STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:5047–5080
Published online 15 June 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2901

TUTORIAL IN BIOSTATISTICS

Frequentist 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, WA 98195, U.S.A.*
[2]*Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, Denver, CO 80262, U.S.A.*
[3]*Department of Statistics, University of California, Irvine, CA 92697, U.S.A.*

SUMMARY

Group sequential stopping rules are often used as guidelines in the monitoring of clinical trials in order to address the ethical and efficiency issues inherent in human testing of a new treatment or preventive agent for disease. Such stopping rules have been proposed based on a variety of different criteria, both scientific (e.g. estimates of treatment effect) and statistical (e.g. frequentist type I error, Bayesian posterior probabilities, stochastic curtailment). It is easily shown, however, that a stopping rule based on one of these criteria induces a stopping rule on all other criteria. Thus, the basis used to initially define a stopping rule is relatively unimportant so long as the operating characteristics of the stopping rule are fully investigated. In this paper we describe how the frequentist operating characteristics of a particular stopping rule might be evaluated to ensure that the selected clinical trial design satisfies the constraints imposed by the many different disciplines represented by the clinical trial collaborators.

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

1. INTRODUCTION

Clinical trials represent experimentation in human volunteers. The objective of clinical trial design is therefore to find a procedure that ensures scientific credibility while at the same time protecting the safety of human subjects. The ethical constraints associated with this latter concern will include both the individual ethics related to the subjects enrolled on the clinical trial, as well as the group ethics related to the population of people who might benefit from the rapid adoption of a new treatment or preventive strategy. A major tool used to address the scientific, ethical, and efficiency

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

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

Received 8 November 2004
Accepted 13 March 2007

issues raised by the conduct of clinical trials is a properly selected group sequential stopping rule (see, for instance, the texts by Jennison and Turnbull [1] or Whitehead [2]).

During the conduct of a clinical trial, it is now common for the accruing data to be monitored repeatedly in order that patients on the study not be unnecessarily given a treatment known (or credibly demonstrated) to be inferior, and that new, beneficial treatments be adopted as rapidly as possible. The goal is often to allow the early termination of the clinical trial as soon as there is high confidence in a decision about whether to alter standard clinical practice. However, if such repeated analysis of accruing data is allowed to alter the sampling scheme for the study, many of the commonly used statistical analyses of trial results can be greatly affected. As the scientific credibility of the study is most often dependent upon the statistical precision of the estimates of treatment effect, it is of paramount importance that any stopping rule be taken into account during the planning of the study.

The impact that a stopping rule can have on statistical operating characteristics can be illustrated in the context of a randomized, double-blind, placebo-controlled clinical trial of an antibody to endotoxin in the treatment of Gram-negative sepsis. Sepsis is a disease state in which an overwhelming bacterial infection has led to major organ dysfunction. Gram-negative bacteria are a class of bacteria that were initially identified based on their failure to exhibit a particular stain on microscopic examination. Many such bacteria produce a particular poison called endotoxin, and endotoxin has been shown to produce many of the characteristic signs and symptoms of sepsis. In the early 1990s, a number of clinical trials investigated whether treating patients with antibodies to endotoxin might improve the survival of these severely ill patients. In one of these trials, patients with proven Gram-negative sepsis were randomly assigned to receive a single dose of antibody to endotoxin or placebo. The primary endpoint for the trial was to be the 28-day mortality probability, 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.

At the recommendation of the U.S. Food and Drug Administration (FDA), the actual clinical trial was conducted using a group sequential stopping rule. In this manuscript, we describe the general process used by the study sponsors and the independent Data Safety Monitoring Board (DSMB) in selecting a stopping rule for the study. As is typically the case, this process was highly iterative and included a broad spectrum of collaborators. Selection of the clinical trial design involved scientists (basic scientists, clinical researchers, epidemiologists, and biostatisticians), clinicians (bringing expertise in the treatment of the disease and potential adverse effects of the treatments(s) as well as their perspective regarding clinically meaningful treatment benefit), trial sponsors (including financial and marketing perspectives), ethicists, patient advocates, and regulatory agencies. Each of these categories of collaborators had a slightly different set of criteria that were of greatest interest when selecting a clinical trial design, and thus the evaluation of candidate stopping rules involved consideration of a variety of operating characteristics.

In using the sepsis trial as an example, we focus on the evaluation of frequentist operating characteristics for group sequential designs in settings where the treatment effect is measured as the difference in binomial proportions. These same methods have immediate applicability across a wide range of commonly used statistical models including the comparison of means, geometric means, medians, binomial odds, or constant hazard ratios. The key features of such models are that the measure of treatment effect is constant over the course of the study, and that the test of treatment efficacy is based on a statistic estimating treatment effect. We note that some of the methods used for the estimation of operating characteristics under time-invariant measures of treatment effects do not easily generalize to situations in which treatment effects vary over time

DOI: 10.1002/sim

(e.g. the use of the hazard ratio in nonproportional hazards survival data), but we address these issues in a separate manuscript [3]. We also defer to another manuscript a discussion of evaluating Bayesian operating characteristics [4].

In Section 2, we provide notation for the statistical data analysis model used in the sepsis clinical trial that serves as an example for the remainder of this paper. We then provide notation for a stopping rule in Section 3 and review the computation of the sampling density for common test statistics. Section 4 follows with a discussion of the role of frequentist considerations in evaluating a clinical trial design and considers, in turn, each of the several frequentist operating characteristics that might be examined while selecting a stopping rule for a particular trial. We conclude in Section 5 with a discussion of the impact that a careful evaluation of stopping rules has on the need to consider alternative strategies for defining data monitoring plans.

2. NOTATION AND CLASSICAL FIXED SAMPLE TRIAL DESIGN FOR THE SEPSIS CLINICAL TRIAL

The sepsis clinical trial introduced in the previous section was designed to compare 28-day mortality probabilities between groups of patients who received antibody to endotoxin and groups of patients who received placebo. The trial design was based on a data analysis model which used a large sample approximation to the sampling distribution for the 28-day mortality probability on each arm. Notationally, we let X_{ki} be an indicator that the ith patient on the kth treatment arm ($k=0$ for placebo, $k=1$ for antibody) died in the first 28 days following randomization. Thus, $X_{ki}=1$ if the ith 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 probability on the kth treatment arm. Supposing the accrual of N subjects on each treatment arm, asymptotic arguments suggest that $\hat{p}_k=\sum_{i=1}^{N} X_{ki}/N$ is approximately normally distributed with mean p_k and variance $p_k(1-p_k)/N$. We use the difference in 28-day mortality probabilities $\theta=p_1-p_0$ as the measure of treatment effect. In a study which accrues N subjects per arm we therefore have an approximate distribution for the estimated treatment effect $\hat{\theta}=\hat{p}_1-\hat{p}_0$ of

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

When using probability models in which the statistical information grows in direct proportion to sample size, standard formulas for sample size calculation describe the interrelationship between sample size, statistical size and power, and an alternative hypothesis according to

$$n=\frac{\delta_{\alpha\beta}^2 V}{\Delta^2} \tag{2}$$

where n is the sample size on each treatment arm which provides statistical power β to detect a treatment effect Δ using a level α hypothesis test. In this formula, V is the variance contributed by a single sampling unit (e.g. a patient accrued to each of the treatment arms), and $\delta_{\alpha\beta}$ is the alternative which is detected with statistical power β using a standardized level α trial design (e.g. a design appropriate for a study having only one sampling unit accrued).

DOI: 10.1002/sim

In the setting of the sepsis trial, Δ would represent the difference $\theta = p_1 - p_0$ in 28-day mortality probabilities, and $V = p_1(1-p_1) + p_0(1-p_0)$ is the contribution to the variance of $\hat{\theta}$ from a single sampling unit consisting of a patient accrued to each treatment arm. In a fixed sample study using an asymptotically normally distributed test statistic, the standardized alternative for which a one-sided level α test is detected with statistical power β is $\delta_{\alpha\beta} = z_{1-\alpha} + z_\beta$, where $z_p = \Phi^{-1}(p)$ is the pth quantile of a standard normal distribution having cumulative distribution function $\Phi(z)$. Using this formula and assuming the variability of the estimate under the design alternative hypothesis of $p_0 = 0.30$ and $p_1 = 0.23$, we calculate that accruing 1700 patients ($N = 850$ per arm) yields statistical power of 0.907 in a one-sided level 0.025 hypothesis test of the null hypothesis $H_0 : p_0 = p_1$.

As is customary in the setting of tests of binomial proportions, at the time of data analysis the actual test statistic will not use the hypothesized 28-day mortality probabilities and will instead estimate a common mortality probability $\hat{p}$ 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 probabilities 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 probability under the null hypothesis is estimated by

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

Because of the need to estimate the variability of the estimate of treatment effect, the critical value for the estimate of treatment effect which corresponds to rejection of the null hypothesis will depend on the observed mortality probabilities. 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 (i.e. if $\widehat{\mathrm{Var}}(\hat{\theta}) = 0.3871/N$), the null hypothesis would be rejected if the absolute difference in 28-day mortality probabilities 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 $\hat{\theta} \leqslant -0.0418$).

3. DEFINITION OF STOPPING RULES

3.1. General form of a stopping rule

In the general case, a stopping rule is defined for a schedule of analyses occurring at times $t_1, t_2, \ldots, t_J$, which may be random. Often, the analysis times are in turn defined according to the statistical information available at each analysis. Because many statistical models have statistical information proportional to the sample size accrued to the study, such an approach is equivalent to defining the sample sizes $N_1, N_2, \ldots, N_J$ at which the analyses will be performed. For $j = 1, \ldots, J$, we calculate a specified test statistic T_j based on observations available at time t_j. The outcome space for T_j is then partitioned into stopping set $\mathcal{S}_j$ and continuation set C_j. Starting with $j = 1$, the clinical trial proceeds by computing T_j, and if $T_j \in \mathcal{S}_j$, the trial is stopped. Otherwise, T_j is in the continuation set C_j, and the trial gathers additional observations until time t_{j+1}. By choosing $C_J = \emptyset$, the empty set, the trial must stop at or before the Jth analysis.

DOI: 10.1002/sim

As noted by Kittelson and Emerson [5], all of the most commonly used group sequential stopping rules are included if we consider continuation sets of the form $\mathscr{C}_j = (a_j, b_j] \cup [c_j, d_j)$ such that $-\infty \leqslant a_j \leqslant b_j \leqslant c_j \leqslant d_j \leqslant \infty$. Quite often, these boundaries are interpreted as the critical values for a decision rule. For instance, in a clinical trial comparing two active treatments A and B, test statistics less than a_j might correspond to decisions for the superiority of treatment A, test statistics exceeding d_j might correspond to decisions for the inferiority of treatment A, and test statistics between b_j and c_j might correspond to decisions for approximate equivalence between the two treatments. The appropriateness of such interpretation will of course depend upon the boundaries and the statistical precision afforded by the trial design. However, it should be noted that it is possible to dissociate the stopping boundaries from any particular decision, and this has in fact been used when dealing with frequentist inference in the presence of imprecisely defined stopping rules [6, 7]. That is, the stopping rule first and foremost defines a sampling distribution. The efficiency of that sampling distribution to answer any particular question will depend upon how reasonable (scientifically, ethically, statistically) those boundaries are when interpreted as a statistical decision rule.

Particular families of group sequential designs correspond to parameterized boundary functions which relate the stopping boundaries at successive analyses according to the proportion of statistical information accrued and the hypothesis rejected by the boundary. For instance, letting Π_j represent the proportion of the maximal statistical information available at the jth analysis (e.g. $\Pi_j = N_j/N_J$ for the most commonly used analytic models), then for some specified parametric function $f_d()$, the boundary function for the upper boundary might be given by $d_j = f_d(\theta_d, \Pi_j)$, where θ_d is the hypothesis rejected when $T_j > d_j$. Furthermore, many of the group sequential design families previously described can be expressed in a parameterization which has $d_j = f(\theta_d, g(\Pi_j; A_d, P_d, R_d, G_d))$ with boundary shape function

$$g(\Pi; A, P, R, G) = (A + \Pi^{-P}(1 - \Pi)^R)G$$

where parameters A, P, and R are typically specified by the user to attain some desired level of conservative behavior at the earliest analyses, and critical value G might be found in an iterative search to attain some specified operating characteristics (e.g. frequentist type I error and power) when the stopping rule is to be used as the basis of a decision rule [5]. The way in which the boundary shape function is combined with the boundary hypothesis will depend upon the exact form of the test statistic. For instance, in the unified family [5], the boundaries are expressed on the scale of the maximum likelihood estimate (MLE), and the boundary hypothesis is merely a shift of the boundary shape function, so

$$d_j = \theta_d + g(\Pi_j; A_d, P_d, R_d, G_d)$$
$$a_j = \theta_a + g(\Pi_j; A_a, P_a, R_a, G_a)$$

In other families, the formulas for the stopping boundaries may involve the boundary hypotheses in a complicated, nonlinear fashion.

As discussed in the next section, however, stopping boundaries defined for one test statistic induce stopping boundaries for all other statistics commonly used in specifying stopping rules. Thus, it is largely immaterial how the stopping rule is initially defined, so long as the operating characteristics of the stopping rule are adequately evaluated.

DOI: 10.1002/sim

3.2. Choice of test statistic

A number of equivalent test statistics are commonly used in the definition of a stopping rule. In the context of the sepsis trial introduced in Section 1, suppose that at the jth analysis we had accrued $N_{0j} = N_{1j} = N_j$ subjects to the placebo and antibody arms, respectively, and that the random variables measuring the corresponding observed number of patients dying within 28 days were $Y_{0j} = \sum_{i=1}^{N_{0j}} X_{0i}$ and $Y_{1j} = \sum_{i=1}^{N_{1j}} X_{1i}$. For the instance in which we observe $Y_{0j} = y_{0j}$ and $Y_{1j} = y_{1j}$, we might consider any of the following test statistics as the basis for the definition of the stopping rule. As previously discussed in part by a number of authors [8–10], these scales can be viewed as monotonic transformations of each other.

1. *Partial sum statistic*: $S_j = s_j = y_{1j} - y_{0j}$, which represents the difference in the number of deaths between the two arms. The partial sum statistic was used for the definition of stopping rules by Whitehead and Stratton [11], Emerson and Fleming [12], and Pampallona and Tsiatis [13]. An O'Brien–Fleming [14] boundary rejecting a null hypothesis of no treatment effect is constant on the scale of this statistic.
2. *Crude estimate of treatment effect*: $\hat{\theta}_j = s_j/N_j = \hat{p}_{1j} - \hat{p}_{0j} = y_{1j}/N_{1j} - y_{0j}/N_{0j}$. The crude estimate of treatment effect was used for the definition of stopping rules by Kittelson and Emerson [5].
3. *Normalized Z statistic*: $Z_j = z_j = (\hat{\theta}_j - \theta_0)/\mathrm{se}(\hat{\theta}_j)$ where $\mathrm{se}(\hat{\theta}_j)$ is typically estimated by $\widehat{\mathrm{se}}(\hat{\theta}_j) = \sqrt{\hat{p}(1-\hat{p})(1/N_{0j} + 1/N_{1j})}$ as described in Section 2. The normalized Z statistic was used for the definition of stopping rules by Wang and Tsiatis [15]. A Pocock [16] boundary rejecting a null hypothesis $H_0 : \theta = \theta_0$ of no treatment effect is constant on the scale of this statistic.
4. *Fixed sample P-value statistic*: $P_j = \Phi(z_j)$, which would represent the lower one-sided P-value if the observed data had been gathered in a fixed sample study. In clinical trial designs which allow for early stopping, however, this scale does not represent a true P-value and is therefore not easily interpreted. Nevertheless, based on the findings of Pocock [16], this statistic is of some use when implementing a group sequential stopping rule derived using asymptotic theory. In that research it was found that the statistical properties of such stopping rules were relatively robust when used with fixed sample P values computed for statistics having other distributions (e.g. the t distribution).
5. *Error spending statistic*: An error spending statistic can be defined for any of the four boundaries based on an arbitrary hypothesized value for the true treatment effect. For instance, if a group sequential stopping rule were defined for the partial sum statistic and the observed value of the test statistic at the jth analysis were $S_j = s_j$, a lower type I error spending statistic defined for the null hypothesis $H_0 : \theta = \theta_0$ would have

$$E_{aj} = \frac{1}{\alpha_L}\left(\Pr\left[S_j \leqslant s_j, \bigcap_{k=1}^{j-1} S_k \in C_k | \theta = \theta_0\right] + \sum_{\ell=1}^{j-1} \Pr\left[S_\ell \leqslant a_\ell, \bigcap_{k=1}^{\ell-1} S_k \in C_k | \theta = \theta_0\right]\right)$$

where α_L is the lower type I error of the stopping rule defined by

$$\alpha_L = \sum_{\ell=1}^{J} \Pr\left[S_\ell \leqslant a_\ell, \bigcap_{k=1}^{\ell-1} S_k \in C_k | \theta = \theta_0\right]$$

DOI: 10.1002/sim

Similar transformations can be defined for the type II error and, in a two-sided test, for both upper and lower type I and II errors [17]. The error spending scale is used for the computation of stopping boundaries based on the methods of Lan and DeMets [18], Pampallona *et al.* [19], and Chang *et al.* [20].

6. *Bayesian posterior probabilities*: Bayesian posterior probabilities are greatly preferred by those statisticians who regard the formal incorporation of prior information as the most appropriate means to quantify uncertainty in estimates. The Bayesian posterior probability of some hypothesis, say $\theta<\theta_1$, can be computed using a prior distribution for θ and conditioning on the observed statistic $S_j = s_j$. We consider a robust approach to such Bayesian inference based on a coarsening of the data by using the asymptotic distribution of a nonparametric estimate of treatment effect [21]. That is, rather than the exact binomial distributions for the two arms of the sepsis trial, we use the approximate normal distribution given in equation (1) for the difference in 28-day mortality probabilities. In the case of a computationally convenient conjugate normal prior $\theta \sim \mathrm{N}(\zeta, \tau^2)$, we can define a Bayesian posterior probability statistic by computing the approximate posterior probability that the null hypothesis $H_0 : \theta \geqslant \theta_0$ is false

$$B_j(\zeta, \tau^2, \theta_0) = \Pr(\theta \leqslant \theta_0 | S_j = s_j) = \Phi\left(\frac{\theta_0[N_j\tau^2 + V] - \tau^2 s_j - V\zeta}{\sqrt{V}\tau\sqrt{N_j\tau^2 + V}}\right)$$

where $V = p_0(1 - p_0) + p_1(1 - p_1)$ and $\Phi(z)$ is the cumulative distribution function for a standard normal random variable. A special case that is of occasional interest is the noninformative (improper) prior corresponding to the limit as $\tau^2 \to \infty$.

7. *Conditional power statistics*: The futility of continuing a study is often measured by computing a statistic which represents the conditional probability that the test statistic at the final (Jth) analysis would exceed the threshold for declaring statistical significance, where we condition on the observed statistic $S_j = s_j$ at the jth analysis and assume some particular value for the true treatment effect θ. For instance, when considering whether to stop a clinical trial due to the futility of obtaining results which would change clinical practice, we might define a conditional power statistic using a threshold a_J defined for the partial sum statistic. Such a threshold would represent the critical value for declaring statistical significance at the Jth analysis. We might compute a conditional power statistic using an alternative hypothesis $H_1 : \theta = \theta_1$ as

$$C_j(a_J, \theta_1) = \Pr(S_J < a_J \mid S_j = s_j; \theta = \theta_1) = \Phi\left(\frac{a_J - s_j - \theta_1 + [N_J - N_j]}{\sqrt{V(N_J - N_j)}}\right)$$

Alternatively, a conditional power statistic might use the current best estimate of the treatment effect $\hat{\theta}_j$ in place of θ_1. Futility measures such as these have been proposed for use when stopping a clinical trial early is to be based on stochastic curtailment [22, 23].

DOI: 10.1002/sim

8. *Predictive probability statistics*: A Bayesian approach similar to the stochastic curtailment procedures would consider the Bayesian predictive probability that the test statistic would exceed some specified threshold at the final analysis. This statistic uses a prior distribution and the observed data to compute a posterior distribution for the treatment effect parameter at the jth analysis. Then, using the sampling distribution for the as yet unobserved data and integrating over the posterior distribution, the predictive distribution of the test statistic at the final analysis can be computed. Using the coarsened Bayesian approach described above, we might compute a predictive probability statistic analogous to the conditional power statistic as

$$H_j(a_J, \zeta, \tau^2) = \int \Pr(S_J < a_J | S_j = s_j, \theta)\, p(\theta | S_j = s_j)\, \mathrm{d}\theta$$
$$= \Phi\left(\frac{N_J[N_j\tau^2 + V][a_J - s_j/N_j] + V[N_J - N_j][s_j/N_j - \zeta]}{\sqrt{V[N_J - N_j][N_J\tau^2 + V][N_j\tau^2 + V]}} \right)$$

Again, the case of a noninformative prior is of special interest. When we consider taking the limit as $\tau^2 \to \infty$, using Xiong's sequential conditional probability ratio test [24] to test the null hypothesis $H_0 : \theta = \theta_0$ results in constant boundaries on the predictive probability statistic scale [10].

The main point to be made in considering all of these different statistics is that in every case the statistic is a straightforward (if not always simple) function of the partial sum statistic $S_j = s_j$. Hence, definition of a stopping boundary on one scale automatically induces a stopping boundary on each of the other scales. Computation of the operating characteristics of a particular design can then be effected using the approach of Armitage *et al.* [25] to compute the sampling density for the partial sum statistic.

3.3. Computation of sampling density

Given a stopping rule with continuation sets C_j for $j = 1, \ldots, J$ defined on the partial sum statistic scale, we define group sequential test statistics (M, S) as $M = \min\{1 \leqslant j \leqslant J : S_j \notin C_j\}$ and $S = S_M$. Using the normal approximation for the distribution of independent increments of information $S_j - S_{j-1}$ and for convenience introducing the notation $n_j = N_j - N_{j-1}$ with $N_0 = 0$, the sampling density for observation $(M = m, S = s)$ is then recursively defined as

$$p(m, s; \theta) = \begin{cases} f(m, s; \theta) & s \notin \mathscr{C}_m \text{ and} \\ 0 & \text{else} \end{cases}$$

where the (sub)density function $f(j, s; \theta)$ is recursively defined as

$$f(1, s; \theta) = \frac{1}{\sqrt{n_1 V}} \phi\left(\frac{s - n_1\theta}{\sqrt{n_1 V}} \right)$$
$$f(j, s; \theta) = \int_{\mathscr{C}_{j-1}} \frac{1}{\sqrt{n_k V}} \phi\left(\frac{s - u - n_k\theta}{\sqrt{n_k V}} \right) f(k-1, u; \theta)\, \mathrm{d}u, \quad j = 2, \ldots, m$$

with $\phi(x) = \mathrm{e}^{-x^2/2}/\sqrt{2\pi}$ denoting the density for the standard normal distribution.

DOI: 10.1002/sim

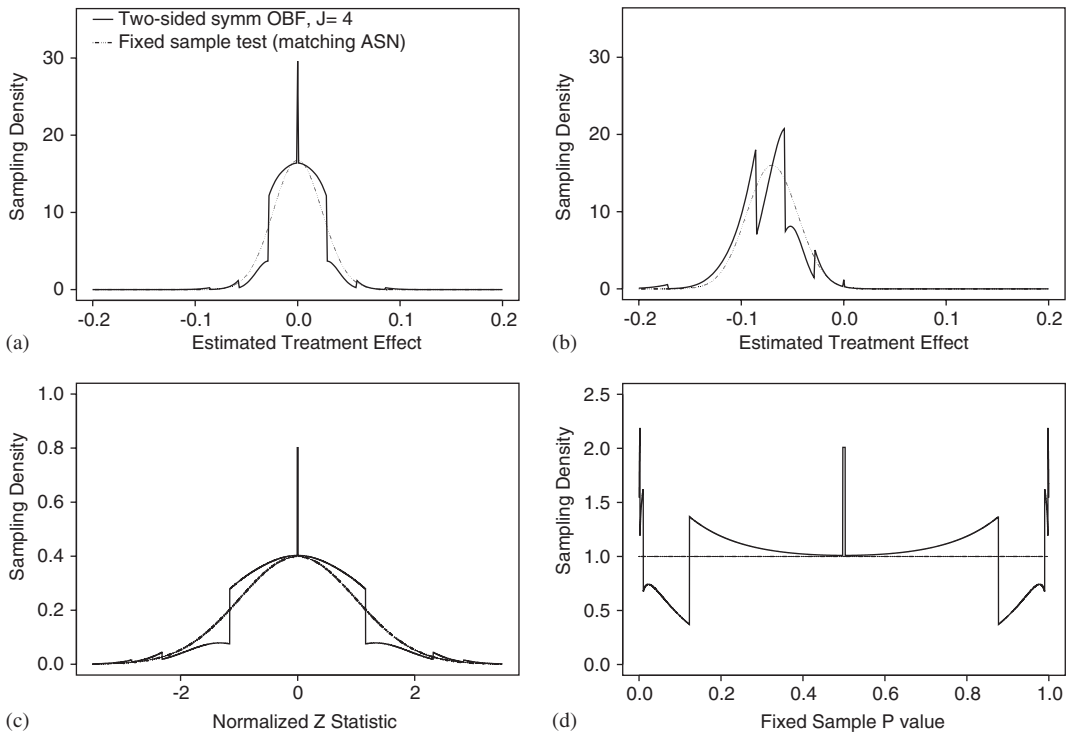

Figure 1. Comparison of sampling densities for the maximum likelihood estimate (MLE) of treatment effect, normalized Z statistic, and fixed sample P-value statistic under group sequential and fixed sample stopping rules. We consider a two-sample comparison of binomial proportions in which treatment effect parameter θ is the difference in 28-day mortality probabilities (treatment minus placebo). The group sequential stopping rule is based on a level 0.05 two-sided symmetric stopping rule having O'Brien–Fleming boundary relationships, and the sample size is a total of 1700 subjects (850 per arm). The fixed sample designs used for comparisons are constructed to have the same maximal sample size as the average sample size (ASN) of the group sequential stopping rule under either the null hypothesis ((a), (c), and (d)) or the alternative hypothesis (b): (a) MLE under $H_0 : \theta = 0$; (b) MLE under $H_1 : \theta = -0.07$; (c) Z statistic under $H_0 : \theta = 0$; and (d) fixed sample P-value under $H_0 : \theta = 0$.

It should be noted that the distribution of the group sequential statistic (M, S) is greatly altered by the use of a stopping rule. For instance, using large sample theory in a fixed sample comparison of binomial proportions across two groups, we know the estimate of treatment effect is approximately normally distributed as noted above in equation (1) and, under the null hypothesis of no treatment effect, the normalized Z statistic has the standard normal distribution and the fixed sample P-value statistic has a uniform distribution. These distributions are displayed in Figure 1 along with the corresponding distributions of those statistics when a similarly powered level 0.05 two-sided test with O'Brien–Fleming boundary relationships and a maximum of four analyses defines the sampling scheme. Using the stopping rule, the distribution of the estimated treatment effect is multimodal with jump discontinuities corresponding to the stopping boundaries at each of the analyses. The spike resembling a hypodermic needle results from the fact that this stopping rule allows early stopping at the second analysis with a decision for the null only if the estimated

DOI: 10.1002/sim

treatment effect is in a very narrow range. The parameter measuring treatment effect is clearly not a shift parameter, as the shape of the density under an alternative hypothesis is markedly different from that under the null hypothesis. We also see that under the null hypothesis the normalized Z statistic has a distribution that is far from the standard normal and that the fixed sample P-value statistic is not uniformly distributed between 0 and 1. The impact of these findings is that while standard frequentist statistical measures based on the sampling distribution can be used, they are obviously not the same as would be obtained when using the normal sampling distributions which apply to the case of fixed sample studies.

4. EVALUATION OF CLINICAL TRIAL DESIGNS

As is clear from Section 3, the scale used to define a stopping rule is unimportant, as a stopping rule on one scale induces a stopping rule on every other scale. Perhaps less clear are the reasons that so many different scales have been used in the historical development of sequential methods. Indeed, some of these scales have been developed primarily for convenience in defining methods or performing statistical computations. In this section, we discuss the use of these various scales and other operating characteristics in the evaluation of candidate study designs during the planning of a clinical trial.

In a fixed sample study in which all data are accrued prior to any analysis, reference to the operating characteristics of the test is usually taken to mean the size (the false-positive probability or type I error), the power curve (the true-positive probability or one minus the type II error computed as a function of the true treatment effect), and the sample size required for the clinical trial. In the presence of a stopping rule, however, there are additional features of the study design that are typically examined. For instance, the sample size accrued during the study is now a random variable, and hence summary statistics for that distribution might be of interest. Similarly, various collaborators in clinical trial design might be interested in the probability of stopping at each of the analyses, the estimates of treatment effect that will correspond to particular decisions, the precision of those estimates, statistical inference (both frequentist and Bayesian) that will be made when stopping at the various analyses, and the probability that a decision made at an early analysis might conflict with the decision which would have been made if the trial had continued to obtain the maximal sample size.

Frequentist measures of statistical evidence 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. Although frequentist inference by no means enjoys universal acceptance [26, 27], 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. Thus no matter how the stopping rule is derived, it is still currently of interest to establish the frequentist operating characteristics of the sequential sampling plan. As noted above, the sequential nature of the sampling plan does not greatly affect the types of operating characteristics examined, but it does greatly affect the way those operating characteristics are computed (e.g. special numerical integration routines are generally required) and summarized (e.g. the sample size requirements are described using

DOI: 10.1002/sim

summary measures of the distribution). Below we discuss the evaluation of a clinical trial design with respect to each of the frequentist operating characteristics we find most useful when selecting a stopping rule. Evaluation of Bayesian operating characteristics [4] and issues related to stochastic curtailment (conditional power and Bayesian predictive power) [28] are discussed in companion manuscripts.

4.1. General procedure and reference designs

The general process of selecting a stopping rule is an iterative one in which initially some operating characteristics are specified by some subset of the clinical trial collaborators, usually basic and clinical scientists and biostatisticians. The resulting clinical trial designs are then evaluated with respect to a wider variety of operating characteristics in consultation with a broader collection of collaborators representing logistical, financial, ethical, and regulatory perspectives. The candidate stopping rules considered might vary with respect to the hypotheses tested (including whether one-sided or two-sided tests are used), the level of statistical significance, statistical power with which specific alternatives are detected, and the degree of conservatism employed when stopping at the earliest analyses. Operating characteristics examined will include the stopping boundaries on one or more scales, the power of the test as a function of the true treatment effect (which includes as a special case the type I error as the power under the null hypothesis), sample size requirements (which can only be described probabilistically when early stopping of the trial is possible), the probability of early termination at each of the analyses as a function of the true treatment effect, statistical inference—both frequentist and Bayesian—which would be made at each of the stopping times, and various measures of the futility of continuing the trial at each stopping time. In this process, the sensitivity of those operating characteristics to various design assumptions about variability of measurement, number and timing of analyses, and patterns of data accrual are examined.

In the sepsis trial considered as an example in this paper, the trial collaborators considered a wide variety of stopping rules and criteria.

1. The trial was originally designed as a fixed sample, two-sided level 0.05 hypothesis test.
2. When the FDA requested a monitoring plan, the sponsors chose an *ad hoc* stopping rule involving a single interim analysis approximately halfway through the study. This rule suggested early termination of the study only if an interim analysis using fixed sample methods would reject the null hypothesis of no treatment effect in favor of an alternative corresponding to harm from the antibody treatment with a one-sided P-value of 0.05.
3. When a DSMB was formed, attention shifted to focus on a one-sided level 0.025 hypothesis test for this placebo-controlled study. Several stopping rules based on varying numbers of analyses with both Pocock and O'Brien–Fleming boundaries were used to familiarize the study sponsor and DSMB with the general behavior of stopping rules with respect to possible statistical inference, study power, and statistical efficiency. A one-sided symmetric test with O'Brien–Fleming boundary shape function [12] was found to satisfy the safety criteria desired by the DSMB without decreasing the power of the study when the maximum sample size was left unchanged from the fixed sample design. In particular, the conservatism of an O'Brien–Fleming efficacy boundary was considered desirable in light of the need to gain additional safety data as well as a need to provide credible evidence of effect when data from this trial would be combined with more equivocal data from previous trials. The O'Brien–Fleming

DOI: 10.1002/sim

futility boundary was judged acceptable for a single interim analysis occurring with half the maximal sample size: such a boundary would recommend stopping for futility if the crude estimate of treatment effect were in the wrong direction.

4. When the operating characteristics of this stopping rule were discussed with the study sponsors, the sponsors were impressed by the efficiency gains afforded by inclusion of a less conservative futility rule than they had considered. The sponsors were also somewhat surprised to find that the crude estimate of treatment effect that would correspond to rejection of the null hypothesis at the final analysis was as low as it was. In the design of the fixed sample study, apparently the critical value had only been expressed on the normalized Z statistic and fixed sample P-value scales—scales that relate to the strength of statistical evidence, but provide no information about scientific relevance. Thus, a sample size was selected which the sponsor's medical director now regarded as overpowered: the trial might continue until it could declare a statistically significant improvement in 28-day mortality with an estimated treatment effect that would not likely be judged clinically important and thus not of economic interest to the sponsor. In light of this newer information, the sponsors were interested in exploring stopping rules which retained an O'Brien–Fleming efficacy boundary, but used a less conservative futility boundary. Criteria examined in the selection of the ultimate stopping rule for the study included the power curve, the expected sample size curve, conditional power, and predictive probabilities.
5. When the accrual to the study turned out to be much slower than initially anticipated, the schedule of interim analyses was changed from a single interim analysis conducted after half the patients' data were available to a schedule including three interim analyses in addition to the final analysis.
6. After conducting the first interim analysis, the sponsors briefly considered decreasing the maximal sample size for the study due to the slow accrual and the fact that there was some concern that the clinical trial was perhaps still overpowered relative to the treatment effect which would correspond to an economically viable treatment. Argument against such a modification, however, was that it might cloud dealings with the FDA by raising suspicions of data-driven decisions. When the conditions under which the study would continue to the final sample size were examined, the sponsors decided that the existing stopping rule adequately protected their financial concerns. In reaching this decision, they examined the estimated treatment effect which would cause the trial to continue past the third analysis, as well as the probability of stopping at or before the accrual of 75 per cent of the maximal sample size as a function of hypothesized treatment effects.

Ultimately, the trial was stopped at the second of four planned analyses with a decision that the treatment was not sufficiently beneficial as to warrant adoption into standard medical practice. That is, while the best estimate of treatment effect suggested very slightly better 28-day survival on the antibody arm than the placebo arm, the magnitude of that observed effect was not typical of what might be reasonably observed when the true treatment effect was an absolute increase of 0.07 in 28-day survival, nor was it likely that a different decision would have obtained if the trial continued to the planned maximal sample size of 1700 subjects (850 per arm).

In the discussion of the operating characteristics which follows, we will use comparisons similar to (but not exactly the same as) those explored by the collaborators in that study. We thus consider the following stopping rules. In all cases, we consider level 0.025 one-sided hypothesis tests appropriate for testing a null hypothesis $H_0 : \theta \geqslant 0$ *versus* the lesser alternative

DOI: 10.1002/sim

$H_1 : \theta \leqslant -0.07$. The variability of the estimate of treatment effect was assumed to be that which would occur if the 28-day mortality were 30 per cent on the placebo arm and 23 per cent on the antibody arm.

1. *Fixed.Sample*: A fixed sample study with 1700 subjects (850 per arm) providing 90.66 per cent power to detect the alternative H_1.
2. *SymmOBF.2*, *SymmOBF.3*, *SymmOBF.4*: One-sided symmetric stopping rules with O'Brien–Fleming boundary relationships [12] having a total of two, three, and four equally spaced analyses, respectively, and a maximal sample size of 1700 subjects.
3. *SymmPoc.2*, *SymmPoc.3*, *SymmPoc.4*: One-sided symmetric stopping rules with Pocock boundary relationships [12] having a total of two, three, and four equally spaced analyses, respectively, and a maximal sample size of 1700 subjects.
4. *SymmOBF.Power*, *SymmPoc.Power*: One-sided symmetric stopping rules with O'Brien–Fleming and Pocock boundary relationships, respectively, with a total of four equally spaced analyses and providing 90.66 per cent power to detect the alternative H_1.
5. *Futility.5*, *Futility.8*, *Futility.9*: One-sided stopping rules from the unified family [5] with a total of four equally spaced analyses, with a maximal sample size of 1700 subjects, and having O'Brien–Fleming lower (efficacy) boundary relationships and upper (futility) boundary relationships corresponding to boundary shape parameters $P = 0.5$, 0.8, and 0.9, respectively. 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 boundary relationships.
6. *Fixed.Power*: A fixed sample study which provides the same power to detect H_1 as the *Futility.8* trial design.

4.2. Stopping boundaries

When monitoring a clinical trial, a DSMB is typically presented with interim results which then need to be compared to the stopping boundaries. It is important that the DSMB not be surprised by the conditions under which a particular stopping rule suggests that a trial might continue or stop early. Furthermore, because frequentist statistical significance must take into account the sampling scheme which led to the observed data, it is also important that the data analysts understand the conditions under which the DSMB might recommend termination of the trial. We have found in practice that a failure to examine both the scientific and statistical relevance of the stopping boundaries in detail may result in the DSMB making recommendations that do not adhere to the stopping rule defined in the study protocol. It is, to our mind, of paramount importance that the stopping boundary at each analysis be considered as the stopping rule is selected. As noted in Section 3.2, there are a number of scales on which the boundaries can be examined, though it is our feeling that not all such scales are equally useful when designing a trial.

4.2.1. On the scale of the crude estimate of treatment effect. We find it most useful to consider stopping boundaries on the scientifically relevant scale of the estimated treatment effect. The DSMB will often have prior biases regarding observed effects that are not compatible with continuing a trial. Furthermore, the sponsor often needs to consider the minimal estimated treatment effect that will correspond to statistically significant results. Too low an estimated effect, even though it might be compatible with true effects that are much larger, may not be economically viable,

DOI: 10.1002/sim

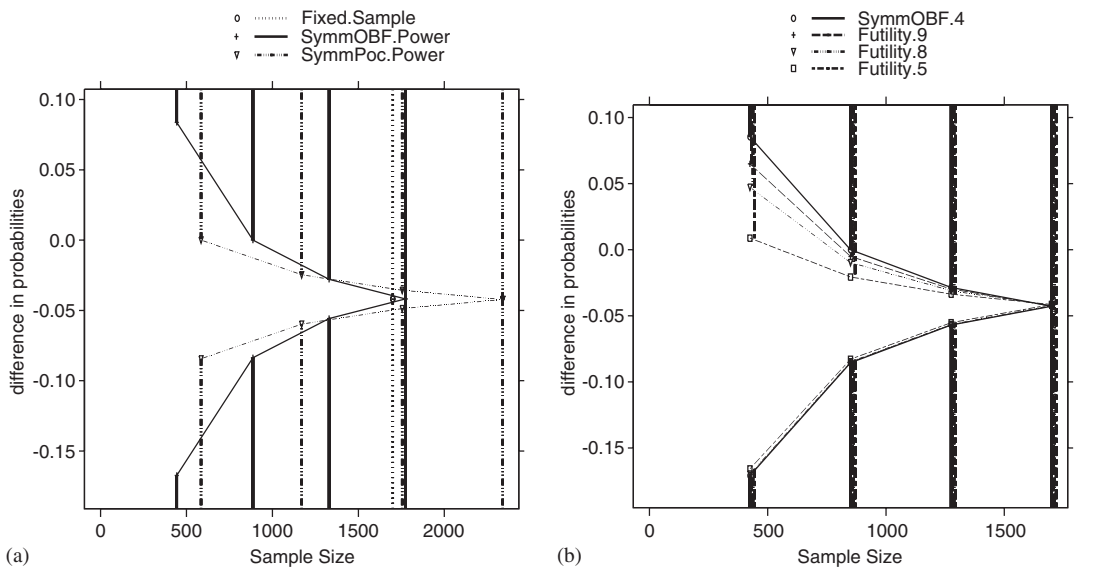

Figure 2. Stopping boundaries on the scale of the crude estimate of treatment effect: (a) stopping boundaries for a fixed sample design (*Fixed.Sample*) and one-sided symmetric tests with O'Brien–Fleming (*SymmOBF.Power*) and Pocock (*SymmPoc.Power*) boundary relationships, all having type I error of 0.025 under the null hypothesis $H_0 : \theta \geqslant 0$ and power 0.9066 to detect the alternative hypothesis $H_1 : \theta \leqslant -0.07$ and (b) stopping boundaries for level 0.025 one-sided stopping rules for a maximum of 1700 subjects (850 per arm) and having O'Brien–Fleming efficacy (lower) boundary relationships and various levels of conservatism for the futility (upper) boundary relationships.

because it is the estimate of treatment effect (as opposed to the true parameter value) which will motivate the eventual use of a new intervention.

We note that the crude (fixed sample analysis equivalent) estimate is biased and more variable than the adjusted estimates properly used for inference following the conclusion of the study [29]. Thus for scientific purposes, we prefer the frequentist inferential evaluation discussed in Section 4.7. However, because the crude estimates are the basis for the unified family [5], and because the DSMB will generally be provided such point estimates, we do find it useful to examine the boundaries on this scale.

By graphing the stopping boundaries *versus* the sample size (or statistical information) available at each analysis, we can see both the degree of conservatism employed at the earliest analyses and the worst case sample size requirements for the study. In Figure 2(a) we display the stopping boundaries for the *Fixed.Sample*, *SymmOBF.Power*, and *SymmPoc.Power* stopping rules, all of which use the same level of significance and have the same power to detect the alternative hypothesis. It should be noted that the true stopping boundaries are the vertical lines displayed in the figure; the thinner lines connecting the endpoints of those vertical lines are included only to aid the visualization of the boundary relationships.

From Figure 2(a) we see that the O'Brien–Fleming boundaries are markedly more conservative than the Pocock boundaries at the earliest analyses, and this results in the O'Brien–Fleming having smaller maximal sample size requirements (approximately a 4.3 per cent increase over the fixed sample design) than the Pocock design (approximately a 37.6 per cent increase over the fixed sample

DOI: 10.1002/sim

design). It should be noted that when graphed on this scale, upper and lower boundaries having the same boundary shape function (e.g. either O'Brien–Fleming or Pocock boundary relationships) have the same degree of curvature, but when graphed on other scales (such as the normalized Z statistic scale) these upper and lower boundaries will not appear to have the same shape. It is, of course, also possible to define asymmetric boundaries as shown in Figure 2(b). These four designs all have a maximal sample size of 1700 subjects and use O'Brien–Fleming efficacy (lower) boundary shape functions, though they differ in the boundary shape function used for the futility (upper) boundary, ranging from an O'Brien–Fleming futility boundary (*SymmOBF.4*) to a Pocock futility boundary (*Futility.5*). It can be seen that altering the futility boundary has only minimal effects on the efficacy boundary.

As an aid in the evaluation of a clinical trial design, the stopping boundaries can also be presented in tabular form as in Table I, which provides the stopping boundaries for the fixed sample design (*Fixed.Sample*), the one-sided symmetric design (*SymmOBF.4*), and the stopping rule ultimately used in the sepsis trial *Futility.8*. From this table, we see that while the modification of the futility boundary does affect the O'Brien–Fleming efficacy boundary, the difference is 0.001 on the scale of the crude estimate of treatment effect.

Table I. Stopping boundaries for *Fixed.Sample*, *SymmOBF.4*, and *Futility.8* stopping rules on several scales.

		Stopping boundaries			
Analysis time	Sample size	Crude estimate of treatment effect	Normalized Z statistic	Fixed sample P-value (lower)	Cumulative type I or II error spent
Fixed.Sample stopping boundary					
Eff 1	1700	−0.042	−1.960	0.02500	0.02500
Fut 1	1700	−0.042	−1.960	0.02500	0.02500
SymmOBF.4 stopping boundary					
Eff 1	425	−0.171	−4.007	0.00003	0.00003
Eff 2	850	−0.086	−2.833	0.00231	0.00232
Eff 3	1275	−0.057	−2.313	0.01036	0.01118
Eff 4	1700	−0.043	−2.003	0.02258	0.02500
Fut 1	425	0.086	2.003	0.97742	0.00003
Fut 2	850	0.000	0.000	0.50000	0.00232
Fut 3	1275	−0.029	−1.157	0.12372	0.01118
Fut 4	1700	−0.043	−2.003	0.02258	0.02500
Futility.8 stopping boundary					
Eff 1	425	−0.170	−3.976	0.00004	0.00004
Eff 2	850	−0.085	−2.811	0.00247	0.00248
Eff 3	1275	−0.057	−2.295	0.01086	0.01171
Eff 4	1700	−0.042	−1.988	0.02342	0.02500
Fut 1	425	0.047	1.108	0.86611	0.00085
Fut 2	850	−0.010	−0.321	0.37408	0.00591
Fut 3	1275	−0.031	−1.258	0.10425	0.01489
Fut 4	1700	−0.042	−1.988	0.02342	0.02500

Note: Cumulative error spending functions refer to the type I error spending function for the efficacy boundary and to the type II error spending function for the futility boundary.

DOI: 10.1002/sim

At the design stage, such boundaries should generally be regarded as merely illustrative of the way the chosen stopping rule might behave. During the conduct of the clinical trial, the variability of the data might be markedly different than that estimated prior to collecting any data, and logistical considerations may dictate a different schedule of analyses than originally anticipated, each of which may cause marked changes in the stopping boundaries at particular analyses. The boundaries shown in Table I for the proposed stopping rule were based on variance estimates corresponding to the alternative hypothesis of a 28-day mortality probability of 30 per cent on placebo and 23 per cent on the antibody arm. Similar tables were also examined under hypotheses corresponding to lower event probabilities (and hence less variable data and greater power to detect an absolute difference in mortality of 7 per cent) or higher event probabilities. Despite the sponsor's original intent of performing only one interim analysis, the boundaries are examined under the assumption of four equally spaced analyses in order to assess the thresholds that would result from the chosen boundary shape function if that single analysis occurred earlier or later than anticipated. Such a situation might arise when study accrual is markedly slower or faster than planned. For instance, in the presence of slower accrual than planned, the interim analysis might occur after only 425 subjects had been accrued to the study. In this setting, the O'Brien–Fleming boundary shape function would suggest early termination for futility only if the absolute difference in mortality probabilities were 0.086 in the wrong direction—a difference that seemed too large to the DSMB. The futility boundary shape function for the *Futility.8* stopping rule, on the other hand, would allow early termination for futility when there was as much as a 0.047 difference in mortality probabilities favoring the placebo arm. Without examining how the boundaries might be configured at earlier or later analyses, a stopping rule with undesirable criteria might be selected.

4.2.2. On the scale of the normalized Z statistic and the fixed sample P-value. As noted above, we prefer to display stopping boundaries on a scale that is scientifically (as opposed to merely statistically) relevant, although some authors and software packages display boundaries preferentially on the normalized Z statistic (e.g. EaSt [30]) or partial sum scale (e.g. PEST [31]). It should be noted that when plotting boundaries on scales other than the estimate scale, stopping rules with similar boundary shape functions for all boundaries will not necessarily appear to have the same shape due to the nonlinear effect of the hypothesis shift on those scales. That is, a one-sided symmetric design with, say, O'Brien–Fleming boundary relationships for both the efficacy and futility boundaries will appear to have the same shape for both boundaries when they are plotted on the scale of the estimated treatment effect, but not when they are plotted on the scales of the partial sum statistic or the normalized Z statistic.

Furthermore, a failure to report the scientifically relevant boundaries to the study sponsors and collaborating clinical researchers may lead to the design of underpowered or overpowered studies. This was in fact the case during the conduct of the sepsis trial used as an example here. When the DSMB was investigating the operating characteristics of potential stopping rules, the sponsor management and medical directors were surprised to find out the observed treatment effect which would be statistically significant in the fixed sample study proposed in their original protocol. The estimated treatment effect which would be judged statistically significant was markedly less extreme than what the sponsor felt would be necessary to generate sufficient market penetrance. However, as they had not been informed of the critical boundaries for the estimate of treatment effect, they had accepted a trial design that involved a larger sample size than might otherwise have been acceptable (see Section 4.7 for further discussion of this issue).

DOI: 10.1002/sim

Of course, even when boundaries are desired on a statistically relevant scale, it is not the case that the normalized Z statistic or the fixed sample P-value statistic necessarily provide useful information at the study design stage. As was shown in Figure 1, when data are gathered using a stopping rule, the null distribution of the normalized Z statistic does not correspond to any well-known distribution. Furthermore, it is not true that in the presence of interim analyses the critical value at the final analysis will correspond to a normalized Z statistic more extreme than that for a fixed sample statistic. Instead, the value will depend upon the relative early conservatism of the efficacy and futility boundaries. Less conservative efficacy boundaries will tend to result in a more extreme final analysis critical value for the Z statistic, and less conservative futility boundaries will tend to result in a less extreme final analysis critical value. When the upper and lower boundaries have symmetric boundary shape functions, the final analysis critical value will tend to be more extreme than that for a fixed sample test, and even in the case of the mildly asymmetric *Futility.8* boundaries this is the case: the critical value of -1.988 at the final analysis is more extreme than the fixed sample critical value of -1.96. However, for the *Futility.5* stopping rule with O'Brien–Fleming efficacy boundaries and Pocock futility boundaries, the critical value for the Z statistic at the last of four equally spaced analyses is -1.943, less extreme than the fixed sample critical value.

Similar problems exist with the interpretation of the fixed sample P-value statistic, which does not correspond to a true P-value, because it is not uniformly distributed under the null hypothesis. From Table I we can compare the efficacy boundaries for *Futility.8* on the fixed sample P-value scale to those same boundaries on the error spending scale, the latter of which can be considered a true P-value when using the analysis time ordering of the outcome space [32–36] (see Section 4.6). With the exception of the first analysis, there is no agreement between the fixed sample P-value statistic and the error spending statistic, although the difference between the two numbers is small in the case of an O'Brien–Fleming boundary. However, it is in fact possible to devise stopping rules corresponding to a level 0.025 test which would have an efficacy bound on the P-value scale equal to an arbitrary number between 0 and 1 for some analysis. We also note that the futility boundaries under the fixed sample P-value scale are not at all comparable to the error spending scale, as the former are defined for the null hypothesis and the latter for an alternative hypothesis. In the case of Table I, the error spending scale for the error spending function is defined for the alternative for which the clinical trial design provides 97.5 per cent power, although other alternatives could have been used.

Thus, while the normalized Z statistic and P-value scales are of use when implementing stopping boundaries (standard statistical routines can be used to compute Z statistics, and those statistics can be compared to the threshold for early stopping), we find that many people become confused in their attempts to gain intuition about stopping boundaries when using these scales. In particular, many clinical trialists (and indeed experienced statisticians) erroneously believe that the fixed sample P-value scale will bear some consistent relation to an error spending function as the type I or II errors are varied, which is not the case.

4.2.3. On the scale of the error spending function. Lan and DeMets [18] described a method for computing group sequential rules on the basis of the error spending scale. Such a scale has some statistical relevance in that it defines an increment of the total type I error which is to be 'spent' at each analysis. Such a scale also has been proposed for the implementation of group sequential tests when the schedule of analyses is random [5, 19]. However, at the time of study design we find that the nonlinearities inherent in this scale mean that little relevant scientific or statistical

DOI: 10.1002/sim

insight is gained at the design stage. Testimony to the fact that the operating characteristics of error spending functions are little understood is perhaps provided by noting that a number of authors [1, 18, 37] have described error spending functions which they erroneously claim would mimic the error spending functions of O'Brien–Fleming or Pocock boundaries. In fact, there is in general no single error spending function which corresponds to either of these boundary shape functions. Instead, as the schedule of analyses is varied, O'Brien–Fleming and Pocock boundaries will correspond to different error spending functions. Similarly, as the magnitude of the type I or II statistical error varies, the error spending functions for these boundary relationships will vary quite markedly.

To illustrate this behavior, the top panels in Figure 3 display the proportion of error spent at each of five equally spaced analyses for one-sided symmetric designs with O'Brien–Fleming

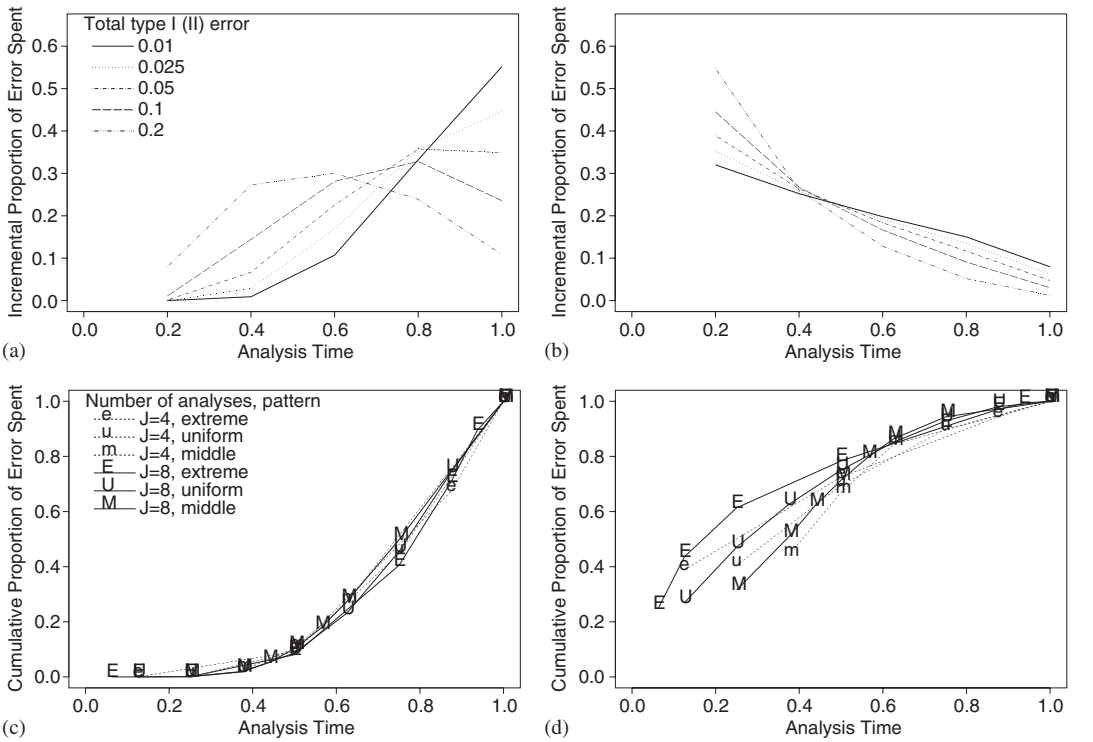

Figure 3. Actual error spending functions for one-sided symmetric stopping rules having O'Brien–Fleming (left panels) or Pocock (right panels) boundary relationships. Top panels present incremental proportion of type I (or II) error spent at each of five analyses for selected values of total type I (or II) error. Bottom panels present cumulative type I (or II) error spent as a function of number of analyses ($J = 4$ or 8) and pattern of analyses: extreme (eight analyses at 0.0625, 0.125, 0.25, 0.5, 0.75, 0.875, 0.9375, 1.0 proportion of maximal sample size), uniform (eight analyses at 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1.0 proportion of maximal sample size), or middle (eight analyses at 0.25, 0.375, 0.4375, 0.5, 0.5625, 0.625, 0.75, 1.0 proportion of maximal sample size), with patterns for four analyses corresponding to the second, fourth, sixth, and eighth analysis time of the pattern with eight analyses: (a) O'Brien–Fleming boundary relationships; (b) Pocock boundary relationship; (c) O'Brien–Fleming boundary relationships; and (d) Pocock boundary relationship.

DOI: 10.1002/sim

(Figure 3(a)) and Pocock (Figure 3(b)) boundary shape functions for type I or II errors that range between 0.01 and 0.20 (or power ranges between 0.99 and 0.80). Clearly evident are marked differences in the error spending functions. In the case of O'Brien–Fleming designs, a negligible proportion of the error is spent at the earliest two analyses when the type I or II error is 0.01 (power is 0.99), while for a type I or II error of 0.20 (power of 0.80) approximately 8 per cent of the error is spent at the first analysis and an additional 27 per cent of the error is spent at the second analysis. For the Pocock designs, the error spent at the first analysis ranges from 32 per cent for a type I or II error of 0.01 to 55 per cent for a type I or II error of 0.20.

The bottom panels of Figure 3 display the cumulative proportion of error spent at each analysis for six one-sided level 0.025 group sequential stopping rules under different patterns of interim analyses. The patterns differ according to the number of analyses ($J = 4$ or 8) and whether the analyses tend to be clustered to the extremes (early and late), uniform, or clustered to the middle. For an O'Brien–Fleming boundary relationship, the schedule of analyses has the greatest impact on the error spending function at the later analyses. Specifically, the cumulative proportion of error spent when 75 per cent of the statistical information has been accrued is 0.45 for $J = 4$ uniformly spaced analyses and 0.40, 0.46, and 0.50 for $J = 8$ analyses clustered to the extremes, uniformly spaced, or clustered to the middle, respectively. With Pocock boundary relationships, the schedule of analyses has the greatest impact on the error spent at the earlier analyses. The cumulative proportion of error spent when 25 per cent of the statistical information has been accrued is 0.40 for $J = 4$ uniformly spaced analyses and 0.61, 0.47, and 0.31 for $J = 8$ analyses clustered to the extremes, uniformly spaced, or clustered to the middle, respectively.

The question then arises whether it is more useful to name families of designs according to the similarity of their boundary shape functions on the sample mean scale (e.g. the O'Brien–Fleming, Pocock, and triangular design families) or the error spending scale. A particular nomenclature is probably most useful if it identifies designs having similar operating characteristics. For instance, when testing at intervals of equal statistical information, Pocock boundary functions tend to be approximately optimal (have lowest ASN) under the alternative hypothesis [12, 15], while triangular tests tend to be approximately optimal under an alternative intermediate to the null and alternative hypotheses. Examples of such similarity of behavior with respect to specific operating characteristics can typically also be found for families of group sequential designs which share the same error spending function, though, as demonstrated in Figure 3, they obviously cannot correspond in all cases to a family defined on the sample mean scale.

4.2.4. On other scales. As noted in Section 3.2, Bayesian posterior probabilities, conditional power, and predictive probabilities can be considered transformations of one another simply by inverting the stopping boundaries defined on some other scale. However, as these statistics can have broader application, we discuss these scales (and the corresponding operating characteristics) in companion papers to the current manuscript [4, 28].

4.3. Frequentist type I error and power

The most commonly used definition for statistical evidence against a null hypothesis is to consider the probability of falsely rejecting the null hypothesis. In fact, regulatory agencies often use this criterion as a *de facto* standard for strength of evidence that will be attained in a clinical trial design. Thus, when specifying a group sequential stopping rule, clinical trialists most often constrain the

DOI: 10.1002/sim

type I error associated with a decision boundary to some prescribed level, typically 0.05 for a two-sided test and 0.025 for a one-sided test.

Similarly, it is often the case that the sample size to be used in a clinical trial is determined by computing the sample size that will allow estimation of the treatment effect with specified precision (often according to the width of a 95 per cent confidence interval) or that will allow a decision to reject the null hypothesis to be made with high probability (e.g. 80, 90, 95, or 97.5 per cent statistical power) when a specific alternative hypothesis is true. This criterion of statistical power is of particular interest from a scientific standpoint: it describes the probability that the clinical trial will discriminate between the two viable scientific hypotheses represented by the null and alternative hypotheses. Hence, basic scientists, clinical researchers, epidemiologists, and biostatisticians often focus on the statistical power of the study to detect a hypothesis representing the minimal treatment effect which is of clinical importance.

When a clinical trial has been designed according to a given type I error and the power of the trial to detect a specific alternative, these standard operating characteristics are known to be satisfied. However, even in this setting, the probability with which the clinical trial design will detect other hypotheses is not constrained. Furthermore, when a stopping rule is defined based on stochastic curtailment criteria, Bayesian posterior probabilities, or some arbitrary criteria, neither the type I error nor the statistical power at any alternative is constrained *a priori*. Thus, it is of great interest to examine the power curve: the probability that the null hypothesis will be rejected, expressed as a function of the true treatment effect. Of particular interest when examining the power curve are (1) the type I error (the probability that the null hypothesis would be falsely rejected when the null hypothesis is in fact true), (2) the statistical power to reject the null hypothesis under specific hypothesized values for the true treatment effect that are of particular scientific interest, and (3) the alternative that would cause the null hypothesis to be rejected with high confidence.

The last of these three criteria is of particular scientific importance, because it defines the frequentist interpretation of a negative clinical trial (i.e. one with results which correspond to a failure to reject the null hypothesis). If a one-sided hypothesis test is conducted using a type I error of, say, 0.025, then using the same criteria for statistical evidence, a negative study can be regarded as rejecting the alternative hypothesis for which the clinical trial design provides statistical power of 97.5 per cent power. That is, if a 'design alternative' is defined according to the 97.5 per cent power criterion, one can with 95 per cent confidence regard that the clinical trial will perfectly discriminate between the null and alternative hypothesis: with probability 1 a 95 per cent confidence interval will not contain both the null and alternative hypotheses.

We note that the alternative for which the trial provides 50 per cent power can also be used as an approximate measure of the smallest *estimated* treatment effect which will be judged statistically significant. While this final stopping boundary on the scale of estimated treatment effect is better examined explicitly (see Section 4.6), the fact that it corresponds approximately to the 50 per cent power point (the correspondence will be exact if the stopping boundaries treat the null and alternative hypotheses symmetrically) does allow specification of designs which satisfy constraints related to the minimal estimated treatment effect that will be judged statistically significant in a frequentist hypothesis test.

Numerically integrating the sampling density for the group sequential test statistic allows straightforward computation of the power curve for any group sequential design as a function of a hypothesized true treatment effect. It is often convenient to calculate separate power curves for each boundary when a null hypothesis is tested against both upper and lower alternatives. Figure 4

DOI: 10.1002/sim

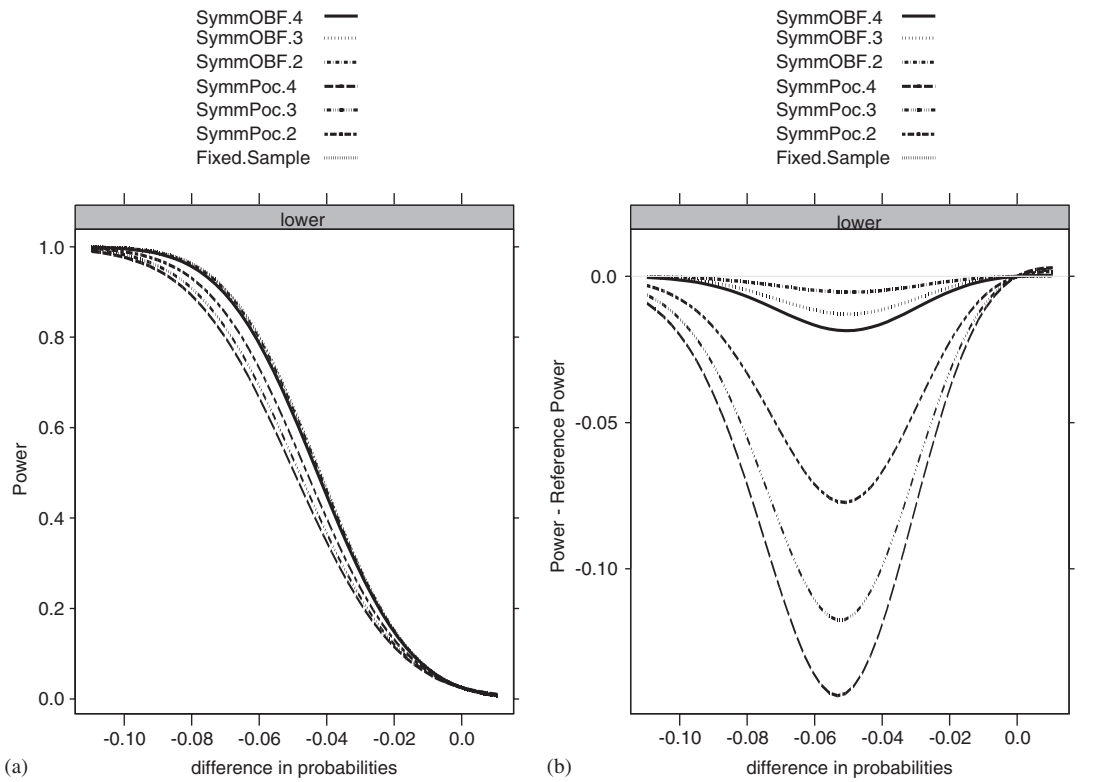

Figure 4. Power curves and difference in power relative to a fixed sample design for a fixed sample design (*Fixed.Sample*) and one-sided symmetric tests with O'Brien–Fleming (*SymmOBF.J*) and Pocock (*SymmPoc.J*) boundary relationships with $J = 2$, 3, or 4 analyses. All designs have type I error of 0.025 under the null hypothesis $H_0 : \theta \geqslant 0$ and a maximal sample size of 1700 subjects (850 per arm).

displays power curves for some stopping rules considered in the design of the sepsis trial. In this figure we compare the effect of increasing the number of interim analyses on the statistical power when the maximal sample size is maintained at 1700 patients (850 per arm). Rather than display the absolute power curve as in Figure 4(a), we often find it most convenient to display the power relative to some reference design. When absolute power curves are displayed, our eyes have a tendency to focus on the distance separating power curves in the direction perpendicular to the curves, rather than focussing on the vertical separation. By 'de-trending' the power curves displayed in Figure 4(a) (i.e. by plotting differences in power relative to some standard design), we are better able to judge the magnitude of any gains or loss in power. Thus, in Figure 4(b), we examine the loss of power relative to a fixed sample clinical trial for several stopping rules which vary in the number of interim analyses. With the O'Brien–Fleming boundary relationships considered in this figure, we see relatively little loss of power: a one-sided symmetric design with O'Brien–Fleming relationships and a total of four equally spaced analyses [12] loses at most 0.019 power (from 66.3 to 64.4 per cent) relative to a fixed sample analysis with the same maximal sample size. Had less conservative boundary relationships been chosen, a more substantial loss of

DOI: 10.1002/sim

Table II. Comparison of alternatives for which three candidate stopping rules have prescribed power and the power to detect specified alternatives, along with the average sample size accrued at study termination.

	Fixed.Sample stopping rule		*SymmOBF.4* stopping rule		*Futility.8* stopping rule	
Power	Difference in proportions	Average sample size	Difference in proportions	Average sample size	Difference in proportions	Average sample size
0.800	−0.060	1700	−0.061	1316	−0.062	1283
0.900	−0.069	1700	−0.071	1236	−0.071	1211
0.950	−0.077	1700	−0.079	1162	−0.080	1141
0.975	−0.084	1700	−0.086	1099	−0.087	1079
Difference in proportions	Power	Average sample size	Power	Average sample size	Power	Average sample size
0.000	0.025	1700	0.025	1099	0.025	987
−0.050	0.649	1700	0.631	1376	0.624	1331
−0.070	0.907	1700	0.895	1242	0.889	1222
−0.085	0.978	1700	0.974	1103	0.971	1092

power would have been noted: a one-sided symmetric design with Pocock boundary relationships would have a worst case loss of power of 0.143 (from 70.4 to 56.0 per cent).

Table II compares the power of the *Fixed.Sample*, *SymmOBF.4*, and *Futility.8* stopping rules under specific hypotheses and provides the alternative hypotheses for which the various designs have prescribed statistical power. From this table it is apparent that the introduction of either of these stopping rules has relatively minimal impact on the statistical power of the study. This in turn means that the introduction of either of these stopping rules has relatively little effect on the scientific interpretation of a failure to reject the null hypothesis: under the assumed event probability used in planning this study, and using a confidence level of 95 per cent as the statistical criterion for evidence, a failure to reject the null can be interpreted as a rejection of a treatment effect of -0.084 using the *Fixed.Sample* design, a rejection of a treatment effect of -0.086 using the *SymmOBF.4* design, and a rejection of a treatment effect of -0.087 using the *Futility.8* design. We note that this difference in rejected alternatives is due to the fact that the maximal sample size was not increased when a stopping rule was introduced. With an increase in the maximal sample size, we can maintain the magnitude of the alternative that is rejected by a failure to reject the null hypothesis.

4.4. Sample size distribution

The sample size which provides the desired precision to detect a specified alternative can be computed for a wide variety of probability models according to equation (2). That formula applies equally well to fixed sample or group sequential clinical trials when comparing means, proportions, odds, or hazard functions. The value of the standardized alternative $\delta_{\alpha\beta}$ is, however, specific to the exact stopping rule used.

In a fixed sample clinical trial, if the sample size is chosen to attain some prespecified statistical power, one of the first operating characteristics considered is whether obtaining that sample size is feasible logistically and financially. Clinical trial collaborators also have to consider whether the

DOI: 10.1002/sim

sample size would provide credible scientific evidence. For instance, if the sample size is too small, the clinical trial may not detect important, but relatively rare, adverse outcomes: a clinical trial in which n patients are treated without observing a particular adverse outcome (e.g. myocardial infarction) would correspond to an upper 95 per cent confidence bound for that adverse outcome event probability of approximately $3/n$. Other settings in which a sample size might be too small would include a situation in which results from the planned trial would ultimately be combined with results of previously conducted trials when seeking regulatory approval for a new treatment. In this setting, we must evaluate whether the additional data from the new trial would contribute a meaningful level of new information, or whether any results from this trial would be overwhelmed by the prior data.

In the presence of data collected using a stopping rule, the actual sample size obtained during the conduct of a clinical trial is a random variable with a distribution that depends on the magnitude of the true treatment effect—a dependence that is, of course, behind the ethical motivation for interim analyses: we want to use fewer patients when one treatment is markedly inferior to another or not sufficiently superior to warrant further investigation. Thus, when examining the sample size requirements of a particular clinical trial design, we will be interested in summary measures of the probability distribution for the sample size. The maximal sample size will be of interest for the feasiblity of accrual, just as it is in a fixed sample trial. Examination of the curves for the average sample size (ASN = average sample number) and various quantiles of the sample size distribution provides some indication of the values that might reasonably be attained under various hypotheses. In Figure 5(a) we compare the average and 75th percentile of the sample size distributions for the stopping rules considered in Figure 2(a), all of which had the same type I error and same power to detect the design alternative $H_1 : \theta = -0.07$. From these plots we see that although the introduction of interim analyses required inflation of the maximal sample size, both the O'Brien–Fleming and Pocock boundaries are more efficient than the fixed sample size study in terms of the average sample size. It is of particular interest to note that over the entire range of treatment effects examined in these graphs (alternatives for which the power ranges between 0.01 and 0.99) the Pocock design is on average more efficient than the O'Brien–Fleming design, despite the much greater maximal sample size (37.6 *versus* 4.3 per cent increase over the fixed sample study). For much more extreme hypothesized levels of treatment effect (i.e. values which would tend to cause early termination of the study at the first analysis with high probability), the O'Brien–Fleming design would tend to be more efficient than the Pocock design due to the earlier timing of the first analysis. By examining the curve for the 75th percentile of the sample size distribution, it can also be seen that there is at least 25 per cent probability that an O'Brien–Fleming design would recommend continuing to the final analysis when the true treatment effect is between -0.016 and -0.068, and that there is never for any hypothesized value of the treatment effect as much as a 25 per cent probability that the Pocock design would recommend continuing to the final analysis.

Figure 5(b) displays similar graphs of the average and 75th percentile of the sample size distribution for the group sequential stopping rules considered for the futility boundary in the sepsis trial. From this figure it can be seen that substantially smaller numbers of patients would be accrued on average as the futility (upper) boundary becomes successively less conservative. Of course, because the maximal sample size does not differ among these stopping rules, the power curves will vary. Therefore, the ultimate selection of a stopping rule involved simultaneous graphical comparisons of the ASN curves and the respective power curves (not shown here, but analogous to those shown in Figure 4) in order to judge the acceptability of trade-offs between the loss of power and gains in average efficiency.

DOI: 10.1002/sim

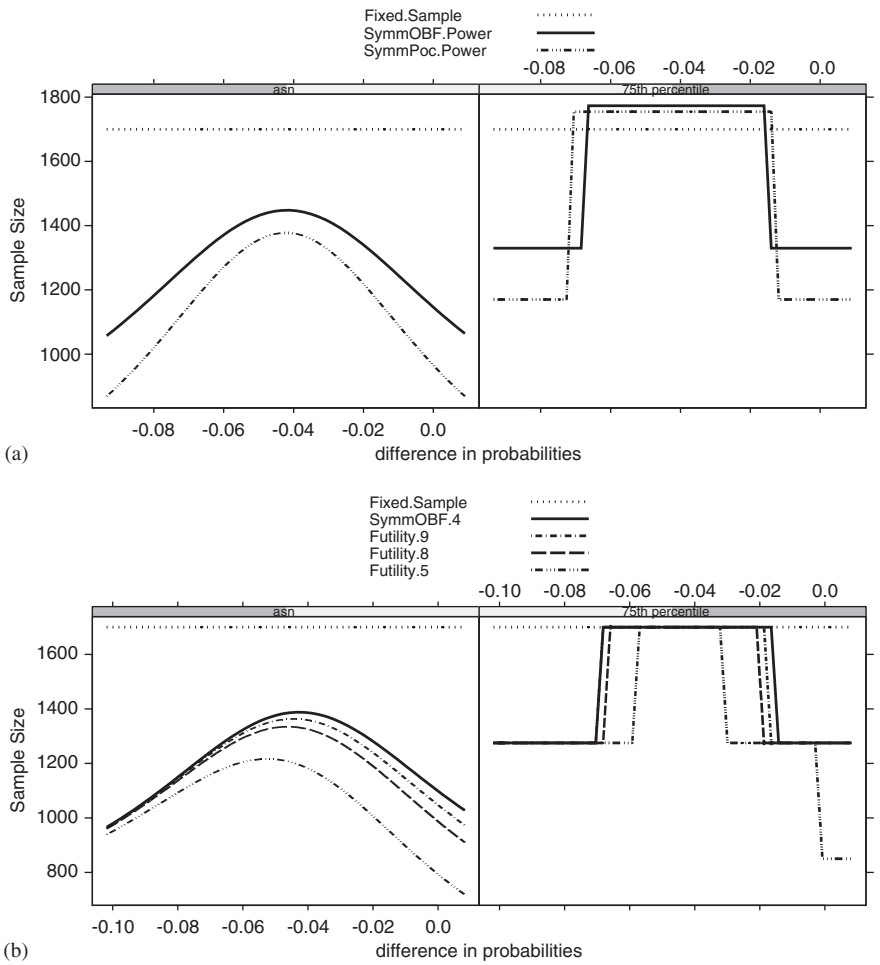

Figure 5. Average and 75th percentile of sample size distribution as a function of the hypothesized treatment effect: (a) sample size distribution for a fixed sample design (*Fixed.Sample*) and one-sided symmetric tests with O'Brien–Fleming (*SymmOBF.Power*) and Pocock (*SymmPoc.Power*) boundary relationships, all having type I error of 0.025 under the null hypothesis $H_0 : \theta \geqslant 0$ and power 0.9066 to detect the alternative hypothesis $H_1 : \theta = -0.07$ and (b) sample size distribution for level 0.025 one-sided stopping rules for a maximum of 1700 subjects (850 per arm) and having O'Brien–Fleming efficacy (lower) boundary relationships and various levels of conservatism for the futility (upper) boundary relationships.

In the actual trial, the study sponsors noted that if the null hypothesis of no beneficial treatment effect is true, the use of the *Futility.8* boundary results in an expected sample size that is 10.2 per cent lower than that using the one-sided symmetric design with O'Brien–Fleming boundary relationships, which had been deemed adequate for patient safety by the DSMB. The sponsors felt that such a gain in average efficiency made up for the worst case loss of power of 0.007 (from 74.4 per cent with the *SymmOBF.4* design to 73.7 per cent with the *Futility.8* design when $\theta = -0.057$). On the other hand, although the use of the even less conservative *Futility.5* stopping rule would

DOI: 10.1002/sim

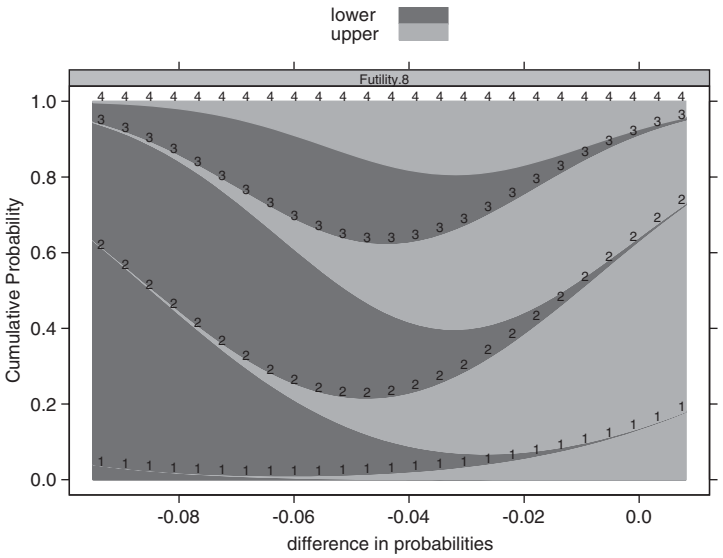

Figure 6. Cumulative stopping probabilities at each analysis as a function of hypothesized treatment effect. The one-sided level 0.025 stopping rule to test the null hypothesis $H_0 : \theta \geqslant 0$ has a maximum of four equally spaced analyses with an O'Brien–Fleming efficacy (lower) boundary shape function and a futility (upper) boundary shape function corresponding to $P = 0.8$ in the unified family of group sequential designs.

provide a 27.8 per cent decrease in the average sample size if the null hypothesis is true, they felt that the worst case loss of power of 0.033 (from 77.2 per cent with the *SymmOBF.4* design to 73.9 per cent with the *Futility.5* design when $\theta = -0.059$) was more of a change in power than was acceptable.

4.5. Stopping probabilities

When more detail about the stopping behavior of the group sequential trial design is desired, the probability of stopping at each analysis time can be examined as a function of the hypothesized true treatment effect. Figure 6 displays the cumulative stopping probability at each analysis *versus* the true treatment effect. In this figure, the shading indicates the probability with which a decision at stopping will be made for the alternative hypothesis (i.e. when the test statistic is less than the lower boundary) or the null hypothesis (i.e. when the test statistic is greater than the upper boundary). Thus, from this figure we can see that under the *Futility.8* stopping rule, when the true treatment effect corresponds to a difference in 28-day mortality probabilities of $\theta = -0.06$, the probability of stopping at or before the third analysis is approximately 0.75, and as the shading below that curve is generally the darker color, the predominant decision will be one to reject the null hypothesis.

Viewing stopping probability curves such as this was important in convincing the sponsor management to refrain from modifying the maximal sample size during the course of the study. From the stopping probability curves it could be seen that the probability of the study continuing past the third analysis was less than 0.25 unless the true treatment effect corresponded to $\theta < -0.02$.

DOI: 10.1002/sim

Furthermore, by examining the stopping boundaries on the scale of the crude estimate of treatment effect (Table I), it could be seen that the stopping rule would only recommend continuing past the third analysis if the observed difference in 28-day mortality probabilities were between -0.031 and -0.057, a situation which the sponsor management decided would look promising enough to invest in the larger sample size.

4.6. Frequentist inference at the stopping boundaries

The ultimate goal of a clinical trial is to be able to make inference about treatment effect. Design of a clinical trial is essentially directed toward guaranteeing adequate precision of that inference. In order to ensure the scientific and statistical credibility of the study results, it is therefore important at the design stage to examine the statistical inference that would be reported if the study were to be terminated early. Of particular interest is whether estimates of treatment effect would indeed be extreme enough (i.e. corresponding to clinically important differences) and precise enough (i.e. having statistically ruled out competing scientific hypotheses) to convince the scientific community that action should be taken with less precision in the estimates than would be available with a larger sample size. When using frequentist inference, we typically consider point estimates of treatment effect with small bias and mean-squared error, and we consider the precision of such estimates using 95 per cent confidence intervals. Strength of evidence against a null hypothesis is often quantified by the P-value—the probability that results as or more extreme than those actually obtained would be observed when the null hypothesis is true. These same frequentist measures are possible in the setting of group sequential stopping rules, though as noted in Section 3, the calculation of the estimates, confidence intervals, and P values must use the correct sampling distribution. Emerson and Fleming [29] compare the statistical properties of several frequentist methods of providing such estimates, and no single method clearly dominates all others. While the general approach we use in evaluation of a clinical trial design is equally applicable for all inferential methods, for purposes of illustration in this paper we use a single method based on our preference.

In the computation of P values and confidence intervals adjusted for the stopping rule, we must first decide upon an ordering of the outcome space. As noted by Emerson [38], this situation is in fact no different than that exists when making inference in many fixed sample settings, including the two sample comparison of binomial proportions in which the chi-squared test and Fisher's exact test provide different orderings of the outcome space. In the group sequential setting, the two most commonly used orderings are the analysis time ordering, investigated by Armitage [32], Siegmund [33], Jennison and Turnbull [35], and Tsiatis *et al.* [36], among others, and the sample mean ordering, investigated by Duffy and Santner [39] and Emerson and Fleming [29]. Orderings based on the likelihood ratio [40–45] have also been investigated but are not yet as widely implemented in standard statistical software.

As noted in Section 4.2.2, for those group sequential stopping rules for which it is defined, the analysis time ordering corresponds to the type I error spending function, thus providing some interpretation for the error spending function. However, confidence intervals derived under the analysis time ordering tend to be wider than those derived under the sample mean ordering [29]. As the sample mean ordering can also be applied to all stopping rules (including four boundary designs such as the double triangular test [11]), we prefer to use that ordering for inference following a group sequential test despite its dependence on knowing the timing of all interim analyses. Under the sample mean ordering, observed clinical trial results are judged more extreme

DOI: 10.1002/sim

Table III. Frequentist inference which would be reported for observed results corresponding exactly to the stopping boundaries for the *Fixed.Sample*, *SymmOBF.4*, and *Futility.8* stopping rules.

		Stopping boundaries				
Analysis time	Sample size	Crude estimate of treatment effect	Adjusted estimate of treatment effect	Adjusted P-value (lower)	Adjusted 95 per cent confidence interval	
					Lower	Upper
Fixed.Sample stopping boundary						
Eff 1	1700	−0.042	−0.042	0.02500	−0.084	0.000
Fut 1	1700	−0.042	−0.042	0.02500	−0.084	0.000
SymmOBF.4 stopping boundary						
Eff 1	425	−0.171	−0.163	0.00003	−0.224	−0.087
Eff 2	850	−0.086	−0.080	0.00241	−0.130	−0.025
Eff 3	1275	−0.057	−0.054	0.01234	−0.096	−0.007
Eff 4	1700	−0.043	−0.043	0.02500	−0.086	0.000
Fut 1	425	0.086	0.077	0.97653	0.001	0.139
Fut 2	850	0.000	−0.006	0.40112	−0.061	0.044
Fut 3	1275	−0.029	−0.031	0.06715	−0.079	0.010
Fut 4	1700	−0.043	−0.043	0.02500	−0.086	0.000
Futility.8 stopping boundary						
Eff 1	425	−0.170	−0.161	0.00004	−0.223	−0.085
Eff 2	850	−0.085	−0.079	0.00259	−0.129	−0.024
Eff 3	1275	−0.057	−0.055	0.01291	−0.096	−0.006
Eff 4	1700	−0.042	−0.044	0.02500	−0.087	0.000
Fut 1	425	0.047	0.038	0.84581	−0.037	0.101
Fut 2	850	−0.010	−0.017	0.26282	−0.071	0.034
Fut 3	1275	−0.031	−0.035	0.05297	−0.082	0.008
Fut 4	1700	−0.042	−0.044	0.02500	−0.087	0.000

Note: Adjusted estimates of treatment effect are based on the bias adjusted mean, and P values and confidence intervals are computed using the ordering based on the maximum likelihood estimate.

according to the magnitude of the crude estimate of treatment effect (the MLE in this case). The sample mean ordering would also seem to have benefits similar to the likelihood ratio ordering in obtaining pivotal results in the presence of time-varying treatment effects [45], although this is not truly a concern in this clinical trial.

In the process of evaluating group sequential designs, it is useful to consider the inference associated with outcomes which correspond exactly to the stopping boundaries. Clearly, if such outcomes are scientifically and statistically convincing, more extreme results would also be acceptable. Such inference can be tabulated (Table III) or displayed graphically to allow better visualization of the effect of early termination on the reported point estimates and precision of inference. For instance, from Table III, we see that with the *Futility.8* stopping rule, early stopping for efficacy is possible after accruing 425 patients if the treatment group has a mortality probability 0.17 less than that for the placebo group (e.g. 28-day mortality of 35 per cent on the placebo arm and 18 per cent on the antibody arm). The adjusted statistical inference reported in a scientific journal would be for a treatment effect of an absolute 16.1 per cent improvement in 28-day survival (95 per cent CI 8.5–22.3 per cent improvement in 28-day survival, $P = 0.00004$). The sponsor and DSMB must consider whether such results would in fact be both scientifically important and

DOI: 10.1002/sim

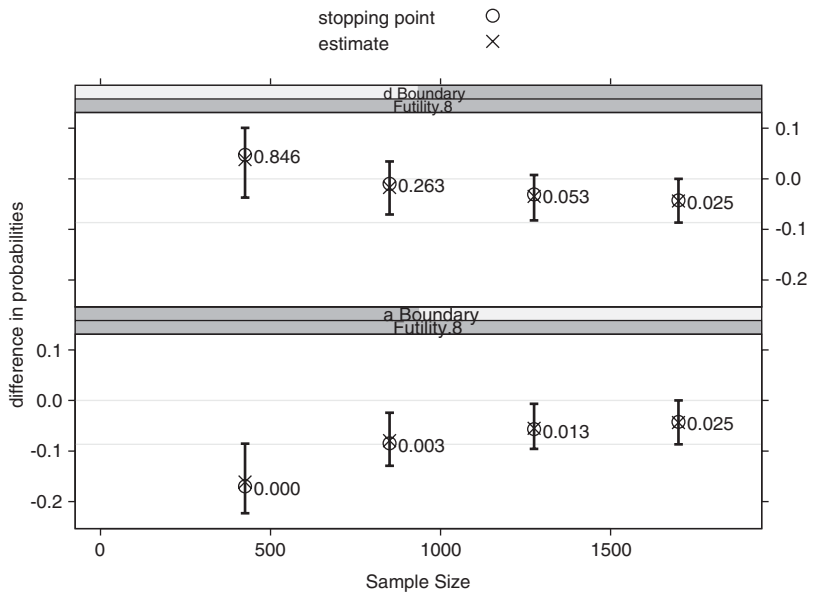

Figure 7. Display of estimates, confidence intervals, and P values for observed trial results which correspond exactly to the stopping boundaries of a one-sided level 0.025 stopping rule to test the null hypothesis $H_0 : \theta \geqslant 0$ and having a maximum of four equally spaced analyses, an O'Brien–Fleming efficacy (lower) boundary shape function, and a futility (upper) boundary shape function corresponding to $P = 0.8$ in the unified family of group sequential designs. Inference for the futility (upper) boundary is displayed in the upper panel, and inference for the efficacy (lower) boundary is displayed in the lower panel. All estimates, confidence intervals, and P values are adjusted for the stopping rule. Horizontal lines correspond to the null hypothesis $\theta = 0$ and the alternative hypothesis $\theta = -0.087$ for which the group sequential trial has power 97.5 per cent.

statistically credible. In particular, they must consider how such data might be viewed in conjunction with previous trials. Clearly, if those results would not present compelling evidence in favor of the use of the antibody in clinical settings, then further study of the treatment is necessary. In the actual trial, the sponsor and DSMB did agree that such extreme results would in fact be convincing evidence of treatment effect.

Figure 7 displays such hypothetical inference for the stopping boundaries of the *Futility.8* stopping rule. The top and bottom panels display the adjusted point estimates (bias adjusted mean as described by Whitehead [46]) and the sample mean ordering based 95 per cent confidence intervals and P values for hypothetical results which correspond to the futility (upper) and efficacy (lower) boundaries, respectively. Also displayed for reference are horizontal lines corresponding to the null hypothesis $\theta = 0$ and the alternative hypothesis $\theta = -0.087$ for which the clinical trial design provides 97.5 per cent power. As might be expected, confidence intervals corresponding to the stopping boundaries at the earliest analyses are markedly wider (less precise) than those that might be obtained at the later analyses when more data would be available. From this plot we see the extreme conservatism of the O'Brien–Fleming efficacy boundary. At the first analysis, we would stop the study early with a decision for efficacy only if we could with high confidence rule out that the treatment effect was less extreme than an alternative far beyond that which we

DOI: 10.1002/sim

considered in the design of the trial (i.e. the 95 per cent confidence interval not only excludes the null hypothesis, but also excludes an alternative corresponding to an absolute improvement in 28-day survival of 0.087). On the other hand, the futility boundary is less conservative as evidenced by the fact that although results which would cause termination have ruled out a markedly beneficial effect of treatment, they have not established with high confidence that the treatment might have some small beneficial effect (i.e. the 95 per cent confidence interval corresponding to results at the futility stopping boundary includes the null hypothesis of $\theta = 0$.)

4.7. Probability of obtaining economically important evidence of treatment effect

In recent consulting experiences, we have been asked about design operating characteristics that are primarily of economic importance. When considering the ability for a new treatment to penetrate clinical practice, marketers are keenly aware of the fact that merely demonstrating that a treatment confers some benefit over placebo may not be sufficient to generate sales of the treatment. Clinicians will base their prescription of the treatment on the estimate of benefit conferred by the treatment. Hence, in addition to the statistically credible results demanded by regulatory agencies, it is important that the trial conclude with clinically important estimates of treatment effect.

One approach to addressing this concern is to ensure that when viewed on the scale of the estimated treatment effect, the stopping boundary for declaring statistical significance is more extreme than any such clinically important threshold. Such an approach may not be most efficient, however. There may be situations in which it is economically useful to continue a study when an interim analysis shows an estimate slightly less extreme than a clinically important threshold in the hopes that additional data may cause the final estimate to exceed the threshold.

For instance, in the sepsis trial used as an example in this manuscript, the sponsors came to worry that the clinical trial was overpowered in that although an estimated treatment effect of -0.043 would be judged statistically significant, it was not felt that an expensive treatment that showed such a small difference would see much use in clinical practice. The medical director of the clinical trial felt instead that clinicians would not be likely to prescribe the new treatment unless the trial estimated a 6 per cent or greater improvement in 28-day mortality.

This then raised the question of whether a better trial design would be one in which the smallest statistically significant effect corresponded to a estimated treatment effect of -0.06. Using the *Futility.8* stopping rule, a maximal sample size of 850 results in a critical value of -0.06 at the final analysis. Thus, we can compare the 'power to obtain an economically important estimate' (i.e. an estimated treatment effect more extreme than -0.06) for the two choices of the maximal sample size under various alternatives. Planning on a maximal sample size of 1700 with the *Futility.8* boundary shape functions results in economically important estimates with probability 0.844, 0.564, 0.244, and 0.062 when the true treatment effect is $\theta = -0.08, -0.06, -0.04$, and -0.02, respectively. Planning on a maximal sample size of 850 with those same boundary shape functions results in economically important estimates with probability 0.730, 0.488, 0.250, and 0.094 when the true treatment effect is $\theta = -0.08, -0.06, -0.04$, and -0.02, respectively. Thus, we see that it can be advantageous to continue a study past the point at which the minimal economically important estimate would be the critical value.

When obtaining such an attractive estimate is of major importance, it might also be advisable to consider a stopping rule which does not allow early stopping with an estimate just slightly smaller than that economically important threshold. For instance, in the *Futility.8* design with a maximal sample size of 1700 subjects, the efficacy boundary at the third analysis

DOI: 10.1002/sim

corresponds to a crude estimate of treatment effect of -0.0566. We might thus want to consider a stopping boundary which is modified at the third analysis to stop for efficacy only if the crude estimate of the treatment effect were more extreme than -0.06. Such a constrained stopping rule can be obtained as described by Burington and Emerson [47]. Using this modified stopping rule, the probability of obtaining a crude estimate more extreme than -0.06 is 0.868, 0.589, 0.256, and 0.065 when the true treatment effect is $\theta = -0.08, -0.06, -0.04$, and -0.02, respectively.

Bayesian predictive probabilities 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. The use of Bayesian predictive probabilities is discussed in further detail in a companion manuscript [4] which more generally considers the evaluation of Bayesian operating characteristics of group sequential designs.

5. SUMMARY

In this manuscript we have described a number of frequentist criteria by which clinical trial stopping rules can be evaluated. In our approach, it is generally immaterial on which scale the stopping rule is defined. Instead the emphasis is on examining the behavior of the stopping rule with respect to scientific measures of treatment effect, statistical measures of precision, and ethical and economic measures of efficiency. Our typical evaluation of a clinical trial design will compare a number of candidate designs with respect to:

1. The scientific measures of treatment effect which will correspond to early termination for futility and/or efficacy.
2. The sample size requirements as described by the maximal sample size and summary measures of the sample size distribution (e.g. mean, 75th percentile) as a function of the hypothesized treatment effect.
3. The probability that the trial would continue to each analysis as a function of the hypothesized treatment effect.
4. The frequentist power to reject the null hypothesis as a function of the hypothesized treatment effect, with the type I error corresponding to the power under the null hypothesis.
5. The frequentist inference (adjusted point estimates, confidence intervals, and P values) which would be reported were the trial to stop with results corresponding exactly to a boundary.
6. The frequentist power to obtain a point estimate above some economically relevant threshold.

A common starting point for such an evaluation is the fixed sample design which uses the same maximal sample size that is logistically feasible. In our experience, logistical and financial considerations are most often the limiting factor in the clinical trial design, and thus much of the focus in clinical trial design shifts to questions of the trade-offs between loss of power and gains of average efficiency when using a stopping rule. Thus, the impact of variations in the schedule of interim analyses and the early conservatism of the stopping boundary are considered relative to the achievable fixed sample design. Owing to their popularity in previous clinical trials, O'Brien and Fleming [14] and/or triangular [11] boundary relationships are usually among those considered in early exploration of stopping boundaries, and design parameters are then modified

DOI: 10.1002/sim

to achieve better adherence to the other operating characteristics desired by the clinical trial collaborators.

In this manuscript we have restricted attention to those frequentist operating characteristics we regard most informative about the relative costs and benefits associated with particular sequential sampling plans when using a time-invariant measure of treatment effect. In a separate manuscript we discuss ways in which Bayesian operating characteristics can be explored and concisely reported [4] in a similar setting.

Noticeably absent from our discussion are measures related to stochastic curtailment, such as conditional power or Bayesian predictive probabilities. While we often find that some of the collaborators on a clinical trial will ask questions related to the possibility that a trial stopped early would have proceeded to the opposite decision at the final analysis, as discussed in a separate manuscript, we find that such measures are not in general helpful in choosing an appropriate stopping rule [28]. Thus, when these situations arise, we demonstrate the conflicting answers that arise from the varied approaches to stochastic curtailment: conditional power under different hypotheses and predictive power under different priors. We then use comparisons of the stopping rule and a fixed sample test to show the trade-offs between unconditional power and average sample size. It has been our experience that no given group of collaborators has ever again asked about stochastic curtailment measures.

The importance of carefully evaluating a clinical trial design cannot be overstated. In large phase III clinical trials, much time and money are invested in the scientific investigation of a new treatment. It is not uncommon that per patient costs run into the tens of thousands of dollars, and that a clinical trial would be conducted over a five-year period or longer. It is crucial that all collaborators understand the operating characteristics of the trial in order that there be no surprises regarding the conditions under which a stopping rule would or would not suggest early termination of the study. As demonstrated in this manuscript, it is relatively straightforward to consider all such issues related to the primary endpoint of the trial at the time the stopping rule is chosen, though due to the complexity of the sampling density for the group sequential test statistic, specialized computer software is necessary to perform many of the computations. Fortunately, however, commercially available software packages can be used to compute all [48] or some [30, 31] of these operating characteristics. (Annotated S+SeqTrial code used to generate all of the tables and figures in this manuscript are available from the authors.)

When implementing group sequential stopping rules, it is often the case that some of the assumptions used in the design of the clinical trial no longer hold. In particular, the actual schedule of interim analyses may differ from that specified when evaluating the trial designs and the estimates of data variability (variance of the measurements or the event probability) may have been incorrect. Fortunately, methods have been described for the flexible implementation of group sequential stopping rules defined on any of the boundary scales in such a way as to maintain the type I error [47]. Such flexible methods not only define the ways in which a stopping rule can adapt to changes in the schedule of analyses, but also the ways in which the maximal sample size for the study can be adapted to maintain the statistical power to detect an important alternative hypotheses—the so-called ‘information-based monitoring.’

With careful evaluation of stopping rules and methods for flexibly implementing those rules under changing circumstances, there seems little reason to resort to less efficient adaptive designs such as those based on using conditional power to redesign a study [49] or Fisher’s ‘self-designing clinical trial’ [50]. The most frequently cited motivation for using such adaptive designs include the possibility that at an interim analysis a clinical trialist might observe treatment effects that were

DOI: 10.1002/sim

promising, but not statistically significant, and thus want to continue the clinical trial to obtain a larger sample size. Of course, as noted in this manuscript, by examining the stopping boundary on the scale of the estimated treatment effect, all such possibilities can truly be considered at the design stage, and there is no real need to accommodate adaptive designs based solely on the estimate of the primary measure of treatment effect. Additionally, if conditions external to the trial suggest a change in the clinical or economic importance of particular alternative hypotheses or estimates of treatment effect, redesign of the clinical trial can proceed without materially affecting the relevant appropriate type I error, because in that setting the factors affecting the redesign of the trial are not based on the trial results. This then argues that there is no real need for using adaptive designs. Furthermore, there are distinct disadvantages of the adaptive methods, most notably those related to the loss of statistical efficiency [51].

Our belief is that the proper prior specification of a sampling plan that addresses the scientific, ethical, and economic concerns of the clinical trial collaborators is the best approach to clinical trial design. Examination of the operating characteristics can ensure that the chosen stopping rule provides an acceptable compromise between such competing issues as power and sample size. Methods that allow flexible implementation of those stopping rules are adequate to address the commonly encountered settings in which limited modifications of sample size are indicated.

Finally, we note that the methods discussed in this manuscript are appropriate for settings in which the measure of treatment effect does not vary with time. It is often the case, however, that a given treatment might have a delayed effect within individuals or that the effect of treatment might wear off over time. Special issues not discussed here arise in such settings. For instance, when using weighted logrank statistics in the presence of survival data exhibiting nonproportional hazards, in order to evaluate the frequentist operating characteristics discussed here one must consider (among other things):

1. The formulation of alternatives at which operating characteristics are to be evaluated.
2. The rate of information growth of the test statistic for appropriately timing interim analyses.
3. The changing censoring distribution across interim analyses and its impact on the asymptotic distribution of the test statistic under alternatives.

Gillen and Emerson [45, 52, 53] consider the information growth of weighted logrank statistics, the computation of adjusted P values following a group sequential test, and the effects of changing censoring distributions on the performance of weighted logrank statistics under nonproportional hazards treatment effects. In a further extension to the evaluation paradigm presented here, Gillen and Emerson [3] describe one general approach to the evaluation of clinical trial designs in the setting of nonproportional hazards.

ACKNOWLEDGEMENT

This research was supported by NIH grant HL69719.

REFERENCES

1. Jennison C, Turnbull BW. *Group Sequential Methods With Applications to Clinical Trials*. CRC Press: Boca Raton, FL, 2000.
2. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Wiley: New York, 1997.

DOI: 10.1002/sim

3. Gillen DL, Emerson SS. Evaluating a group sequential design in the setting of nonproportional hazards. *UW Biostatistics Working Paper Series*. Working Paper 307, 2007. http://www.bepress.com/uwbiostat/paper307
4. Emerson SS, Kittelson JM, Gillen DL. Bayesian evaluation of group sequential clinical trial designs. *Statistics in Medicine* 2007; **26**:1431–1449.
5. Kittelson JM, Emerson SS. A unifying family of group sequential test designs. *Biometrics* 1999; **55**:874–882.
6. Emerson SS, Banks PLC. Interpretation of a leukemia trial stopped early. *Case Studies in Biometry*. Wiley-Interscience: Somerset, NJ, 1994; 275–299.
7. Emerson SS. Stopping a clinical trial very early based on unplanned interim analyses: a group sequential approach. *Biometrics* 1995; **51**:1152–1162.
8. Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: a review and commentary. *Statistical Science* 1990; **5**:299–317.
9. Whitehead J. A unified theory for sequential clinical trials. *Statistics in Medicine* 1999; **18**:2271–2286.
10. Betensky RA. Alternative derivations of a rule for early stopping in favor of H_0. *The American Statistician* 2000; **54**(1):35–39.
11. Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions (corr: V39, p. 1137). *Biometrics* 1983; **39**:227–236.
12. Emerson SS, Fleming TR. Symmetric group sequential test designs. *Biometrics* 1989; **45**:905–923.
13. Pampallona S, Tsiatis AA. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference* 1994; **42**:19–35.
14. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
15. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 1987; **43**:193–199.
16. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–200.
17. Emerson SS. S + seqtrial technical overview. *Technical Report*, Insightful Corporation, Seattle, WA, 2003.
18. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
19. Pampallona S, Tsiatis AA, Kim KM. Spending functions for the type I and type II error probabilities of group sequential tests. *Technical Report*. Department of Biostatistics, Harvard University, 1995.
20. Chang MN, Hwang IK, Shih WJ. Group sequential designs using both type I and type II error probability spending functions. *Communications in Statistics*, *Part A—Theory and Methods* [*Split from*: @*J*(*CommStat*)] 1998; **27**:1323–1339.
21. Pratt JW, Raiffa H, Schlaifer R. *Introduction to Statistical Decision Theory*. MIT Press: Cambridge, MA, 1995.
22. Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. *Sequential Analysis* 1982; **1**:207–219.
23. Demets DL, Lan KKG. An overview of sequential methods and their application in clinical trials. *Communications in Statistics*, *Part A—Theory and Methods* [*Split from: @J*(*CommStat*)] 1984; **13**:2315–2338.
24. Xiong X. A class of sequential conditional probability ratio tests. *Journal of the American Statistical Association* 1995; **90**:1463–1473.
25. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A*, *General* 1969; **132**:235–244.
26. Berry DA. A case for Bayesianism in clinical trials (Disc: p. 1395–1404). *Statistics in Medicine* 1993; **12**: 1377–1393.
27. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials (Disc: p. 387–416). *Journal of the Royal Statistical Society*, *Series A*, *General* 1994; **157**:357–387.
28. Emerson SS, Kittelson JM, Gillen DL. On the use of stochastic curtailment in group sequential clinical trials. *UW Biostatistics Working Paper Series*. Working Paper 243, 2005. http://www.bepress.com/uwbiostat/paper243
29. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika* 1990; **77**:875–892.
30. The Cytel Software Corp. *EaSt*. The Cytel Software Corp.: Cambridge, MA, 2000.
31. *PEST* (*Planning and Evaluation of Sequential Trials*). The MPS Research Unit, The University of Reading, Reading, U.K., 2000.
32. Armitage P. Restricted sequential procedures. *Biometrika* 1957; **44**:9–56.
33. Siegmund D. Estimation following sequential tests. *Biometrika* 1978; **65**:341–350.
34. Fairbanks K, Madsen R. *P* values for tests using a repeated significance test design. *Biometrika* 1982; **69**:69–74.
35. Jennison C, Turnbull BW. Confidence intervals for a binomial parameter following a multistage test with application to Mil-std 105d and medical trials. *Technometrics* 1983; **25**:49–58.

DOI: 10.1002/sim

36. Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. *Biometrics* 1984; **40**:797–803.
37. Kim K, Demets DL. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 1987; **74**:149–154.
38. Emerson SS. Statistical packages for group sequential methods. *The American Statistician* 1996; **50**:183–192.
39. Duffy DE, Santner TJ. Confidence intervals for a binomial parameter based on multistage tests. *Biometrics* 1987; **43**:81–93.
40. Chang MN, O'Brien PC. Confidence intervals following group sequential tests. *Controlled Clinical Trials* 1986; **7**:18–26.
41. Rosner GL, Tsiatis AA. Exact confidence intervals following a group sequential trial: a comparison of methods. *Biometrika* 1988; **75**:723–729.
42. Chang MN. Confidence intervals for a normal mean following a group sequential test. *Biometrics* 1989; **45**: 247–254.
43. Chang MN, Gould AL, Snapinn SM. p-Values for group sequential testing. *Biometrika* 1995; **82**:650–654.
44. Cook TD. p-Value adjustment in sequential clinical trials. *Biometrics* 2002; **58**:1005–1011.
45. Gillen DL, Emerson SS. A note on P-values under group sequential testing and nonproportional hazards. *Biometrics* 2005; **61**(2):546–551.
46. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika* 1986; **73**: 573–581.
47. Burington BE, Emerson SS. Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics* 2003; **59**:770–777.
48. *S + SeqTrial*. Insightful Corporation: Seattle, WA, 2002.
49. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**: 1315–1324.
50. Fisher LD. Self-designing clinical trials. *Statistics in Medicine* 1998; **17**:1551–1562.
51. Tsiatis AA, Mehta CR. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**:367–378.
52. Gillen DL, Emerson SS. Information growth in a family of weighted logrank statistics under repeated analyses. *Sequential Analysis* 2005; **24**(1):1–22.
53. Gillen DL, Emerson SS. Non-transitivity in a class of weighted logrank statistics under non-proportional hazards. *Statistics and Probability Letters* 2007; **77**:123–130.

DOI: 10.1002/sim