cancer clinical trials. Statistics in Medicine, 16, 17911802.

[7] Johnson, V.E. (2005). Bayes Factors Based on Test Statistics. Journal of the Royal Statistical Society, Series B, 67, 689-701.

[8] Johnson, V.E. (2008). Properties of Bayes Factors Based on Test Statistics. to appear in Scandinavian Journal of Statistics; included in other material.

[9] Johnson, V.E. and Rossell, D. (2008). Non-Local Prior Densities for Objective Bayes Hypothesis Tests. In submission and available at *www.bepress.com/mdandersonbiostat/paper42/.*

[10] O'Hagan, A. (1995), "Fractional Bayes factors for model comparison," *Journal of the Royal Statistical Society, Series B*, **57**, 99-118.

[11] O'Hagan, A. (1997), "Properties of intrinsic and fractional Bayes factors," *Test*, **6**, 101-118.

[12] Moreno, E., Bertolino, F. and Racugno, W. (1998), "An intrinsic limiting procedure for model selection and hypotheses testing," *Journal of the American Statistical Association*, **93**, 1451-1460.

[13] Simon, R. (1989) Optimal Two-Stage Designs for Phase II Clinical Trials Controlled Clinical Trials, 10, 1-10.

[14] Simon, R. (1999) Bayesian design and analysis of active control clinical trials. Biometrics, 55, 484-487.

[15] Tan, S.-B. and Machin, D. (2002), Bayesian two-stage designs for phase II clinical trials. Statistics in Medicine, 21, 1991-2012.

[16] Thall, P. and Simon, R. (1994). A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. Controlled Clinical Trials, 15, 463-481.

[17] Thall PF, Wooten LH, Tannir NM. 2005. Monitoring Event Times in Early Phase Clinical Trials: Some Practical Issues. Clinical Trials, 2, 467-478.

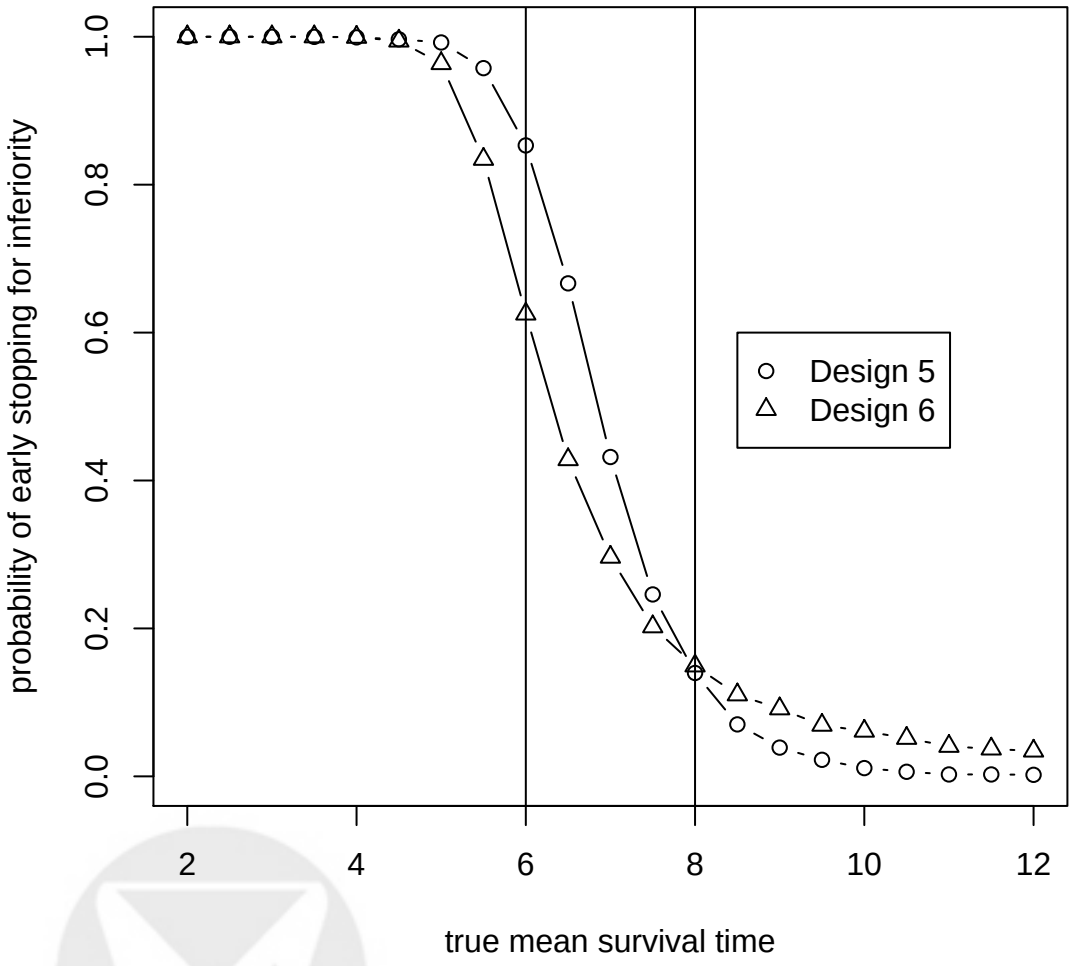

Figure 5: Probabilities of concluding that a trial agent is inferior to a standard treatment, when the standard treatment produces a mean survival time of 6 months.

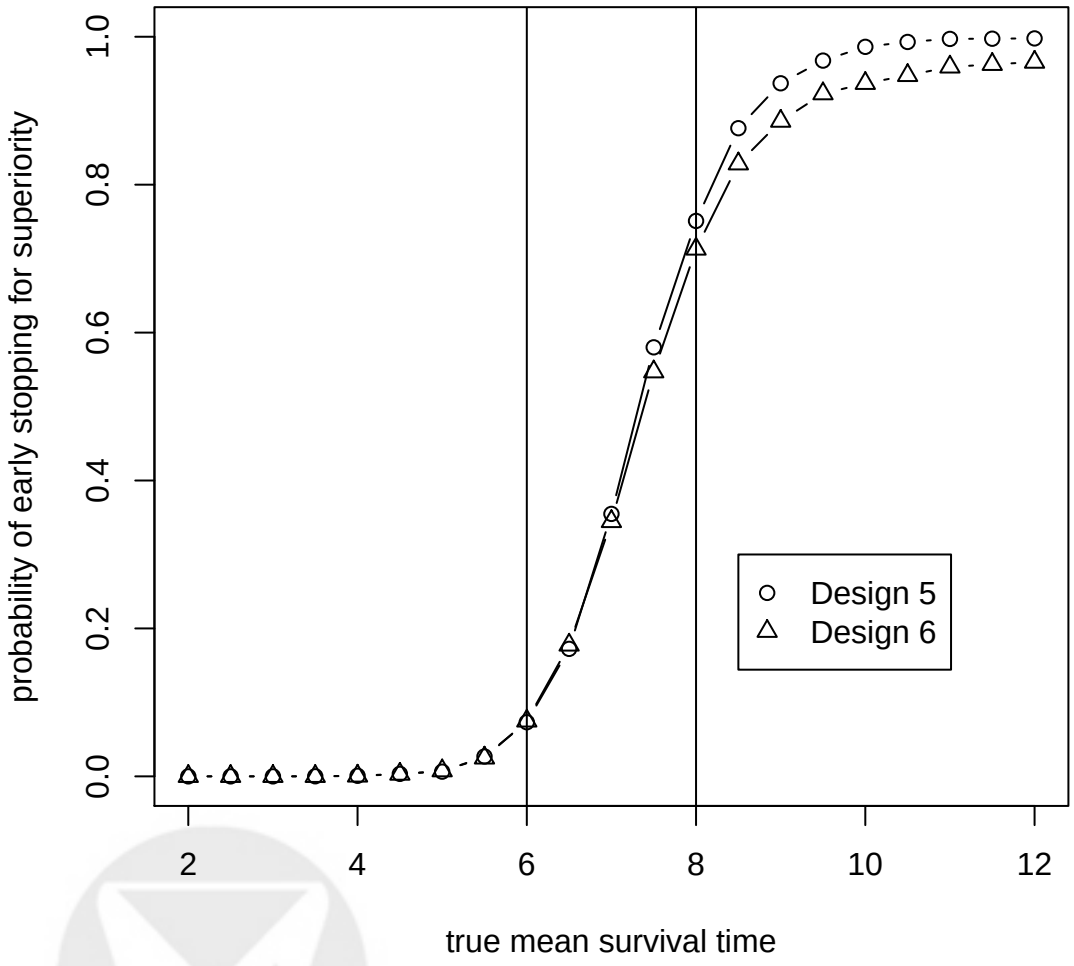

Figure 6: Probabilities of concluding that a trial agent is superior to a standard treatment, when the standard treatment produces a mean survival time of 6 months.

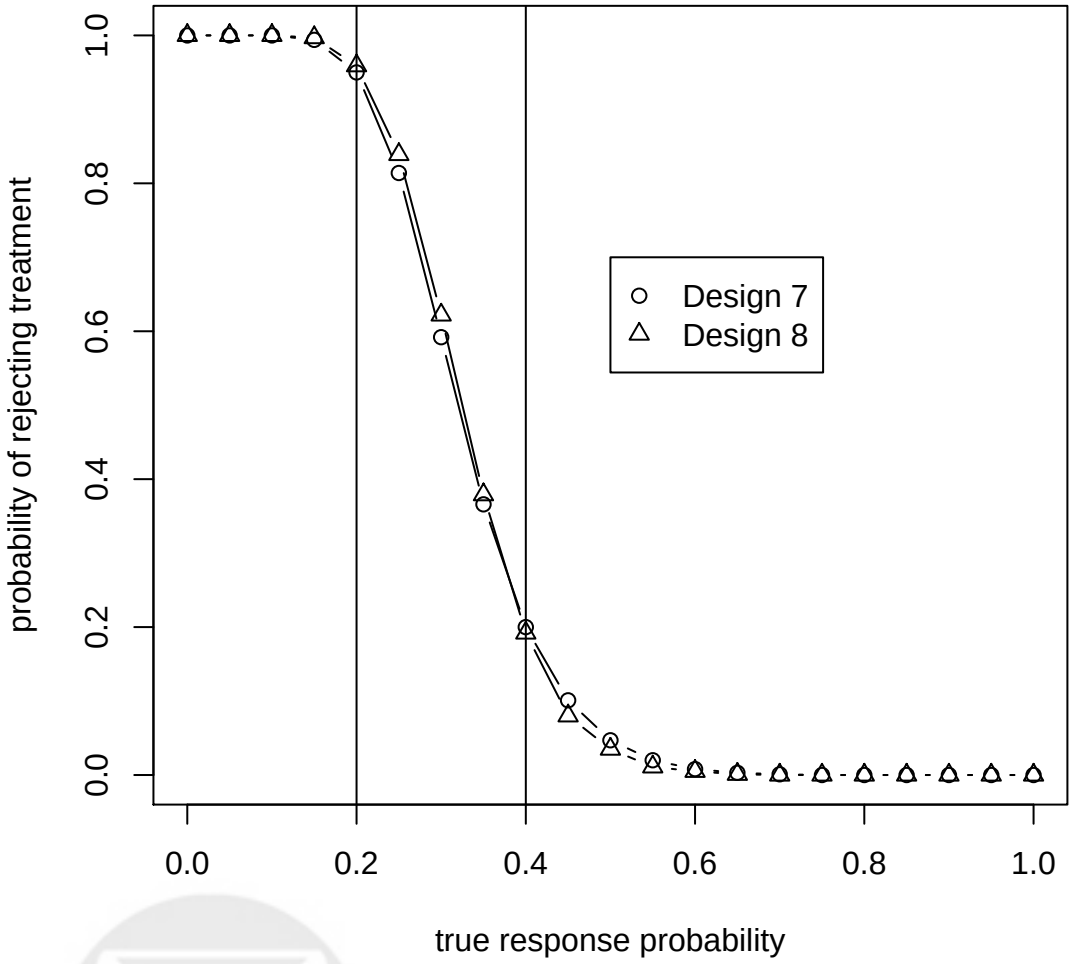

Figure 7: Probabilities of concluding that a trial agent is inferior to a standard treatment, when the standard treatment has success probability 0.2. The curve for the Simon two-stage design is denoted by Design 7, while the curve corresponding to the Bayesian hypothesis test is denoted by Design 8.

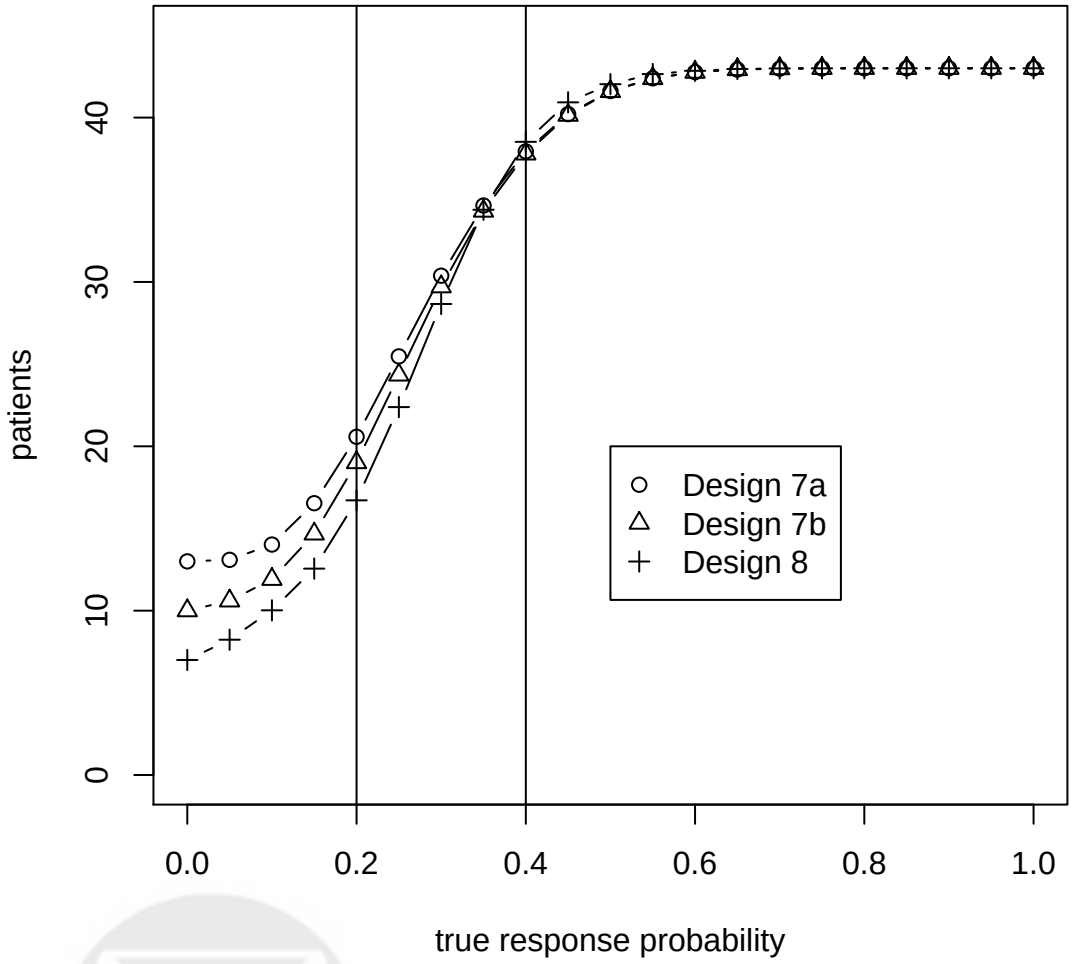

Figure 8: Number of patients treated as the true probability of response varies when the standard treatment has success probability 0.2. The curve labeled Design 8 represents the operating characteristics of the Bayesian design, while curves labeled Design 7a and 7b correspond to the operating characteristics of the naïve and modified Simon designs, respectively.

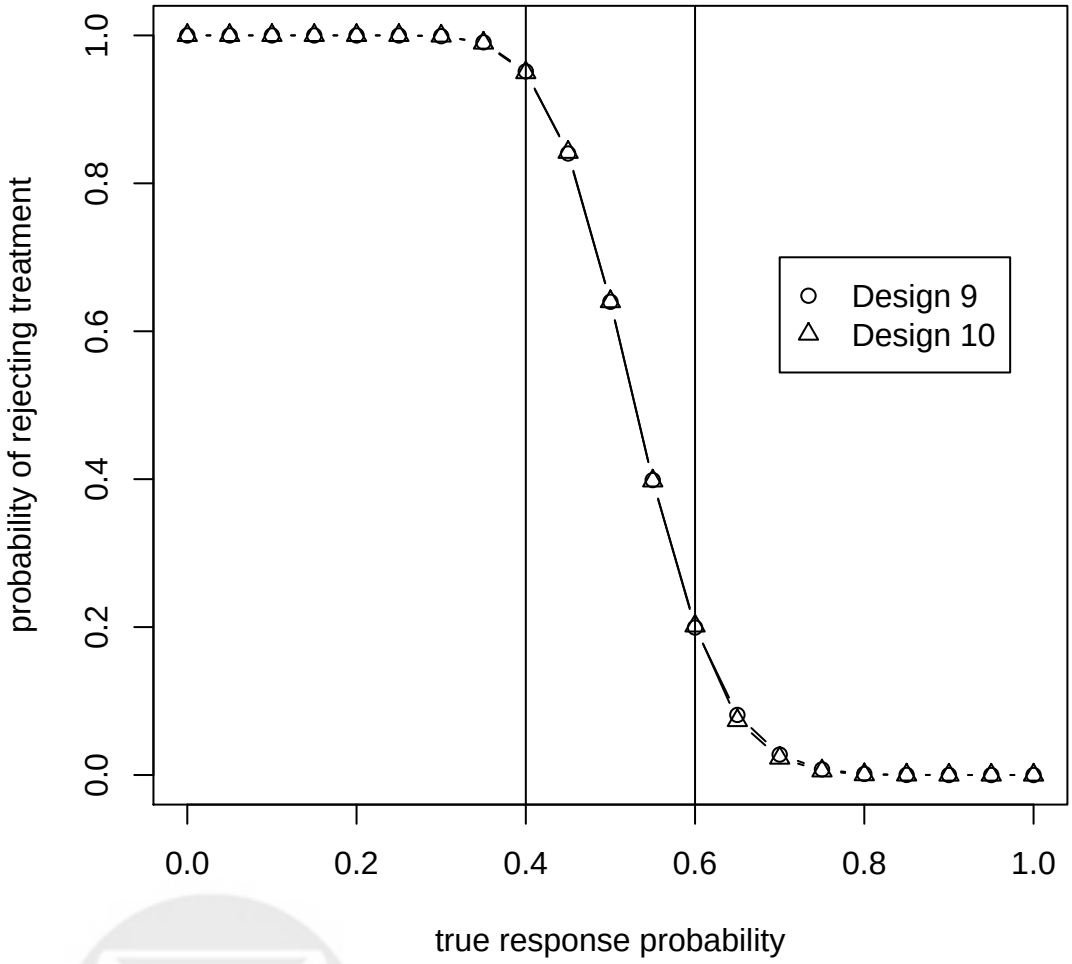

Figure 9: Probabilities of concluding that a trial agent is inferior to a standard treatment, when the standard treatment has success probability 0.4 and the alternative hypothesis is 0.6. The curve for the Simon two-stage design is denoted by Design 9, while the curve corresponding to the Bayesian hypothesis test is denoted by Design 10.

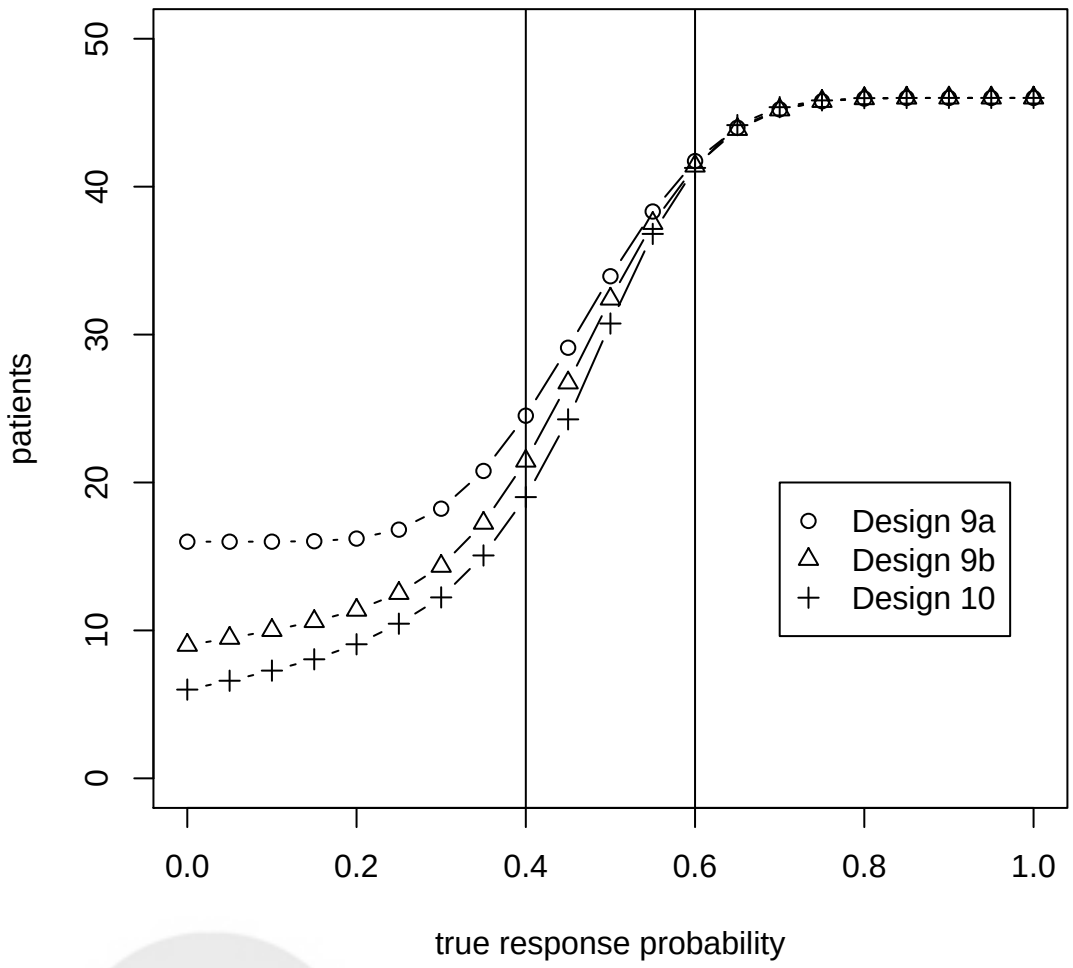

Figure 10: Number of patients treated as the true probability of response varies when the standard treatment has success probability 0.4 and the alternative hypothesis is 0.6. The curve labeled Design 10 represents the operating characteristics of the Bayesian design, while curves labeled Design 9a and 9b correspond to the operating characteristics of the naïve and modified Simon designs, respectively.

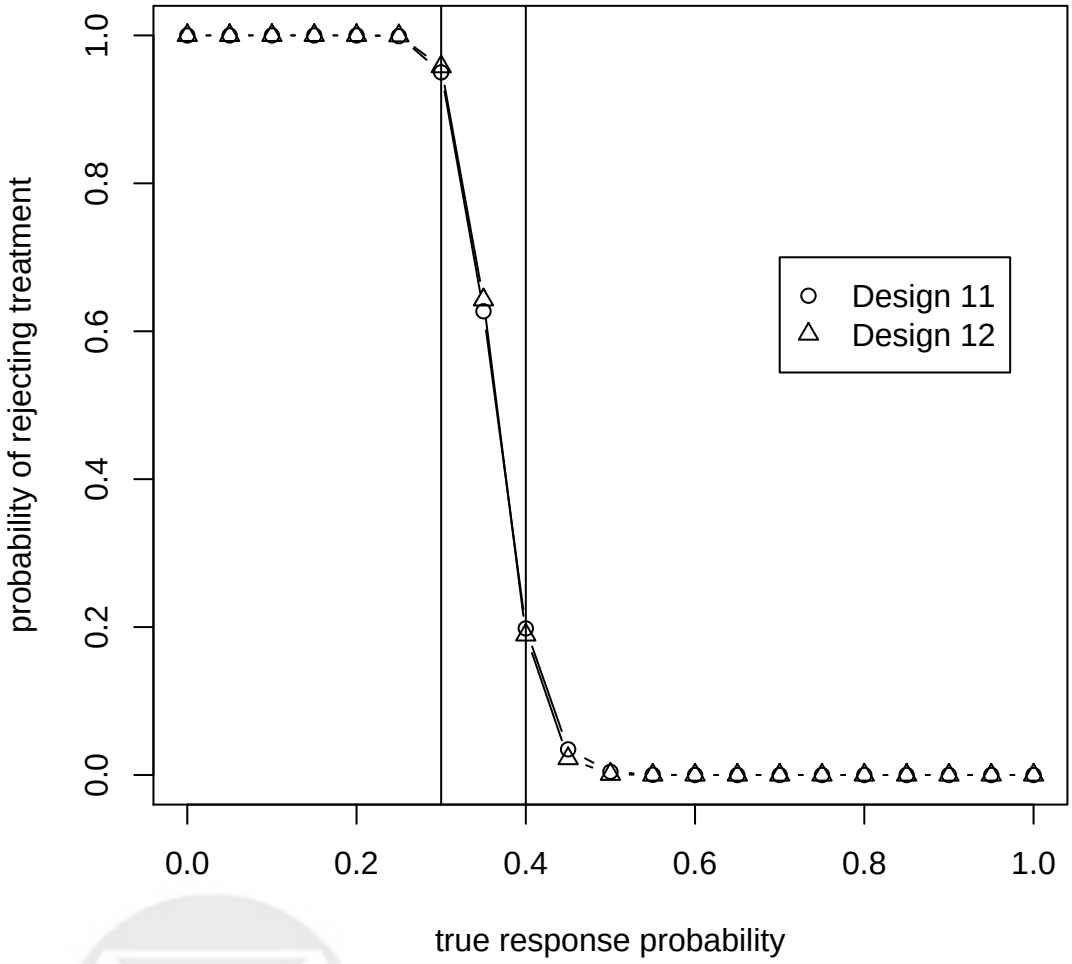

Figure 11: Probabilities of concluding that a trial agent is inferior to a standard treatment, when the standard treatment has success probability 0.3 and the alternative hypothesis is 0.4. The curve for the Simon two-stage design is denoted by Design 11, while the curve corresponding to the Bayesian hypothesis test is denoted by Design 12.

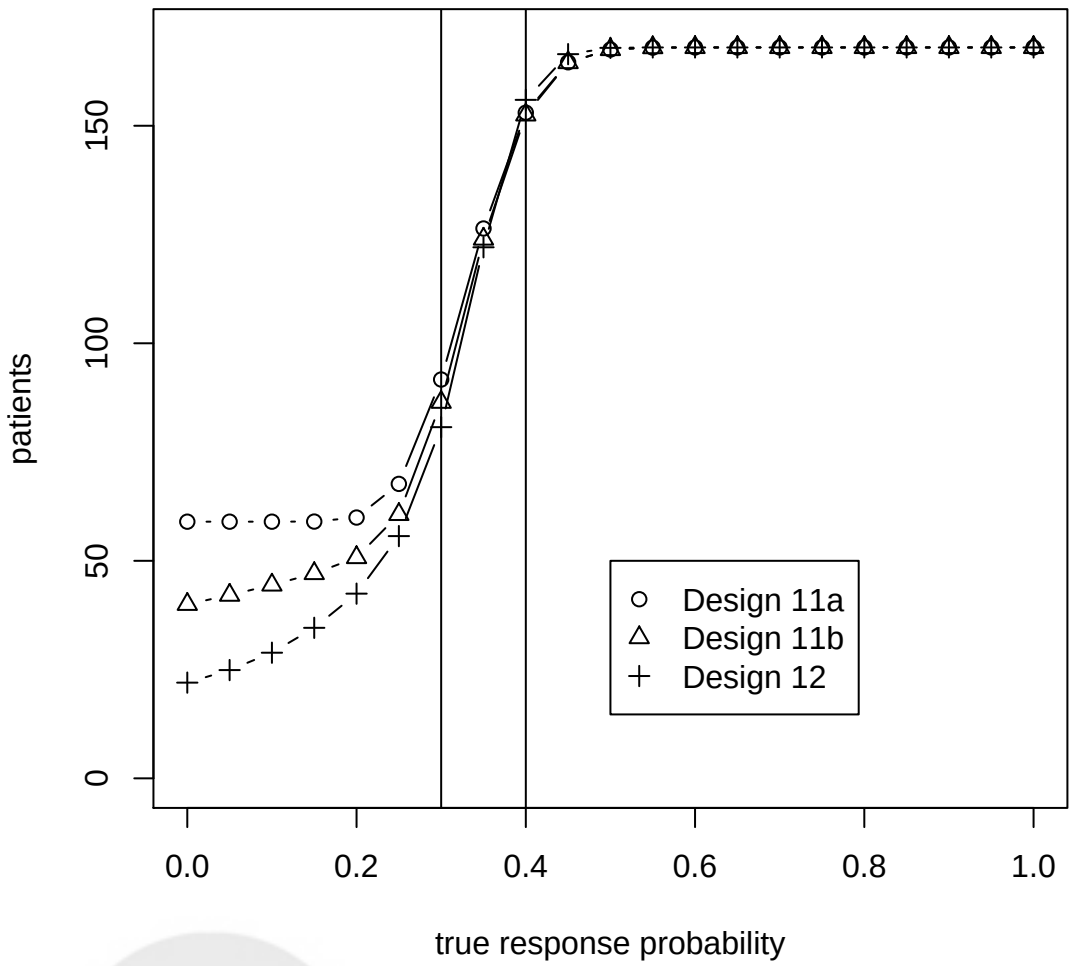

Figure 12: Number of patients treated as the true probability of response varies when the standard treatment has success probability 0.3 and the alternative hypothesis is 0.4. The curve labeled Design 12 represents the operating characteristics of the Bayesian design, while curves labeled Design 11a and 11b correspond to the operating characteristics of the naïve and modified Simon designs, respectively.