UW Biostatistics Working Paper Series

Year 2005 *Paper* 243

On the Use of Stochastic Curtailment in Group Sequential Clinical Trials

Scott S. Emerson
University of Washington

John M. Kittelson
University of Colorado

Daniel L. Gillen
University of California, Irvine

On the Use of Stochastic Curtailment in Group Sequential Clinical Trials

Abstract

Many different criteria have been proposed for the selection of a stopping rule for group sequen- tial trials. These include both scientific (e.g., estimates of treatment effect) and statistical (e.g., frequentist type I error, Bayesian posterior probabilities, stochastic curtailment) measures of the evidence for or against beneficial treatment effects. Because a stopping rule based on one of those criteria induces a stopping rule on all other criteria, the utility of any particular scale relates to the ease with which it allows a clinical trialist to search for sequential sampling plans having de- sirable operating characteristics. In this paper we examine the use of such measures as conditional power and predictive power in the definition of stopping rules, especially as they apply to decisions to terminate a study early for "futility". We illustrate that stopping criteria based on stochastic curtailment are relatively difficult to interpret on the scientifically relevant scale of estimated treat- ment effects, as well as with respect to commonly used statistical measures such as unconditional power. We further argue that neither conditional power nor predictive power adhere to the standard optimality criteria within either the frequentist or Bayesian data analysis paradigms. Thus when choosing a stopping rule for "futility", we recommend the definition of stopping rules based on other criteria and careful evaluation of the frequentist and Bayesian operating characteristics that are of greatest scientific and statistical relevance.

1 Introduction

Sequential stopping rules are often used in clinical trials to address efficiency and ethical issues that arise in human experimentation. Group sequential stopping rules which maintain desired frequentist operating characteristics (e.g., type I and II error rates) were first described for situations in which early termination of the clinical trial was considered when interim results were so extreme as to suggest a beneficial effect of a new treatment.[1, 2] However, it is now also quite common for clinical trialists to choose stopping rules which allow for early stopping when all clinically important beneficial treatment effects have been credibly eliminated. Such boundaries are often referred to as "futility" boundaries, because they are meant to identify those settings in which it is futile to continue the clinical trial: The results of the clinical trial are unlikely to lead to adoption of the new therapy, and no further useful information will be obtained by continuing the study.

The statistical and clinical trials methodology literature is replete with alternative criteria to be used for the specification of a stopping rule, including stopping boundaries defined for the efficient score [3, 4, 5], normalized Z statistic and/or fixed sample P value [1, 6], crude estimate of the treatment effect [7], error spending functions [8, 9], Bayesian posterior probabilities [10, 11], conditional power [12, 13], and predictive power [14]. In companion papers to this manuscript, we have discussed the 1:1 correspondence between these various stopping boundary scales, arguing that the criterion used to define a stopping boundary is less important than the evaluation of the frequentist [15] and Bayesian [16] operating characteristics associated with it. In neither of those papers, however, did we address the evaluation of the stochastic curtailment measures of conditional and predictive power. This omission was purposeful.

When collaborating on a sequential clinical trial design, we often find that some of our collaborators will ask questions related to whether a trial stopped early with one decision would have proceeded to the opposite conclusion at the final analysis. When computed in the setting of early stopping with a failure to reject the null hypothesis, the probability of such a reversal of decisions is often regarded as a measure of the "futility" of continuing the trial: If there is only a low probability that the trial would obtain results allowing rejection of the null hypothesis, then it might

seem futile to continue collecting data.

Our response to such questions is to 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 discuss the foundational inconsistencies with stochastic curtailment measures under either frequentist or Bayesian paradigms and present alternative measures of the "futility" of continuing a clinical trial based on the tradeoffs 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. In this paper, we amplify on this presentation in the context of the sepsis clinical trial used as the example in the companion papers.

In section 2, we provide a brief review of the scientific setting and basic statistical design of the clinical trial. In section 3 we discuss statistical paradigms which might be used as a basis for a decision to terminate a clinical trial early. We present the correspondence between stopping rule thresholds defined for the efficient score, the crude estimate of treatment effect, conditional power, and predictive power. We then illustrate some of the difficulties that arise when using stochastic curtailment as a criterion for a stopping rule. We conclude in section 4 with some general comments regarding the alternatives to stochastic curtailment measures.

2 Example Used for Illustration

We illustrate our approach in the context of a randomized, double-blind, placebo-controlled clinical trial of an antibody to endotoxin in the treatment of gram-negative sepsis. Details of the scientific setting and the clinical trial design are provided in the companion paper [15].

2.1 Notation and Sample Size

Briefly, a maximum of 1,700 patients with proven gram-negative sepsis were to be randomly assigned in a 1:1 ratio to receive a single dose of antibody to endotoxin or placebo. The primary endpoint for the trial was to be the 28 day mortality rate, which was anticipated to be 30% in the placebo treated patients and was hoped to be 23% in the patients receiving antibody. Notationally, we let X_{ki} be an indicator that the i-th patient on the k-th treatment arm (k=0 for placebo, k= 1 for antibody) died in the first 28 days following randomization. Thus $X_{ki} = 1$ if the i-th patient on treatment arm k dies in the first 28 days following randomization, and $X_{ki} = 0$ otherwise. We are interested in the probability model in which the random variables X_{ki} are independently distributed according to a Bernoulli distribution $\mathcal{B}(1, p_k)$, where p_k is the unknown 28 day mortality rate on the k-th treatment arm. We use the difference in 28 day mortality rates $\theta = p_1 - p_0$ as the measure of treatment effect.

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

$$\hat{\theta} \;\dot{\sim}\; \mathcal{N}\left(\theta, \frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}\right). \tag{1}$$

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

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

where the common mortality rate under the null hypothesis would be estimated by

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

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).

In the setting of the sepsis trial, Δ would represent the difference $\theta = p_1 - p_0$ in 28 day mortality rates, and $V = p_1(1-p_1) + p_0(1-p_0)$ would be 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 p-th 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 level 0.025 hypothesis test of the null hypothesis $H_0 : p_0 = p_1$.

In a fixed sample study, the 1,700 subjects (850 per arm) provide statistical power of 0.9066 to detect the design alternative of $\theta = -0.07$ when the control group's 28 day mortality rate is 30%. If the estimated variability of $\hat{\theta}$ at the conclusion of such a trial were to agree exactly with the variance used in the sample size calculation, the null hypothesis would be rejected in a frequentist

hypothesis test if the absolute difference in 28 day mortality rates showed that the mortality on the antibody arm was at least .0418 lower than that on the placebo arm (i.e., we would reject H_0 if and only if $\hat{\theta} \leq -0.0418$). The precision provided by the planned sample size can also be characterized by the hypotheses that can be discriminated by a 95% confidence interval. For instance, a clinical trial result corresponding to the greatest observed treatment effect which still does not allow rejection of the null hypothesis (so $\hat{\theta}$ just greater than -0.0418) would allow a 95% confidence interval for θ of -0.084 to 0.000. Hence, in such a fixed sample clinical trial a failure to reject the null hypothesis could with 95% confidence be viewed as ruling out as much as an 8.4% improvement in 28 day mortality.

2.2 Definition of Stopping Rules

Stopping rules are introduced into clinical trial design in order to allow early termination of a trial when the ultimate decision is known with high confidence. Such a stopping rule defines the conditions under which accrual of new information will be halted. Typically, the conditions for early stopping are defined in the context of some statistic estimating the scientific measure of treatment effect or the statistical measures of our confidence in some decision.

Notationally, a stopping rule is defined for a schedule of analyses occurring at sample sizes N_1, N_2, ..., N_J, where we define N_j as the total number of observations accrued by the time of the jth analysis. For $j = 1, \ldots, J$, we calculate a statistic T_j based on the first N_j observations. Common choices for T_j include the maximum likelihood estimate $\hat{\theta}_j$, a normalized Z statistic based on the null hypothesis, a P value, Bayesian posterior probabilities, Bayesian predictive probabilities, or conditional power. 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 test statistic 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 observations until the available sample size is N_{j+1}. By choosing $C_J = \emptyset$, the empty set, the trial must stop at or before the J-th analysis.

As noted in the companion paper [15], for this placebo controlled trial, it seems reasonable

to restrict attention to stopping rules having at most two boundaries, i.e., stopping rules with continuation sets of the form $\mathcal{C}_j = (a_j, d_j)$ such that $-\infty \leq a_j \leq d_j \leq \infty$. In the example sepsis trial, the test statistic was defined such that interim results which were less than a_j would be suggestive of a truly beneficial treatment, and hence this lower boundary is often referred to as the "efficacy boundary" for the candidate stopping rules. Similarly, interim results which exceeded d_j would be suggestive of a treatment which was not as efficacious as hoped for. For reasons described more fully in the next section, this upper boundary is referred to as the "futility boundary".

Particular families of group sequential designs correspond to parameterized boundary functions which relate the stopping boundaries for some specified statistic T_j at successive analyses according to the proportion of statistical information accrued and some hypothesized treatment effect. For instance, letting Π_j represent the proportion of the maximal statistical information available at the j-th 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 some hypothesis of relevance to the computation of that boundary (e.g., the hypothesis rejected when $T_j > d_j$, the null or alternative hypothesis, or the current best estimate of the treatment effect). 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 [7]. The way in which the boundary shape function is combined with the boundary hypothesis will depend upon the exact form of the test statistic, and contrasting the intuitive appeal of some of the different approaches is the major topic of this paper. However, as discussed in the companion papers, 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.

For the purposes of our illustration, we consider several of the stopping rules actually considered during the design of the sepsis clinical trial. As this paper focuses primarily on the choice of "futility" boundary for the sequential sampling plan, we will restrict attention to fixed sample designs and stopping rules having an O'Brien-Fleming "efficacy" boundary combined with several candidate "futility" boundaries. Using the nomenclature from the companion paper [15], we consider level 0.025 one-sided hypothesis tests appropriate for testing a null hypothesis $H_0 : \theta \geq 0$ versus the lesser alternative $H_1 : \theta \leq -0.07$. The variability of the estimate of treatment effect was assumed to be that which would occur if the 28 day mortality were 30% on the placebo arm and 23% on the antibody arm. Specific futility stopping boundaries examined reflect a spectrum of strategies for defining such boundaries.

3 Criteria for Early Decisions Against Efficacy

Our goal in this paper is to contrast two alternative approaches to selection of "futility boundaries": decision theoretic and stochastic curtailment. These two approaches differ primarily in the way they use the "boundary hypothesis"– the hypothesized treatment effect used to compute a stopping boundary. In the decision theoretic approach, the futility stopping boundary can be parameterized by criteria for rejection of the boundary hypothesis. This is the approach used in such families as the triangular and double triangular test [3], the symmetric designs [4], the asymmetric designs of Pampallona and Tsiatis [5], the unified family [7], error spending families which consider type II as well as type I error [9, 17], and families defined for Bayesian posterior probabilities. In the stochastic curtailment approach, the futility stopping boundary considers the conditional or Bayesian predictive probability that a final study result would correspond to rejection of the null. The magnitude of such a probability depends of course on some hypothesized treatment effect (or distribution for the treatment effect in the case of Bayesian inference), and the boundary hypothesis

is used as that hypothesized effect. In order to make clear these distinctions, we first review the statistical basis for frequentist clinical trial design, and then describe in more detail each of the two approaches.

3.1 Frequentist Clinical Trial Design

The most common paradigm for clinical trial design is based on classical frequentist hypothesis testing. Treatment effect is measured by some parameter θ, which is typically some comparison (difference or ratio) of summary measures from probability distributions (e.g., means, medians, proportions exceeding some threshold, time averaged hazards). The user specifies a null hypothesis $H_0 : \theta = \theta_0$ corresponding to a treatment having no effect. Ideally, the sample size is then chosen to provide sufficient precision to be able to reject the null hypothesis with high power when some "design alternative" $H_1 : \theta = \theta_1$ is true, where θ_1 would represent some minimal treatment effect that is clinically important. "Sufficient precision" is typically taken to mean that the trial would have high probability of rejecting the null hypothesis under the design alternative, with choices of power in the range of 80% to 97.5% being common. Alternatively, sample size can be chosen according to the width of, say, a 95% confidence interval– an approach that corresponds exactly to a choice of power of 97.5%. In practice, however, logistical constraints are often the limiting factor, and our ability to accrue patients becomes a major criterion in the definitions of the "minimal treatment effect that is clinically important" and "sufficient precision".

No matter whether the values of θ_1 and statistical power are chosen purely on scientific and clinical grounds or whether the logistical constraints are the dominating factor, any given clinical trial design can be viewed as an experiment to discriminate between hypotheses. This was illustrated in section 2.1 using the fixed sample design for the sepsis trial. In that clinical trial design, a 0.025 level of significance was chosen for rejection of the null hypothesis, and it was desired to have a 90% chance of obtaining statistically significant results when the difference in 28 day mortality rates was $\theta = -0.07$. However, as also noted in section 2.1, the sample size of 1700 subjects was not sufficient to discriminate with 95% confidence between the null hypothesis and the "design alternative" of

-0.07. Instead, it is possible that an estimated treatment effect of $\hat{\theta} = -0.0417$ might be observed, with a failure to reject the null hypothesis with a P value of 0.0253 just greater than 0.025 and a 95% confidence interval for θ of -0.0835 to 0.0001. Such a confidence interval has clearly not ruled out the design alternative of -0.07, although it does necessarily rule out the alternative $\theta = 0.0837$ for which the study had 97.5% power.

3.2 Decision Theoretic Approach

The decision theoretic approach to a futility stopping boundary chooses thresholds for early termination of a study according to rejection of the alternative to be discriminated from the null hypothesis. The approach here is to define the futility stopping boundary in a manner that is exactly analogous to that used for early stopping with a decision for efficacy. Thus, a clinical trial is stopped early for futility when the data provides sufficient evidence that the alternative hypothesis is not true, with some allowance for conservatism at the earliest analyses.

A number of equivalent test statistics are commonly used in the definition of an efficacy stopping rule for a one-sided test of a lesser hypothesis. In the context of the sepsis trial introduced in section 2.1, suppose that at the j-th 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 $Y_{1j} = y_{1j}$, for a one-sided hypothesis test of a lesser hypothesis an efficacy stopping boundary in the unified family of group sequential designs [7] rejects $H_0 : \theta \geq \theta_0$ with (lower) type I error α if $\hat{\theta}_j \leq a_j^{(\hat{\theta})}$, where

$$a_j^{(\hat{\theta})} = \theta_0 - (A_a + \Pi_j^{-P_a}(1 - \Pi_j)_a^R)G_a$$

for suitably chosen design parameters A_a, P_a, R_a, and G_a. The analogous approach to a futility stopping boundary thus rejects $H_1 : \theta \leq \theta_1$ with (upper) type II error β if $\hat{\theta}_j \geq d_j^{(\hat{\theta})}$, where

$$d_j^{(\hat{\theta})} = \theta_1 + (A_d + \Pi_j^{-P_d}(1 - \Pi_j)_d^R)G_d$$

for suitably chosen design parameters A_d, P_d, R_d, and G_d. In particular, the design parameters are chosen such that $a_J^{(\hat{\theta})} = d_J^{(\hat{\theta})}$ to force stopping at the Jth analysis and such that

$$\begin{aligned} \alpha &= \sum_{\ell=1}^{J} Pr\left[\hat{\theta}_\ell \leq a_\ell^{(\hat{\theta})}, \bigcap_{k=1}^{\ell-1} a_k^{(\hat{\theta})} < \hat{\theta}_k < d_k^{(\hat{\theta})} | \theta = \theta_0\right] \\ \beta &= \sum_{\ell=1}^{J} Pr\left[\hat{\theta}_\ell \geq d_\ell^{(\hat{\theta})}, \bigcap_{k=1}^{\ell-1} a_k^{(\hat{\theta})} < \hat{\theta}_k < d_k^{(\hat{\theta})} | \theta = \theta_1\right], \end{aligned}$$

to obtain the desired type I and type II errors. Note that the choice $\alpha = \beta$ results in the same statistical criteria to be used in rejecting the null and alternative hypotheses, and with such a choice the discrimination between the null and alternative hypotheses is exactly equivalent to inference based on a $100(1-2\alpha)\%$ confidence interval for θ.

These stopping boundaries could also be converted to a number of equivalent boundary scales suitable for comparing to other test statistics:

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 [3], Emerson and Fleming [4], and Pampallona and Tsiatis [5]. An O'Brien-Fleming [2] boundary rejecting a null hypothesis of no treatment effect is constant on the scale of this statistic. Conversion of the unified family stopping boundary to this scale results in rejection of H_0 if $S_j \leq a_j^{(S)}$ and rejection of H_1 if $S_j \geq d_j^{(S)}$ where $a_j^{(S)} = N_j a_j^{(\hat{\theta})}$ and $d_j^{(S)} = N_j d_j^{(\hat{\theta})}$.

2. *Normalized Z statistic*: $Z_j = z_j = (\hat{\theta}_j - \theta_0)/se(\hat{\theta}_j) = \sqrt{N_j}(\hat{\theta}_j - \theta_0)/\sigma$ where $se(\hat{\theta}_j) \equiv \sigma/\sqrt{N_j}$ is typically estimated as described in section 2 using $\hat{\sigma} = \sqrt{2\hat{p}(1-\hat{p})}$ when sample sizes are equal on the two treatment arms. The normalized Z statistic was used for the definition of stopping rules by Wang and Tsiatis [6]. A Pocock [1] boundary rejecting a null hypothesis $H_0 : \theta = \theta_0$ of no treatment effect is constant on the scale of this statistic. Conversion of the unified family stopping boundary to this scale results in rejection of H_0 if $Z_j \leq a_j^{(Z)}$ and rejection of H_1 if $Z_j \geq d_j^{(Z)}$ where $a_j^{(Z)} = \sqrt{N_j}(a_j^{(\hat{\theta})} - \theta_0)/\sigma$ and $d_j^{(Z)} = \sqrt{N_j}(d_j^{(\hat{\theta})} - \theta_0)/\sigma$. While the

normalized Z statistic Z_j is defined in the form used for rejecting the null hypothesis, stopping boundaries could have been defined equally easily for rejecting the alternative hypothesis as

$$Z_j^* = \sqrt{N_j}(\hat{\theta}_j - \theta_1)/\sigma = Z_j - \sqrt{N_j}(\theta_1 - \theta_0)/\sigma,$$

thus showing the parallels between the form of the efficacy and futility boundaries.

3. *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 [1], 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 invariant when used with fixed sample P values computed for statistics having other distributions (e.g., the t distribution). Conversion of the unified family stopping boundary to this scale results in rejection of H_0 if $P_j \leq a_j^{(P)}$ and rejection of H_1 if $P_j \geq d_j^{(P)}$ where $a_j^{(P)} = \Phi(\sqrt{N_j}(a_j^{(\hat{\theta})} - \theta_0)/\sigma)$ and $d_j^{(P)} = \Phi(\sqrt{N_j}(d_j^{(\hat{\theta})} - \theta_0)/\sigma)$.

4. *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 j-th analysis were $S_j = s_j$, the type I error spending statistic defined for the null hypothesis $H_0 : \theta = \theta_0$ is

$$E_{aj} = \frac{1}{\alpha}\left(Pr\left[S_j \leq s_j,\ \bigcap_{k=1}^{j-1} a_k^{(S)} < S_k < d_k^{(S)} \mid \theta = \theta_0 \right] + \sum_{\ell=1}^{j-1} Pr\left[S_\ell \leq a_\ell^{(S)},\ \bigcap_{k=1}^{\ell-1} a_k^{(S)} < S_k < d_k^{(S)} \mid \theta = \theta_0 \right] \right),$$

and the type II error spending statistic defined for the alternative hypothesis $H_1 : \theta = \theta_1$ is

$$E_{dj} = \frac{1}{\beta}\left(Pr\left[S_j \geq s_j,\ \bigcap_{k=1}^{j-1} a_k^{(S)} < S_k < d_k^{(S)} \mid \theta = \theta_0 \right] + \sum_{\ell=1}^{j-1} Pr\left[S_\ell \geq d_\ell^{(S)},\ \bigcap_{k=1}^{\ell-1} a_k^{(S)} < S_k < d_k^{(S)} \mid \theta = \theta_0 \right] \right).$$

The error spending scale is used for the computation of the stopping boundaries using the methods of Lan and DeMets [8], Pampallona, Tsiatis, and Kim [9], and others [17, 18].

Computation of the probabilities used for this scale generally requires recursive numerical integration as described by Armitage, McPherson, and Rowe [19].

5. *Bayesian posterior probabilities*: The decision theoretic approach can also be used when Bayesian posterior probabilities are the basis for rejection of hypotheses. This approach is discussed in detail in the companion paper on Bayesian evaluation of group sequential stopping rules [20].

3.3 Stochastic Curtailment Approach

In the stochastic curtailment approach to a futility stopping boundary, the criterion for early stopping is based on a measure of the probability that the null hypothesis would eventually be rejected at the final analysis. This approach includes computations of frequentist conditional power and Bayesian predictive power.

Conditional power is the frequentist conditional probability that the test statistic at the final (J-th) analysis would exceed the threshold for declaring statistical significance, where we condition on the observed statistic $S_j = s_j$ at the j-th analysis and assume some particular value for the true treatment effect θ [21]. The conditional power at the j-th analysis is computed by noting that under the independent increment structure of information accrual, the test statistic at the final (J-th) analysis is a weighted average of the analogous test statistic at the j-th analysis and an increment of information accrued between the j-th analysis and the (J-th) analysis. We condition on the observed results at the interim analysis, and we compute the sampling distribution of the as yet unobserved increment under some presumption of the true treatment effect θ. Common choices for the hypothesized value of θ to use in these calculations are the null hypothesis $\theta = \theta_0$ (especially when considering interim results that lead to early stopping for efficacy), the alternative hypothesis $\theta = \theta_1$ (especially when considering interim results that lead to early stopping for futility), or the current crude estimate of treatment effect $\theta = \hat{\theta}_j$.

These computations most often ignore any effect of a stopping rule on the sampling density for

the data observed to date, and presume a distribution corresponding to a fixed sample study. 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 an efficacy threshold $a_J^{(S)}$ defined for the partial sum statistic. Such a threshold would represent the critical value for declaring statistical significance at the J-th analysis. Using large sample results and an alternative hypothesis $H_1 : \theta = \theta_1$, we might compute conditional power as

$$\begin{aligned} C_j(a_J^{(S)}, \theta_1) &= Pr(S_J \leq a_J^{(S)} \mid S_j = s_j; \theta = \theta_1) \\ &= \Phi\left(\frac{a_J^{(S)} - s_j - (N_J - N_j)\theta_1]}{\sigma\sqrt{(N_J - N_j)}}\right). \end{aligned}$$

The statistic C_j based on this conditional power can be used as a basis for a futility stopping rule if we stop the clinical trial for futility when $C_j \geq d_j^{(C)}$ for a suitable set of thresholds $d_j^{(C)}$ for $j = 1, \ldots, J-1$. (Note that unlike the statistics described for the decision theoretic approach where high values of the test statistic were suggestive of futility in the one-sided test of a lower alternative, a *low* value for the conditional power statistic will tend to lead to early stopping for futility).

It is often the case that futility rules based on conditional power use a constant threshold across successive analyses, with values of $d_j^{(C)} = 0.10$ or 0.20 chosen by many users. However, as with stopping rules based on other statistics, a boundary shape function can be used to describe thresholds that might make it easier to stop for futility as the statistical information accrues. It is also possible to describe futility stopping rules that are based on conditional power statistics that use different values of θ at the different analyses. For instance, a conditional power statistic might use the current best estimate of the treatment effect $\hat{\theta}_j$ [22] or the lower limit of, say, a fixed sample 95% confidence interval for θ computed at the jth analysis. Futility measures based on conditional power have been proposed for use when stopping a clinical trial early is to be based on stochastic curtailment [21, 12], as well as for adaptive redesign of a clinical trial [23].

It should be clear that there is a 1:1 correspondence between stopping rules defined under the decision theoretic approach and those defined using stochastic curtailment, because the conditional

power statistic defined above is a monotonic transformation of the partial sum statistic. It is thus of some interest to examine how stochastic curtailment futility boundaries relate to designs within the unified family of group sequential stopping rules. For instance, the futility boundary within the unified family will be of the form $d_j^{(S)} = N_j(\theta_1 + (A_d + \Pi_j^{-P_d}(1-\Pi_j)^{R_d})G_d)$, with an efficacy boundary of the form $a_j^{(S)} = N_j(\theta_0 - (A_a + \Pi_j^{-P_a}(1-\Pi_j)^{R_a})G_a)$. The constraint that $a_J^{(S)} = d_J^{(S)}$ dictates that the threshold for statistical significance at the final analysis is $a_J^{(S)} = N_J(\theta_0 - (A_a + 0^{R_a})G_a) = N_J(\theta_1 + (A_d + 0^{R_d})G_d)$. Inserting these formulas into the formula for the conditional power with $\Pi_j = N_j/N_J$ and $s_j = d_j^{(S)}$ yields

$$C_j(a_J^{(S)}, \theta_1) = \Phi\left(\frac{N_J 0^{R_d} G_d + N_J A_d G_d(1-\Pi_j) - N_J \Pi_j^{-P_d+1}(1-\Pi_j)^{R_d}}{\sigma\sqrt{N_J - N_j}}\right),$$

which is constant across analyses (i.e., independent of j) if $A_d = 0$, $P_d = 1$, and $R_d = 0$. In that case, which corresponds to an O'Brien-Fleming boundary shape function, $C_j(a_J^{(S)}, \theta_1) = 0.5$ at each analysis. This also suggests that no other useful member of the unified family will correspond to a constant conditional power when computed under a single alternative.

We can also examine the conditional power futility rule when computed under the maximum likelihood estimate of treatment effect at each analysis. In that case, at the jth analysis, we compute $C_j(a_J^{(S)}, \hat{\theta}_j)$, with $\hat{\theta}_j = \theta_1 + (A_d + \Pi_j^{-P_d}(1-\Pi_j)^{R_d})G_d$. Again using $\Pi_j = N_j/N_J$ and $s_j = d_j^{(S)}$ we find

$$C_j(a_J^{(S)}, \theta = \hat{\theta}_j) = \Phi\left(\frac{\sqrt{N_J}(0^{R_d} G_d - \Pi_j^{-P_d}(1-\Pi_j)^{R_d}}{\sigma\sqrt{1-\Pi_j}}\right),$$

which is constant across analyses (i.e., independent of j) if $P_d = 0$, and $R_d = 0.5$. In that case, the constant conditional power threshold will vary with the choice of A_d. There is no member of the unified family of group sequential stopping boundaries that corresponds to a constant conditional power computed using the lower bound $\hat{\theta}_j - z_{0.975}\sigma/\sqrt{N_j}$ of a fixed sample 95% confidence interval at the jth analysis.

In Table 1, we explore the relationships between stopping boundaries derived from the stochastic curtailment and decision theoretic approaches in more detail for seven group sequential designs

defined for the setting of the sepsis trial: a total of 1700 subjects used to compare 28 day mortality in a level 0.025 one-sided test of a lesser alternative. All designs considered in Table 1 assume four equally spaced analyses and have O'Brien-Fleming boundary shape functions for the efficacy boundary. The futility boundaries considered include

1. *SymmOBF.4:* O'Brien-Fleming boundary shape function as parameterized in the unified family ($A_d = 0$, $P_d = 1$, $R_d = 0$) with $\beta = 0.025$ to detect an alternative of $\theta_1 = -0.0855$.

2. *Futility.8:* A boundary shape function as parameterized in the unified family ($A_d = 0$, $P_d = 0.8$, $R_d = 0$) with $\beta = 0.025$ to detect an alternative of $\theta_1 = -0.0866$. (This was the stopping rule ultimately chosen for the sepsis clinical trial.)

3. *Futility.tri:* Triangular test [3] boundary shape function as parameterized in the unified family ($A_d = 1$, $P_d = 1$, $R_d = 0$) with $\beta = 0.025$ to detect an alternative of $\theta_1 = -0.0889$.

4. *Cond.07.20:* Stopping for futility if the conditional power to detect $\theta_1 = -0.07$ is less than 0.20.

5. *Cond.Est.20:* Stopping for futility if the conditional power to detect $\theta = \hat{\theta}_j$ at the jth analysis is less than 0.20. (This design can also be parameterized in the unified family as $A_d = 3.866$, $P_d = 0$, $R_d = 0.5$ with $\beta = 0.025$ to detect an alternative of $\theta_1 = -0.1091$.)

6. *Cond.Est.10:* Stopping for futility if the conditional power to detect $\theta = \hat{\theta}_j$ at the jth analysis is less than 0.10. (This design can also be parameterized in the unified family as $A_d = 2.267$, $P_d = 0$, $R_d = 0.5$ with $\beta = 0.025$ to detect an alternative of $\theta_1 = -0.1028$.)

7. *Cond.LowCI.20:* Stopping for futility if the conditional power to detect $\theta = \hat{\theta}_j - 1.96\sigma/\sqrt{N_j}$ at the jth analysis is less than 0.20.

For each of these seven designs, we present the frequentist inference (bias adjusted estimate, along with confidence intervals and P values computed using the sample mean ordering [24]) corresponding to the futility stopping boundaries at the jth analysis for $j = 1, 2, 3$, along with the

conditional power computations when assuming $\theta_1 = -0.07$, $\theta_1 = -0.0855$, $\theta_1 = \hat{\theta}_j$ (the current MLE), and $\theta_1 = \hat{\theta}_j - 1.96\sigma/\sqrt{N_j}$ (the lower bound of the current fixed sample confidence interval).

Table 1: Frequentist inference (bias adjusted estimate, confidence intervals and P values computed using the sample mean ordering [24]) corresponding to the futility stopping boundaries at the jth analysis for $j = 1, 2, 3$, along with the conditional power computations when assuming $\theta_1 = -0.07$, $\theta_1 = -0.0855$, $\theta_1 = \hat{\theta}_j$ (the current MLE), and $\theta_1 = \hat{\theta}_j - 1.96\sigma/\sqrt{N_j}$ (the lower bound of the current fixed sample confidence interval). Each design assumes four equally spaced analysis after 425, 850, 1275, and 1700 subjects have been accrued to the study.

Design	Analysis Time:j	BAM	Crude 95% CI	P Value	Presumed True Treatment Effect (θ) .0855	.0700	$\hat{\theta}_j$	$\hat{\theta}_{\text{low}(0.95)}$
SymmOBF.4	1	0.077	(0.001, 0.139)	0.977	0.500	0.265	0.000	0.000
	2	-0.006	(-0.060, 0.044)	0.401	0.500	0.304	0.002	0.191
	3	-0.031	(-0.079, 0.010)	0.067	0.500	0.358	0.091	0.252
Futility.8	1	0.038	(-0.037, 0.101)	0.846	0.704	0.462	0.000	0.072
	2	-0.017	(-0.071, 0.034)	0.263	0.634	0.432	0.015	0.417
	3	-0.035	(-0.082, 0.008)	0.053	0.582	0.438	0.142	0.281
Futility.tri	1	0.019	(-0.055, 0.082)	0.697	0.793	0.575	0.000	0.333
	2	-0.026	(-0.080, 0.025)	0.161	0.748	0.561	0.059	0.655
	3	-0.039	(-0.087, 0.005)	0.040	0.681	0.543	0.231	0.326
Cond.07.20	1	0.092	(0.016, 0.153)	0.990	0.416	*0.200*	0.000	0.000
	2	0.003	(-0.051, 0.053)	0.541	0.371	*0.200*	0.000	0.066
	3	-0.025	(-0.072, 0.017)	0.113	0.316	*0.200*	0.025	0.198
Cond.Est.20	1	-0.035	(-0.109, 0.016)	0.083	0.951	0.846	*0.200*	0.995
	2	-0.037	(-0.109, 0.009)	0.056	0.864	0.721	*0.200*	0.868
	3	-0.040	(-0.109, 0.005)	0.039	0.671	0.532	*0.200*	0.372
Cond.Est.10	1	-0.029	(-0.102, 0.026)	0.140	0.932	0.806	*0.100*	0.983
	2	-0.032	(-0.102, 0.016)	0.090	0.800	0.628	*0.100*	0.751
	3	-0.037	(-0.102, 0.008)	0.054	0.534	0.391	*0.100*	0.302
Cond.LowCI.20	1	0.024	(-0.049, 0.088)	0.756	0.748	0.515	0.000	*0.200*
	2	-0.011	(-0.064, 0.043)	0.376	0.504	0.307	0.003	*0.200*
	3	-0.028	(-0.074, 0.016)	0.111	0.315	0.199	0.024	*0.200*

From Table 1, we immediately see that there is a wide range of conditional power values as we vary the assumptions about the true treatment effect for the same futility stopping rule. Furthermore, it is also evident that some of the *ad hoc* rules commonly proposed for futility rules (e.g., *Cond.07.20* which suggests early stopping for futility only if the conditional power computed under the design alternative is less than 20%) are markedly more conservative than the O'Brien-Fleming boundary, which is well-known for its extreme conservatism. On the other hand, other such *ad hoc* rules (e.g., *Cond.Est.20* which suggests early stopping for futility if the conditional power computed under the current MLE is less than 20%) is so liberal as to cause substantial loss of precision, as evidenced by the width of the 95% confidence interval: At the third futility analysis, the point estimates at the *Futility.8* stopping boundary and the *Cond.Est.20* stopping boundary

are comparable, but the 95% confidence interval is 24% wider for the *Cond.Est.20* stopping rule.

In Table 1, surprisingly high conditional powers are sometimes associated with conservative stopping boundaries. Part of this seeming paradox can be explained by considering whether the assumptions used to compute the various conditional powers are relevant to the current state of knowledge. Table 2 presents such conditional probability values for both the futility and efficacy boundaries of the *Futility.8* stopping rule actually used in the sepsis trial. For each analysis time, we consider the conditional probability that an observation corresponding exactly to the threshold for early stopping might eventually lead to a test statistic at the final analysis which would allow rejection of the null hypothesis. That is, according to the stopping boundaries presented in Table 2, the null hypothesis is to be rejected if the crude estimate for the difference in 28 day mortality rates (treatment minus comparison) is -0.042 or less at the final analysis (when 1700 subjects' data is available). That stopping rule also suggests that after observing data on the first 425 subjects, a crude estimate for the difference in mortality rates of 0.047 or greater would lead to early termination of the study for futility. From Table 2, we see that if the alternative hypothesis of a difference of 28 day mortality rates of -0.07 is true, then upon observing a difference of 0.047 on the first 425 subjects, there is still a 46.2% chance that the next 1275 subjects' data would be such that the crude estimate of treatment difference would be less than -0.042 at the final analysis. On the other hand, if the null hypothesis of a true difference in mortality of 0.00 were true, there is only a 0.2% chance that the data yet to be accrued, when combined with the observed crude estimate of 0.047 at the first analysis, would result in a crude estimate of the difference in mortality rates less than -0.042 at the final analysis. Of course, if a clinical trial obtains results corresponding to the futility boundary at the first analysis, it might not be reasonable to assume either the null or alternative hypothesis. Thus some clinical trialists would consider computing the conditional probability of obtaining significant results at the final analysis under the assumption that the true difference in 28 day mortality rates corresponds to the crude estimate obtained at the current analysis, i.e., for an observed value corresponding exactly to the futility boundary at the first analysis, calculate the conditional probability of achieving a statistically significant result at the final analysis under the assumption that the true difference in mortality is 0.047. From Table

Table 2: Stopping probabilities and stochastic curtailment measures of the conditional probability of rejecting the null hypothesis at the final analysis for the *Futility*.8 stopping rule. Using this stopping rule, after 1700 subjects have been accrued to the study, a trial result corresponding to an absolute difference in 28 day mortality rates less than -0.0424 would be judged statistically significant at the 0.025 level. Conditional probabilities are computed assuming a true value of θ corresponding to the current crude estimate of treatment effect, the null $H_0 : \theta \geq 0$, and the alternative $H_1 : \theta = -0.07$.

		Probability of Exceeding Stopping At First Time			Conditional Power $\Pr(\hat{\theta}_J \geq -0.0424 \mid \hat{\theta}_j, \theta)$		
Analysis Time	Crude MLE of Treatment Effect	Alternative $\theta = -.07$	Null $\theta = 0$	Current MLE	Alternative $\theta = -.07$	Null $\theta = 0$	Current Estimate
Efficacy (lower) boundary							
1:N=425	-0.170	0.010	0.000	0.500	0.998	0.500	1.000
2:N=850	-0.085	0.302	0.002	0.477	0.990	0.500	0.998
3:N=1275	-0.057	0.400	0.009	0.331	0.950	0.500	0.907
4:N=1700	-0.042	0.178	0.013	0.200	–	–	–
Futility (upper) boundary							
1:N=425	0.047	0.003	0.134	0.500	0.462	0.002	0.000
2:N=850	-0.010	0.021	0.496	0.413	0.432	0.006	0.015
3:N=1275	-0.031	0.040	0.271	0.271	0.438	0.036	0.142
4:N=1700	-0.042	0.047	0.074	0.175	–	–	–

2, we see that under this assumption that the true difference is equal to the current crude estimate, the conditional power of the study is less that 0.05%. Similar interpretations can be applied to trial results which correspond to the futility boundary at the second analysis. Thus, if we observe a crude estimate of mortality rate difference of -0.01 after accruing data on 850 subjects, the conditional probability of a statistically significant result at the final analysis is 43.2% if the true treatment effect is the alternative of -0.07, 0.6% if the true treatment effect is the null hypothesis of 0.00, and 1.5% if the true treatment effect corresponds to the current crude estimate of -0.01.

None of these conditional power calculations are entirely satisfactory in and of themselves, because each is assuming a single value for the unknown treatment effect, and that assumption may or may not be appropriate. This problem is highlighted when we consider the conditional power calculations for the futility boundary under the assumption of the alternative. For instance, it seems somewhat surprising to see that when calculating the conditional power under the alternative hypothesis of a true difference in mortality rates of -0.07, the *Futility.8* stopping boundary for futility corresponds to surprisingly high values of conditional power– much higher than the 10% or 20% values commonly quoted by clinical trialists using conditional power criteria to define such

stopping boundaries. A conditional power of, say, 46.2% must be reconciled with the fact that the *Futility.8* boundary was chosen in the actual clinical trial, because it did not result in a very marked loss of statistical power for the alternative hypotheses of greatest interest, nor did adoption of that stopping rule greatly affect the alternatives for which the clinical trial has prescribed levels of power [15]. This seeming paradox is resolved when we consider whether assuming the alternative hypothesis when such results have been obtained is reasonable.

To address this issue, we also present in Table 2 the probability of stopping the clinical trial at each of the analyses under the corresponding presumed true treatment effects. It should be noted that while the stopping probabilities under the null hypothesis of $\theta = 0$ and the design alternative of $\theta = -0.07$ sum to 1.0, those given under the assumption that the current MLE is correct do not. This is because in the column corresponding to the current MLE, a different treatment effect is presumed for each row of the table. Immediately apparent from these stopping probabilities is the fact that it is often the case that presuming the null or alternative hypothesis is true is often quite unreasonable for some stopping boundaries. For instance, if $\theta = -0.07$, the probability of stopping at the first analysis with a decision for futility is 0.003. From Table 1, we see that at that boundary, the 95% confidence interval for θ is from -0.037 to 0.101. These results would argue that a conditional power computation based on a presumed treatment effect of $\theta = -0.07$ (which has been ruled out with high confidence) was largely irrelevant. The 46.2% conditional probability or a reversed decision under this presumption is a negligible number of actual trials. This point is examined further using simulations.

For each of the hypotheses used to compute the conditional power at the stopping boundaries, Table 3 presents the results of one million clinical trials simulated under either the null ($\theta = 0.00$) or design alternative ($\theta = -0.07$) hypotheses. Included in this table is the estimated probability that the trial might exceed either the futility or efficacy boundaries at each analysis. From this table we see that the probability that the trial might stop at the first analysis with a crude estimate of the difference in mortality rates of 0.047 or more is 0.31% when the true difference is -0.07 and 13.51% when the true difference is 0.00. Similarly, the probability that the trial might continue past the first analysis and then stop at the second analysis with a crude estimate greater (more

positive) than -0.01 is 2.24% and 50.34% when the true difference is -0.07 and 0.00, respectively. Again, the extremely low probability of stopping for futility at the first analysis when the true treatment effect is -0.07 would, in a frequentist sense, argue against the relevance of a conditional power calculation computed under that hypothesis.

It is thus clear that stopping a trial for futility might be quite reasonable despite there being a high conditional power of reversing the decision at the planned final analysis of a continued trial. This is illustrated further in Table 3, which also explicitly considers the probability of conflicting decisions being made at interim analyses and a planned final analysis. Rather than focusing on trial results occurring exactly on the stopping boundaries, we present for each stopping boundary, both the conditional and the unconditional probabilities that the decision made at an interim analysis would not agree with the decision made in a fixed sample design. It should be noted, however, that when comparing a group sequential design to a fixed sample test in this way, we must consider the differences in power and sample size. That is, compared to a fixed sample test with the same maximal sample size, a group sequential test has less power. On the other hand, when compared to a fixed sample test having the same power, the group sequential test uses fewer subjects on average. In an attempt to isolate the value of conditional power as a futility measure, we compute the probability of reversed decisions relative to a fixed sample design which either has the same maximal sample size (1700 subjects), has the same sample size as the worst case expected sample size (the worst case ASN for *Futility.8* is 1336 subjects when the true treatment effect is -0.047), or has the same power to detect the alternative of a true treatment effect of -0.07 (1598 subjects). We note that when comparing efficiency of statistics, the usual comparison is that between two statistics providing the same type I error and the same power to detect some alternative. Hence, the comparison based on matched power is perhaps the most theoretically relevant of these fixed sample tests.

One million clinical trials are simulated under the null hypothesis of a true difference in mortality rates of 0.00 and the alternative hypothesis of a true difference of -0.07. Under each hypothesis, we count the number of studies which stop at each analysis for futility and efficacy. We then compute the percentage of those stopped studies which would have had a reverse statistical decision made

Table 3: Probabilities that trial decisions at an interim analysis might disagree with that obtained in a fixed sample analysis. One million clinical trials with a total of 1700 patients were simulated under the null and alternative hypotheses. Test statistics were computed at four equally spaced analyses, and the *Futility*.8 stopping rule was used to make decisions regarding early stopping. Analyses of each simulated data set were also performed at sample sizes corresponding to the level 0.025 fixed sample tests having the same sample size as the worst case ASN of the *Futility*.8 stopping rule (N= 1336), having the same power (N= 1598), and having the same maximal sample size as the stopping rule (N=1700). For each interim analysis, the empirical probability of stopping for efficacy or futility was computed, along with the unconditional probability that the stopping rule would dictate early stopping with one decision and the fixed sample test would result in the opposite decision. Also presented is the conditional probability of reverse decisions defined as the proportion of trials stopped at a given analysis which would have a reverse decision in the fixed sample test.

			Same Worst Case ASN (N=1336)		Same Power Under the Alternative (N=1598)		Same Maximal Sample Size (N=1700)	
Analysis Time	Crude Est of Trt Effect	Stopping Probability	Cond	Uncond	Cond	Uncond	Cond	Uncond
			Null Hypothesis : $\theta = 0$					
Efficacy (lower) boundary								
1: N= 425	-0.170	0.0000	0.3103	0.0000	0.4483	0.0000	0.4483	0.0000
2: N= 850	-0.085	0.0024	0.2242	0.0005	0.3477	0.0008	0.3751	0.0009
3: N=1275	-0.057	0.0091	0.0228	0.0002	0.2446	0.0022	0.3066	0.0028
4: N=1700	-0.042	0.0132	0.5538	0.0073	0.2431	0.0032	0.0000	0.0000
Total		0.0247	0.3258	0.0080	0.2539	0.0063	0.1498	0.0037
Futility (upper) boundary								
1: N= 425	0.047	0.1351	0.0003	0.0000	0.0007	0.0001	0.0009	0.0001
2: N= 850	-0.010	0.5034	0.0004	0.0002	0.0014	0.0007	0.0017	0.0009
3: N=1275	-0.031	0.2619	0.0000	0.0000	0.0040	0.0010	0.0063	0.0017
4: N=1700	-0.042	0.0749	0.1040	0.0078	0.0596	0.0045	0.0159	0.0012
Total		0.9753	0.0082	0.0080	0.0065	0.0063	0.0039	0.0038
			Alternative Hypothesis : $\theta = -.07$					
Efficacy (lower) boundary								
1: N= 425	-0.170	0.0091	0.0012	0.0000	0.0015	0.0000	0.0014	0.0000
2: N= 850	-0.085	0.2984	0.0035	0.0011	0.0033	0.0010	0.0029	0.0009
3: N=1275	-0.057	0.4010	0.0020	0.0008	0.0091	0.0037	0.0088	0.0035
4: N=1700	-0.042	0.1791	0.3909	0.0700	0.0940	0.0168	0.0000	0.0000
Total		0.8877	0.0810	0.0719	0.0242	0.0215	0.0050	0.0044
Futility (upper) boundary								
1: N= 425	0.047	0.0031	0.1927	0.0006	0.3484	0.0011	0.4038	0.0012
2: N= 850	-0.010	0.0224	0.0794	0.0018	0.2534	0.0057	0.3283	0.0073
3: N=1275	-0.031	0.0391	0.0003	0.0000	0.1722	0.0067	0.2764	0.0108
4: N=1700	-0.042	0.0478	0.1974	0.0094	0.1819	0.0087	0.0719	0.0034
Total		0.1123	0.1052	0.0118	0.1974	0.0222	0.2032	0.0228

in a fixed sample study conducted after accruing either 1336 (for the study with the same worst case ASN), 1598 (for the same power study), or 1700 (for the same maximal sample size study). These conditional probabilities should correspond only approximately to the conditional power calculations given in Table 2, because in Table 3 we consider studies which exceed the stopping boundaries in addition to those which stop with results exactly on the boundary. We also present the unconditional probabilities of reversed decisions, which are equal to the conditional probability times the stopping probability.

From Table 3 we see that under the alternative hypothesis of a true treatment effect of -0.07, the probability of stopping for futility at the first analysis is approximately 0.003. Of those trials that

would stop early for futility in this manner, approximately 40.4% would correspond instead to a decision for efficacy in a fixed sample analysis conducted with data from 1700 subjects. This number differs somewhat from the conditional power of 46.2% reported in Table 2 in part because the latter number conditions on results observed exactly on the boundary at the first analysis while the value in Table 3 includes trial results that exceeded the futility boundary by some amount and thus would tend to have a lower probability of being reversed. The actual impact of this relatively high conditional probability is quite slight however: As shown in the column for unconditional power of a reversed decision for this fixed sample test, a 40.4% reversal rate for studies stopped for futility at the first analysis corresponds to 0.12% of all studies. Table 3 also shows that approximately 19.3% of studies stopped for futility at the first interim analysis would be expected to correspond to a decision for efficacy if a fixed sample study continued to accrue 1336 subjects, although such reversal of the decision represents only 0.06% of all possible outcomes under the alternative. Similarly, while 34.8% of those trials stopping for futility at the first analysis do not agree with the result which would have been reported in a fixed sample study with 1598 subjects, the actual proportion of studies with such a reversal is quite small at 0.11%. Clearly, there would be minimal impact on the unconditional power when using futility rules which correspond to these seemingly high thresholds for conditional power.

This then highlights one problem with the use of conditional power arguments: A high conditional power may correspond to a neglible proportion of trials overall, and a lower conditional power may correspond to a higher proportion of trials overall. For instance, though the conditional probability of reversing a decision for efficacy at the first analysis is approximately 44.8% under the null hypothesis when considering a 1700 subject fixed sample study, this pertains to less than 0.005% of the one million simulated trials. On the other hand, the conditional probability under the alternative hypothesis of reversing a decision for futility at the third analysis is lower at 27.6% but pertains to 1.1% of the one million simulated trials.

A further foundational problem with the use of conditional power is apparent when a group sequential design is compared to a fixed sample design having the same power to detect the alternative hypothesis. When the group sequential design *Futility.8* is compared to a fixed sample design

having 1598 subjects, both the type I error (0.025) and the power (0.975) agree between the two studies. This can be seen from Table 3 by noting that the total proportion of trials corresponding to a futility decision in the group sequential test and an efficacy decision in the fixed sample test is 0.0063 under the null hypothesis and 0.0222 under the alternative hypothesis–identical (except for random sampling in the simulations) to the proportion of trials corresponding to an efficacy decision in the group sequential test and a futility decision in the fixed sample test under the respective hypotheses (0.0063 under the null and 0.0215 under the alternative). When holding the treatment effect constant, there is of course no reason to prefer making a mistake with one sample over another. In the case of the null hypothesis, ever deciding for efficacy is an error, and trading an earlier erroneous decision for a later erroneous decision (or vice versa) is of no consequence on the error rates when only the behavior of the test under the null is considered. Instead, the usual frequentist paradigm is to consider which error made under the null hypothesis will lead to a more powerful and/or efficient test. Because conditional power arguments are based solely on considering tradeoffs between decisions made under the same hypothesis, they cannot accurately predict the impact of a stopping rule on statistical power or efficiency. (These latter concerns are adequately addressed by evaluating the impact of a stopping rule on the power curve and the ASN curve relative to various fixed sample designs, as illustrated in Table 5 below.)

As noted above, a portion of the seeming paradox between conditional power calculations and the more relevant unconditional power and efficiency considerations is due to the use of (at times) unreasonable assumptions in the calculation of conditional power. The use of the current MLE and/or the lower bound of confidence intervals to calculate conditional power as shown in Table 1 was an effort to address this problem. Another approach to avoid basing calculations on untenable assumptions uses a Bayesian paradigm.

The use of Bayesian prior distributions to obtain predictive probabilities addresses some, but not all, of the problems identified with conditional power. In this approach, the observed data is used to update some prior distribution for the treatment effect, and then the predictive distribution of the result at the final analysis is obtained by integrating over the posterior distribution of the treatment effect parameter. These predictive probabilities have a distinct advantage over the

conditional probabilities in that the predictive probabilities take into account both prior notions of the likely values for the true treatment probability and the evidence in the data for the true value.

The Bayesian predictive probability is the probability that the test statistic would exceed some specified threshold at the final analysis, using a prior distribution and the observed data to compute a posterior distribution for the treatment effect parameter at the j-th analysis. We consider a robust approach to Bayesian inference based on a coarsening of the data by using the asymptotic distribution of a nonparametric estimate of treatment effect [25] as described more fully in our companion paper on Bayesian evaluation of group sequential designs [16]. That is, rather than the exact binomial distributions for the two arms of the sepsis trial, we use the approximate normal distribution for the estimated difference in 28 day mortality rates. In the case of a compuatationally convenient conjugate normal prior $\theta \sim N(\zeta, \tau^2)$, at the jth analysis we can define an approximate Bayesian posterior distribution for the true treatment effect θ conditioned on the observation $\hat{\theta}_j$ as

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

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

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

We might therefore compute a predictive probability statistic analogous to the conditional power statistic as

$$\begin{aligned} H_j(a_J^{(\hat{\theta})}, \zeta, \tau^2) &= \int Pr(\hat{\theta}_J < a_J^{(\hat{\theta})} \mid S_j = s_j, \theta)\, p(\theta \mid S_j = s_j)\, d\theta \\ &= \Phi\left(\frac{[\Pi_j\tau^2 + \sigma^2/N_J][a_J^{(\hat{\theta})} - \hat{\theta}_j] + [1-\Pi_j][\hat{\theta}_j - \zeta]\sigma^2/N_J}{\sqrt{[1-\Pi_j][\tau^2 + \sigma^2/N_J][\Pi_j\tau^2 + \sigma^2/N_j]\sigma^2/N_J}}\right). \end{aligned}$$

The case of a noninformative (although improper) prior is of special interest. When we consider

taking the limit as $\tau^2 \to \infty$, the predictive probability statistic becomes

$$H_j(a_J^{(\hat{\theta})}, \zeta, \tau^2 = \infty) \quad = \quad \Phi\left(\frac{(a_J^{(\hat{\theta})} - \hat{\theta}_j)\sqrt{\Pi_j}}{\sqrt{[1-\Pi_j]\sigma^2/N_J}}\right).$$

As with the conditional power statistic, we can examine the relations between stopping rules defined based on thresholds for predictive power and stopping boundaries defined using the decision theoretic approach of the unified family. On the MLE scale, the futility boundary within the unified family will be of the form $d_j^{(\hat{\theta})} = \theta_1 + (A_d + \Pi_j^{-P_d}(1-\Pi_j)^{R_d})G_d$, with an efficacy boundary of the form $a_j^{(\hat{\theta})} = \theta_0 - (A_a + \Pi_j^{-P_a}(1-\Pi_j)^{R_a})G_a$. The constraint that $a_J^{(\hat{\theta})} = d_J^{(\hat{\theta})}$ dictates that the threshold for statistical significance at the final analysis is $a_J^{(\hat{\theta})} = \theta_0 - (A_a + 0^{R_a})G_a = \theta_1 + (A_d + 0^{R_d})G_d$. Inserting these formulas into the formula for the conditional power with $\Pi_j = N_j/N_J$ and $\hat{\theta}_j = d_j^{(\hat{\theta})}$ yields

$$H_j(a_J^{(\hat{\theta})}, \zeta, \tau^2) = \Phi\left(\frac{[\Pi_j\tau^2+\sigma^2/N_J][0^{R_d}-\Pi_j^{-P_d}(1-\Pi_j)^{R_d}]G_d+[1-\Pi_j][\theta_1+(A_d+\Pi_j^{-P_d}(1-\Pi_j)^{R_d})G_d-\zeta]\sigma^2/N_J}{\sqrt{[1-\Pi_j][\tau^2+\sigma^2/N_J][\Pi_j\tau^2+\sigma^2/N_j]\sigma^2/N_J}}\right),$$

which is in general dependent upon j, suggesting that no useful member of the unified family of stopping rules corresponds to a constant threshold on the Bayesian predictive probability scale for an arbitrary prior. However, for a noninformative prior, the statistic on a unified family futility stopping boundary becomes

$$H_j(a_J^{(\hat{\theta})}, \zeta, \tau^2) \quad = \quad \Phi\left(\frac{[\Pi_j^{0.5}0^{R_d} - \Pi_j^{-P_d+0.5}(1-\Pi_j)^{R_d-0.5}]G_d}{\sqrt{\sigma^2/N_J}}\right),$$

which is constant across analyses (i.e., independent of j) if $P_d = 0.5$, and $R_d = 0.5$. In that case, the constant conditional power threshold will vary with the choice of A_d. Such a boundary also corresponds to Xiong's [26] sequential conditional probability ratio test.

In Table 4, we explore the relationships between stopping boundaries derived from the predictive probability and decision theoretic approaches in more detail for the *SymmOBF.4*, *Futility.8*, and

Futility.tri stopping rules defined above, as well as for five stopping rules defined by Bayesian predictive probabilities. Again, all group sequential designs are defined for the setting of the sepsis trial: a total of 1700 subjects used to compare 28 day mortality in a level 0.025 one-sided test of a lesser alternative. All designs considered in Table 4 assume four equally spaced analyses and have O'Brien-Fleming boundary shape functions for the efficacy boundary. The futility boundaries all correspond to early stopping if the predictive probability of a significant result at the final analysis is less than 10%, though they differ according to the location (ζ) and spread (τ^2) of the prior distribution for θ. The rules considered include some of those considered during the planning of the sepsis clinical trial: [16]

1. *Pred.Dogm.Opt:* A highly dogmatic (prior SD $\tau = 0.015$), optimistic prior (prior mean $\zeta = -0.09$).
2. *Pred.Dogm.Pess:* A highly dogmatic (prior SD $\tau = 0.015$), pessimistic prior (prior mean $\zeta = 0.02$).
3. *Pred.Vague.Opt:* A vague (prior SD $\tau = 0.15$), optimistic prior (prior mean $\zeta = -0.09$).
4. *Pred.Vague.Pess:* A vague (prior SD $\tau = 0.15$), pessimistic prior (prior mean $\zeta = 0.02$).
5. *Pred.Consensus:* The sponsor's consensus prior (prior SD $\tau = 0.04$, prior mean $\zeta = -0.04$).
6. *Pred.Noninform:* A noninformative prior (prior SD $\tau = \infty$). (This design can also be parameterized in the unified family as $A_d = 1.77$, $P_d = 0.5$, $R_d = 0.5$ with $\beta = 0.025$ to detect an alternative of $\theta_1 = -0.0906$.)

For each of these designs, we present in Table 4 the frequentist inference (bias adjusted estimate, along with confidence intervals and P values computed using the sample mean ordering [24]) corresponding to the futility stopping boundaries at the jth analysis for $j = 1, 2, 3$, along with the Bayesian predictive power when assuming prior distributions corresponding to the dogmatic optimistic ($\zeta = -0.09$, $\tau = 0.015$), vague optimistic ($\zeta = -0.09$, $\tau = 0.15$), sponsor's consensus ($\zeta = -0.04$, $\tau = 0.04$), dogmatic pessimistic ($\zeta = 0.02$, $\tau = 0.015$), vague pessimistic ($\zeta = 0.02$, $\tau = 0.15$), and noninformative ($\tau = \infty$) priors.

Table 4: Stochastic curtailment measures of the predictive probability of rejecting the null hypothesis at the final analysis. Predictive probabilities are computed under each of the prior distributions defined above.

	Analysis				Prior Distribution On θ					
Design	Time:j	BAM	Crude 95% CI	P Value	Dogm/ Opt	Vague/ Opt	Cons	Dogm/ Pess	Vague/ Pess	Noninf
SymmOBF.4	1	0.077	(0.001, 0.139)	0.977	0.301	0.000	0.003	0.000	0.000	0.000
	2	-0.006	(-0.060, 0.044)	0.401	0.343	0.026	0.033	0.001	0.021	0.023
	3	-0.031	(-0.079, 0.010)	0.067	0.393	0.130	0.129	0.019	0.118	0.124
Futility.8	1	0.038	(-0.037, 0.101)	0.846	0.536	0.011	0.028	0.000	0.007	0.008
	2	-0.017	(-0.071, 0.034)	0.263	0.487	0.070	0.079	0.003	0.057	0.063
	3	-0.035	(-0.082, 0.008)	0.053	0.476	0.184	0.182	0.031	0.169	0.177
Futility.tri	1	0.019	(-0.055, 0.082)	0.697	0.658	0.037	0.066	0.001	0.025	0.028
	2	-0.026	(-0.080, 0.025)	0.161	0.624	0.147	0.155	0.009	0.125	0.135
	3	-0.039	(-0.087, 0.005)	0.040	0.584	0.271	0.266	0.056	0.252	0.262
Pred.Dogm.Opt	1	0.125	(0.049, 0.186)	0.999	0.100	0.000	0.000	0.000	0.000	0.000
	2	0.018	(-0.037, 0.068)	0.737	0.100	0.001	0.002	0.000	0.001	0.001
	3	-0.018	(-0.065, 0.023)	0.185	0.100	0.013	0.014	0.001	0.011	0.012
Pred.Vague.Opt	1	-0.003	(-0.075, 0.061)	0.474	0.755	0.100	0.133	0.003	0.073	0.081
	2	-0.024	(-0.082, 0.028)	0.188	0.549	0.100	0.109	0.005	0.083	0.090
	3	-0.034	(-0.085, 0.011)	0.072	0.343	0.100	0.101	0.014	0.090	0.095
Pred.Consensus	1	0.005	(-0.067, 0.070)	0.569	0.713	0.068	0.100	0.002	0.048	0.054
	2	-0.023	(-0.078, 0.029)	0.206	0.531	0.091	0.100	0.005	0.076	0.082
	3	-0.033	(-0.082, 0.011)	0.074	0.341	0.100	0.100	0.014	0.090	0.095
Pred.Dogm.Pess	1	-0.062	(-0.141, -0.007)	0.013	0.985	0.765	0.727	0.100	0.709	0.742
	2	-0.041	(-0.117, 0.045)	0.154	0.918	0.522	0.528	0.100	0.482	0.501
	3	-0.035	(-0.111, 0.053)	0.207	0.698	0.335	0.344	0.100	0.313	0.323
Pred.Vague.Pess	1	-0.010	(-0.081, 0.054)	0.394	0.788	0.134	0.165	0.004	0.100	0.111
	2	-0.026	(-0.086, 0.025)	0.162	0.582	0.119	0.128	0.007	0.100	0.108
	3	-0.035	(-0.088, 0.010)	0.067	0.364	0.111	0.111	0.016	0.100	0.105
Pred.Noninform	1	-0.008	(-0.079, 0.057)	0.422	0.776	0.121	0.153	0.004	0.090	0.100
	2	-0.026	(-0.085, 0.026)	0.173	0.567	0.110	0.119	0.006	0.092	0.100
	3	-0.035	(-0.087, 0.011)	0.069	0.354	0.105	0.106	0.015	0.095	0.100

From Table 4, we immediately see that, as with conditional power, for any given futility stopping rule there is a wide range of predictive power values as we vary the assumptions about the true treatment effect (i.e., vary the prior distribution for θ). It is evident that seemingly conservative futility thresholds for predictive power can be either markedly more conservative or less conservative than the O'Brien-Fleming boundary, and thus result in sampling plans with greatly varying efficiency. *A priori*, we find it difficult to guess the loss of frequentist power that might result from implementing particular futility rules based on predictive power.

We further note that when used for stochastic curtailment, the Bayesian predictive probability has many of the same foundational issues as the conditional power measures. From a Bayesian perspective, it would make the most sense to base scientific decisions on the posterior probability

of hypotheses and a credible interval for the parameter θ measuring treatment effect. There does not seem to be a good Bayesian rationale for basing decisions about early stopping on predictions of whether a future analysis would meet any particular standard, much less a frequentist standard. While frequentists might be interested in using the Bayesian predictive probability to predict whether statistical significance (or some other decision criterion) would be attained at some future analysis (such an approach does account for the variability in the data at an interim analysis), in this use of predictive probabilities, the foundational issues described for the conditional power are still present: The predictive probabilities do not take into account the tradeoffs between the relative likelihood of particular outcomes under the null and alternative hypotheses.

In order to use stochastic curtailment measures such as conditional power or predictive power as a stopping criterion, it is clear that we must account for the diversity of estimates arising from making different assumptions about the prior distribution of the treatment effect. Should we use the frequentist approach placing all emphasis on a single hypothesis (and if so, which hypothesis), or should we use the Bayesian approach based on a prior distribution for the true treatment effect parameter (and if so, which prior)? Our feeling is that if some single such measure must be used, the Bayesian predictive probability based on a noninformative prior provides a reasonable standard approach. We do note that when using the coarsened Bayes approach, the sensitivity of the predictive probability to the choice of prior can be displayed in contour plots as described in our paper on the Bayesian evaluation of group sequential stopping plans. But even with such a sensitivity analysis, it is not at all clear when a predictive power is sufficiently low to warrant early termination of a study for reasons of futility.

We have found that the best criterion for establishing whether any particular threshold for either conditional power or predictive power is reasonable is based on tradeoffs between efficiency (ASN) and power. Table 5 presents this information for each of the stopping rules considered above. Each of the stopping rules were based on a maximal sample size of 1700 subjects, and hence any introduction of a stopping rule will tend to decrease the power to detect a given alternative relative to the fixed sample test. It was this sort of information that was used by the sponsor and DSMB in the actual sepsis trial as they chose the *Futility.8* stopping rule: Although that stopping rule led to

a slight decrease in power relative to the symmetric O'Brien-Fleming rule (loss of power of 0.007, 0.006, and 0.002 when $\theta = -0.05$, -0.07, and -0.0855, respectively), such a small loss of power was judged acceptable given the approximate 10% gain in average efficiency when the null hypothesis is true. It is worth noting that all of the stopping rules based on stochastic curtailment statistics (conditional power or predictive power) used futility stopping thresholds of 10-20%, though they had markedly different unconditional power and efficiency operating characteristics.

Table 5: Posterior probabilities of hypotheses for trial results corresponding to stopping boundaries of *Futility*.8 stopping rule with four equally spaced analyses after 425, 850, 1275, and 1700 subjects have been accrued to the study. Posterior probabilities are computed based on optimistic, the sponsor's consensus, and pessimistic centering of the priors using three levels of assumed information in the prior. The variability of the likelihood of the data corresponds to the alternative hypothesis: event rates of 0.30 in the control group and 0.23 in the treatment group.

	θ=0		θ=-0.05		θ=-0.07		θ=-0.0855	
Design	Power	ASN	Power	ASN	Power	ASN	Power	ASN
SymmOBF.4	0.025	1099	0.631	1376	0.895	1242	0.975	1099
Futility.8	0.025	987	0.624	1331	0.889	1222	0.972	1088
Futility.tri	0.025	883	0.610	1266	0.876	1187	0.965	1069
Cond.07.20	0.025	1182	0.636	1419	0.899	1260	0.977	1107
Cond.Est.20	0.025	623	0.543	1023	0.797	1024	0.907	964
Cond.Est.10	0.025	677	0.571	1110	0.828	1086	0.928	1006
Cond.LowCI.20	0.025	1033	0.633	1386	0.896	1248	0.975	1102
Pred.Dogm.Opt	0.025	1290	0.638	1450	0.900	1269	0.978	1111
Pred.Vague.Opt	0.025	843	0.616	1281	0.879	1196	0.965	1075
Pred.Consensus	0.025	883	0.621	1306	0.884	1210	0.969	1083
Pred.Dogm.Pess	0.025	489	0.386	687	0.602	726	0.742	727
Pred.Vague.Pess	0.025	803	0.609	1248	0.871	1177	0.960	1064
Pred.Noninform	0.025	818	0.612	1261	0.874	1185	0.962	1068

4 Summary

When clinical trialists are first confronted with the use of a stopping rule, it is quite typical that they worry about the possibility that decisions made at the interim analysis might be different from those which would have been reached if the trial had continued to accrue the full sample size. Indeed, some researchers have suggested that such considerations are at times the ones which should drive the selection of a stopping rule [12, 13]. While we find the operating characteristics discussed in our companion papers [15, 16] much more relevant, the persistence of questions about the futility of continuing a study often dictates that these properties be evaluated. In demonstrating the ways that measures of futility can be evaluated, we highlighted the reasons that we believe they can be less useful (at best) or misleading (at worst). Specifically, we find that 1) the dependence of the

stochastic curtailment calculations on a presumed treatment effect leads to a confusing array of statistics on which a stopping decision might be based, 2) the nonlinear relationship between the conditional or predictive power calculations, the probability of stopping at a given analysis, and the unconditional power functions means that naive users often choose conditional or predictive power thresholds that are suboptimal with respect to their treatment of scientific, ethical, and efficiency issues, 3) conditional or predictive power alone, from a statistical foundations viewpoint, does not address either frequentist or Bayesian optimality criteria, and 4) consideration of tradeoffs between unconditional power and efficiency is sufficient to ensure adequate treatment of futility concerns. In particular, we do not find any particular advantage in the adaptive redesign of a clinical trial based on stochastic curtailment issues. Careful evaluation of stopping rules and information based implementation procedures can handle most of the situations where uncertainty exists about the imprecision of estimates of treatment effects.

Acknowledgements

This research was supported by NIH grant HL69719.

References

[1] Stuart J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191–200, 1977.

[2] Peter C. O'Brien and Thomas R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556, 1979.

[3] John Whitehead and Irene Stratton. Group sequential clinical trials with triangular continuation regions (corr: V39 p1137). *Biometrics*, 39:227–236, 1983.

[4] Scott S. Emerson and Thomas R. Fleming. Symmetric group sequential test designs. *Biometrics*, 45:905–923, 1989.

[5] Sandro Pampallona and Anastasios A. Tsiatis. 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*, 42:19–35, 1994.

[6] Samuel K. Wang and Anastasios A. Tsiatis. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43:193–199, 1987.

[7] John M. Kittelson and Scott S. Emerson. A unifying family of group sequential test designs. *Biometrics*, 55:874–882, 1999.

[8] K. K. Gordan Lan and David L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70:659–663, 1983.

[9] S. Pampallona, A. Tsiatis, and K. Kim. Spending functions for the type i and type ii error probabilities of group sequential tests. 1995.

[10] David J. Spiegelhalter, Laurence S. Freedman, and Mahesh K. B. Parmar. Bayesian approaches to randomized trials (Disc: p387-416). *Journal of the Royal Statistical Society, Series A, General*, 157:357–387, 1994.

[11] Daniel F. Heitjan. Bayesian interim analysis of phase II cancer clinical trials. *Statistics in Medicine*, 16:1791–1802, 1997.

[12] David L. Demets and K. K. Gordon Lan. An overview of sequential methods and their application in clinical trials. *Communications in Statistics, Part A – Theory and Methods [Split from: @J(CommStat)]*, 13:2315–2338, 1984.

[13] Barry R. Davis and Robert J. Hardy. Repeated confidence intervals and prediction intervals using stochastic curtailment. *Communications in Statistics, Part A – Theory and Methods [Split from: @J(CommStat)]*, 21:351–368, 1992.

[14] D. J. Spiegelhalter and L. S. Freedman. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5:1–13, 1986.

[15] Scott S. Emerson, John M. Kittelson, and Daniel L. Gillen. Frequentist evaluation of group sequential designs. *UW Biostatistics Working Paper Series. http://www.bepress.com/uwbiostat*, 2004.

[16] Scott S. Emerson, John M. Kittelson, and Daniel L. Gillen. Bayesian evaluation of group sequential designs. *UW Biostatistics Working Paper Series. http://www.bepress.com/uwbiostat*, 2004.

[17] Irving K. Hwang, Weichung J. Shih, and John S. de Cani. Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine*, 9:1439–1445, 1990.

[18] Myron N. Chang, Irving K. Hwang, and Weichung J. Shih. 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)]*, 27:1323–1339, 1998.

[19] P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A, General*, 132:235–244, 1969.

[20] Scott S. Emerson, John M. Kittelson, and Daniel L. Gillen. On the use of stochastic curtailment in group sequential clinical trials. *UW Biostatistics Working Paper Series. http://www.bepress.com/uwbiostat*, 2004.

[21] K. K. Gordon Lan, Richard Simon, and Max Halperin. Stochastically curtailed tests in long-term clinical trials. *Sequential Analysis*, 1:207–219, 1982.

[22] K. K. Gordon Lan and Janet Wittes. The B-value: A tool for monitoring data. *Biometrics*, 44:579–585, 1988.

[23] Michael A. Proschan and Sally A. Hunsberger. Designed extension of studies based on conditional power. *Biometrics*, 51:1315–1324, 1995.

[24] Scott S. Emerson and Thomas R. Fleming. Parameter estimation following group sequential hypothesis testing. *Biometrika*, 77:875–892, 1990.

[25] John W. Pratt, Howard Raiffa, and Robert Schlaifer. *Introduction to statistical decision theory.* MIT Press, 1995.

[26] Xiaoping Xiong. A class of sequential conditional probability ratio tests. *Journal of the American Statistical Association*, 90:1463–1473, 1995.