Group Sequential Clinical Trials: A Classical Evaluation of Bayesian Decision-Theoretic Designs

Roger J. LEWIS and Donald A. BERRY*

Bayesian decision-theoretic designs for a clinical trial comparing two treatments for a disease with binary outcomes are developed and evaluated. The probability of successful outcome with treatment i is denoted by p_i, $i = 1, 2$, and prior knowledge regarding each p_i is assumed to follow a beta distribution. The p_i are assumed to be independent. To facilitate comparison with frequentist clinical trial designs, we take a hypothesis-testing approach. The null hypothesis is $\delta < \delta_0$, and the alternative hypothesis is $\delta > 0$, where δ_0 is the minimum treatment effect sought by the trial and $\delta = p_2 - p_1$ is the true treatment difference. We use a simple terminal loss function reflecting the hypothesis-testing goal of the trial, and the total cost of the trial is the final sample size plus the terminal loss function. The stopping and decision rules that minimize the expectation of the total cost are determined by backward induction. Monte Carlo simulation is used to compare Bayesian and frequentist error rates and mean sample sizes of these Bayesian designs with one-tailed classical group-sequential designs of Pocock and O'Brien–Fleming. As expected, the Bayesian decision-theoretic designs have smaller mean costs than the classical designs. More surprising, when the magnitude of the terminal loss function is chosen to yield frequentist error rates similar to those for classical designs, the mean sample sizes of the Bayesian designs are usually smaller.

KEY WORDS: Bayesian decision theory; Clinical trials; Group-sequential methods; Interim analysis.

1. INTRODUCTION

The design of many modern clinical trials includes a provision for interim data analyses. The function of such interim analyses is to stop a trial early should accumulating data suggest a clear difference between the treatments. If there is a clear difference, continuing the trial would unnecessarily expose some patients in the trial to the less efficacious therapy and delay applying the results to patients outside the trial.

A variety of frequentist statistical procedures for interim analysis have been suggested (Geller and Pocock 1987; Kim and DeMets 1987; Lan and DeMets 1983; O'Brien and Fleming 1979; Pocock 1977). These procedures have the characteristic that they control the overall type I error rate, usually keeping it at .05. The possibility of stopping at an interim point in the trial "spends" a portion of the type I error rate and, in essence, the different classical designs are different plans for allocating the spending of the type I error risk.

The Pocock design (Pocock 1977) uses the same nominal significance level at each interim analysis. For example, if the overall α is .05 and there are two interim analyses and one final analysis (and strictly speaking, one is testing a hypothesis regarding the mean of observations sampled from a normal distribution), then each of the three nominal significance levels is .0221 (Pocock 1977). The O'Brien–Fleming design (O'Brien and Fleming 1979) is more conservative (i.e., spending less of the error risk) than the Pocock design at the early analyses, but less conservative (i.e., allocating more of the error risk) in the later and final analyses, should the trial not be stopped earlier. In the example given earlier, the three nominal significance levels for a O'Brien–Fleming design would be .0005, .0141, and .451 (Geller and Pocock 1987; O'Brien and Fleming 1979).

A larger number of analyses gives more opportunities for early stopping and will decrease the mean sample size should the treatment effect be very large. On the other hand, increasing the number of analyses can actually increase the expected number of patients required for the trial under the null hypothesis, because the nominal significance levels must be adjusted downward to maintain the overall type I error rate. This "penalty" for additional interim analyses decreases the probability of stopping at any particular early analysis. Furthermore, when a larger number of analyses are used, the maximum sample size must be adjusted upward to maintain a given power, because the terminal significance level is decreased as well.

The optimum number of interim analyses generally depends on the uncertainty in the size of the treatment effect. Using a probability density function (pdf) to model the uncertainty in the true treatment effect, the optimum number of interim analyses for a classical group-sequential clinical trial was investigated by McPherson (McPherson 1982). McPherson found between two and ten interim analyses led to a minimum mean sample size, with the larger number of analyses being best when there is great uncertainty in the magnitude of the treatment effect. In situations analogous to most clinical trials, two to five interim analyses would be appropriate.

A Bayesian approach can also be used for interim data analyses. The prior pdf for the treatment effect is continually updated using the Bayes theorem and, at any time, an interim analysis may be conducted by using the current pdf. The likelihood principle implies that interpretation of the data does not depend on the number of inspections or on the stopping rule for the trial. Thus no "penalty" is paid for frequent interim analyses (Berry 1985, 1987). This advantage of the Bayesian approach can be realized whether a simple

* Roger J. Lewis is Assistant Professor, Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, CA, 90509. Donald A. Berry is Professor, Institute for Statistics and Decision Sciences, Duke University, Durham, NC 27706. The authors thank A. F. Parlow for the use of the VAX network at Harbor-UCLA Medical Center. A portion of this work was supported by a grant from the Emergency Medicine Foundation and The Society for Academic Emergency Medicine to Roger J. Lewis. Donald A. Berry was supported by National Science Foundation Grant DMS 8911548 and Public Health Service Grant R01 HS06475-01.

Journal of the American Statistical Association
December 1994, Vol. 89, No. 428, Theory and Methods

stopping rule based on the current pdf is used (Berry 1989; Freedman and Speigelhalter 1989; Lewis 1990; Lewis and Wears 1993) or if a more complex decision-analysis approach is taken (Berry and Ho 1988; Berry, Wolff, and Sack 1992; Lewis and Berry 1992).

Bayesian decision-theoretic designs for clinical trials have been suggested for various types of loss functions. Frequently, the goal considered is to maximize the expected therapeutic benefit over a patient horizon (Anscombe 1963; Berry et al. 1992). Bayesian designs have also been suggested as competitors of the frequentist group sequential approach (Berry 1985), but from a Bayesian perspective.

Bayesian methods have had essentially no impact on the design of actual clinical trials. One reason for this is that frequentist ideas dominate in biostatistical practice and previous discussions have ignored frequentist hypothesis testing characteristics.

In this article we approach the design of clinical trials from an unusual Bayesian perspective. Namely, we use a loss function that explicitly addresses the classical problem of hypothesis testing. The result is a class of Bayesian designs that can be compared with classical group sequential designs on frequentist grounds. We choose parameters of the loss function to yield classical type I and type II error rates comparable with particular Pocock and O'Brien–Fleming designs. We show that the power functions for the Bayesian designs are comparable with those of the frequentist designs. Moreover, we show that the mean sample sizes for trials with Bayesian designs are generally smaller than for the frequentist designs.

It is well known that the Bayesian philosophy regarding interim analysis is very different from the frequentist philosophy (Berry 1987). In particular, in the Bayesian approach the stopping rule is not relevant for drawing an eventual inference (Berger and Berry 1988). But we want to stress that here we use the Bayesian approach only for developing designs. Moreover, we do so using a particular prior probability distribution of the unknown parameters. Once we have a design, we evaluate its characteristics from the frequentist perspective.

2. DESIGN OF THE BAYESIAN TRIAL

We take a Bayesian decision-theoretic approach to the design of a clinical trial comparing two treatments for a disease with a binary outcome, loosely termed success and failure. The true rates of success are denoted by p_i, $i = 1, 2$, and the difference in efficacy is $\delta = p_2 - p_1$. The null hypothesis is $\delta < \delta_0$, where δ_0 is the minimum (positive) difference in efficacy sought by the trial. The alternative hypothesis is $\delta > 0$. These are overlapping hypotheses.

Patients are enrolled into the trial in randomized permuted blocks of size $2N$, with N patients in each block receiving each therapy. After the jth block of patients has been studied, jN patients have been allocated to the ith treatment, with s_i successes and $f_i = jN - s_i$ failures. The trial may be terminated after any complete block, and the null hypothesis accepted (A) or rejected (R) in favor of the alternative hypothesis.

The terminal loss function, $L(\delta, \text{action})$, expresses the hypothesis testing focus of the trial and is shown graphically in Figure 1. It is defined by

$$\begin{aligned} L(\delta, \mathrm{A}) &= 0 \quad \text{if } \delta \le \delta_0 \\ &= K \quad \text{if } \delta > \delta_0 \end{aligned}$$

and

$$\begin{aligned} L(\delta, \mathrm{R}) &= K \quad \text{if } \delta < 0 \\ &= 0 \quad \text{if } \delta \ge 0. \end{aligned} \tag{1}$$

This loss function implies a "zone of indifference" that extends from 0 to δ_0. If the true difference in efficacy lies in this region, then no terminal decision loss is associated with accepting or rejecting the null hypothesis. This is equivalent to the "range of equivalence" used by Freedman and Spiegelhalter in their non-decision–theory approach to the development of Bayesian stopping rules (Freedman, Lowe, and Macaskill 1984; Freedman and Spiegelhalter 1983, 1989).

The unit of cost is the cost of enrolling a patient into the study. This cost is constant throughout the trial. The total cost for a trial terminated after the jth block of patients is the sample size plus the terminal loss, or $2jN + L(\delta, \text{action})$. The constant K in the terminal loss function [Eq. (1)] is the cost of committing a classical error, either type I or type II. A large value of K implies that the accuracy of the final conclusion is important relative to the cost of patient enrollment. Thus clinical trial designs developed with larger values of K will have larger average sample sizes and lower error rates.

In most decision analysis problems, the loss function is chosen to represent as closely as possible the true costs incurred for various actions and true states of nature. Here, because we wish to compare the Bayesian decision-theoretic trial designs directly with classical group-sequential designs, values of K will be used that are found empirically to yield classical error rates similar to those commonly used in classical designs.

The prior pdf's for p_i are $\pi_i = \text{beta}(a_i, b_i)$, $i = 1, 2$, where the beta pdf is defined as usual,

$$\text{beta}(p \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}.$$

Because of the conjugate nature of beta distributions, the p_i continue to have beta distributions and remain independent throughout the trial. The state or pattern of information is given by the number of blocks of patients treated, j, and the

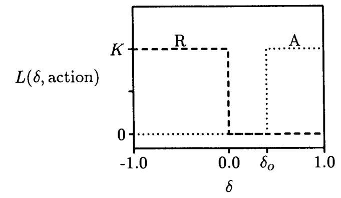

Figure 1. The Terminal Decision Loss Function for the Action to Accept (A) or Reject (R) the Null Hypothesis $\delta < \delta_0$, as a Function of the True Value of δ.

parameters of these beta distributions (x_i, y_i). If the state of information is (j, x_1, y_1, x_2, y_2), then the state after the next block is $(j + 1, x_1 + \Delta s_1, y_1 + \Delta f_1, x_2 + \Delta s_2, y_2 + \Delta f_2)$, where Δs_i and Δf_i are the number of successes and failures observed with the ith treatment in that block.

After any block j, the overall cost of stopping the trial is $W_{\text{stop}}(j, x_1, y_1, x_2, y_2) = 2jN + \min[EL(\delta, \text{A}), EL(\delta, \text{R})]$, where the expectation E is with respect to the current distribution of δ, which is given by $\pi_\delta = \int \pi_2(p_1 + \delta) \pi_1(p_1)\, dp_1$.

We will consider truncated designs, in that the trial will be stopped after at most M blocks of patients. For the designs shown here, the value of M will be large enough so that the truncated design is in fact also the optimum nontruncated design.

Determining a value of M large enough to ensure that we will obtain the optimal nontruncated design during backward induction is straightforward. As the trial progresses, the pdf of δ becomes narrower and narrower until the cost of enrolling an additional block of patients is greater than the cost savings achieved by any possible decrease in $EL(\delta, \text{action})$. When this point is reached, the inequality $W_{\text{stop}}(j + k, \ldots) > W_{\text{stop}}(j, \ldots)$ holds for all possible trial results and positive integral values of k, and thus the optimum action for any result in block j is to terminate the trial. The minimum value of j for which this inequality is strictly true is the value of M.

To initiate the backward induction (Berger 1985; DeGroot 1970), we determine the costs associated with each possible trial outcome after the Mth block of patients. Because a terminal action, either to accept or reject the null hypothesis, will be optimal should sampling continue through block M, the cost at that block is just $W_{\text{stop}}(M, x_1, y_1, x_2, y_2)$ as defined previously. The possible values of x_i, y_i are restricted by $(x_1, y_1, x_2, y_2) = (a_1 + s_1, b_1 + f_1, a_2 + s_2, b_2 + f_2)$, where $s_1 + f_1 = s_2 + f_2 = MN$.

The remainder of the function W is defined and calculated recursively, starting at $j = M$ and proceeding backward to $j = 0$. The fundamental backward induction equation is $W(j, x_1, y_1, x_2, y_2) = \min[W_{\text{stop}}(j, x_1, y_1, x_2, y_2), W_{\text{cont}}(j, x_1, y_1, x_2, y_2)]$ for $j = M - 1, M - 2, \ldots, 0$, where W_{stop} is as defined previously and $W_{\text{cont}}(j, x_1, y_1, x_2, y_2) = EW(j + 1, x_1 + \Delta s_1, y_1 + \Delta f_1, x_2 + \Delta s_2, y_2 + \Delta f_2)$. Here the expectation E refers to the predictive distribution of the numbers of successes, Δs_i, and failures, $\Delta f_i = N - \Delta s_i$, in block $j + 1$. This predictive distribution is given by the product of two beta-binomial distributions (Ferguson 1967),

$$\prod_{i=1}^{2} \binom{N}{\Delta s_i} \left(\frac{\Gamma(x_i + y_i)\Gamma(x_i + \Delta s_i)\Gamma(y_i + \Delta f_i)}{\Gamma(x_i)\Gamma(y_i)\Gamma(x_i + y_i + N)} \right).$$

As the function W is calculated, the action that results in the minimum W for each possible interim and final trial result is recorded. This pattern of actions is the stopping and decision rule for the trial. There may be one or more blocks j, with $j < M$, in which stopping is optimal for all possible trial results. Thus the maximum sample size for an optimal Bayesian design may be less—in some cases, much less—than $2NM$.

The final calculation in this backward induction gives $W(0, a_1, b_1, a_2, b_2)$. In performing the actual calculations, the contributions of the terminal loss function and the number of patients enrolled to the function $W(0, a_1, b_1, a_2, b_2)$ are kept separate. This allows the determination of the expected sample size for the trial design and, when the terminal loss contribution is divided by K, the expected overall error rate.

3. NUMERICAL SIMULATIONS

Characteristics of Bayesian clinical trial designs, defined as earlier, were determined by Monte Carlo simulation. Two different types of simulations were performed, differing in the distributions of the true values of the p_i. In "Bayesian" simulations, the true values of the p_i used for each simulated trial were sampled from a beta pdf and, in each case, the beta pdf used was beta(a_i, b_i). Thus there was exact agreement between the prior pdf assumed and the true distribution of the p_i. The mean sample size under these conditions is denoted by $\bar{n}$, and the rate of type I and type II errors is termed the "Bayesian error rate."

In the second set of simulations, the p_i were given fixed values, allowing the determination of classical error rates. Classical type I error rates were determined with $p_1 = p_2 = .50$ and are denoted by α. The mean sample size under these conditions is $\bar{n}_\alpha$. Classical type II error rates were determined with $p_1 = .50 - \delta_0/2$ and $p_2 = .50 + \delta_0/2$ and are denoted by β. The corresponding mean sample size is $\bar{n}_\beta$.

The mean cost of a Bayesian trial $\bar{c}$ is $\bar{n}$ plus ($K \times$ Bayesian error rate), and, for comparison, the same definition was used for the mean cost of a classical group-sequential design, despite the fact that the constant K has no relevance to the frequentist. The value of K used to determine the mean cost of a classical design was that used to create the corresponding Bayesian design.

One-tailed classical group-sequential clinical trials of Pocock (Pocock 1977) and O'Brien–Fleming (O'Brien and Fleming 1979) were simulated as well. Two interim analyses and one final analysis were used, with the overall $\alpha = .05$. Group sizes ($2N$) were chosen by trial and error to yield classical type II error rates as close as possible to the values of β observed with the corresponding Bayesian designs.

4. RESULTS

Eight Bayesian decision-theoretic trial designs were developed. For purposes of identification, these are labeled BD1–BD8. The parameters of these designs are given in Table 1.

Table 2 shows the theoretical and observed Bayesian error rates and mean sample sizes for the designs defined in Table 1. In general, there is good agreement between the theoretical characteristics of the trials and the results of the Monte Carlo simulations. As demonstrated by the design pairs BD3–BD5 and BD4–BD6, the overall error rate decreases and the mean sample size increases as the value of K is increased.

Two pairs of trial designs, BD5–BD7 and BD6–BD8, illustrate the effect of the block size N on trial characteristics. Trials with larger block sizes, and fewer interim analyses, have larger mean sample sizes and lower error rates.

Table 1. Bayesian Decision-Theoretic Clinical Trial Designs

			Priors			
Design	δ_0	*K*	π_1	π_2	*2N*	*Max n*
BD1	.40	2,000	beta(1,1)	beta(1,1)	32	96
BD2	.40	2,000	beta(2,2)	beta(2,2)	32	96
BD3	.20	6,000	beta(1,1)	beta(1,1)	90	450
BD4	.20	6,000	beta(2,2)	beta(2,2)	90	450
BD5	.20	12,000	beta(1,1)	beta(1,1)	90	540
BD6	.20	12,000	beta(2,2)	beta(2,2)	90	540
BD7	.20	12,000	beta(1,1)	beta(1,1)	32	640
BD8	.20	12,000	beta(2,2)	beta(2,2)	32	640

Each trial designed using constant beta(1,1) priors is followed in Tables 1 and 2 by a design that is identical except that it uses beta(2,2) priors, which are weakly peaked around $p_i = .50$. When beta(2,2) priors are used, suggesting that the p_i are more likely to be close to each other, then the average sample size and the Bayesian error rate both increase.

Table 3 shows the classical error rates and mean sample sizes of the Bayesian trial designs. Bayesian designs BD1, BD3, and BD7 were chosen for direct comparison with classical designs, in part because their classical error rates approximate those used in real clinical trials and in part because of their use of constant priors. The results of these comparisons are shown in Table 4. As expected, the Bayesian designs have substantially lower mean costs $\bar{c}$ than the classical designs. It is more interesting to compare the mean sample sizes for the classical simulations, in which the values of the p_i were fixed. In most cases, the Bayesian designs have a lower mean sample size than the corresponding classical designs, coupled with identical or lower classical error rates.

The sample size comparisons shown in Table 4 can be more easily understood by examining the dependence of the mean sample size on the p_i for the different trial designs. The left panels of Figure 2 show the mean sample size of design BD7 and the corresponding classical designs, as a function of the p_i. For all values of the p_i studied, the Bayesian design had a lower mean sample size than the classical designs. The right panels of Figure 2 show the operating characteristics of the same trial designs, also as a function of the p_i. In general, the operating characteristics are very similar, which is not surprising given that the values of K and the sample sizes of the classical designs were chosen to give similar classical type I and type II error rates.

When treatment 2 has a very large positive effect relative to treatment 1, the mean sample size of the Bayesian design BD7 is much smaller than those of the classical designs. One might suspect, however, that this effect occurs largely because the Bayesian design allows more interim analyses than the classical designs, and thus gives opportunities for very early stopping if treatment 2 appears markedly superior. Although this effect is a true advantage of the Bayesian approach, because there is no "penalty" for frequent interim analyses, it is also interesting to compare Bayesian and classical designs that use the same number of interim analyses.

Figure 3 shows the mean sample size and operating characteristic of Bayesian design BD1 and the corresponding classical designs, as a function of the p_i. In this case all designs use up to two interim analyses and one final data analysis. As shown in the left panels, the Bayesian design usually has a smaller mean sample size than the O'Brien–Fleming design. Furthermore, when $p_1 > p_2$, the Bayesian design has a much lower mean sample size than either classical design. Yet when $p_1 < p_2 - \delta_0$, the Pocock design has the lowest sample size. This situation corresponds to the one entry in Table 4 in which a classical design has a lower mean sample size than the Bayesian design. As shown in the right panels of Figure 3, the operating characteristics of the three trial designs are similar.

5. EXAMPLE

Although this article emphasizes the Bayesian approach for designing clinical trials, the method presented is also applicable to other types of studies. As an example, we consider data from an animal experiment evaluating a medical treatment being considered for use in humans (Niemann, Cairns, Sharma, and Lewis 1992). The study compared two approaches to the resuscitation of dogs who have undergone experimentally induced cardiac arrest and was designed to detect an increase in the proportion of animals successfully resuscitated from 20% to 60% with the experimental therapy. The standard therapy, presently used in humans, is to immediately deliver an electric shock (i.e., defibrillation) to a patient whose heart is not pumping blood because of disorganized electrical activity. Unfortunately, often the patient has had no blood circulation for some time, and the biochemical environment in the cardiac muscle is not conducive to successful defibrillation. This leads to a low survival rate. The experimental therapy, which is to give a large dose of

Table 2. Comparison of Theoretical and Observed Characteristics of Bayesian Decision-Theoretic Clinical Trial Designs

					Theoretical		*Observed*	
Design	δ_0	*K*	*Priors*	*2N*	*Error rate*	$\bar{n}$	*Error rate*	$\bar{n}$
BD1	.40	2,000	beta(1,1)	32	.00257	37.6	.00246	37.6
BD2	.40	2,000	beta(2,2)	32	.00386	38.5	.00391	38.5
BD3	.20	6,000	beta(1,1)	90	.00219	106.3	.00211	106.2
BD4	.20	6,000	beta(2,2)	90	.00333	114.8	.00298	114.4
BD5	.20	12,000	beta(1,1)	90	.00123	114.3	.00118	114.6
BD6	.20	12,000	beta(2,2)	90	.00180	127.8	.00161	127.8
BD7	.20	12,000	beta(1,1)	32	.00144	75.8	.00128	76.9
BD8	.20	12,000	beta(2,2)	32	.00210	94.1	.00203	94.0

Table 3. Monte Carlo Simulations of Bayesian Designs: Classical Characteristics

					Classical			
Design	δ_0	K	Priors	$2N$	α	$\bar{n}_\alpha$	β	$\bar{n}_\beta$
BD1	.40	2,000	beta(1,1)	32	.039	42.1	.054	44.3
BD2	.40	2,000	beta(2,2)	32	.027	38.3	.093	46.2
BD3	.20	6,000	beta(1,1)	90	.051	152.0	.060	156.7
BD4	.20	6,000	beta(2,2)	90	.047	145.7	.064	155.3
BD5	.20	12,000	beta(1,1)	90	.034	177.8	.036	178.8
BD6	.20	12,000	beta(2,2)	90	.029	178.8	.036	188.2
BD7	.20	12,000	beta(1,1)	32	.035	155.9	.040	161.4
BD8	.20	12,000	beta(2,2)	32	.034	152.3	.042	162.1

epinephrine and perform cardiopulmonary resuscitation (CPR) prior to defibrillation, is aimed at producing some cardiac blood flow, with the hope that the heart may then be more responsive to electric shock therapy.

The original study of Niemann et al. (1992) used a classical fully sequential design that allowed trial termination after 28 animals. A completely randomized design was used, creating imbalances in the two arms during the study. Because our method is for blocked data, we have rearranged the data within treatment groups to be consistent with blocks of size four, two animals on each of the two treatments. The order of the animals within each treatment group is unchanged. The data are shown in Table 5.

We used a Bayesian design with $K = 750$, π_i = beta(2,2), $N = 2$, and $\delta_0 = .40$. The corresponding classical type I error rate is .05 and the power is .826, found using Monte Carlo simulation. Using the optimal Bayesian decision rule, the trial terminates after block seven, with rejection of the null hypothesis. The probability that $\delta > 0$, based on the final pdf for δ, is .982. From a classical viewpoint, the results are statistically significant (P value < .05), had the original study used a Bayesian design.

Suppose that the study had been designed using a three-look Pocock design with $\alpha = .05$ and a power of .80. The required block size would be approximately 18 (9 in each treatment group), for a maximum trial size of 54 (Geller and Pocock 1987). Thus the first interim analysis would occur between blocks 4 and 5, when two of nine treatment 1 animals and six of nine treatment 2 animals had survived. The resulting one-tailed P value is .077 using Fisher's exact test. The nominal α for the Pocock design at each interim analysis is .022, and the actual result would not be statistically significant. The second interim analysis would occur after 36 animals had been studied. This is more than were used in the actual trial.

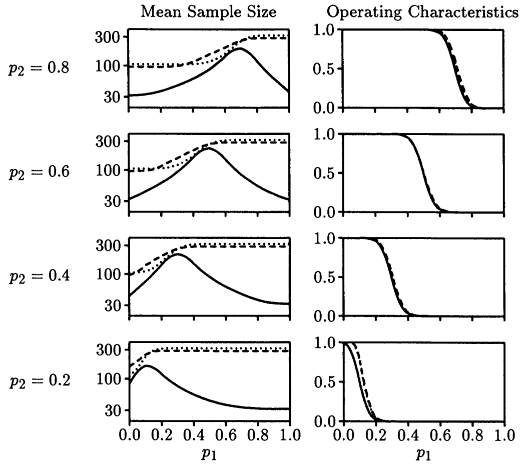

Figure 2. Mean Sample Sizes (Left Panels) and Operating Characteristics (Right Panels) of the Bayesian Clinical Trial Design BD7 and Two Classical Clinical Trial Designs, as Shown on the Bottom Three Lines of Table 4. (——— Bayesian BD7; --- O'Brien–Fleming; · · · Pocock.)

If a three-look O'Brien–Fleming design had been used with $\alpha = .05$ and a power of .80, then the first interim analysis would have occurred after 16 animals, again with a nonsignificant P value (.066). The second interim analysis would occur after 32 animals. So this is a case in which the Bayesian design indeed has a smaller sample size than either the Pocock or the O'Brien–Fleming designs.

As shown in this example, a trial terminated using a Bayesian decision rule can be analyzed from either a Bayesian or classical viewpoint. In either case, the Bayesian design tends to allow for earlier trial termination than either the Pocock or the O'Brien–Fleming group-sequential methods.

6. DISCUSSION

The Bayesian decision-theoretic approach to the design of a group-sequential clinical trial is potentially superior to the

Table 4. Comparison of Bayesian and Classical Group-Sequential Designs

				Bayesian			Classical			
Design type	δ_0	$2N$	Max n	Error rate	$\bar{n}$	$\bar{c}$	α	$\bar{n}_\alpha$	β	$\bar{n}_\beta$
Pocock	.40	24	72	.00250	60.9	65.9	.055	70.3	.054	39.7
OBF	.40	20	60	.00196	54.4	58.3	.055	59.6	.061	45.5
BD1	.40	32	96	.00246	37.6	42.5	.038	42.1	.054	44.3
Pocock	.20	90	270	.00218	204.7	217.7	.051	264.4	.081	158.5
OBF	.20	80	240	.00135	192.2	200.3	.043	238.7	.082	181.8
BD3	.20	90	450	.00211	106.2	118.9	.051	152.0	.060	156.7
Pocock	.20	106	318	.00131	237.7	253.4	.045	312.3	.042	174.8
OBF	.20	96	288	.00072	227.1	235.7	.043	286.7	.042	211.8
BD7	.20	32	640	.00128	76.9	92.3	.035	155.9	.040	161.4

NOTE: OBF = O'Brien–Fleming.

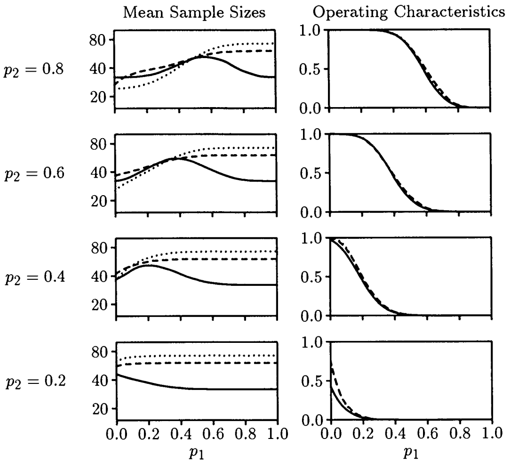

Figure 3. Mean Sample Sizes (Left Panels) and Operating Characteristics (Right Panels) of the Bayesian Design BD1 and Two Classical Clinical Trial Designs, as Shown on the Top Three Lines of Table 4. (——— Bayesian BD1; --- O'Brien–Fleming; · · · Pocock.)

classical approach for several reasons. First of all, from a Bayesian viewpoint there is no "penalty" in error risk or sample size for frequent interim analyses. The effect of incorporating many interim analyses into the design is illustrated by the comparison of the first panels of Figures 2 and 3. The sample size savings of the Bayesian design over the classical designs is explained in part by the additional interim analyses allowed with the Bayesian approach. Second, even when a "one-tailed" hypothesis testing approach is taken, a Bayesian trial may be terminated early when the test treatment appears ineffective, whereas in a classical one-tailed group sequential trial, sampling must continue through the final block before the null hypothesis can be accepted.

This second point, regarding the inability of the one-tailed classical design to accept the null hypothesis during an early interim analysis, is shown diagramatically in Figure 4. The acceptance (A), rejection (R), and continuation (C) regions for an early interim analysis in a Bayesian design and a corresponding classical design are shown. For the classical design, no degree of relative ineffectiveness on the part of treatment 2 will lead to early termination.

It is interesting to consider the degree to which a classical stochastic curtailment approach might overcome this latter limitation in the one-tailed classical design. Stochastic curtailment allows the trial to terminate when the probability of rejecting the null hypothesis, under the alternative hypothesis and given the results already observed, is sufficiently low (Halperin, Lan, Ware, Johnson, and DeMets 1982; Lan, Simon, and Halperin 1982). This allows termination on the basis of futility, analogous to the Bayesian termination that occurs in the acceptance region of Figure 4.

The assumption of independence of the p_i used in the present designs leads to a simplification in both the determination of the optimal Bayes stopping rule and in the Monte Carlo simulations used to verify the Bayesian characteristics of these designs. In many clinical situations, however, this assumption of independence may not be reasonable. To estimate the effect of dependence of the p_i on the Bayesian characteristics of these trial designs, which themselves are based on the assumption of independence, Monte Carlo simulations could be performed, with the p_i drawn from an appropriate bivariate distribution. For example, p_1 could be sampled from a beta distribution and p_2 determined from p_1 and the log odds ratio, θ, where $\theta = \log[p_2(1 - p_1)/p_1(1 - p_2)]$ and θ is itself sampled from some distribution (Carlin 1992).

Table 5. Data From Niemann et al. (1992)

Block (j)	jN	S_1	S_2
1	2	1	2
2	4	1	4
3	6	2	5
4	8	2	6
5	10	3	6
6	12	3	8
7	14	3	9

Although the effect of dependence of the p_i on the Bayesian characteristics of these trial designs may be of theoretical interest, the frequentist characteristics presented here were determined by Monte Carlo simulations using fixed values for the p_i. The independence assumption does not alter the interpretation of these observed frequentist characteristics.

Explicitly including dependence of the p_i in the trial design itself, with a resulting change in the optimal Bayes stopping rule, is more difficult. One approach would be to use the (p_1, θ) parameterization for the treatment efficacies mentioned previously, rewrite the loss function in terms of p_1 and θ, and determine the necessary expected losses using Gibbs sampling. Although the framework is more complex than that used in this study, the use of Gibbs sampling should result in a tractable numerical problem (Smith and Roberts 1993).

Because of frequentist concerns regarding repetitive testing, one might expect the Bayesian designs that use frequent interim analyses to exhibit inflated type I error rates. Although each interim analysis carries with it some risk of a

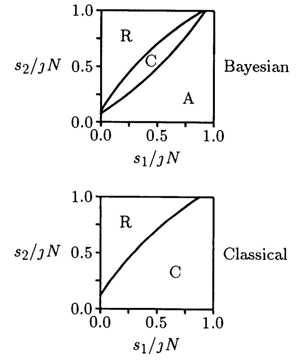

Figure 4. Qualitative Diagrams of the Shapes of the Continuation and Termination Regions of the Bayesian and Classical Trial Designs, for an Early Block j. Although the Bayesian design is "one-tailed" in that it cannot demonstrate that $p_1 > p_2$, trial termination may occur when $s_1/jN \gg s_2/jN$ because the expected cost of continuing the trial is greater than the expected cost of termination. No such acceptance region A exists for a one-tailed classical design until the final group of patients has been enrolled.

type I error, the value of K can be adjusted upward to reduce the overall classical error rate of the Bayesian designs when frequent interim analyses are used. When trial designs with similar classical error rates are compared using strictly frequentist criteria, the Bayesian designs perform as well or better than the classical designs.

The Bayesian designs presented here allow interpretation of the final results along either Bayesian or frequentist lines. To the Bayesian, the designs have the advantages of minimizing the expectation of the total cost and allowing the direct calculation of the pdf for the difference in efficacy, δ. To the frequentist, the designs have well-characterized classical type I and type II error rates and in most cases lead to a reduction in the mean sample size relative to commonly used classical group-sequential designs.

Heitjan and colleagues (Heitjan, Houts, and Harvey 1992) examined some classical two-look group-sequential designs from a decision-theoretic point of view. This is a complementary approach to that taken here, in which we have examined decision-theoretic designs from a classical viewpoint.

7. CONCLUSIONS

Bayesian decision-theoretic clinical trial designs have been developed and compared, using strictly frequentist criteria, with the classical group sequential designs of Pocock and O'Brien–Fleming. The Bayesian designs usually result in a smaller mean sample size, especially when frequent interim analyses are used. When the treatment effect is positive and very large, or if the treatment effect is negative, the sample size savings can be quite substantial. Even when compared on purely frequentist criteria, the Bayesian designs perform similarly or better than commonly used classical group-sequential clinical trial designs.

[Received September 1992. Revised August 1993.]

REFERENCES

Anscombe, F. J. (1963), "Sequential Medical Trials," *Journal of the American Statistical Association,* 58, 365–383.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis,* New York: Springer-Verlag.

Berger, J. O., and Berry, D. A. (1988), "The Relevance of Stopping Rules in Statistical Inference" (with discussion), in *Statistical Decision Theory and Related Topics IV* (Vol. 1), eds. J. O. Berger and S. Gupta, New York: Springer-Verlag, pp. 29–72.

Berry, D. A. (1985), "Interim Analysis in Clinical Trials: Classical vs. Bayesian Approaches," *Statistics in Medicine,* 4, 521–526.

——— (1987), "Interim Analysis in Clinical Trials: The Role of the Likelihood Principle," *The American Statistician,* 41, 117–122.

——— (1989), "Monitoring Accumulating Data in a Clinical Trial," *Biometrics,* 45, 1197–1211.

Berry, D. A., and Ho, C. H. (1988), "One-Sided Sequential Stopping Boundaries for Clinical Trials: A Decision-Theoretic Approach," *Biometrics,* 44, 219–227.

Berry, D. A., Wolff, M. C., and Sack, D. (1992), "Public Health Decision Making: A Sequential Vaccine Trial" (with discussion), in *Bayesian Statistics 4,* eds. J. M. Bernardo, J. O. Berger, A. P. Dalwid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 79–92.

Carlin, J. B. (1992), "Meta-Analysis for 2 × 2 Tables: A Bayesian Approach," *Statistics in Medicine,* 11, 141–158.

DeGroot, M. H. (1970), *Optimal Statistical Decisions,* New York: McGraw-Hill.

Ferguson, T. S. (1967), *Mathematical Statistics: A Decision-Theoretic Approach,* New York: Academic Press.

Freedman, L. S., Lowe, D., and Macaskill, P. (1984), "Stopping Rules for Clinical Trials Incorporating Clinical Opinion," *Biometrics,* 40, 575–586.

Freedman, L. S., and Spiegelhalter, D. J. (1983), "The Assessment of Subjective Opinion and Its Use in Relation to Stopping Rules for Clinical Trials," *The Statistician,* 32, 153–160.

——— (1989), "Comparison of Bayesian With Group-Sequential Methods for Monitoring Clinical Trials," *Controlled Clinical Trials,* 10, 357–367.

Geller, N. L., and Pocock, S. J. (1987), "Interim Analyses in Randomized Clinical Trials: Ramifications and Guidelines for Practitioners," *Biometrics,* 43, 213–223.

Halperin, M., Lan, K. K. G., Ware, J. H., Johnson, N. J., and DeMets, D. L. (1982), "An Aid to Data Monitoring in Long-Term Clinical Trials," *Controlled Clinical Trials,* 3, 311–323.

Heitjan, D. F., Houts, P. S., and Harvey, H. A. (1992), "A Decision-Theoretic Evaluation of Early Stopping Rules," *Statistics in Medicine,* 11, 673–683.

Kim, K., and DeMets, D. L. (1987), "Design and Analysis of Group Sequential Tests Based on the Type I Error Spending Rate Function," *Biometrika,* 74, 149–154.

Lan, K. K. G., and DeMets, D. L. (1983), "Discrete Sequential Boundaries for Clinical Trials," *Biometrika,* 70, 659–663.

Lan, K. K. G., Simon, R., and Halperin, M. (1982), "Stochastically Curtailed Tests in Long-Term Clinical Trials," *Communications in Statistics, Sequential Analysis,* 1, 207–219.

Lewis, R. J. (1990), "Sequential Bayesian Analysis of Clinical Trials," *Annals of Emergency Medicine,* 19, 493.

Lewis, R. J., and Berry, D. A. (1992), "A Comparison of Bayesian and Classical Group-Sequential Clinical Trial Designs," *Annals of Emergency Medicine,* 21, 641.

Lewis, R. J., and Wears, R. L. (1993), "An Introduction to the Bayesian Analysis of Clinical Trials," *Annals of Emergency Medicine,* 22, 1328–1336.

McPherson, K. (1982), "On Choosing the Number of Interim Analyses in Clinical Trials," *Statistics in Medicine,* 1, 25–36.

Niemann, J. T., Cairns, C. B., Sharma, J., and Lewis, R. J. (1992), "Treatment of Prolonged Ventricular Fibrillation: Immediate Countershock Versus High-Dose Epinephrine and CPR Preceding Countershock," *Circulation,* 85, 281–287.

O'Brien, P. C., and Fleming, T. R. (1979), "A Multiple Testing Procedure for Clinical Trials," *Biometrics,* 35, 549–556.

Pocock, S. J. (1977), "Group-Sequential Methods in the Design and Analysis of Clinical Trials," *Biometrika,* 64, 191–199.

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society,* Ser. B, 55, 3–23.