STATISTICS IN MEDICINE
*Statist. Med.* 2007; **26**:1431–1449
Published online 25 October 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2640

# Bayesian evaluation of group sequential clinical trial designs

Scott S. Emerson[1,*,†], John M. Kittelson[2] and Daniel L. Gillen[3]

[1]*Department of Biostatistics, Box 357232, University of Washington, Seattle, Washington 98195-7232, U.S.A.*
[2]*Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, Denver, Colorado 80262, U.S.A.*
[3]*Department of Statistics, University of California, Irvine, California 92697, U.S.A.*

SUMMARY

Clinical trial designs often incorporate a sequential stopping rule to serve as a guide in the early termination of a study. When choosing a particular stopping rule, it is most common to examine frequentist operating characteristics such as type I error, statistical power, and precision of confidence intervals (*Statist. Med.* 2005, in revision). Increasingly, however, clinical trials are designed and analysed in the Bayesian paradigm. In this paper, we describe how the Bayesian operating characteristics of a particular stopping rule might be evaluated and communicated to the scientific community. In particular, we consider a choice of probability models and a family of prior distributions that allows concise presentation of Bayesian properties for a specified sampling plan. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: interim analyses; operating characteristics; Bayesian; stopping rules; sample size

## 1. INTRODUCTION

Clinical trial data are often monitored repeatedly during the conduct of a study in order to address the efficiency and ethical issues inherent in human experimentation. Decisions regarding the continuation of the study are typically guided by a group sequential stopping rule that specifies the conditions under which the clinical trial results might be judged sufficiently convincing to allow early stopping. In a companion paper to this manuscript [1], we consider the evaluation of a clinical trial design with respect to frequentist operating characteristics such as type I error, statistical power, sample size requirements, estimates of treatment effect that correspond to early termination, and precision of confidence intervals. Increasingly, however, there has been much

*Correspondence to: Scott S. Emerson, Department of Biostatistics, Box 357232, University of Washington, Seattle, Washington 98195-7232, U.S.A.
†E-mail: semerson@u.washington.edu

Contract/grant sponsor: NIH; contract/grant number: HL69719

*Received 8 November 2004*
*Accepted 23 May 2006*

interest in the design and analysis of clinical trials under a Bayesian paradigm and multiple authors have recently discussed the role and implementation of Bayesian methods in monitoring clinical trials [2–4]. We thus turn our attention to the evaluation of Bayesian operating characteristics for a clinical trial design.

In considering the Bayesian approach, we again take the stance that the derivation of the stopping rule is relatively unimportant. That is, in the companion paper, we demonstrate the 1:1 correspondence between stopping rules defined for frequentist statistics (on a variety of scales) and Bayesian statistics for a specified prior. Hence, our focus in this paper will be on the computation and presentation of Bayesian operating characteristics for a specified stopping rule. The primary issues to be addressed will be the selection of suitable probability models and families of prior distributions that will allow a standard, concise communication of design precision and statistical inference. In particular, our interest is in providing Bayesian inference in the context of a probability model similar to that which would be assumed in the most common frequentist analyses. For the purposes of brevity, we will consider only a single stopping rule in our illustration, although in practice we would compare the Bayesian operating characteristics among several candidate stopping rules in much the same way that frequentist operating characteristics were compared across stopping rules in the companion paper.

In illustrating our approach to evaluating Bayesian operating characteristics, we will appeal to the same example as used in the companion paper. In Section 2, we provide a brief review of the scientific setting and basic statistical design of the clinical trial. Then in Section 3 we present the general Bayesian paradigm and the non-parametric, 'coarsened Bayesian' approach we adopt here, along with a discussion of the choice of prior distributions. In Section 4 we present a general scheme for presenting the Bayesian operating characteristics in a relatively concise manner. We conclude in Section 5 with some general comments regarding the practical use of the proposed approach to Bayesian evaluation of clinical trial designs.

## 2. EXAMPLE USED FOR ILLUSTRATION

We illustrate our approach in the context of a randomized, double-blind, placebo-controlled clinical trial of an antibody to endotoxin in the treatment of gram-negative sepsis. Details of the scientific setting and the clinical trial design are provided in the companion paper [1].

### *2.1. Notation and sample size*

Briefly, a maximum of 1700 patients with proven gram-negative sepsis were to be randomly assigned in a 1:1 ratio to receive a single dose of antibody to endotoxin or placebo. The primary endpoint for the trial was to be the 28 day mortality rate, which was anticipated to be 30 per cent in the placebo treated patients and was hoped to be 23 per cent in the patients receiving antibody. Notationally, we let $X_{ki}$ be an indicator that the $i$th patient on the $k$th treatment arm ($k=0$ for placebo, $k=1$ for antibody) died in the first 28 days following randomization. Thus $X_{ki}=1$ if the $i$th patient on treatment arm $k$ dies in the first 28 days following randomization, and $X_{ki}=0$ otherwise. We are interested in the probability model in which the random variables $X_{ki}$ are independently distributed according to a Bernoulli distribution $\mathscr{B}(1, p_k)$, where $p_k$ is the unknown 28 day mortality rate on the $k$th treatment arm. We use the difference in 28 day mortality rates $\theta = p_1 - p_0$ as the measure of treatment effect.



DOI: 10.1002/sim

Supposing the accrual of $N$ subjects on each treatment arm, a frequentist analysis of clinical trial results was to be based on asymptotic arguments that suggest that $\hat{p}_k = \sum_{i=1}^{N} X_{ki}/N$ is approximately normally distributed with mean $p_k$ and variance $p_k(1-p_k)/N$. In a study that accrued $N$ subjects per arm we therefore have an approximate distribution for the estimated treatment effect $\hat{\theta} = \hat{p}_1 - \hat{p}_0$ of

$$\hat{\theta} \dot{\sim} \mathcal{N}\left(\theta, \frac{p_1(1-p_1) + p_0(1-p_0)}{N}\right) \tag{1}$$

As is customary in the setting of tests of binomial proportions, at the time of data analysis the actual frequentist test statistic will estimate a common mortality rate $\hat{p}$ under the null hypothesis of no treatment effect. Thus, if at the time of data analysis $n_0$ and $n_1$ patients had been accrued to the placebo and treatment arms, respectively, and the respective observed 28 day mortality rates were $\hat{p}_0$ and $\hat{p}_1$, the test statistic used to test the null hypothesis would be

$$Z = \frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\hat{p}(1-\hat{p})((1/n_1) + (1/n_0))}}$$

where the common mortality rate under the null hypothesis is estimated by

$$\hat{p} = \frac{n_1\hat{p}_1 + n_0\hat{p}_0}{n_0 + n_1}$$

In a fixed sample study using a one-sided level 0.025 test, the 1700 subjects (850 per arm) provide statistical power of 0.9066 to detect the design alternative of $\theta = -0.07$ when the control group's 28 day mortality rate if 30 per cent. If the estimated variability of $\hat{\theta}$ at the conclusion of such a trial were to agree exactly with the variance used in the sample size calculation, the null hypothesis would be rejected in a frequentist hypothesis test if the absolute difference in 28 day mortality rates showed that the mortality on the antibody arm was at least 0.0418 lower than that on the placebo arm (i.e. we would reject $H_0$ if and only if $\hat{\theta} \leqslant -0.0418$).

### 2.2. *Definition of stopping rules*

A stopping rule is defined for a schedule of analyses occurring at sample sizes $N_1, N_2, \ldots, N_J$. For $j = 1, \ldots, J$, we calculate treatment effect estimate $\hat{\theta}_j$ based on the first $N_j$ observations. The outcome space for $\hat{\theta}_j$ is then partitioned into stopping set $\mathcal{S}_j$ and continuation set $\mathcal{C}_j$. Starting with $j = 1$, the clinical trial proceeds by computing test statistic $\hat{\theta}_j$, and if $\hat{\theta}_j \in \mathcal{S}_j$, the trial is stopped. Otherwise, $\hat{\theta}_j$ is in the continuation set $\mathcal{C}_j$, and the trial gathers observations until the available sample size is $N_{j+1}$. By choosing $\mathcal{C}_J = \emptyset$, the empty set, the trial must stop at or before the $J$th analysis.

For the purposes of our illustration, we consider the stopping rule actually used in the sepsis clinical trial. Using the nomenclature from the companion paper [1], the stopping rule *Futility.8* is a level 0.025 one-sided stopping rules from the unified family [5] having O'Brien–Fleming lower (efficacy) boundary relationships and upper (futility) boundary relationships corresponding to boundary shape parameters $P = 0.8$. In this parameterization of the boundary shape function, parameter $P$ is a measure of conservatism at the earliest analyses. $P = 0.5$ corresponds to Pocock boundary shape functions, and $P = 1.0$ corresponds to O'Brien–Fleming



DOI: 10.1002/sim

Table I. Posterior probabilities of hypotheses for trial results corresponding to stopping boundaries of *Futility.8* stopping rule with four equally spaced analyses after 425, 850, 1275, and 1700 subjects have been accrued to the study.

| | Efficacy (lower) boundary | | | | Futility (upper) boundary | | | |
|---|---|---|---|---|---|---|---|---|
| | Posterior probability of beneficial treatment effect $\Pr(\theta \leqslant 0 \mid X)$ | | | | Posterior probability of insufficient benefit $\Pr(\theta \geqslant -0.087 \mid X)$ | | | |
| Analysis time | Crude est of trt effect | Optimistic $\zeta = -0.09$ | Sponsor's consensus $\zeta = -0.04$ | Pessimistic $\zeta = 0.02$ | Crude est of trt effect | Optimistic $\zeta = -0.09$ | Sponsor's consensus $\zeta = -0.04$ | Pessimistic $\zeta = 0.02$ |
| *Dogmatic prior*: $\tau = 0.015$ | | | | | | | | |
| 1: $N = 425$ | −0.170 | 1.000 | 1.000 | 0.524 | 0.047 | 0.795 | 1.000 | 1.000 |
| 2: $N = 850$ | −0.085 | 1.000 | 1.000 | 0.523 | −0.010 | 0.824 | 1.000 | 1.000 |
| 3: $N = 1275$ | −0.057 | 1.000 | 1.000 | 0.522 | −0.031 | 0.836 | 1.000 | 1.000 |
| 4: $N = 1700$ | −0.042 | 1.000 | 1.000 | 0.521 | −0.042 | 0.842 | 1.000 | 1.000 |
| *Consensus prior*: $\tau = 0.040$ | | | | | | | | |
| 1: $N = 425$ | −0.170 | 1.000 | 1.000 | 0.991 | 0.047 | 0.981 | 0.999 | 1.000 |
| 2: $N = 850$ | −0.085 | 1.000 | 0.998 | 0.974 | −0.010 | 0.976 | 0.997 | 1.000 |
| 3: $N = 1275$ | −0.057 | 0.999 | 0.993 | 0.955 | −0.031 | 0.970 | 0.994 | 1.000 |
| 4: $N = 1700$ | −0.042 | 0.998 | 0.987 | 0.936 | −0.042 | 0.963 | 0.991 | 0.999 |
| *Non-informative prior*: $\tau = \infty$ | | | | | | | | |
| 1: $N = 425$ | −0.170 | 1.000 | 1.000 | 1.000 | 0.047 | 0.999 | 0.999 | 0.999 |
| 2: $N = 850$ | −0.085 | 0.998 | 0.998 | 0.998 | −0.010 | 0.995 | 0.995 | 0.995 |
| 3: $N = 1275$ | −0.057 | 0.989 | 0.989 | 0.989 | −0.031 | 0.988 | 0.988 | 0.988 |
| 4: $N = 1700$ | −0.042 | 0.977 | 0.977 | 0.977 | −0.042 | 0.981 | 0.981 | 0.981 |

Posterior probabilities are computed based on optimistic, the sponsor's consensus, and pessimistic centring of the priors using three levels of assumed information in the prior. The variability of the likelihood of the data corresponds to the alternative hypothesis: event rates of 0.30 in the control group and 0.23 in the treatment group.

boundary relationships. The choice $P = 0.8$ is thus intermediate between those two, and tends to be fairly similar to a triangular test stopping boundary. Table I presents the stopping boundaries on the scale of the crude estimate of treatment effect $\hat{\theta}_j$ for four equally spaced analyses.

## 3. BAYESIAN PARADIGM

In the Bayesian paradigm, we consider a joint distribution $p(\theta, X)$ for the treatment effect parameter $\theta$ and the clinical trial data $X$. The marginal distribution $p_\theta(\theta)$ is commonly termed the 'prior' distribution for the treatment effect parameter, because it represents the information about $\theta$ prior to (in the absence of) any knowledge of the value of $X$. From a clinical trial, we observe data $X = x$ and base inference on the conditional distribution $p_{\theta|X}(\theta|X = x)$, which is commonly termed the 'posterior' distribution. As with frequentist inference, we are interested in point and interval estimates of a treatment effect, a measure of strength of evidence for or against particular hypotheses, and perhaps a binary decision for or against some hypothesis. Commonly used Bayesian



DOI: 10.1002/sim

inferential quantities include:

1. *Point estimates of treatment effect* that are summary measures of the posterior distribution such as the posterior mean ($E(\theta|X=x)$), the posterior median ($\theta_{0.5}$ such that $\Pr(\theta \leqslant \theta_{0.5}|X=x) \geqslant 0.5$ and $\Pr(\theta \geqslant \theta_{0.5}|X=x) \geqslant 0.5$), or the posterior mode ($\theta_m$ such that $p_{\theta|X}(\theta_m|X=x) \geqslant p_{\theta|X}(\theta|X=x)$ for all $\theta$).
2. *Interval estimates of treatment effect* that are computed by finding two values $(\theta_L, \theta_U)$ such that $\Pr(\theta_L \leqslant \theta \leqslant \theta_U|X=x) = 100(1-\alpha)$ per cent. Various criteria can be used to define such 'credible intervals':
   (a) the central $100(1-\alpha)$ per cent of the posterior distribution of $\theta$ is defined by finding some $\Delta$ such that $\theta_L = \hat{\theta} - \Delta$ and $\theta_U = \hat{\theta} + \Delta$ provides the desired coverage probability, where $\hat{\theta}$ is one of the Bayesian point estimates of $\theta$;
   (b) the interquantile interval is defined by defining $\theta_L = \theta_{\alpha/2}$ and $\theta_U = \theta_{1-\alpha/2}$, where $\theta_p$ is the $p$th quantile of the posterior distribution, i.e. $\Pr(\theta \leqslant \theta_p|X=x) = p$;
   (c) the highest posterior density (HPD) interval is defined by finding some threshold $c_\alpha$ such that the choices $\theta_L = \min\{\theta : p_{\theta|X}(\theta|X=x) > c_\alpha\}$ and $\theta_U = \max\{\theta : p_{\theta|X}(\theta|X=x) > c_\alpha\}$ provide the desired coverage probability. (Note that in the case of a multimodal posterior density, this definition of a HPD interval may include some values of $\theta$ for which the posterior density does not exceed the given threshold. Hence, one could ostensibly define an HPD *region* that was smaller in this setting.)
3. *Posterior probabilities or Bayes factors* [6] *associated with specific hypotheses* that might be used by Bayesians to make a decision for or against a particular hypothesis. For instance, in the sepsis trial example, we might be interested in computing the posterior probability of the null hypothesis $\Pr(\theta \geqslant 0|X=x)$ or the posterior probability of the design alternative $\Pr(\theta \leqslant -0.07|X=x)$.

In specifying a Bayesian probability model, we most often specify the prior distribution $p_\theta(\theta)$ and the likelihood function $p_{X|\theta}(X|\theta)$, rather than specifying the joint distribution $p(\theta, X)$ directly. Upon observation of $X=x$, the posterior distribution is then computed using Bayes rule as

$$p_{\theta|X}(\theta|X=x) = \frac{p_{X|\theta}(X|\theta)\, p_\theta(\theta)}{\int p_{X|\theta}(X|\theta)\, p_\theta(\theta)\, \mathrm{d}\theta}$$

We note that the Bayesian inference presented above is unaffected by the choice of stopping rule, so long as there is no need to consider the joint distribution of estimates across the multiple analyses of the accruing data. That is, so long as one is content to regard inference at each analysis marginally, then the stopping rule used to collect the data is immaterial. However, the expected cost of a clinical trial does depend very much on the stopping rule used, even when Bayesian inference is used as the basis for a decision.

### *3.1. Frequentist versus Bayesian criteria (and the role of the likelihood principle)*

As noted above, much of statistical inference is concerned with quantification of the strength of statistical evidence in support of or against particular hypotheses and with quantification of the precision with which we can estimate population parameters. There are two major categories of such inferential methods: Bayesian and frequentist. Frequentist measures of statistical evidence



DOI: 10.1002/sim

and precision such as the $P$ value and confidence intervals are currently the most commonly used approaches upon which statistical decisions are based, and frequentist optimality criteria for estimators such as bias and mean squared error are perhaps most commonly used for selecting the estimators of treatment effect. However, frequentist inference by no means enjoys universal acceptance [7, 8]. Because frequentist inference merely provides information about the probability of obtaining the observed data under specific hypotheses, it is not truly addressing the question of greatest interest: after observing the data, what is the probability that a treatment is truly beneficial? Bayesian inference answers this latter question by assuming a prior probability distribution for the treatment effect, and then using the data to update that distribution. Many adherents of Bayesian inference note that it, unlike frequentist inference, adheres to the likelihood principle [9]. The likelihood principle states that all information in the data relevant to discriminating between hypotheses is captured by the ratio of the likelihoods under those hypotheses, and that inference that is not based on that ratio is not as relevant.

Our position is that there is no true conflict between frequentist and Bayesian inference. Instead, they merely answer different questions. If we consider that there exists a joint distribution of the parameter $\theta$ and the estimate $\hat{\theta}$ of that parameter, then frequentist inference and Bayesian inference can be viewed as considering different conditional distributions derived from that same joint distribution: frequentist inference considers the conditional distribution $p(X|\theta)$, and Bayesian inference considers the conditional distribution $p(\theta|X)$. It is this view that led us to focus primarily on the evaluation of stopping rules under frequentist or Bayesian frameworks without regard for the original derivation of the stopping rule. We regard that it is the role of statistics to help quantify the strength of evidence used to convince the scientific community of conclusions reached from studies. As we believe that reasonable people might demand evidence demonstrating results that would not typically be obtained under any other hypothesis, our job as statisticians is to try to answer whether such results have been obtained. These frequentist criteria were addressed in the companion paper on the frequentist evaluation of stopping rules. We similarly believe that in a Bayesian framework where one demands evidence that overpowers his/her prior beliefs, we should also address those questions. It is this second situation that leads to the focus on Bayesian evaluation of clinical trial designs in this paper. In adopting this position of accepting both frequentist and Bayesian inferential measures, we are clearly taking the stance that the likelihood principle is not the only guiding principle of all statistics. That is, if one wishes to address the frequentist criteria for evidence it is necessary to account for the sampling distribution.

As sequelae of this philosophy of using both frequentist and Bayesian inference to address different standards of proof within the same setting, it would seem most appropriate to use the same probability model in each approach. This philosophy also argues that it is never sufficient to use any single prior distribution for the population parameters when providing Bayesian inference. We address these issues in greater detail in the following sections.

*3.1.1. Coarsened Bayesian approach.* As noted above, Bayesian inference is based on the conditional distribution of the treatment effect parameter $\theta|X = x$. Frequentist inference, on the other hand, considers the conditional distribution of the data $X|\theta$. These two approaches to statistical inference are complementary when the same probability model $p(\theta, X)$ is used for all inference. There is, however, a tendency for frequentists to interpret their inference in a distribution-free manner, while the overwhelming majority of Bayesian analyses are fully parametric.

The use of parametric analyses (and, indeed, most commonly used semi-parametric analyses) seems inappropriate in the scientific setting of most clinical trials. Most often, the scientific



DOI: 10.1002/sim

question to be addressed when investigating a new treatment is whether the treatment results in a tendency toward higher *or* lower values for some clinical outcome (but not both). Because of this primary focus on the central tendency (or location) of some probability distribution, we might choose measures of treatment effect based on the mean, median, proportion or odds above some clinically relevant threshold, or the probability that a randomly chosen treated subject would have an outcome larger than a randomly chosen subject receiving the control treatment (the *de facto* treatment effect parameter tested with a Wilcoxon rank sum test). A parametric model in such a setting corresponds to making assumptions more detailed than needed in order to address the question. For instance, in the case of inference about the mean outcome, use of a parametric model is tantamount to admitting that we do not know how the treatment affects the first moment of the distribution of outcomes, but imagining that we do know how it affects the variance, skewness, kurtosis, and all higher moments of the probability distribution. With the exception of independent binary responses, such an assumption would seem illogical based on the current state of knowledge at the start of a clinical trial, and it would also seem unlikely that the effect of a treatment on a population would truly be such that an assumption of this type might hold. Instead, it is quite likely that unidentified subgroups would be either less or more susceptible to a treatment. In that setting, a treatment that has some effect on the primary outcome would tend to have a mixture distribution.

The foundational problem with parametric analyses is also present in those semi-parametric probability models that assume that a finite dimensional parameter of interest and an infinite dimensional nuisance parameter related to a single population's distribution provides full information about the distribution of outcomes in *every* population. For instance, when comparing means or medians, some data analysts consider a semi-parametric location-shift model in which the control group has some unknown distribution of outcomes (i.e. $\Pr(X_{0i} \leqslant x) = F_0(x)$ arbitrary), but that the distribution of outcomes in the treatment group is known to have the shape shifted higher or lower (i.e. $\Pr(X_{1i} \leqslant x) = F_0(x - \theta)$). Similarly, the proportional hazards model commonly used in survival analysis allows an arbitrary distribution for the control population, with a relationship between the control and treatment groups that would have $\Pr(X_{1i} \geqslant x) = [\Pr(X_{0i} \geqslant x)]^{\theta}$. In such models, the semi-parametric distributional assumption can be quite strong, and departures from the assumption can adversely affect the statistical inference.

Clearly, when a treatment has some effect on the outcome, failure to have the correct parametric or semi-parametric model will mean that the probability statements associated with statistical inference (e.g. $P$ values, coverage probabilities, posterior means, unbiasedness) will not hold. In some cases, however, frequentist testing can proceed, because it is based only on knowing the distribution of the outcome under the null hypothesis. If the null hypothesis to be tested is that the treatment has no effect on outcome whatsoever, then the null distribution of the estimated treatment effect can be approximated by pooling all of the data (as is done with the asymptotic test of two binomial proportions and the Wilcoxon rank sum test), by using only the control group's data (which, though not commonly done, could in some cases result in more powerful tests than the other approaches described here), or by assuming some semi-parametric model that reduces to equivalence of distributions under the null (as is commonly done in the $t$-test for equal variances and proportional hazards models). In the latter instance, pooled estimates of the nuisance null variance (in the case of the $t$-test presuming equal variances) or the baseline survivor distribution (in the case of proportional hazards model) can be used, because under the null hypothesis they would be estimated correctly. The fact that any of these standard error estimates might be incorrect under alternative distributions is immaterial, because by hypothesis that can only happen when the



DOI: 10.1002/sim

null hypothesis is false. A problem does arise, however, when it comes to scientific interpretation of a significant test. For instance, it is easily shown that the $t$-test presuming equal variances and the Wilcoxon rank sum test will reject the null hypothesis with probability greater than the nominal type I error in some cases where the treatment affects the variance of the outcome measurements without affecting the mean, median, or probability that a randomly chosen treated individual has a measurement larger than a randomly chosen individual in the control group. Because, as noted above, most scientific questions are more closely related to central tendencies of the distribution of outcome, this would suggest that the use of such semi-parametric models may be misleading even in the frequentist testing.

Fortunately, however, there are robust (distribution-free) probability models that do allow inference about the most commonly used measures of treatment effect. Koprowicz *et al.* [10] note that so long as asymptotically normally distributed non-parametric estimates of treatment effect are used along with correct modelling of how the standard errors of those estimates vary under all alternatives, robust inference is possible. In fact, it is common for frequentist analyses to be based on non-parametric estimates of treatment effect parameters. Modification of the standard error estimates (e.g. using the $t$-test for unequal variance rather than the $t$-test for equal variance) then provides robust inference about the treatment effect in large samples. Such an approach can also be used for robust Bayesian inference as explored by Pratt *et al.* [11], Boos and Monahan [12], Monahan and Boos [13], and Koprowicz *et al.* [10]. In the clinical trial setting, using such an approach allows frequentist and Bayesian inference to be based on the same probability model—a condition not easily duplicated when non-parametric Bayesian inference is based on Dirichlet process priors.

We relax some of the more restrictive assumptions of a parametric or semi-parametric model and adopt a robust approach here. A non-parametric estimator $\hat{\theta}_N = t(X = (X_1, \ldots, X_N))$ consistent for the treatment effect parameter $\theta$ is viewed as a coarsening of the data. In a wide variety of settings, the non-parametric estimator can be shown to be approximately normally distributed with large sample sizes. Hence, we consider the Bayesian paradigm based on a joint distribution $p(\theta, \hat{\theta})$ with marginal (prior) distribution $p_\theta(\theta)$ for the treatment effect parameter and approximate likelihood based on an asymptotic distribution $\hat{\theta}_N \dot{\sim} \mathcal{N}(\theta, V(\theta)/N)$. Hence, the posterior distribution for $\theta|\hat{\theta}_N$ is computed according to

$$\hat{p}_{\theta|\hat{\theta}_N}(\theta|\hat{\theta}_N) = \frac{\sqrt{\frac{N}{V(\theta)}}\phi\left(\frac{(\hat{\theta}_N - \theta)}{\sqrt{V(\theta)/N}}\right) p_\theta(\theta)}{\int \sqrt{\frac{N}{V(\theta)}}\phi\left(\frac{(\hat{\theta}_N - \theta)}{\sqrt{V(\theta)/N}}\right) p_\theta(\theta)\,\mathrm{d}\theta}$$

Such a coarsening of the data has little effect on the efficiency of Bayesian inference when the non-parametric estimator is in fact a sufficient statistic for the data. In that case, the only loss of efficiency is from using the approximate normal distribution of the sufficient statistic rather than the exact distribution [10].

*3.1.2. Choice of prior distributions.* Bayesian inference can depend heavily on the choice of prior distribution $p_\theta(\theta)$ for the treatment effect parameter. For that reason, Bayesian inferential procedures have sometimes been criticized because it is not clear how the prior distribution should be selected for any particular problem. When relevant prior data are available, it would



DOI: 10.1002/sim

seem most sensible to use that prior data to derive a prior distribution. Most often, however, the exact relevance of data from pilot studies is unclear due to changing inclusion/exclusion criteria, changing definitions of the study treatment, and changing standards of ancillary care. Hence, the prior distribution is probably best regarded as a subjective probability measuring an individual's prior belief about the treatment effect.

Much has been written in the statistical literature about methods of eliciting priors from a consensus of experts, as well as the need to consider a range of priors that cover both 'pessimistic' and 'optimistic' priors [8, 14]. While examining inference specific to a single 'expert' prior is indeed often of interest, we regard the sensitivity analysis approach as the more important one. We believe that many Bayesian data analysts' failure to do so in the past is, at least in part, responsible for the lack of greater penetrance of Bayesian methods into the applied clinical trials literature. That is, the purpose of scientific experimentation is to present to the scientific community (for early phase trials) and larger clinical community (for phase III studies leading to the adoption of new treatments) credible evidence for or against specific hypotheses. As a general rule, the investigators collaborating on a particular clinical trial are likely to be more optimistic about that new treatment's benefits than the typical member of the scientific community. For instance, Carlin and Louis [15] describe a setting in which the likelihood function suggested a harmful treatment, but several clinicians had put no prior mass on the possibility of a negative treatment effect. A Bayesian analysis incorporating only their prior would likely be too biased for general utility. Furthermore, the role of a prior is buried too deeply in the computations of a posterior distribution to allow a reader to assess the impact of a different prior on the types of Bayesian inference routinely provided. Hence a standard of presentation is needed that will convey trial results for a wide spectrum of assumed prior distributions.

The approach we take is to use a spectrum of normal prior distributions specified by their mean and standard deviation. Such a choice has several advantages, as well as disadvantages. The primary advantage is that a normal prior is specified entirely by two parameters, greatly reducing the dimension of the space of prior distributions to be considered, while still covering a very broad range of choices for the prior. Furthermore, means and standard deviations are commonly used and understood by many researchers. The mean of the prior distribution will in some sense measure the optimism or pessimism in the prior, and the standard deviation of the prior distribution can be regarded as a measure of dogmatism in those prior beliefs, with lower standard deviations for the prior indicative of more dogmatic beliefs about the benefit or lack of benefit (harm) of the treatment *a priori*. Many researchers are also quite familiar with common properties of the 'bell-shaped' curve, such as the fact that approximately 95 per cent of the measurements are within two standard deviations of the mean.

This approach is probably the most obvious standard in a setting in which investigators have not truly characterized their entire prior distribution, but do have ideas of a prior mean and standard deviation for the treatment effect parameter. In this setting, using a normal prior will tend to underestimate the amount of information in an individual's true prior. The normal distribution is known to maximize entropy over the class of all priors having the same first two moments [16]. This property suggests that in our approach we will in some sense use the least informative of all distributions that could reasonably reflect an individual's true prior. This is beneficial to the extent that many researchers tend to voice priors that are too dogmatic for the available prior information or, indeed, their actions. This is potentially deleterious if an individual's well-based prior is more informative than the normal prior with the same mean and standard deviation. In the latter case, a Bayesian analysis with a normal prior may suggest that the data has overwhelmed



DOI: 10.1002/sim

an investigator's initial belief, when it in fact has not. This latter problem can be ameliorated somewhat by the investigator considering other normal priors that are either more informative (i.e. priors having lower standard deviations), more extreme (i.e. priors having means that are further from the observed value of $\hat{\theta}$), or both.

The choice of normal priors also offers a computational advantage when the distribution of the treatment effect estimate does not exhibit a mean–variance relationship, i.e. when $V(\theta)$ is constant. In this case, the normal distribution is the conjugate prior for the asymptotic distribution of $\hat{\theta}|\theta$, and the posterior distribution is then known to be normal as well. This computational advantage is increasingly less important, however, with the advent of advanced computational methods such as Markov chain, Monte Carlo which can be used to obtain samples from the posterior distribution for the general case when $V$ depends upon $\theta$.

Under the probability model $\hat{\theta}_N|\theta \dot{\sim} \mathcal{N}(\theta, V/N)$ (so no mean–variance relationship) and a prior distribution $\theta \sim \mathcal{N}(\zeta, \tau^2)$, the posterior distribution for the treatment effect is

$$\theta|\hat{\theta}_N \dot{\sim} \mathcal{N}\left(\frac{(1/\tau^2)\zeta + (N/V)\hat{\theta}_N}{(1/\tau^2) + (N/V)}, \frac{1}{(1/\tau^2) + (N/V)}\right)$$

It is often useful to measure the variance of the prior distribution as the 'effective sample size' in the prior information. That is, if $N_0 = V/\tau^2$, the prior distribution is as informative as the posterior distribution from a Bayesian analysis of a sample of $N_0$ subjects and an initially non-informative prior. Similarly, for a study having sample size $N$, the ratio $V/(N\tau^2)$ measures the statistical information presumed in the prior relative to the statistical information contributed by the new data.

## 4. EVALUATION OF STOPPING RULES

The Bayesian evaluation of stopping rules proceeds much the same as for the evaluation of stopping rules with respect to frequentist inference. The major difference relates to the magnitude of the results that need to be presented. When evaluating a stopping rule with respect to frequentist inference, we present the estimate, confidence interval, and $P$ value for clinical trial results corresponding to the stopping boundaries. For Bayesian inference, we must consider how that inference is affected by the choice of prior.

We illustrate Bayesian evaluation of a stopping rule in the context of the sepsis clinical trial example. For ease of presentation, in the example presented here, we suppress the mean–variance relationship inherent in the binomial probability model. Hence, in evaluating the design, we use a variance $V = 0.7742$, which corresponds to the average variance contributed by each observation under the design alternative of 30 per cent 28 day mortality on the control arm and 23 per cent 28 day mortality on the antibody arm.

We specify a pessimistic prior in order to judge whether trial results corresponding to a decision for efficacy (i.e. below the lower stopping boundary) are so strong as to convince a person who believed the new treatment was not efficacious, or even was harmful. Similarly we specify an optimistic prior in order to judge the strength of evidence when trial results correspond to a decision that the new treatment is not sufficiently efficacious to warrant further study. We also consider a prior representing the consensus of opinion of trial collaborators and consultants to the trial sponsor.



DOI: 10.1002/sim

A pessimistic prior for this trial might be centred at a prior mean of $\zeta = 0.02$, suggesting that treatment with the antibody provides harm to the population of treated patients. An optimistic prior, on the other hand, might be centred on a prior mean of $\zeta = -0.09$, representing a treatment effect greater than the 7 per cent absolute improvement in 28 day mortality used in the sample size computation. The strength of optimism or pessimism in the prior is also affected by the standard deviation of the prior distribution for the treatment effect. For instance, a choice of $\tau = 0.015$ is relatively dogmatic, because it suggests that accrual of the full sample size of 1700 subjects provides only half the amount of information about the treatment effect as is already included in the prior. As $\tau$ becomes very large, the prior is increasingly less informative about the magnitude of the treatment effect.

When this phase III sepsis study was conducted, preliminary data was available from several phase II and phase III studies that showed some promising trends toward benefit, especially in some important subgroups. Given this preliminary data, it would seem quite reasonable to base a prior for $\theta$ on the analysis results from those previous studies. Several factors mitigate against doing this blindly, however. First, although it was scientifically plausible that greater treatment effect might be seen in the subgroups identified in the earlier studies, identification of the subgroups of greatest interest were in fact based on some *post hoc* data-driven analyses. Because Bayesian analyses are no more able than frequentist analyses to handle the multiple comparison issues inherent in such data dredging, it is wise to discount the results from the earlier studies in anticipation of some 'regression to the mean'. Also, the inclusion/exclusion criteria were modified for the planned study, so it would be inappropriate to assume that the statistical information in the prior should reflect the full sample size previously exposed to the antibody. A reasonable subjective prior based on the preliminary studies might then consider a prior mean for $\theta$ of $\zeta = -0.04$, and the statistical information presumed in the prior might correspond to the prior distribution having a standard deviation of $\tau = 0.04$, which suggests that the information in the data from 1700 subjects is approximately 3.5 times the information that is presumed in the prior.

The general idea is then to present the inference that would be made if the study were to stop with specific results. Table I presents the posterior probability of a beneficial treatment (i.e. the posterior probability that $\theta<0$) for trial results corresponding exactly to the efficacy boundary and, for trial results corresponding to the futility boundary, the posterior probability that the treatment is not sufficiently beneficial to warrant further study (i.e. the posterior probability that $\theta>-0.087$, which is the alternative for which the planned study has power 0.975). The priors considered in Table I include the optimistic, sponsor's consensus, and pessimistic priors centred at $\zeta = -0.09$, $-0.04$, and 0.02, respectively, with standard deviations corresponding to a high level of dogmatism, the sponsor's consensus, and non-informative $\tau = 0.015$, 0.04, and $\infty$, respectively. This table illustrates the impact that choice of prior can have on Bayesian inference.

For instance, the *Futility.8* design ultimately selected as the stopping rule for the clinical trial would suggest that the trial would be stopped for efficacy (i.e. the trial results suggest a true benefit due to treatment with antibody) if the estimate of treatment effect were $\hat{\theta} = -0.087$ after 850 subjects (425 per arm) had been accrued to the study. With a dogmatic prior centred on a pessimist's belief that the treatment was truly harmful, the posterior probability of a true benefit due to treatment is only 0.523 after obtaining such results. Equally dogmatic priors centred at the sponsor's consensus prior belief or an optimist's prior belief of treatment effect suggest near certain posterior probabilities of treatment benefit. As less dogmatic priors corresponding to the



DOI: 10.1002/sim

level of information in the sponsor's consensus prior are used, those posterior probabilities are 0.974, 0.998, and 1.000. When non-informative (flat) priors are used as the basis for Bayesian inference, identical values are obtained for the pessimistic, sponsor's consensus, and optimistic centring of the priors: a posterior probability of a beneficial treatment of 0.998. This latter result is exactly 1 minus the fixed sample $P$ value for this trial outcome (see Table I in the companion paper on frequentist evaluation [1])—a correspondence that is obtained whenever a flat prior is used for Bayesian inference.

The difference between the conservatism of the efficacy and futility boundaries for this trial is evident in the Bayesian posterior probabilities corresponding to stopping at each analysis. The *Futility.8* design would suggest that the trial would be stopped for futility (i.e. the trial results do not suggest sufficient benefit from the treatment to warrant further study of the antibody in this clinical trial) if the estimate of treatment effect were $\hat{\theta} = -0.0097$ after 850 subjects (425 per arm) had been accrued to the study. Under the sponsor's consensus prior, the posterior probability $\Pr(\theta > -0.0866|X) = 0.997$ is slightly less certain than that when stopping for efficacy.

Although such a prior fairly accurately reflects the prior beliefs elicited from the study sponsors during the selection of the stopping rule, as noted above, it is important to also report the result of Bayesian analyses for a spectrum of priors. For instance, one logical strategy would be to evaluate the futility boundary under an optimistic prior, because the burden of proof for establishing futility should be to convince the optimists. Similarly, the efficacy boundary might be evaluated under a pessimistic prior. It is interesting to note from Table I that the futility boundary is somewhat paradoxically anticonservative from the viewpoint of a researcher with the optimistic, dogmatic prior: stopping for futility at the first analysis confers less certainty about an insufficiently effective treatment than does the stopping boundary at later analyses.

The optimistic and pessimistic priors presented in Table I were chosen rather arbitrarily, and thus may not be relevant to some of the intended audience for the published results of a clinical trial. We thus can also present contour plots of Bayesian point estimates (posterior means), lower and upper bounds of 95 per cent credible intervals, and posterior probabilities of the null and alternative hypotheses for a spectrum of prior distributions.

The mean and standard deviation of the prior distribution is probably the most scientifically relevant, while being statistically concise, way to specify the normal prior, as both the mean and standard deviation are in the units of the treatment effect parameter itself. For this reason, when displaying Bayesian inferential statistics as a function of the prior in a contour plot, we label the $x$-axis by the mean of the normal prior and the $y$-axis by the prior standard deviation. However, there are additional scales on which it is at times useful to characterize the prior distribution. In the case of the prior mean for the treatment effect parameter $\theta$, it is often useful to consider the power of the designed clinical trial to detect a particular value of $\theta$. This is particularly the case when the stopping rule corresponds to a frequentist trial design in which the study has been adequately powered to detect the minimal treatment effect judged to be of clinical importance. We thus find it useful at the top of the contour plot to provide readers with an alternative labelling of the $x$-axis according to the power of a frequentist test based on the stopping rule.

Similarly, at the right side of the contour plot, we provide readers with an alternative scale for the standard deviation of the prior distribution. When using a conjugate normal prior distribution with a normal sampling distribution for the estimated treatment effect, the mean of the posterior



DOI: 10.1002/sim

distribution for $\theta|\hat{\theta}$ is a weighted average of the prior mean and $\hat{\theta}$, where the weights are the statistical information in the prior (equal to $1/\tau^2$, where $\tau$ is the standard deviation of the prior distribution) and the statistical information in the approximate likelihood for $\hat{\theta}$ (equal to $N/V(\theta)$). It may therefore be of interest to consider the ratio of information presumed in the prior to the total information expected if the study were to continue until the maximal sample size were accrued. On the right axis of the contour plot we display for selected values of $\tau$ the value of $N_J/(V(\theta)\tau^2)$. As that ratio goes to 0 (equivalently, as $\tau$ becomes large), the prior is tending to be 'non-informative' in that all information in the posterior distribution is derived from the data rather than from the prior. Bayesian inference with non-informative priors result in similar point estimates, confidence intervals, and statistical decisions as derived under frequentist procedures, though the interpretation of those inferential measures remains different.

As a general rule, we might expect that priors that contain more information than would be present in the maximal sample size for the study would not be very reasonable for the study investigators as a whole. (Why would they bother doing the study if the complete data were unable to sway their opinion?) Nonetheless, other members of the scientific community may indeed have such dogmatic priors.

Lacking any other particular criteria by which to choose the range of means $\zeta$ and standard deviations $\tau$ for the prior distribution that should be considered, we choose values of prior mean $\zeta$ corresponding to the values of $\theta$ for that the frequentist test based on the stopping rule would have power between 0.001 and 0.999, and we choose values of the standard deviation $\tau$ corresponding to prior information ranging between one-twentieth to four times as much information as would be present in the data should the study continue until 1700 subjects were accrued. Figure 1 presents contours of the Bayesian point estimate of treatment effect that would be reported under the spectrum of priors described above. From the plot we see that use of the sponsor's consensus prior of $\theta \sim \mathcal{N}$ ($\zeta = -0.04$, $\tau^2 = 0.04^2$) would suggest that upon observing a difference in proportions of $-0.0097$, the mean, median, and mode of the posterior distribution is $-0.021$. Note that an individual who before the study was quite sure the treatment worked and thus had a prior of $\theta \sim \mathcal{N}$ ($\zeta = -0.08$, $\tau^2 = 0.015^2$) would use the posterior mean of $-0.067$ as the point estimate of treatment effect. Of course, such an individual was assuming a prior that was twice as informative as would be provided by data on 1700 subjects, as indicated by the dotted horizontal line.

Similar contour plots can be constructed for the lower and upper bounds for the 95 per cent credible interval, as well as the posterior probability of the null and design alternative hypotheses. At the end of a clinical trial, contour plots such as these can be used in a report of Bayesian inference, thereby allowing readers to assess the credibility of the results across a wide range of priors. At the design phase, when we are trying to assess the Bayesian operating characteristics of the entire stopping rule, such plots can be shown for results that correspond to each of the stopping boundaries. Figure 2 shows such a plot for a stopping rule with boundaries the same shape as those for *Futility.8*, but having only two analyses: one interim analysis halfway through the study (i.e. after 425 patients accrued to each arm; Figure 2(a)) and the final analysis when 1700 subjects (850 per arm; Figure 2(b)) have been accrued. The contours in each panel are the posterior probability that the decision made in favour of efficacy by the frequentist stopping rule is correct. Thus for the efficacy boundary that rejects (with 97.5 per cent confidence) the null hypothesis $H_0 : \theta \geqslant 0$, we display the posterior probability $\Pr(\theta<0|(M, \hat{\theta}_{N_M}))$ for each of the efficacy boundaries, where $M$ is the analysis index and $\hat{\theta}_{N_M}$ is the difference in proportions that corresponds to the efficacy



DOI: 10.1002/sim

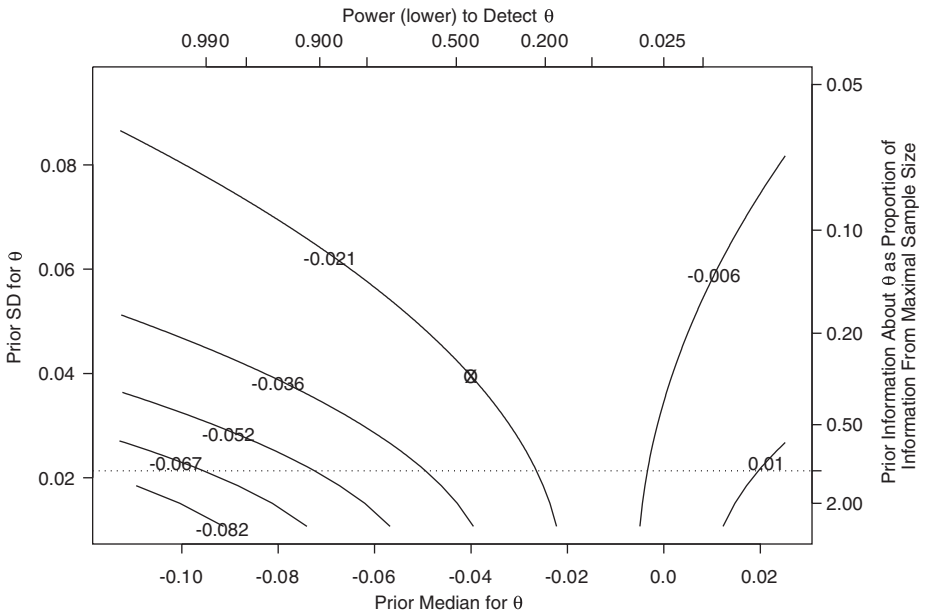


Figure 1. Contours of posterior mean (median) conditional on observing an estimated treatment effect of −0.0097 after accruing 850 subjects (425 per arm). Contours are displayed for normal prior distributions as a function of the mean (median) and standard deviation of the prior distribution, with the prior corresponding to the sponsor's consensus prior indicated by an *X*. Top axis displays the power of the hypothesis test corresponding to stopping rule *Futility.8*, and the right-hand axis displays the prior information in the prior as a proportion of the statistical information in a sample size of 1700 subjects (850 per arm).

boundary at that analysis. For reference, Figure 2(c) displays the contours for the prior probability that $\theta<0$.

From Figure 2, we see that the sponsor's consensus prior corresponded to a $P(\theta>-0.085)$ slightly less than 0.9. Stopping for futility at the interim analysis corresponds to a posterior probability that $\theta>-0.085$ between 0.99 and 0.999. It can also be seen that for all but the most dogmatic of prior beliefs that the treatment had a marked benefit, stopping for futility at the interim analysis corresponds to a posterior probability greater than 0.9 that the treatment does not provide as much as a 0.085 improvement in 28 day mortality. The efficacy boundary could be evaluated in a similar manner.

It is also possible to consider Bayesian prior distributions when evaluating stopping probabilities and sample size distributions. Figure 3 displays the contours of the predicted sample size for the *Futility.8* boundary for a spectrum of prior distributions. For this plot however, we find it useful to use truncated normal prior distributions. As the standard deviation of a normal prior increases, increasing prior probability is placed on very extreme values of $\theta$. In a stopping rule with two boundaries, as the treatment effect $\theta$ gets very large or very small, the trial will tend to stop at the first analysis with probability approaching 1. Thus non-informative priors will all tend to suggest that the predicted sample size is that which corresponds to the timing of the first analysis. To downplay such effects, for contour plots of the expected sample size we truncate the normal



DOI: 10.1002/sim

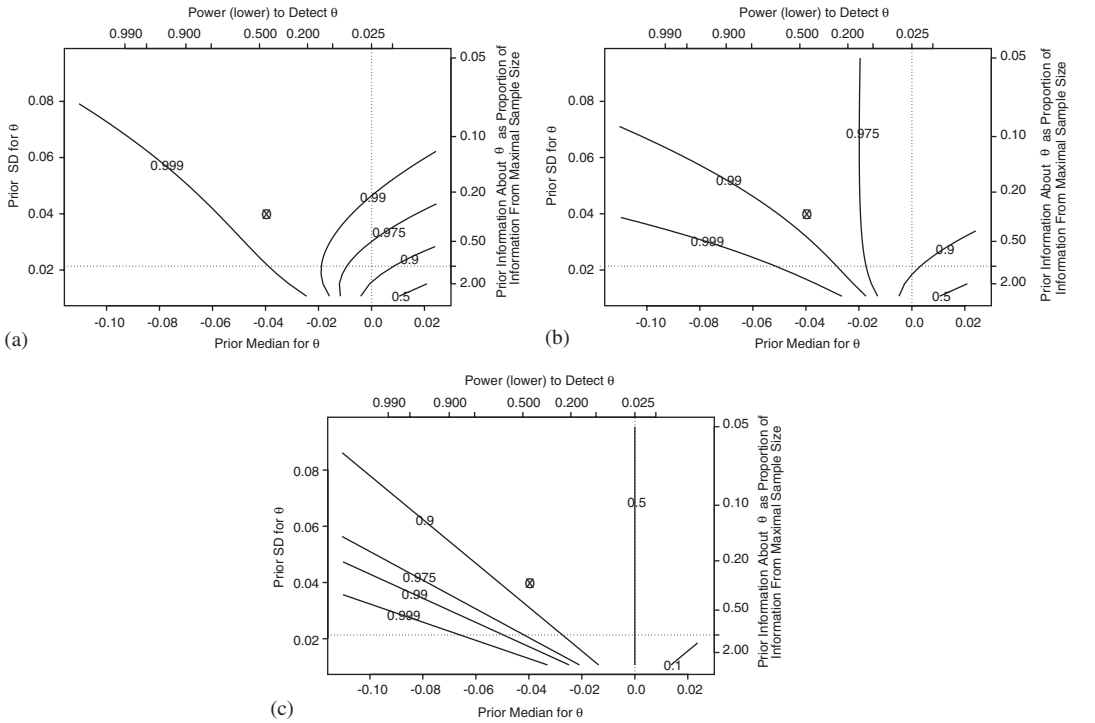


Figure 2. Contours of posterior probabilities at the efficacy boundary for a stopping rule having a maximum of 2 analyses and boundary shapes similar to the *Futility.8* stopping rule. Panel (a) considers the posterior probability that the decision reached at the first interim analysis (with 850 subjects accrued, 425 per arm) will be correct, $\Pr(\theta<0|M=1, T=-0.084)$; panel (b) considers the posterior probability of a correct decision at the final analysis (with 1700 subjects accrued, 850 per arm), $\Pr(\theta<0|M=2, T=-0.042)$; and panel (c) considers the prior probability that the decision to reject the corresponding hypothesis will be correct, $\Pr(\theta < 0)$ (prior). In each case, the prior corresponding to the sponsor's consensus prior indicated by an $X$. Top axis displays the power of the hypothesis test corresponding to the stopping rule, and the right hand axis displays the prior information in the prior as a proportion of the statistical information in a sample size of 1700 subjects (850 per arm).

distribution at specified values of $\theta$, and renorm the prior so that it integrates to 1. For Figure 3 we arbitrarily truncated the prior distribution at the values of $\theta$ for which the *Futility.8* stopping rule had power of 0.001 and 0.999. From this figure we see that under such a truncation of the sponsor's consensus prior the average sample size accrued to the study is between 1150 and 1200. We note that as the standard deviation of the prior approaches 0, the lower limits of the contour plot should correspond to an ASN curve for the stopping rule (see our companion paper [1] dealing with frequentist evaluation of a stopping rule).

We can also use Bayesian methods to address the issue of economically important estimates of treatment effect in a manner analogous to that using frequentist methods [1]. When obtaining such an attractive estimate is of major importance, Bayesian predictive probabilities [17] may also be of use in judging whether it is useful to continue a clinical trial in the hopes of obtaining an economically attractive estimate of treatment effect.



DOI: 10.1002/sim

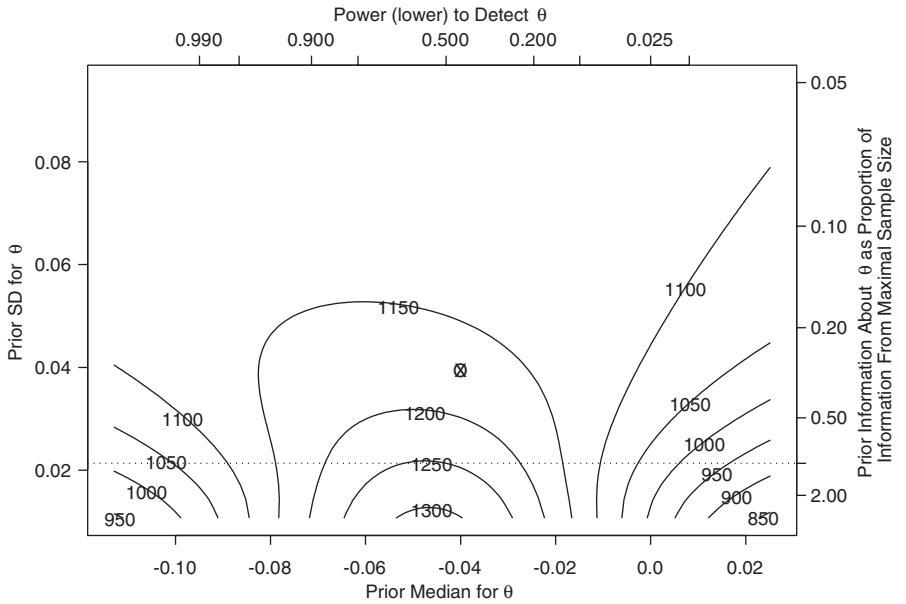


Figure 3. Contours of expected sample sizes for the *Futility.8* stopping rule based on truncated normal priors. Contours are displayed for truncated normal prior distributions as a function of the mean (median) and standard deviation of the normal distribution, with truncation at the values corresponding to the 0.001 and 0.999 power points for the stopping rule. The prior corresponding to the sponsor's consensus prior is indicated by an $X$. Top axis displays the power of the hypothesis test corresponding to the stopping rule, and the right-hand axis displays the prior information in the untruncated prior as a proportion of the statistical information in a sample size of 1700 subjects (850 per arm).

The Bayesian predictive probability is the probability that the estimate would exceed some specified threshold at a particular analysis. The computation uses the updated distribution of the treatment effect parameter at the $j$th analysis (i.e. the posterior distribution conditioned on the observed data at the $j$th analysis) along with the sampling distribution of the as yet unobserved data. Using the coarsened approach based on the approximate normal distribution for the estimated difference in 28 day mortality rates and the computationally convenient conjugate normal prior $\theta \sim \mathrm{N}(\zeta, \tau^2)$, at the $j$th analysis we can define an approximate Bayesian posterior distribution for the true treatment effect $\theta$ conditioned on the observation $\hat{\theta}_j$ as

$$\theta|\hat{\theta}_j \dot{\sim} \mathcal{N}\left(\frac{\hat{\theta}_j\tau^2 + \zeta\sigma^2/N_j}{\tau^2 + \sigma^2/N_j}, \frac{\tau^2\sigma^2/N_j}{\tau^2 + \sigma^2/N_j}\right)$$

Then, using the sampling distribution for the as yet unobserved data and integrating over the posterior distribution, the predictive distribution for the estimate $\hat{\theta}_J$ at the final analysis is

$$\hat{\theta}_J|\hat{\theta}_j \dot{\sim} \mathcal{N}\left(\frac{(\tau^2 + \sigma^2/N_J)\Pi_j\hat{\theta}_j + (1-\Pi_j)\zeta\sigma^2/N_J}{\Pi_j\tau^2 + \sigma^2/N_J}, \frac{(1-\Pi_J)(\tau^2 + \sigma^2/N_J)\sigma^2/N_J}{\Pi_j^2(\Pi_J\tau^2 + \sigma^2/N_J)}\right)$$



DOI: 10.1002/sim

We might therefore compute a predictive probability of an economically attractive treatment estimate less than some threshold $c$ as

$$\int \Pr(\hat{\theta}_J < c | S_j = s_j, \theta)\, p(\theta | S_j = s_j)\, \mathrm{d}\theta$$
$$= \Phi\left( \frac{[\Pi_j \tau^2 + \sigma^2/N_J][c - \hat{\theta}_j] + [1 - \Pi_j][\hat{\theta}_j - \zeta]\sigma^2/N_J}{\sqrt{[1 - \Pi_j][\tau^2 + \sigma^2/N_J][\Pi_j \tau^2 + \sigma^2/N_j]\sigma^2/N_J}} \right)$$

The case of a non-informative (although improper) prior is of special interest. When we consider taking the limit as $\tau^2 \to \infty$), the predictive probability becomes

$$\int \Pr(\hat{\theta}_J < c | S_j = s_j, \theta)\, p(\theta \mid S_j = s_j)\, \mathrm{d}\theta = \Phi\left( \frac{(c - \hat{\theta}_j)\sqrt{\Pi_j}}{\sqrt{[1 - \Pi_j]\sigma^2/N_J}} \right)$$

For instance, for a result corresponding to a crude estimate of treatment effect of $-0.0566$ at the third analysis, the predictive probability of obtaining a crude estimate of treatment effect less than $-0.06$ at the final analysis is 35.0 per cent under the sponsor's consensus prior and 39.0 per cent under a non-informative prior. In either case, such high probabilities of obtaining a more economically viable estimate of treatment effect may be enough to warrant modifying the stopping rule to avoid early termination at the third analysis with a crude estimate between $-0.0566$ and $-0.06$. On the other hand, had the minimal economically viable estimate been $-0.08$, the predictive probabilities of obtaining such an estimate at the fourth analysis after observing $-0.0566$ at the third analysis are 1.92 and 2.86 per cent under the sponsor's consensus and non-informative priors, respectively, and such low probabilities would likely be regarded as support for a clinical trial design based on a smaller maximal sample size.

## 5. CONCLUSION

In this manuscript we have described a general approach to the evaluation of a stopping rule with respect to its Bayesian operating characteristics. Such evaluation is complementary to the exploration of frequentist operating characteristics described in our companion paper [1]. Using a prior distribution for the parameter measuring treatment effect, the Bayesian evaluation of the clinical trial design might typically include the following analogues to the frequentist evaluation criteria:

1. The sample size distribution averaged across a Bayesian prior distribution for the true treatment effect.
2. The Bayesian predictive probability that the trial would continue to each analysis.
3. The Bayesian inference (posterior mean or mode, credible intervals, and posterior probabilities of hypotheses) that would be reported were the trial to stop with results corresponding exactly to a boundary.
4. The Bayesian predictive probability of obtaining a point estimate that exceeds some economically relevant threshold.



DOI: 10.1002/sim

Each of the above criteria will, of course, vary according to the prior distribution assumed for the parameter measuring treatment effect. Because different people will have different priors, it is important that an attempt be made to communicate how the Bayesian operating characteristics will vary as a function of the prior distribution. In this paper, we have suggested that a contour plot would be presented for each of the most important measures. This will not, however, always be feasible for every Bayesian analysis, and thus standards will also have to be adopted for a more parsimonious presentation of Bayesian analyses. Numerical results based on a single consensus prior may be acceptable when there is widespread agreement on the prior. Other times the strength of evidence may be sufficiently strong that one can simply document the most pessimistic prior still consistent with decisions for treatment benefit or the most optimistic prior still consistent with decisions for lack of benefit. In any case, the general 'coarsened Bayesian' approach would argue that a sufficient statistic for a reader to apply his or her own prior would be the estimate $\hat{\theta}$ and the standard error estimate, when that estimator is approximately normally distributed. We note that such an estimator will not in general correspond to the least biased frequentist estimate. Thus, it will generally be the case that the clinical trial results should report both the unadjusted estimator and an estimator adjusted for the stopping rule.

In this paper and its companion paper, we have made the case for frequentist and Bayesian evaluation of clinical trials. We have not addressed other methods such as those based purely on the likelihood principle. Though quantification of statistical evidence via the likelihood principle has received much attention [18–20], we have chosen not to consider these methods in our evaluation of clinical trials. The absence of a discussion on this topic stems from the fact that the definition of statistical evidence via the ratio of likelihoods does not provide any probabilistic interpretation regarding the effect of an experimental treatment. Likelihood ratio values of 8 and 32 have been proposed for distinguishing between weak, moderate and strong evidence in favour of one hypothesis over another [19], yet without a probabilistic interpretation of the value of the likelihood ratio the relevance of these cut-offs is questionable in our minds. Similar problems exist with the construction of support intervals for parameter estimation via the likelihood principle.

## ACKNOWLEDGEMENTS

This research was supported by NIH grant HL69719. We would also like to thank the Editor and three anonymous reviewers whose comments helped to improve this manuscript.

## REFERENCES

1. Emerson SS, Kittelson JM, Gillen DL. Frequentist evaluation of group sequential designs. *Statistics in Medicine* 2005, accepted.
2. Fayers PM, Ashby D, Parmar MKB. Bayesian data monitoring in clinical trials. *Statistics in Medicine* 1997; **16**:1413–1430.
3. Berry DA. Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science* 2004; **19**(1): 175–187.
4. Jennison C, Turnbull BW. *Group Sequential Methods With Applications to Clinical Trials*. CRC Press: Boca Raton, 2000.
5. Kittelson JM, Emerson SS. A unifying family of group sequential test designs. *Biometrics* 1999; **55**:874–882.
6. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1995; **90**:773–795.
7. Berry DA. A case for Bayesianism in clinical trials (Disc: p1395-1404). *Statistics in Medicine* 1993; **12**:1377–1393.
8. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials (Disc: p387-416). *Journal of the Royal Statistical Society*, *Series A*, *General*, 1994; **157**:357–387.



DOI: 10.1002/sim

9. Berry DA. Interim analysis in clinical trials: the role of the likelihood principle (C/R: 88V42 p88-89). *The American Statistician* 1987; **41**:117–122.
10. Koprowicz KM, Emerson SS, Hoff PD. A comparison of parametric and coarsened Bayesian interval estimation in the presence of a known mean–variance relationship. *Technical Report*, Department of Biostatistics, University of Washington, 2003.
11. Pratt JW, Raiffa H, Schlaifer R. *Introduction to Statistical Decision Theory*. MIT Press: Cambridge, MA, 1995.
12. Boos DD, Monahan JF. Bootstrap methods using prior information. *Biometrika* 1986; **73**:77–83.
13. Monahan JF, Boos DD. Proper likelihoods for Bayesian analysis. *Biometrika* 1992; **79**:271–278.
14. Heitjan DF. Bayesian interim analysis of phase II cancer clinical trials. *Statistics in Medicine* 1997; **16**:1791–1802.
15. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall: London, 2000.
16. Cover TM, Thomas JA. *Elements of Information Theory*. Wiley: New York, 1991.
17. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials* 1986; **7**:8–17.
18. Birnbaum A. On the foundations of statistical inference (Com: p307-326). *Journal of the American Statistical Association* 1962; **57**:269–306.
19. Royall RM. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall: London, 1999.
20. Blume JD. Likelihood methods for measuring statistical evidence. *Statistics in Medicine* 2002; **21**(17):2563–2599.





DOI: 10.1002/sim