STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:3270–3296
Published online 11 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2626

Issues in the use of adaptive clinical trial designs

Scott S. Emerson[*, †, ‡]

Department of Biostatistics, University of Washington, Box 357232, Seattle, Washington 9815, U.S.A.

SUMMARY

Sequential sampling plans are often used 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. Group sequential stopping rules are perhaps the most commonly used approaches, but in recent years, a number of authors have proposed adaptive methods of choosing a stopping rule. In general, such adaptive approaches come at a price of inefficiency (almost always) and clouding of the scientific question (sometimes). In this paper, I review the degree of adaptation possible within the largely prespecified group sequential stopping rules, and discuss the operating characteristics that can be characterized fully prior to collection of the data. I then discuss the greater flexibility possible when using several of the adaptive approaches receiving the greatest attention in the statistical literature and conclude with a discussion of the scientific and statistical issues raised by their use. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: group sequential; stopping rules; sequential sampling

0. INTRODUCTION

Properly conducted clinical trials are crucial to the process of evaluating proposed treatments and preventive strategies with respect to safety, efficacy, and effectiveness.

The difficulties inherent in conducting such experimentation in humans are well known. As experiments, the methods must ensure the scientific and statistical credibility of the evidence used to promote the adoption of some new standard of care. With the involvement of humans as subjects, we must protect the safety of the patients participating in the clinical trial (individual ethics), while ensuring the rapid introduction of optimal treatments (group ethics). Additionally, because economic resources are always limited, we prefer the most efficient designs that satisfy the scientific, statistical, and ethical constraints.

[*] Correspondence to: Scott S. Emerson, Department of Biostatistics, University of Washington, Box 357232, Seattle, Washington 9815, U.S.A.

[†] E-mail: semerson@u.washington.edu

[‡] Professor of Biostatistics.

Received 3 March 2006
Accepted 17 May 2006

Sequential sampling methods have become quite commonplace in the conduct of clinical trials in order to address the above issues. In its broadest sense, sequential sampling refers to the idea that results observed in the data accrued to one point in time are used to decide the future sampling plan. These sequential methods can be divided into two broad (and often overlapping) categories: 'prespecified' sequential sampling plans and 'adaptive' sampling plans.

Historically, much attention has been focused on the prespecified sampling designs (e.g. group sequential designs), while the attractiveness of the latter (e.g. sample size re-estimation, dropping treatment arms, changing primary endpoints) has led to many articles in the recent statistical literature. In this paper, I try to highlight some of the differences between these two approaches in how they adhere to the scientific, statistical, and ethical issues inherent in clinical trials. I believe that the most common reasons that adaptive designs are advocated can truly be addressed through careful planning of the clinical trial, and that such careful planning is more in keeping with the scientific method and statistical efficiency desired in the conduct of a clinical trial.

I first review the scientific setting in which clinical trials are conducted in Section 1. I then discuss in Section 2 the 'prespecified' approach to sequential sampling as typified by group sequential clinical trial designs. In Section 3, I describe some of the 'adaptive' approaches that have been proposed in the statistical and clinical trials literature. I then conclude in Section 4 with a discussion of the relative advantages and disadvantages of the adaptive approach. I note that in this presentation, I restrict attention to issues relating to the primary measure of treatment efficacy. In any real clinical trial, there are of course many secondary and safety endpoints that will need to be considered, but they are outside the scope of this paper.

1. STATISTICAL DESIGN OF A CLINICAL TRIAL

Statistics is about science (in its broadest sense), and science is about proving things to people. Statistical design of a clinical trial is about specifying the structure of the experiment in such a way that the resulting data will present convincing evidence as to whether some proposed standard of care is or is not sufficiently efficacious and safe to warrant either its further study (for phase I and II studies) or its adoption into clinical practice (for phase III studies).

Each clinical trial presents different design issues according to the prior experience of the biomedical community with the disease under study, the patient population of greatest interest, and the general characteristics of the investigational treatment. As clinical investigation of a new treatment progresses through the well-known stages of phase I, II, and III trials, the focus shifts from being primarily directed toward the most preliminary measures of safety to more detailed assessment of safety, efficacy, and effectiveness in both the short and long term. These issues affect the scientific hypotheses that the clinical trial is to address. The statistical design of the clinical trial then specifies the implementation of the experiment.

Statistical design of a clinical trial involves a collaboration between statisticians and researchers from a variety of disciplines: epidemiologists interested in risk factors for disease and disease outcomes, basic scientists interested in mechanisms of both disease and treatment effect, clinical scientists interested in the overall health of the patient, ethicists interested in the welfare of both the patients on the current trial and the greater population who will benefit from the rapid introduction of beneficial treatments, health care economists interested in cost effectiveness of new treatments, financial and marketing specialists interested in the costs of the clinical

DOI: 10.1002/sim

trial and potential profits from sales of a new treatment, operational personnel concerned about the logistics of mounting the trial, and regulatory officials charged with evaluating treatment claims and ensuring the safety of treatments approved for use in a wider population. Each of those collaborators has a slightly different set of optimality criteria defining what makes a 'good' design.

Statisticians play a crucial role in maximizing the precision with which the scientific goals of the study can be met, while finding an acceptable compromise among the often competing goals of the collaborators. To that end, statistical planning for the trial is directed towards:

- ensuring that the clinical trial will discriminate between relevant scientific hypotheses in a manner which leads to scientific and statistical credibility of the eventual results;
- protecting the interests of patients on the trial by terminating a trial when the treatment is unsafe or unethical or when a statistically credible decision can be made about a scientifically important result;
- protecting the economic interests of the sponsor by using efficient designs which will lead to regulatory approval and adoption of the new treatment when estimates of treatment effect reflect important improvements; and
- promoting the rapid discovery of new treatments that are safe and beneficial.

The first statistical task in clinical trial design is to refine the scientific hypotheses into statistical hypotheses that can be statistically evaluated. Statisticians will typically consult on the precise specification of:

- the target population as specified by the inclusion and exclusion criteria;
- the intervention as defined by the dose, administration, frequency, and rules for modification in the face of adverse events (where the latter are especially important in the interpretation of an 'intention to treat' analysis); and
- the clinical measures to be used to evaluate safety, efficacy, and/or effectiveness of the treatment.

Scientific issues are the primary concern in the specification of the above elements of the experimental hypotheses, but statistical input is often sought for quality control and quality assurance in the measurement of variables important in patient accrual, treatment, and follow-up. A greater statistical role, however, is played in the statement of the exact hypotheses to be tested: the statistical hypotheses.

In formulating the statistical hypotheses, statisticians usually take the lead in defining a probability model that describes the variability of outcome measurements both within and across the treatment arms. Then, some distributional summary measure (e.g. the mean, the geometric mean, the median, the probability or odds of exceeding some important threshold, the hazard rate for some event) is chosen as the primary basis for the description of treatment effect. The statistical hypotheses are typically stated as a comparison (e.g. difference or ratio) of the summary measure across treatment arms. Although the statistician provides substantial input into the choice of summary measure, the decision should reflect scientific issues first. That is, the summary measure should first capture that aspect of the distribution that is clinically most important (e.g. it may be more important to lower a blood pressure to within a normal range, than it is to merely show that the average blood pressure is lower). When several measures are of equal importance, then the choice should be based on the likelihood with which the treatment will affect the summary measure (e.g. a treatment that is effective only in an unidentified subset may substantially change the mean blood pressure in a treated group, while having relatively little impact on the median). Only when both these scientific criteria are satisfied are such statistical issues as precision of estimates and statistical power of interest in the selection of the summary measure.

However, no matter what distributional summary measure is selected, ensuring the statistical credibility of the results is of great importance to the clinical trial design. As noted above, the purpose of an experiment is to gain information about differences in the distribution of some clinical endpoint across treatment arms. Statistical precision then relates to our ability to estimate the treatment effect in such a way that we can discriminate between possible values of particular interest. This statistical precision can be based on either standard frequentist measures of precision (confidence intervals, type I error, power) or their Bayesian analogues (credible intervals, posterior probabilities of particular hypotheses).

Notationally, we let $\theta \in \Theta$ be the primary measure of effect. For instance, in a clinical trial directed toward a new treatment for hypertension, θ might be the difference between the mean blood pressure on a control arm and that on the arm receiving the new treatment. The parameter space Θ would then represent all possible values for the true treatment effect. In this notation, a statistical hypothesis is merely some subset of Θ. Common choices in frequentist inference might be a null hypothesis $\Theta_0 = (-\infty, \theta_0)$ and a 'design' alternative hypothesis $\Theta_1 = (\theta_1, \infty)$ for suitable choices of $\theta_0 < \theta_1$. In the blood pressure example used above, $\theta_0 = 0$ might be a typical choice for the null, and θ_1 would ideally be chosen to represent some minimal clinically important improvement in blood pressure.

Ultimately, the data from a clinical trial would be used to produce:

- a 'point' estimate $\hat{\theta}$ of θ,
- an 'interval' estimate (θ_L, θ_U) providing a range of possible values for θ that are in some sense reasonably supported by the data, and
- a probability quantifying the evidence in the data for or against particular 'hypotheses' for the value of θ.

In frequentist inference, the point estimate might be chosen to have optimality properties of minimal bias and/or mean squared error, the interval estimate might be a 95 per cent confidence interval, and the strength of evidence against the null hypothesis might be characterized by a P-value. In Bayesian inference, the corresponding evidence about the treatment effect might be summarized by the mode of the posterior density, a 95 per cent credible interval, and the posterior probability of the null or alternative hypotheses.

As noted above, the purpose of the scientific experiment is to discriminate among particular hypotheses about the treatment effect. In the presence of variable response among individuals on the clinical trial, this then generally translates into designing the study in such a way to guarantee adequate precision for the estimates of treatment effect. Frequentist alternatives for ensuring sufficient precision include:

- specifying the power of the study: ensuring that when the 'design' alternative is true, the trial results will correspond to a suitably low P-value with high probability, or
- specifying the precision of the confidence interval: ensuring that the confidence interval computed at the end of the study will not overlap both the null and 'design' alternative hypotheses.

In either case, we typically find a sample size that will allow the desired precision. I note that these two approaches are identical when specifying a one-sided level $\alpha/2$ test with power $1-\alpha/2$ to

DOI: 10.1002/sim

detect the 'design' alternative and when using a $100(1-\alpha)$ per cent confidence interval to specify the precision of the estimate (e.g. performing a one-sided level 0.025 test with 97.5 per cent power and providing inference based on a 95 per cent confidence interval). Furthermore, for any clinical trial design with specified null hypothesis based on a particular choice of θ_0, we can find a θ_1 which will be discriminated from the null by a confidence interval (i.e. every study has power $1-\alpha/2$ against some alternative).

The ethical and efficiency issues inherent in clinical trials are then addressed through the definition of a data monitoring plan. It is now commonplace (and often demanded by funding and regulatory agencies) that accruing data be periodically examined by an independent data monitoring committee (DMC). Using the results of these interim analyses, the DMC might then recommend early termination of the clinical trial for reasons of patient safety, demonstrated efficacy of the new treatment, or futility (demonstrated lack of sufficient efficacy to warrant further study of the new treatment). If instead the DMC recommends continuation of the clinical trial, it may recommend modifications to the study protocol, which changes might include modifying the scientific and/or statistical hypotheses or any other aspect of the data collection process. In addition to using the results of the accruing data, these recommendations may also be based on new information derived from sources external to the study.

The use of interim data to alter the data collection process thus defines a sequential sampling plan. Most sequential sampling plans can be described using the following notation. We let $X_1, X_2, X_3, \ldots$ denote the potential observations. Up to J interim analyses will be conducted at sample size $N_1, N_2, N_3, \ldots, N_J$ (both J and the N_js may be random variables). We specify a statistic T which is used as the basis for decisions to terminate or modify a clinical trial. Without loss of generality, we assume that the distribution of T tends toward larger values as the treatment effect θ is larger. At the jth interim analysis, we compute statistic $T_j = T(X_1, \ldots, X_{Nj})$ using the first N_j observations. Then, for specified stopping boundaries $a_j \leqslant b_j \leqslant c_j \leqslant d_j$ we might make early decisions according to:

- if $T_j \leqslant a_j$, stop with a decision for inferiority of the new treatment relative to control;
- if $b_j \leqslant T_j \leqslant c_j$, stop with a decision for approximate equivalence between the new treatment and control;
- if $d_j \leqslant T_j$, stop with a decision for superiority of the new treatment relative to control;
- otherwise, continue the study with possible modifications to the protocol and/or sampling scheme (which changes might also be specified according to a more complicated set of modification boundaries).

In the general case of a fully adaptive design, the boundaries used for modifying the sampling plan might be determined at any time up to and including the time of analysis. Under this notation, the study definitely terminates the first time $a_j = b_j$ and $c_j = d_j$. I note that although the notation is given for a clinical trial design with four boundaries, designs with two boundaries are easily effected either by setting $b_j = c_j$ to remove the inner boundaries, setting $a_j = b_j = -\infty$ to remove a lower hypothesis test, or setting $c_j = d_j = \infty$ to remove an upper hypothesis test. Similarly, single boundary designs can be obtained either by setting $a_j = b_j = c_j = -\infty$ or by setting $b_j = c_j = d_j = \infty$ for $j < J$ to remove lower or upper stopping boundaries, respectively, at interim analyses. We let M denote the analysis at which the clinical trial data first meet the stopping criteria, and T_M is the value of the statistic at that stopping time. It is easily shown that when T is sufficient for θ in a fixed sample study, (M, T_M) is the sufficient statistic under a sequential sampling plan.

DOI: 10.1002/sim

Issues related to the sequential sampling plan need to be addressed during the design, monitoring, and analysis stages of the clinical trial. At the design stage, we focus attention on sampling plans which satisfy the desired operating characteristics (e.g. type I error, power, sample size requirements). At the monitoring stage, we must define implementations of the sampling plan which account for errors in the assumptions used during clinical trial design. Then, upon termination of the clinical trial, inference must account for the ramifications of the sequential sampling plan on both the scientific and statistical interpretation of study results.

This paper considers the use of sequential sampling plans under the two broad categories of prespecified stopping guidelines and adaptive procedures. A major focus of the paper is to document the extent to which:

1. the class of prespecified clinical trial designs (as typified by group sequential stopping rules) include a sufficiently broad spectrum of designs as to address the types of issues that arise during clinical trial design,
2. the proper evaluation of a prespecified clinical trial design can anticipate and address the scenarios that have been used to motivate adaptive designs, and hence
3. there is little true need to introduce the loss of efficiency and/or difficulty of scientific interpretation that most often accompanies some of the adaptive procedures described in the statistical literature.

By a prespecified stopping plan, I mean a setting in which prior to collection of the data the clinical trial protocol will specify:

- the scientific and statistical hypotheses of interest;
- the statistical criteria for credible evidence (e.g. frequentist or Bayesian inference, confidence levels to be used in constructing intervals, thresholds for decisions for or against particular hypotheses);
- the rule for determining the maximal statistical information to be accrued (e.g. fix the maximal sample size, fix the power to detect the design alternative, fix the calendar date on which the final analysis is performed);
- the randomization scheme;
- the rule for determining the schedule of analyses (e.g. according to sample size, statistical information, or calendar time);
- the rule for determining the conditions for early stopping (e.g. a boundary shape function for the stopping boundaries for a specified test statistic); and
- the inferential methods that will be used when reporting the results of the clinical trial.

The key aspect that makes a design 'prespecified' in my nomenclature is the ability to provide full frequentist statistical inference. Further details of these components of a prespecified stopping plan are given in Section 2, where I also highlight the issues which prespecification was meant to address, as well as the flexibility possible in such a setting.

Adaptive sampling plans, in contrast, have been described in which the interim trial results might be used to change the scientific and statistical hypotheses of interest (e.g. the target patient population, the dose or administration of treatment, the statistical hypotheses to be discriminated), the statistical criteria for credible evidence, the maximal statistical information, randomization ratios, schedule of analyses, and/or conditions for early stopping. In distinguishing the fully adaptive sampling plans, I focus on the degree to which they do not make clear *a priori* why and how a sampling plan might be modified. It is this lack of specification that contributes to

DOI: 10.1002/sim

the difficulty of frequentist inference with such procedures. Specific examples of such adaptive sampling plans are discussed in Section 3.

2. PRESPECIFIED SAMPLING PLANS

Because a major premise of this paper is that prespecified sampling plans allow sufficient flexibility to address the major issues of clinical trial design, I describe below the spectrum of clinical trial designs and specifications which I include in the category of 'prespecified sampling plans'. In particular, I touch upon the ways in which sample size modification can be effected when using a group sequential design.

As noted previously, a sequential sampling plan can be described by the way in which the observed statistic $T_j = T(X_1, \ldots, X_{N_j})$ computed at the jth analysis might be used to dictate the collection of future data. In a prespecified stopping rule, this is generally effected by describing at the time of *clinical trial design* a rule for choosing the thresholds $a_j \leqslant b_j \leqslant c_j \leqslant d_j$ which are to be used to decide for or against early termination of the study.

There are a number of scientific or statistical criteria which might be used as the basis of the stopping rule:

- Early termination might be motivated by the observation of such an extreme estimate $\hat{\theta}$ of the true treatment effect θ that continuation of the clinical trial would seem unethical (due to estimates suggestive of clinically important benefit or harm) or inefficient (perhaps due to an estimate suggestive of a lack of a clinically important benefit). In such a setting, the test statistic T_j used in the definition of the stopping rule might be an estimate of the treatment effect, such as the maximum likelihood estimate or the mode of the Bayesian posterior density.
- Attention might instead focus on the statistical credibility of an estimated treatment effect. In this setting, early termination might be based upon ruling out important scientific hypotheses in such a way as to minimize the probability of making an erroneous decision (e.g. maintaining a low frequentist type I error rate or a high Bayesian posterior probability). In the frequentist setting, the stopping rule might define stopping thresholds appropriate for a normalized Z statistic, a P-value, or a cumulative type I error 'spent' up to the time of the interim analysis. Bayesian approaches might define thresholds for the Bayesian posterior probability of relevant hypotheses.
- Another justification for early termination of a clinical trial might be based on the (frequentist) conditional or (Bayesian) predictive probability of achieving statistically significant results at the final analysis. In this setting, we might consider whether any future data collected in a continued trial would, when combined with the data observed to date in the trial, lead to a test statistic at the final analysis exceeding the threshold for statistical significance with high probability. In computing the probability distribution for the future data, we might assume a magnitude of treatment effect corresponding to one of the scientific hypotheses (either null or alternative) or the current best estimate of treatment effect (e.g. the maximum likelihood estimate or the mode of the Bayesian posterior density computed at the interim analysis). Furthermore, the distribution of the future data can be based solely on the distribution of the future data under the presumed treatment effect (e.g. a frequentist conditional power approach), or it can also take into account the uncertainty in the estimated treatment effect used in the probability computations (e.g. a Bayesian predictive probability approach).

DOI: 10.1002/sim

There exist in the statistical literature approaches to sequential clinical trial design based on each of the above criteria. However, it is easily shown that there is a 1:1 correspondence among stopping rules defined for any two of the above stopping criteria [1–7]. For instance, the unified family of group sequential designs [8] was defined on the scale of maximum likelihood estimate of treatment effect. That family of group sequential designs was specifically designed to include previously described families of designs defined for the partial sum statistic [9–11], the normalized Z statistic or fixed sample P-value [12–14], and the Bayesian predictive power using a non-informative prior (which in turn is equivalent to the sequential conditional probability ratio test [15]). Furthermore, Pampallona *et al.* [16] and Kittelson and Emerson [8] have discussed the implementation of such designs using error spending functions [17]. Emerson *et al.* [18] discuss in detail the Bayesian interpretation of group sequential stopping rules which might have been initially defined in a frequentist framework. Because of these correspondences, we treat the different criteria for early stopping as alternative *boundary scales* on which any given stopping rule can be expressed.

The unified family of group sequential stopping rules can be used to illustrate the high degree of flexibility possible in the choice of a prespecified stopping rule. At the time of study design, a clinical trialist specifies the rules to be used in computing the early stopping boundaries. Such specification identifies:

- *The hypotheses to be discriminated in the clinical trial.* Common settings consider tests of a null hypothesis against either a one-sided upper alternative, a one-sided lower alternative, or a two-sided alternative. The unified family specifies such alternatives through the use of a continuous parameterization of hypothesis shifts that also allow alternatives intermediate to the one- and two-sided alternatives. Figure 1 displays a spectrum of group sequential designs from the unified family. In this figure, the stopping boundaries are displayed on the scale of the estimated treatment effect (y-axis) *versus* the sample size accrued at each analysis. A value of the test statistic corresponding to any point on the vertical lines would suggest early termination of the clinical trial (the horizontal lines connecting the endpoints are merely to depict the shape of the function describing relationships between successive stopping thresholds). Stopping rules in the leftmost column test one-sided hypotheses of a greater alternative, and stopping rules in the rightmost column test two-sided hypotheses. The centre column of designs, which are intermediate to those more common settings, are sometimes of use in the specification of stopping rules in which both non-inferiority and superiority of a new treatment might be of interest or when the sponsor would like to submit the study results to regulatory agencies as a single pivotal clinical trial.
- *The probability with which a decision might be made to erroneously reject the null hypothesis when it is true* (*the type I error*), *and the probability with which the clinical trial will correctly reject the null hypothesis when some particular alternative(s) are true* (*the power*). In two-sided hypothesis tests, error probabilities could be set differently for testing the upper and lower hypothesis tests.
- *The decisions for which early termination of a clinical trial might be entertained.* Depending upon the scientific, ethical, economic, and regulatory setting, it may be appropriate to terminate a study early only in case of decisions for superiority of a new treatment, only in case of decisions for inferiority, only in case of decisions for approximate equality, only in case of decisions for non-inferiorty, only in case of decisions for non-superiority, or in any combination of these possible decisions. In the unified family, appropriate choice of the boundary shape function parameters allows a clinical trialist to move continuously among these choices, as

DOI: 10.1002/sim

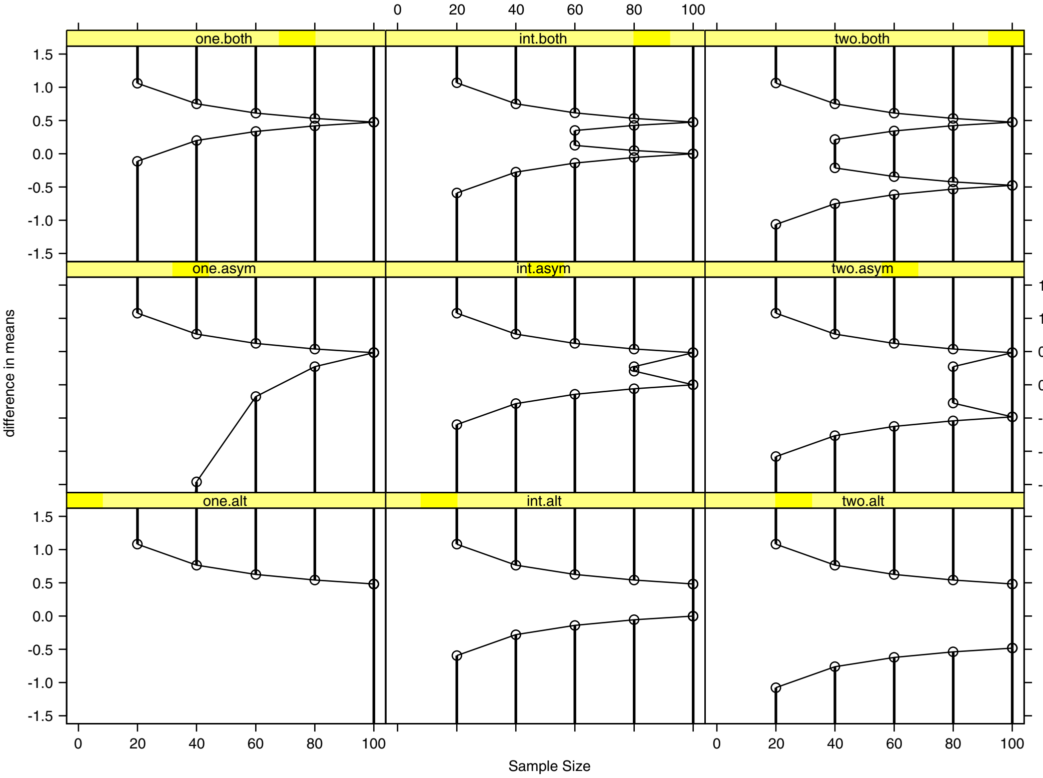

Figure 1. Examples of group sequential designs differing in the type of alternative hypothesis (one-sided, intermediate, and two-sided for the left, centre, and right columns, respectively) and conditions under which early termination is possible (both alternative and null hypotheses symmetrically, more conservative under the null than the alternative, and only under the alternative for the top, middle, and bottom rows, respectively).

depicted in part in Figure 1. The top row of that figure displays group sequential designs which allow early termination for both the null and alternative hypotheses, while the bottom row of that figure displays designs which allow early termination only in the event of a decision for the alternative hypotheses. The middle row demonstrates that such variations of the design can be viewed as merely increasing the early conservatism of the boundary for the null hypothesis until that early stopping boundary is eliminated completely.

- *The degree of early conservatism to be employed in any early termination of the clinical trial.* A wide variety of stopping boundary shapes can be defined which vary in the relative ease with which a clinical trial might be stopped early. In the unified family of designs the early conservatism is controlled through the selection of boundary shape parameters. Figure 2 displays the wide variety of stopping boundaries possible within that family. In this figure, the vertical lines representing the true stopping boundaries have been suppressed, and instead only the lines depicting the varied boundary shape functions have been shown. In Figure 2, all stopping boundaries have the same type I error and the same power to detect the

DOI: 10.1002/sim

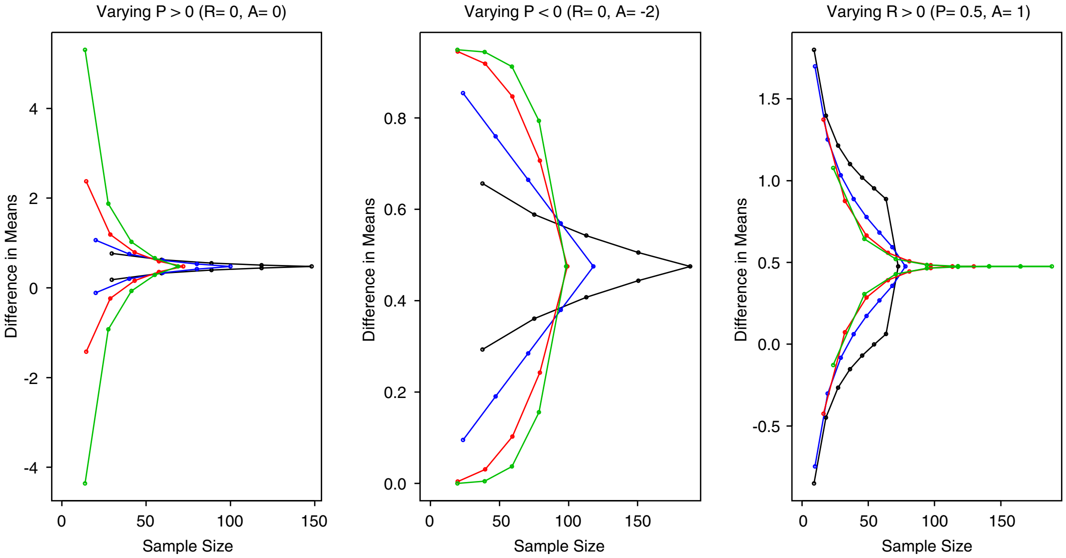

Figure 2. Examples of boundary shape functions possible in the unified family of group sequential designs. Boundaries are plotted on the scale of the maximum likelihood estimate (y-axis) *versus* the sample size at each analysis (scale varies for each panel). All designs have the same type I error and power to detect the same alternative.

same one-sided design alternative. The boundaries differ markedly, however, on the degree of early conservatism, and consequently also differ on the maximal sample size that might be necessary (as shown on the x-axis). This in turn also has impact on the average efficiency of the clinical trial. Stopping rules in the leftmost panel are all in the symmetric family of one-sided designs of Emerson and Fleming [10], which incorporate an extension of the Wang and Tsiatis [14] power family of boundary shape functions, which in turn includes the O'Brien and Fleming [13] and Pocock [12] boundary shape functions as special cases. The other two panels demonstrate other boundary shapes possible under the richer parameterization of the unified family.

A key feature of the above is the prespecification of all these choices prior to the collection and analysis of data. The importance of such prespecification stems from the desire at the end of the trial to provide frequentist inference about the magnitude and statistical precision of estimates of the treatment effect. That is, the most common type of statistical inference used to report the results of clinical trials is frequentist inference involving estimates satisfying frequentist optimality criteria, confidence intervals, and P-values. This frequentist inference is based on the sampling distribution of the estimates of treatment effect, and that sampling distribution is greatly affected by the sampling plan used to gather the data. Figure 3 presents an example of the impact a stopping rule can have on the sampling distribution of a maximum likelihood estimate. Even when using normally distributed data, a two-sided group sequential design with O'Brien–Fleming boundary relationships at four equally spaced analyses yields a null sampling distribution for the estimated treatment effect that is markedly non-normal (upper left panel), and the alternative

DOI: 10.1002/sim

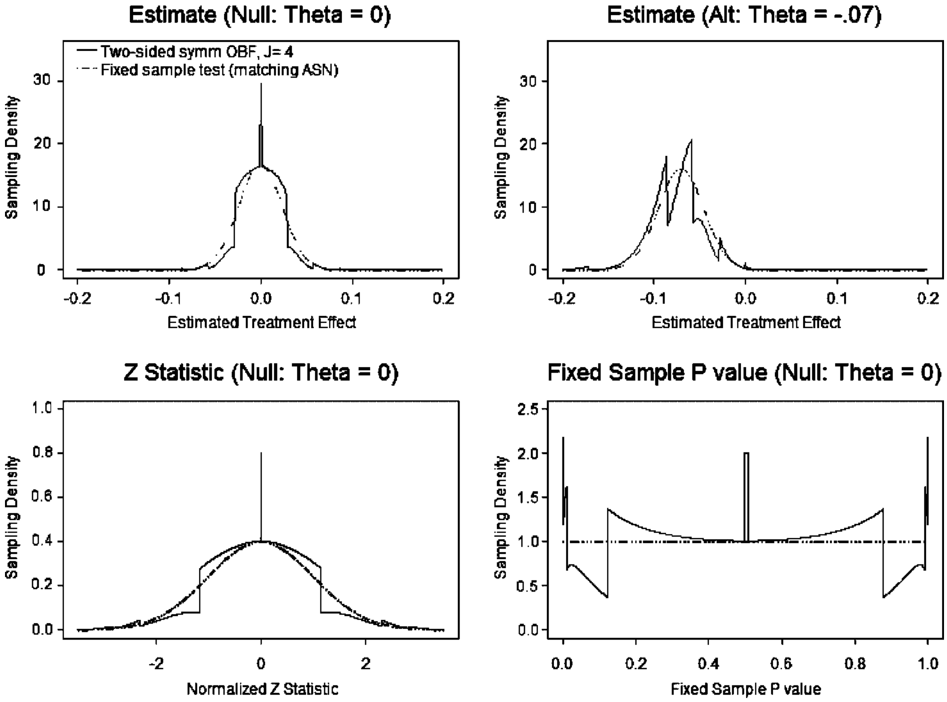

Figure 3. Comparison of sampling distributions from a group sequential test (solid line) and a fixed sample test (broken line) having the same average sample sizes. Sampling distributions are shown for the maximum likelihood estimates under the null (upper left) and alternative (upper right) hypotheses, as well as for the normalized Z statistic (lower left) and fixed sample P-value (lower right) under the null hypothesis.

sampling distribution is not a simple shift of the null distribution (upper right panel). Superimposed on these two panels is the normal distribution expected had a clinical trial been conducted with a single analysis at the same sample size as that expected with the group sequential test. The non-normality of the sampling distribution also means that the normalized Z statistic does not have a standard normal distribution (lower left panel), nor does a 'P-value' computed from that Z statistic truly represent a P-value, because it is not uniformly distributed between 0 and 1 under the null hypothesis (lower right panel).

The fact that a stopping rule affects the distribution of the test statistics and estimates poses no real problem in the presence of a prespecified sampling plan, however. We merely need to use the correct sampling distributions when computing optimal estimates, confidence intervals, or true P-values [19–24]. A number of commercially available statistical software packages are available which will provide correct statistical inference [25–27].

The process of choosing a particular stopping rule is, of course, complicated somewhat by the introduction of the additional dimension of time. Typically a first candidate stopping rule is selected by some subset of the clinical trial collaborative team. Generally, that candidate design will be constrained to have some desired type I error and either prescribed power to detect the 'design alternative' or some maximal sample size. Other operating characteristics of the design

DOI: 10.1002/sim

will then be evaluated, and the design modified to more closely address other optimality criteria. This process is then iterated with an ever widening representation of the disciplines collaborating on the conduct of the clinical trial.

As noted previously, the various collaborators on a clinical trial bring many different areas of expertise and definitions of 'an optimal design'. The types of operating characteristics examined tend to be the same regardless of the exact sampling plan, whether it be fixed sample or involve one or more interim analyses. Emerson *et al.* [5, 18, 28] illustrate the evaluation of clinical trial designs with respect to both frequentist and Bayesian criteria in the context of a case study. The operating characteristics considered include:

- *The frequentist power curve*. This curve describes the probability of rejecting the null hypothesis as a function of the true treatment effect. The value of the power at the null hypothesis is the type I error. Of additional interest might be the power to detect a 'design alternative' of particular scientific importance or the alternative detected with high power meeting some standard of statistical precision (e.g. 97.5 per cent power, if there is interest in describing the hypothesis definitely discriminated from the null by a 95 per cent confidence interval).
- *The sample size requirements*. In a fixed sample study, this represents the number of subjects to be accrued. In a sequential study, however, the sample size accrued at the time of study termination is a random variable. In this latter setting, we might focus on the maximal sample size to be accrued, the expected sample size at study termination (average sample N, or ASN), the median (or some other quantile) of the sample size distribution, or the probability of stopping the trial at each of the interim analyses.
- *The statistical inference which would be reported at each stopping boundary*. A stopping boundary is only appropriate if the clinical trial results will be both scientifically and statistically credible. By examining the outcomes that barely meet the criteria for early termination, the clinical trial collaborators can judge whether the scientific community will in fact be persuaded by the results which might be reported at the end of the study. Such inference will typically include a point estimate, an interval estimate, and some measure indicating the consistency of their results and a hypothesized treatment effect. The value of examining that inference applies equally well to frequentists or Bayesians.
- *The probability of obtaining an economically important estimate of treatment effect*. This analogue of a power curve is of particular importance to industry sponsors of clinical trials.
- *Measures of the 'futility' of continuing the clinical trial.* It is not uncommon for researchers to ask whether decisions which correspond to a candidate stopping rule are predictive of the ultimate decision which would have been reached had the trial continued.

We believe that in evaluating a clinical trial, the 'futility' measures are of less importance owing to their frequent misinterpretation [28]. However, owing to their prominent use in adaptive clinical trials, it is worthwhile to examine in greater detail the correspondence between stopping boundaries on the futility scales and the boundary scales more commonly used to define group sequential stopping rules.

As noted above, the great variety of boundary shapes defined on the scale of the estimate of treatment effect also confers a variety of boundary shapes defined on the other boundary scales. Of particular relevance to a comparison of the more commonly used prespecified stopping rules and the most common implementation of adaptive designs are the correspondences between the estimate boundary scale and the conditional and predictive power scales. For instance, suppose that a one-sided group sequential clinical trial design is such that an observed estimated treatment effect

DOI: 10.1002/sim

of $\hat{\theta} \leqslant a_j$ at the jth analysis were to be judged as so much lacking evidence of effect as to cause early termination of the trial for futility. In the framework of the unified family, such a threshold is defined on the basis of having already rejected (with suitable early conservatism) the hypothesis that the treatment effect is as great as the design alternative. For instance, such 'futility' boundaries with O'Brien–Fleming [13], Pocock [12], or triangular [9] boundary relationships, respectively, would be of the form

$$a_{Oj} = \theta_1 - \frac{N_J}{N_j} G_{\mathrm{Oa}}, \quad a_{Pj} = \theta_1 - \sqrt{\frac{N_J}{N_j}} G_{\mathrm{Pa}}, \quad a_{Tj} = \theta_1 - \left(1 + \frac{N_J}{N_j}\right) G_{\mathrm{Ta}}$$

where θ_1 is the 'design alternative' to be detected with high power, N_j is the sample size accrued by the time of the jth analysis, N_J is the maximal sample size to be accrued by the Jth (final) analysis, and G_{Oa}, G_{Pa}, and G_{Ta} are critical values chosen to achieve the desired type I error and power.

Some authors have instead advocated defining futility boundaries on the basis of the conditional probability of declaring statistical significance at the final analysis. Using that strategy, a clinical trial might be stopped for futility only if

$$\begin{aligned} &\Pr(\hat{\theta}_J \geqslant a_J \mid \hat{\theta}_j = a_j;\ \theta = \theta_*) \leqslant K_j \\ &\quad \Leftrightarrow\ \hat{\theta}_j \leqslant a_j = \frac{N_J}{N_j} a_J - \frac{N_J - N_j}{N_j} \theta_* - \frac{\sqrt{V(N_J - N_j)}}{N_j} \Phi^{-1}(1 - K_j) \end{aligned} \tag{1}$$

where V is the (average) variability contributed by a single sampling unit, and θ_* is the presumed treatment effect under which future data will be sampled. Common values used for θ_* are the null or alternative hypotheses, θ_0 or θ_1, or the current maximum likelihood estimate $\hat{\theta}_j$. The equivalence relationship is based on the typical setting where the estimated treatment effect is presumed to be approximately normally distributed in a fixed sample design. The key point is that a stopping threshold defined for the conditional power scale is just a monotonic transformation of a stopping threshold defined on the scale of the estimated treatment effect. Furthermore, that transformation involves only quantities that are generally presumed independent of the actual treatment effect. (Further clarification of this point is given below in the discussion of sample size re-estimation.)

Similar relationships hold for futility boundaries defined for Bayesian predictive probabilities. For instance, we might use such a scale with a presumed prior distribution $\theta \sim \mathrm{N}(\zeta, \tau^2)$ and take the limit as τ^2 becomes infinite to represent a 'non-informative' prior. In such a setting, a clinical trial might be stopped for futility only if

$$\Pr(\hat{\theta}_J \geqslant a_J \mid \hat{\theta}_j = a_j;\ \theta \sim \mathrm{N}(\zeta, \tau^2 \to \infty)) \leqslant K_j\ \Leftrightarrow\ \hat{\theta}_j \leqslant a_j = a_J - \sqrt{\frac{V(N_J - N_j)}{N_j N_J}} \Phi^{-1}(1 - K_j)$$

where again the equivalence relationship presumes that the estimate of treatment effect would be approximately normally distributed in a fixed sample trial.

DOI: 10.1002/sim

When a fully prespecified stopping rule (see the discussion of exceptions to this setting below) is evaluated with respect to the above evaluation criteria, *prior to collecting any data* we know:

- *The power of the trial to detect arbitrary alternatives*. This includes the alternative for which the study has 97.5 per cent power, which in turn represents the alternative which would definitely be excluded in a trial which fails to reject the null hypothesis.
- *For each interim analysis, the conditions under which the stopping rule would recommend continuation of the study*. Prior to starting the study, we can examine the stopping/continuation boundaries at each analysis on a variety of scales, including:
    - *The scale of the estimated treatment effect*. The DMC can judge whether the results would be compatible with continuation of the trial, and the sponsor can judge the economic importance of such an estimate. (Such judgements, will likely also involve consideration of the statistical precision of those estimates.)
    - *The inference which would be reported if the trial were to stop at each boundary.* The statistical and scientific credibility of the potential results can be evaluated.
    - *The conditional or predictive probability that the trial would have resulted in a different decision had it not stopped.* These 'futility' measures can be computed conditional on having observed results exactly on a stopping boundary and using a variety of hypotheses about the true treatment effect for consideration of any future observations which might be observed.
- *The probability of the study being stopped at each analysis under various alternatives for the true treatment effect.* In conjunction with discussing the conditions (e.g. estimated treatment effect) under which a clinical trial might continue past some sample size, we can discuss the *unconditional* probability that the trial would have such results under each possible value of the true treatment effect.
- *The maximal sample size required under the worst case scenarios, as well as the expected sample size at the time of stopping the trial.*

Further, by comparing several candidate stopping rules, the team collaborating on the clinical trial can consider the relative efficiency of particular designs. For instance, a natural stopping rule to consider in any clinical design setting is that of the fixed sample design—a special case of a group sequential design with a single analysis after all data have been accrued. In such a comparison, the team collaborating on the clinical trial design can consider the *unconditional* power lost by the introduction of a stopping rule when no adjustment is made to the maximal sample size. (I note that if the maximal sample size is appropriately adjusted, there need be no loss of unconditional power through the use of a stopping rule. It is typically the case that introduction of a stopping rule can maintain the type I error under the null and statistical power under the design alternative while greatly decreasing the expected sample size. The worst case sample size, however, will need to be inflated.)

Using the above strategy, the clinical trial design team might be able to make the following statements (given in the context of a trial to compare a new treatment to placebo with respect to 28 day mortality in gram negative sepsis [5]):

- When using a fixed sample level 0.025 one-sided test to detect lower mortality with the new treatment, a sample size of 1700 subjects will provide 90 per cent power to detect the difference between 30 per cent mortality on the placebo arm and 23 per cent mortality

DOI: 10.1002/sim

on the new treatment arm. Such a trial would allow declaration of a statistically significant improvement in survival if the absolute difference between mortality rates was at least 0.042 in favour of the new treatment.

- If we instead use a level 0.025 one-sided symmetric test [10] having Pocock [12] boundary relationships and a maximum of four equally spaced analyses, a maximum of 2340 subjects (an increase of 38 per cent over the fixed sample study) will need to be accrued to have the same power as the fixed sample study. However, it should be noted that:
    - Interim analyses will occur after the accrual of 585, 1170, and 1755 patients.
    - This stopping rule would recommend continuing to the full sample size of 2340 subjects only in the event that the estimated treatment benefit at the third analysis (when 1755 subjects have been accrued) was an observed absolute difference in mortality between 0.036 and 0.049 in favour of the new treatment. An absolute difference in mortality rates of 0.042 at the final analysis (2340 patients) would be judged statistically significant.
    - No matter what the true benefit of treatment, the unconditional probability of continuing past the third analysis is never greater than 16 per cent (which is the case when the true treatment effect is a 0.042 difference in mortality rates), and is generally much lower.
    - The expected sample size at study termination is 1149 subjects when the design alternative (an absolute improvement in mortality of 0.07) is true, 965 subjects when the null hypothesis of no improvement is true, and only 1377 subjects in the worst case of an intermediate hypothesis of 0.042 absolute difference in mortality rates.
    - At the first interim analysis occurring with 292 subjects treated on each arm, a statistically significant result would correspond to a crude observed absolute difference in mortality of 0.084 in favour of the new treatment, which would be reported as a bias adjusted estimate of a 0.076 improvement in mortality (95 per cent CI 0.013–0.130, $P = 0.01$).

To the extent that these (and other) operating characteristics for the symmetric Pocock design satisfy the optimality criteria of the clinical trialists, it is an appropriate stopping rule. If not, then other stopping rules should be explored. (In the actual trial, a very different stopping rule was chosen—see Reference [5]).

In order to be able to determine all of the above operating characteristics exactly, the specification of the stopping rule must include full knowledge of the variability of the individual measurements and the exact schedule of interim analyses (as described by the sample size at each analysis). It is often the case, however, that some aspects of the implementation of a prespecified stopping rule are dependent upon factors that remain unknown up until the time of each interim analysis. Allowing for such flexibility poses no problems providing the estimation of those unknown factors at the time of the interim analyses is independent of the estimated measure of treatment effect. In that setting, we make inference conditional on the sampling plan actually realized. If any modification of the sampling plan is based on ancillary statistics, there will be no introduction of bias or imprecision.

Hence, there is a level of ‘adaptation’ of clinical trial designs that is possible even within the context of my definition of prespecified stopping rules. Below I discuss the way in which a straightforward evaluation of the clinical trial design can proceed when the rules for adaptation are fully prespecified. The two particular cases examined for sample size re-estimation during the conduct of the trial are those due to revised estimates of variability of test statistics or the flexible determination of the number and timing of interim analyses.

Clinical trial design is often based on estimates of the variance of estimates of treatment effect. This can arise, for instance, through uncertainty in the variance contributed by each individual when comparing means using the t test or uncertainty in the control group's event rate when comparing proportions using a chi squared test. Methods have been described in the statistical literature which allow for the implementation of stopping rules in the presence of unknown variability. Such information based monitoring was the recommended implementation of the triangular and double triangular tests [9], and its use in the general setting was reviewed more recently by Mehta and Tsiatis [29]. The appropriateness of such an approach can be motivated by examining the commonly used formulae for sample size estimation.

In a wide variety of statistical models, the maximal number N_J of sampling units needed is estimated by

$$N_J = \frac{\delta^2_{\alpha\beta} V}{\Delta^2}$$

where $1/V$ is the (average) statistical information contributed by a single sampling unit, Δ is the measure of treatment effect in the probability model (usually θ or $\log \theta$) under the alternative hypothesis to be detected with statistical power β in a level α hypothesis test, and $\delta_{\alpha\beta}$ is the design alternative in some standardized version of the test. For instance, in a fixed sample (no interim analyses) two-sided hypothesis test of the difference in means between groups 1 and 2 having equal sample sizes, N_J might be the sample size to be accrued in each group, $\Delta = \theta = \mu_1 - \mu_2$ might be the difference between group means under the design alternative, $V = \sigma_1^2 + \sigma_2^2$ is the sum of the within group variances (statistical information in this case is related to the inverse of the variance of a single observation), and $\delta_{\alpha\beta} = z_{1-\alpha/2} + z_\beta$. In a proportional hazards regression model in which the measure of treatment effect is based on the hazard ratio θ comparing the two equal sized groups, N_J might be the total number of events to be observed in the combined groups, $\Delta = \log \theta$ is the log hazard ratio under the design alternative, $V = 4$ is the inverse of the statistical information contributed by each event, and (as before) $\delta_{\alpha\beta} = z_{1-\alpha/2} + z_\beta$. This same formula can be used in a group sequential test, providing the value of $\delta_{\alpha\beta}$ specific to the selected stopping rule is used.

As noted above, it is often the case that the exact value of V is unknown, in which case the interpretation of our sample size formula needs to be reconsidered. If the sample size is held constant at the value obtained with an estimate of V, we can not know the exact value of Δ in absolute terms. However, we can, for instance, know that our test comparing means does provide the desired power to detect an alternative $\Delta\sqrt{V}$ measured in units of standard deviations of differences between a pair of observations. In this setting of an unmodified sample size, when evaluating the clinical trial design, we need to either express the power curve with respect to this standardized alternative, or to perform a sensitivity analysis of the operating characteristics under different presumptions of the measurement variability.

It is also statistically valid to re-estimate the sample size based on the observed variability at an interim analysis. In this approach, the clinical trial design does not fix the maximal sample size, but instead fixes the maximal statistical information N/V. In this setting, the statistical inference is made conditional on the observed variability, and this does not bias inference about the treatment effect so long as the estimates of V are not correlated with the estimate of treatment effect. In some statistical models, such as the comparison of continuous means and the proportional hazards model, the standard estimate of variability satisfies this criterion. In tests of binomial proportions, however, it is important that the estimate be based on an estimate of the pooled event rate. This issue is addressed again in Section 4.

A second factor that can cause uncertainty in the exact operating characteristics of a sampling plan is that due to flexibly determined number and timing of interim analyses. Frequentist operating characteristics of a sampling plan are specific to the precise schedule of interim analyses as defined by the statistical information available at each analysis. It is common, however, for monitoring of a clinical trial to be scheduled more according to calendar time (e.g. every 6 months) than according to specified increments of statistical information: due to vagaries in the accrual of patients to the trial and/or event rates, the statistical information available on specific calendar dates is not known at the time of clinical trial design.

Methods have been described for the implementation of stopping rules in a manner which allows the type I error and (if desired) type II error probabilities to be maintained when the exact number and timing of interim analyses is not known in advance. In terms of sample size formulas, this merely amounts to an update of the value of $\delta_{\alpha\beta}$, the design alternative in the standardized version of the test. Pampallona *et al.* [16] describe such methods for stopping boundaries defined for type I and II error spending functions, and Burington and Emerson [30] describe the extension of these general methods to stopping boundaries described on arbitrary boundary scales. As before, the statistical validity of these procedures is guaranteed when the factors influencing the number and timing of interim analyses is independent of the estimate of treatment effect. (Proschan *et al.* [31] have explored for some commonly used stopping rules the degree to which the type I error can be affected by data driven changes to the schedule of interim analyses.)

3. ADAPTIVE SAMPLING PLANS

The recent statistical and clinical trial literature includes many papers on methods appropriate for the redesign of a sampling plan during the conduct of a clinical trial. The sampling plan has major impact on the proper computation of frequentist statistical inference, and I have thus focused on the degree to which the sampling plan is known at the time of clinical trial design. Having already discussed in Section 2 those adaptive procedures I consider in keeping with ‘prespecified’ stopping rules, I now turn to those adaptive procedures that present a higher level of scientific and statistical complexity. The major distinction between the adaptive methods described in the previous section and those discussed below is that in this section I consider adaptive designs which use interim estimates of the primary measure of treatment effect in the modification of the sampling plan. It is this ‘data-driven’ aspect that introduces the possibility of bias, imprecision, and/or inefficiency.

The motivation for adaptive re-design of an ongoing clinical trial is varied. Upon observing interim estimates of the treatment effect that are either higher or lower than anticipated, the clinical trialists might want to alter:

- *The scientific hypotheses of interest*. This might include changes to the patient population (e.g. dropping subgroups of patients in which the treatment does not seem to have as great an effect), changes to the treatment (e.g. changing the indications for modifying the treatment in individual patients due to lack of response; changing the dose, administration, or schedule of the treatment for all patients because of observed trends toward better response; dropping one or more treatment arms to focus on those arms with most promising results), or changes to the measure of treatment effect (e.g. changing the method of measuring the primary clinical endpoint to conditions that might favour the treatment group; changing the timeframe over

DOI: 10.1002/sim

which the treatment is evaluated to a timepoint exhibiting more of an effect; shifting focus to a different clinical endpoint entirely).

- *The statistical hypotheses of interest.* This might include changing the hypotheses that are to be statistically discriminated (e.g. re-powering the study to detect an alternative more in keeping with the interim estimate of effect, instead of the alternative originally considered—this will in turn change the value of the alternative for which the study has 97.5 per cent power) or changing the aspect of the distribution used to summarize the population's response to treatment (e.g. changing the summary measure used to compare distributions of response from the mean to, say, the median or proportion exceeding some threshold in order to agree with the maximal difference observed between the interim empirical distributions).
- *The randomization scheme.* For instance, a decision might be made to decrease the probability with which patients are randomized to a treatment arm evidencing less favourable outcomes.
- *The rule for determining the maximal statistical information to be accrued.* This might include a decision to enrol fewer patients than planned if the estimate of treatment effect looks extremely promising or to enrol more patients than planned in order to obtain greater power to declare a less striking effect statistically significant.
- *The rule determining the schedule of analyses.* On the basis of observed results, the frequency of interim analyses might be either increased (e.g. to be able to stop the study earlier for efficacy) or decreased (e.g. to be able to attain a larger sample size and, hence, greater power).
- *The rule for determining the conditions for early stopping.* For instance, there may be impetus to switch to a stopping boundary that is either less or more conservative at the interim analyses, or perhaps to alter which decisions can be made at interim analyses.
- *The inferential methods that will be used when reporting the results of the clinical trial.* For instance, the statistic used to test differences in distributions might be changed (e.g. switching between the chi squared statistic and Fisher's exact test when comparing binomial proportions; switching between the Wald, score, or likelihood ratio tests when using likelihood methods).

I note that there is considerable overlap among several of the above categories. Changing a sample size could be viewed as merely changing a stopping rule. Changing the summary measure used to describe group treatment effects will usually entail a change in the statistic and, hence, inferential methods. Furthermore, I usually argue that the summary measure used to compare distributions across treatment groups should be chosen on scientific, rather than purely statistical, grounds; thus a change in the statistical hypotheses involving a change in the summary measure may also represent a meaningful change in the scientific hypotheses. Similarly, changing the sample size in order to power the study to detect a different alternative is in a decision theoretic sense changing the statistical hypotheses discriminated in the clinical trial.

As noted previously, the major statistical complication introduced by adaptive designs is related to frequentist inference: the sampling distribution of the statistics must be known in order to compute estimates that minimize bias and mean squared error, to produce confidence intervals, and to derive P-values. If the rules for a sampling plan are fully specified in advance, we can compute unconditional frequentist inference, and we can fully evaluate the appropriateness of any particular stopping rule. If some aspect of the sampling plan is modified based on ancillary statistics, we can compute conditional frequentist inference as noted in the previous section. When we are conditioning on statistics whose distribution is independent of the measure of treatment effect and independent of the estimate of treatment effect, there is no bias or imprecision introduced by making conditional rather than unconditional inference. On the other hand, when the sampling

DOI: 10.1002/sim

plan is modified in a *post hoc* manner based on statistics which are related to the measure of treatment effect, frequentist inference which conditions on the realized sampling plan may be both biased and imprecise. It is this setting that therefore requires, and has received, the greatest attention when considering the fully adaptive designs.

Below, I briefly describe four commonly cited examples of approaches to *post hoc* modifications of a sampling plan based on interim estimates of treatment effect. In order to facilitate the comparison of these adaptive designs to the stopping rules described in Section 2, it is useful to augment the notation presented above. In defining the prespecified stopping boundary at the jth analysis, I focused on the statistics based on all data accrued up to the time of that analysis. Thus, at the jth analysis, the estimate of the treatment effect $\hat{\theta}_j = \hat{\theta}(X_1, \ldots, X_{N_j})$ is defined as the estimate based on the cumulative data available at that time, and the normalized Z statistic and the upper one-sided fixed sample P value are transformations of that statistic

$$Z_j = \sqrt{N_j}\frac{(\hat{\theta}_j - \theta_0)}{\sqrt{V}}, \quad P_j = 1 - \Phi(Z_j)$$

However, when adaptively changing a study based on results of interim analysis, it is more convenient to consider the incremental data added between analyses. Hence, I now define N_j^* as the sample size of the groups accrued between the $(j-1)$th and jth analysis, with $N_1^* = N_1$ and $N_j^* = N_j - N_{j-1}$ for $j > 1$. Similarly, the estimate of the treatment effect based on the data accrued between the $(j-1)$th and jth analysis is denoted $\hat{\theta}_j^* = \hat{\theta}(X_{N_{j-1}+1}, \ldots, X_{N_j})$, and I transform this incremental estimate to define the normalized Z statistic and the upper one-sided fixed sample P-value for the data from the jth sampling stage

$$Z_j^* = \sqrt{N_j^*}\frac{(\hat{\theta}_j^* - \theta_0)}{\sqrt{V}}, \quad P_j^* = 1 - \Phi(Z_j^*)$$

Hence, we have the following relationships between the incremental and cumulative statistics:

$$\hat{\theta}_j = \frac{\sum_{k=1}^{j} N_k^* \hat{\theta}_k^*}{N_j}, \quad Z_j = \frac{\sum_{k=1}^{j} \sqrt{N_k^*} Z_k^*}{\sqrt{N_j}}$$

It should be noted that the distribution of the incremental statistics can depend very much on the sampling plan. Suppose that the estimate of treatment effect computed for each group is approximately normally distributed. Then when conditioning on the sample size N_j^* of the jth group,

$$\hat{\theta}_j^* | N_j^* \sim \mathrm{N}\left(\theta, \frac{V}{N_j^*}\right), \quad Z_j^* | N_j^* \sim \mathrm{N}\left(\frac{\theta - \theta_0}{\sqrt{V/N_j^*}}, 1\right), \quad P_j^* | N_j^* \overset{H_0}{\sim} \mathrm{U}(0, 1)$$

The unconditional distributions of the incremental statistics are derived using the standard laws of conditional probabilities, e.g.

$$\Pr(Z_j^* \leqslant z) = \sum_{n=0}^{\infty} \Pr(Z_j^* \leqslant z | N_j^*) \Pr(N_j^* = n)$$

DOI: 10.1002/sim

It is of note that under the null hypothesis, the unconditional distribution of Z_j^* is the standard normal distribution, and P_j^* is unconditionally distributed according to the standard uniform. Furthermore, under the null hypothesis, the incremental statistics $(Z_1^*, \ldots, Z_J^*)$ are totally independent, as are the incremental statistics $(P_1^*, \ldots, P_J^*)$. On the other hand, it is readily seen that even under the null hypothesis, the unconditional distributions of the cumulative statistics Z_j are mixtures of normals when the sample sizes accrued at each stage might vary across replications of the trial. (In the case of a group sequential test in which $N_j^* = 0$ or some prespecified value according to the value of Z_{j-1}^*, the lower left panel of Figure 1 shows the unconditional sampling distribution of Z_M, the value of the normalized statistic when the study terminates, rather than the sampling distribution for the incremental normalized statistic at a particular stage.)

The above distributional considerations form the basis for several approaches to controlling the frequentist type I error when using interim results to modify the sampling plan for a clinical trial. Proschan and Hunsberger [32] consider the case of extending a clinical trial when interim results are suggestive of a treatment effect less than originally hoped for. In that paper, the authors consider a two stage sampling plan in which unplanned modifications of the sampling plan are based on the normalized statistic Z_1 from the first stage, and decisions for or against rejection of the null hypothesis are primarily based on whether the normalized statistic Z_2 (computed from the entire sample) at the second stage exceeds some specified threshold $Z_2 \geqslant a_2^{(Z)}$. The authors demonstrate that when the maximal sample size depends on an estimate of treatment effect from an interim analysis, the unconditional type I error (the experimentwise probability of rejecting the null hypothesis) can be as high as

$$\alpha_{\text{worst}} = 1 - \Phi(a_2^{(Z)}) + \frac{\exp(-(a_2^{(Z)})^2/2)}{4}$$

which is derived by choosing the worst case second stage sample size $N_2^* = N_2^*(z_1)$ such that $N_2^* = \infty$ for $z_1 \leqslant 0$, that $N_2^* = ((a_2^{(Z)}/z_1)^2 - 1)N_1^*$ for $0 \leqslant z_1 \leqslant a_2^{(Z)}$, and $N_2^* = 0$ for $a_2^{(Z)} \leqslant z_1$. Interestingly, when $a_2^{(Z)} = z_{1-\alpha}$ and $\alpha = 0.025$, the maximal type I error of 0.0616 is more than twice the nominal rate, thus indicating that a simple Bonferroni correction would be inadequate in this type of a two stage procedure. Hence, when unplanned sample size modifications are based on interim estimates of the treatment effect, the threshold for statistical significance $a_2^{(Z)}$ at the end of the second stage would need to be chosen to guarantee that a_{worst} in the above equation does not exceed the desired nominal level. In the case when a one-sided level 0.025 test is desired, if the rules for modifying the sample size were not prespecified, the threshold for statistical significance would have to be $a_2^{(Z)} = 2.353$, instead of 1.96.

The above adjustment of the threshold to account for the worst case scenario is necessary whenever the sampling strategy is not fully specified, because frequentist inference demands full knowledge of what would have been done with results that are not observed. If a clinical trialist has not specified actions to be taken under all eventualities, then the adversarial approach to establishing statistical credibility generally requires that the inference be protected against the most unfavourable conditions. Less adjustment, however, is required if a clinical trialist prespecifies constraints on the modification of the sample size in at least some settings. For instance, building on the approach of Gould and Pecore [33], Proschan and Hunsberger consider adaptive sampling plans that: (1) stop the study with a decision not to reject the null hypothesis if $Z_1 \leqslant a_1^{(Z)}$ after observing N_1 subjects, (2) sets the second stage sample size according to an arbitrary function $N_2^* = N_2^*(z_1)$ depending on the results from the first stage, and (3) rejects the null hypothesis only if the normalized statistic Z_2 computed using all the data is greater than some threshold $a_2^{(Z)}$, which might depend upon the interim results z_1 and threshold $a_1^{(Z)}$. If only the threshold $a_1^{(Z)}$ for continuation at the interim analysis is prespecified, the threshold $a_2^{(Z)}$ for statistical significance at the final analysis must account for the possibility that the worst case sample size might be chosen if the study were to continue. On the other hand, those authors also consider the case in which a prespecified conditional error function $A(z_1) = \Pr(\text{rej } H_0 | Z_1 = z_1, \theta = \theta_0)$ satisfies $\int_{-\infty}^{\infty} A(z_1)\phi(z_1)\,\mathrm{d}z_1 = \alpha$. Then, a specification of

$$a_2^{(Z)} = \frac{\sqrt{N_1^*}z_1 + \sqrt{N_2^*}\Phi^{-1}(1 - A(z_1))}{\sqrt{N_1^* + N_2^*}}$$

yields a level α test regardless of how the second stage sample size N_2^* is chosen. The authors then focus on sample size determination in order to satisfy conditional power requirements, thereby potentially altering the hypotheses which can be discriminated by the clinical trial.

More general approaches to adaptive designs are taken by Bauer and Köhne [34] and Fisher [35]. These authors concentrate on the independence of the groups accrued between analyses under the null hypothesis and compute statistics based on either combining P-values (Bauer and Köhne) or normalized Z statistics (Fisher) in such a way as to preserve the experimentwise type I error. The key issue in combining the incremental statistics in an adaptive trial is that the weighting of the contributions from each stage be prespecified. For instance, Bauer and Köhne consider Fisher's method based on the product of J independent P-values. Because $(P_1^*, \ldots, P_J^*)$ are independently distributed standard uniform random variables under the null hypothesis, it follows that under the null $\log P_j^*$ has an exponential distribution with mean 1, and the unweighted sum $\sum_{j=1}^{J} \log P_j^*$ has a gamma distribution $\Gamma(J, 1)$. The experimentwise type I error of an adaptive design can thus be controlled by comparing the unweighted sum of log transformed incremental P-values to the corresponding critical values of the gamma distribution. (As noted by Bauer and Köhne, if the sample size of the jth stage is based on interim estimates of the treatment effect from previous stages, any attempt to weight the incremental P-values by the sample size will result in a different distribution for the weighted sum of log transformed P-values. Furthermore, with any other weighting scheme, the relative weight to be applied to the jth stage must be specified prior to observing data from that stage, or the type I error can be inflated in a manner analogous to that explored by Proschan and Hunsberger [32].) Bauer and Köhne also considered prespecified rules which would allow terminating a study early with decisions either for or against the null hypothesis, where again the null distribution of the test statistic under such a stopping rule is known due to the independence of the incremental P-values under the null hypothesis. Those authors also note the difficulties posed by trying to describe the distribution of the test statistic under alternative hypotheses, because the incremental group P-values do not have a prescribed distribution under the alternative.

Fisher [35] took a similar approach, but instead used a weighted average of the incremental normalized Z statistics from each stage: $T = \sum_{j=1}^{J} a_j Z_j^*$ where $\sum_{j=1}^{J} a_j = 1$ and the value of a_j can be adaptively chosen any time prior to beginning to collect data in the jth stage. Under this adaptive procedure, statistic T has a standard normal distribution under the null hypothesis. (However, because the alternative distribution of the incremental statistics $(Z_1^*, \ldots, Z_J^*)$

does depend on the potentially adaptively chosen sample size, the alternative distribution for T is not well quantified unless the rules for modification of the clinical trial design are totally prespecified.)

Both Bauer and Köhne [34] and Fisher [35] note that their methods can be applied more broadly than sample size estimation. In fact in either case the null distribution of the statistics holds so long as the incremental P value has a null standard uniform distribution and the incremental normalized Z statistic has a null standard normal distribution. This then suggests that adaptive modifications of the treatment, study population, primary endpoint, randomization ratio, etc. can be instituted and a test of some global null hypothesis be performed. We return to the scientific and statistical ramifications of such broad adaptive modification of a clinical trial design in Section 4.

A fourth example of an adaptive clinical trial design is that of the Randomized Play the Winner approach [36]. In this design, the randomization ratio is modified to increase the probability that a future subject would be randomly assigned to the arm with the more favourable results at an interim analysis. For instance, a clinical trial might start with a 1:1 randomization ratio between the new treatment and control arms. Then, after each 'success' on the new treatment arm or 'failure' on the control arm, the odds of a subject being assigned to the new treatment arm might be increased by a factor of r, and after each 'failure' on the new treatment arm or 'success' on the control arm, the odds of a subject being assigned to the new treatment arm might be decreased by a factor of r. As with the previous adaptive schemes, because the sample size for each arm depends on estimates of the treatment effect from an interim analysis, frequentist inference must explicitly adjust for the way in which the observed data might change the sampling distribution for the test statistic. When the rule for changing the randomization ratio is fully prespecified, full frequentist inference is straightforward, though not necessarily simple. When the rule for modifying the randomization ratio is chosen adaptively, a prespecified weighting scheme will again be necessary to ensure the validity of frequentist testing.

4. DISCUSSION: WHAT IS THE COST OF PLANNING NOT TO PLAN?

Adaptive modification of clinical trial design has been the focus of many recent articles in the statistical literature, as well as a number of special conferences such as the one where this paper was delivered. Most often proponents of adaptive designs invoke the uncertainty inherent in clinical trial design as arguments for their methods:

1. We usually select clinical measures of treatment effect from relatively small pilot studies.
2. We often select a target population for the treatment according to a subgroup exhibiting the greatest observed response in a pilot study (and that tends to decrease the relevant sample size in the pilot study even further).
3. We often try to select statistics to summarize the differences between the distributions of outcome across study arms in a way that will maximize the probability of detecting an effect.
4. We often have only imprecise measures of the variability of response within treatment arm.

Clearly, as a larger, confirmatory trial progresses, the additional information gained from a trial might be a better indication of what we should be looking for. However, I believe there are substantial costs associated with a truly adaptive design that need to be carefully considered. As stressed above, the major issues revolve around our ability to provide full inference about the scientifically relevant questions a clinical trial is to address and our ability to produce that inference

DOI: 10.1002/sim

in an appropriately efficient manner. Below I highlight what I believe to be the most salient issues that should be addressed before using an adaptive approach to the conduct of a clinical trial:

- *Adaptively modifying scientific hypotheses.* The methods of Bauer and Köhne [34] and Fisher [35] allow for changes in the target population, some aspects of the treatment, or the primary clinical measure of treatment response while controlling the experimentwise type I error. However, as Bauer and Köhne stress, such modifications to the scientific hypotheses can be problematic. For instance, in a regulatory setting, it is important that both the treatment and the indication for the treatment be clearly identified. Unfortunately, when modifying the scientific hypotheses during the conduct of a trial, we are in essence statistically testing the global null hypothesis that at least one of the treatment regimens considered affects at least one of the proposed clinical outcomes in at least one of the identified target populations. Rejection of that global null hypothesis can hardly be regarded as confirmatory of pilot study results. Furthermore, it is difficult to see how the results of the clinical study can inform clinicians of the proper use of the treatment if it is unclear which exact alternative hypothesis (i.e. which treatment leads to which outcome in which target population) is preferred to the null. (I do note that I do have some sympathy for perhaps allowing some adaptive exclusion of subgroups in a final confirmatory trial, providing the scientific rationale and statistical operating characteristics are carefully considered at the start of the study.)
- *Adaptively modifying statistical hypotheses.* The statistical hypotheses which are discriminated by the clinical trial are materially affected by changes in the summary measure used to compare distributions (e.g. changing from the mean response to median response) or changes in the sample size in order to achieve a different power curve. To the extent that different summary measures might not be of equal scientific importance (e.g. it may be more important to achieve normal glucose levels in diabetes than to merely lower mean glucose levels by some possibly clinically unimportant amount), it can be inappropriate to try to gain more statistical power through a change in the statistic used to compare distributions across treatment arms. On the other hand, when it can be justified by conditions independent of interim estimates of treatment effect, re-powering a study can be quite appropriate. For instance, changes in the availability of alternative or ancillary treatments (including management of toxicities) may change the treatment effect judged clinically important. However, I think that the emphasis placed on re-powering a study based on conditional or predictive power is often misplaced: neither conditional power nor predictive power have good foundations in either frequentist or Bayesian inference [28] and both of those measures can be computed at the time of clinical trial design. It has been my experience that because of the inherently non-linear relationships between conditional power and the commonly used methods of statistical inference, researchers who choose what seem to be intuitively reasonable rules for modifying sample size based on conditional power often end up with a strikingly inefficient or, even worse, unethical sampling plan. This problem can, of course, be avoided by evaluating the sampling plan with respect to all of its operating characteristics, and in such a full evaluation I believe that conditional and predictive power play only relatively minor roles [5, 18].
- *Adaptively modifying randomization ratios.* Play the Winner rules are attractive due to their ability to better address some of the ethical issues related to the individuals participating in the clinical trial: this approach will tend to minimize the number of patients receiving a treatment that is eventually regarded as inferior. However, several caveats are in order. Implementation of such rules can be difficult logistically, especially if treatment response

is only known after some delay. Furthermore, frequentist statistical inference is made much more difficult, because the relative sample sizes on the treatment arms are very much a function of the true treatment effect. Seemingly logical Play the Winner sampling plans have led to great controversy because the ultimate clinical trial results were not convincing to those desiring frequentist inference [37]. A clinical trial which does not provide evidence that is both scientifically and statistically credible can therefore fail to address the group ethics related to speedy discovery of new treatments, even when it is conceptually more beneficial to the patients on the trial. This arises at times due to the fact that the precision of frequentist estimates of treatment effect can be greatly affected by the ratio of group sizes. For instance, when using normally distributed estimates of treatment effects, the most efficient inference about differences across treatment arms is obtained when sample sizes are proportional to the within group standard deviations. However, Play the Winner rules can result in very small sample sizes in one of the treatment arms.

- *Ability to provide full frequentist inference.* The statistical precision of clinical trial results is most often evaluated using frequentist methods. It is the hallmark of such methods that we must be able to describe what would have been done with data that was not actually obtained and at analyses that may never have been performed (this is, in fact, the basis of the criticism of frequentist methods). If hypothesis tests are the only goal, we need only consider the sampling distribution under a null hypothesis. The methods proposed by Proschan and Hunsberger [32], Bauer and Köhne [34], and Fisher [35] all preserved the validity of frequentist hypothesis tests, because the sampling distribution of the respective proposed test statistics was either well characterized, or at least bounded, under the null hypothesis even when decisions about the sampling plan were made 'on the fly' and perhaps only specified for the data actually observed. However, if confidence intervals or information about the bias and mean squared error of treatment effect estimates are desired, such inference requires full prespecification of the sampling plan. Even then, the ease with which frequentist inference can be computed can vary considerably according to the rules for adaptive modification of the clinical trial. For instance, in the setting of adaptively modifying the dose of a study drug, the frequentist estimates of treatment effect at the single dose chosen at an interim analysis may depend on the shape of the entire dose response curve: we would have to consider the probability that other doses might have been chosen for continued sampling, and that will depend upon the response to treatment at those other doses. While one solution to this problem would be to use Bayesian inference exclusively, such an approach does not provide convincing evidence to those members of the audience for the clinical trial results who are satisfied only by evidence satisfying frequentist criteria.
- *Impact on efficiency.* The major motivation for sequential sampling is to reduce the sample size required to address a scientific question ethically and efficiently. When the stopping rules allow adaptive modification of a clinical trial design, it is intuitively obvious that there will be a loss of efficiency: in order to protect against all eventualities, statistical inference must consider the worst case inflation of type I error under adaptive design. Such consideration of the worst case can use the explicit formulation of the upper bound on type I error as given by Proschan and Hunsberger [32], or it can be addressed as proposed by Bauer and Köhne [34] or Fisher [35] through the use of special statistics whose distribution is unaffected by the exact way in which the study design is modified. It is of note that in the latter approaches, neither of the proposed test statistics is based on sufficient statistics. This suggests that there is a loss of efficiency with these approaches even when the rules for modifying the sampling

plan are completely prespecified. This is confirmed by Tsiatis and Mehta [38], who show that there exists a standard group sequential design that is in some sense at least as efficient as any adaptive design. Shi [39], in an unpublished Masters thesis, explores the decreased efficiency of exact inference based on the statistic proposed by Fisher [35] relative to that based on the maximum likelihood estimate when an adaptive sampling plan is prespecified. Of course, it should be noted that statistical efficiency is not the only criterion used to select a stopping rule: the popular O'Brien–Fleming [13] stopping boundaries are not on average the most efficient stopping rule. Instead, that design's popularity is derived from its ability to provide ethical behaviour in the face of extreme benefit of a new treatment, while allowing collection of additional data on secondary endpoints when results are less striking. Hence, the loss of efficiency of the adaptive designs does not automatically exclude their use. Furthermore, Proschan and Hunsberger [32] compare the efficiency of an adaptive design to the sample size requirements of starting an entirely new study to address new statistical hypotheses derived from an interim analysis. That is, they argue that when an interim estimate of treatment effect causes the clinical trialists to re-evaluate the objectives of their clinical trial, it is markedly more efficient to try to salvage the data from the current study than to discard all of that data and start anew.

- *Dissemination of interim results.* When monitoring a clinical trial, it is generally preferred that the study investigators and the general public not be informed of the interim results. Access to the interim estimates of treatment effect and safety is usually limited to a DMC in order to avoid unblinding individual patients and their physicians or otherwise influencing future collection of data. This avoids the possibility that clinicians unschooled in the variability of interim results would over-interpret the findings and modify the types of patients randomized to the study. Were this to happen, serious time trends in the data may result, thus confounding our ability to provide valid scientific and statistical analysis of the results. Adaptive modification of clinical trials presents special issues in this regard. With a prespecified rule for adaptive re-design, it may be the case that the public could gain relatively specific knowledge about the interim estimates of treatment effect. This could happen if the rule for extending the sample size were a 1:1 function of the interim results: knowledge of the final sample size confers knowledge of the estimated treatment effect. (This occurs to a lesser extent with group sequential designs, where the continuation region of a stopping rule places bounds on the interim estimate of treatment effect when a study is not stopped at the interim analysis.) When the rule for modifying the clinical trial is not prespecified, the problem may be even worse: because the modifications were not part of the original clinical trial protocol, Institutional Review Boards and study sponsors would likely need to review the appropriateness of the modifications, and patients may need to sign a revised Informed Consent document. In this process of review, it is difficult to see how the public dissemination of complete results could be avoided.

My bottom line is that adaptive modification of clinical trial designs poses at least as many problems as it is intended to solve. Many of the issues used to motivate adaptive designs can be addressed quite well through careful evaluation of candidate clinical trial designs. Certainly, that evaluation process can include both classic group sequential designs (with the fixed sample design as a special case), as well as sampling plans similar to those described in Section 3. It is my belief that during that process, it will be the rare case that a group sequential stopping rule will not prove adequate.

The really bottom line, then, is perhaps best summed up by Aretha Franklin: 'You better think (think), think about what you're trying to do... .'

REFERENCES

1. Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: a review and commentary. *Statistical Science* 1990; **5**:299–317.
2. Whitehead J. A unified theory for sequential clinical trials. *Statistics in Medicine* 1999; **18**:2271–2286.
3. Betensky RA. Alternative derivations of a rule for early stopping in favor of H0. *The American Statistician* 2000; **54**:35–39.
4. Emerson SS. S+SeqTrial technical overview. *Technical Report*, Insightful Corporation, Seattle, Washington, 2000.
5. Emerson SS, Kittelson JM, Gillen DL. Frequentist evaluation of group sequential designs. *UW Biostatistics Working Paper Series*, http://www.bepress.com/uwbiostat, 2005.
6. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. CRC Press: Boca Raton, FL, 2001.
7. Whitehead J. *The Design and Analysis of Sequential Clinical Trials.* Wiley: New York, 1997.
8. Kittelson JM, Emerson SS. A unifying family of group sequential test designs. *Biometrics* 1999; **55**:874–882.
9. Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions (corr: V39 p1137). *Biometrics* 1983; **39**:227–236.
10. Emerson SS, Fleming TR. Symmetric group sequential test designs. *Biometrics* 1989; **45**:905–923.
11. 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.
12. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–200.
13. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
14. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 1987; **43**:193–199.
15. Xiong X. A class of sequential conditional probability ratio tests. *Journal of the American Statistical Association* 1995; **90**:1463–1473.
16. 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.
17. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
18. Emerson SS, Kittelson JM, Gillen DL. Bayesian evaluation of group sequential designs. *UW Biostatistics Working Paper Series*, http://www.bepress.com/uwbiostat, 2006, in press.
19. 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.
20. Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. *Biometrics* 1984; **40**:797–803.
21. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika* 1986; **73**:573–581.
22. Chang MN. Confidence intervals for a normal mean following a group sequential test. *Biometrics* 1989; **45**: 247–254.
23. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika* 1990; **77**:875–892.
24. Emerson SS, Kittelson JM. A computationally simpler algorithm for an unbiased estimate of a normal mean following a group sequential test. *Biometrics* 1997; **53**:365–369.
25. S+SeqTrial. Insightful Corporation: Seattle, Washington, 2002.
26. PEST (Planning and Evaluation of Sequential Trials). The MPS Research Unit, The University of Reading, Reading, U.K., 2000.
27. EaSt. The Cytel Software Corp.: Cambridge, Massachusetts, 2000.
28. Emerson SS, Kittelson JM, Gillen DL. On the use of stochastic curtailment in group sequential clinical trials. *UW Biostatistics Working Paper Series*, http://www.bepress.com/uwbiostat, 2005.
29. Mehta CR, Tsiatis AA. Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal* 2001; **35**:1095–1112.
30. Burington BE, Emerson SS. Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics* 2003; **59**:770–777.

DOI: 10.1002/sim

31. Proschan MA, Follmann DA, Waclawiw MA. Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics* 1992; **48**:1131–1143.
32. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
33. Gould AL, Pecore VJ. Group sequential methods for clinical trials allowing early acceptance of H0 and incorporating costs. *Biometrika* 1982; **69**:75–80.
34. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041.
35. Fisher LD. Self-designing clinical trials. *Statistics in Medicine* 1998; **17**:1551–1562.
36. Zelen, M. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association* 1969; **64**:131–146.
37. Ware JH. Investigating therapies of potentially great benefit: ECMO. *Statistical Science* 1989; **4**:306–316.
38. Tsiatis AA, Mehta CR. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**:367–378.
39. Shi S. Estimation following self-designing clinical trials. *M.S. Thesis*, Department of Biostatistics, University of Washington, Seattle, Washington, 2003, Unpublished.

DOI: 10.1002/sim