Commentary

Statistics in Medicine

Received 3 March 2011, Accepted 23 March 2011 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4271

Comments on 'Adaptive increase in sample size when interim results are promising: A practical guide with examples'

Scott S. Emerson,[a]*† Gregory P. Levin[a] and Sarah C. Emerson[b]

Keywords: adaptive design; clinical trial; group sequential test; group sequential trial; statistical efficiency

In their paper [1], Drs Mehta and Pocock illustrate the use of a particular approach to revising the maximal sample size of a randomized clinical trial (RCT) by using an interim estimate of the treatment effect. Slightly extending the results of Gao *et al*. [2], the authors define conditions on an adaptive rule such that one can know that the naive statistical hypothesis test that ignores the adaptation is conservative. They then use this knowledge to define an adaptive rule for a clinical trial. In our review of this paper, however, we do not find that such an adaptive rule confers any advantage by the usual criteria for clinical trial design. Rather, we find that the designs proposed in this paper are markedly inferior to alternative designs that the authors do not (but should) consider.

By way of full disclosure, the first author of this commentary provided to the authors a signed referee's report on an earlier version of this manuscript, and that report contained the substance (and most of the detail) of this review. In the comments to the editor accompanying that review, the first author described the dilemma that arose during that review. In essence, the methods described in the manuscript do not seem to us worthy of emulation. But on the other hand, the purpose of case studies in the statistical literature is to present an academic exposition of lessons that can be learned. From years of recreational spelunking, we have noted parallels between research and cave exploration. In both processes, explorers spend their time in the dark exploring the maze of potential leads, most often without a clear idea of where they will end up. Because the overwhelming majority of such leads are dead ends, the most useful companions to have along with you are the ones who will willingly explore the dead ends. However, they rapidly become the least useful companions if they have a tendency to explore the dead ends and then come back and tell you the leads went somewhere. Furthermore, the most important skill that any explorers can have is the ability to recognize when they are back at the beginning, lest they believe that the promising lead took them someplace new and become hopelessly lost. According to these criteria, then, the fact that we would not adopt some approach does not necessarily detract from the importance of a paper to the statistical literature. Instead, a paper's value relates to the extent to which it contributes to our understanding of the methods, which can often be greatly enhanced by identifying dead ends and/or leads that take us back to the beginning.

We note that there are several levels to what could be called the 'recommended approach' in this manuscript. At the topmost level, it can be viewed merely as advocating the use of adaptive designs to assess the likelihood of future futility and efficacy of a clinical trial. But in illustrating that use, the authors seem also to advocate for adaptive methods resulting in sampling distributions that are

[a]*Department of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A.*
[b]*Department of Statistics, Oregon State University, Corvallis, OR, U.S.A.*
**Correspondence to: Scott S. Emerson, Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, U.S.A.*
†*E-mail: semerson@u.washington.edu*

less 'heavy tailed' than analogous fixed sample designs (so that they can safely use naive analytic approaches), and they seem to fall prey to some of the difficulties in interpreting conditional power. We note that if the last two constraints prove to be inadvisable (and we certainly think they are), that does not necessarily establish that use of an adaptive rule in this setting is foolhardy.

However, we do not find that the authors' paper provides sufficient detail for a reader to judge the relative merits of alternative approaches. As we see it, the problem we have faced in the statistical literature about adaptive designs versus group sequential designs is the ever changing optimality criteria. It is like a game of 'Whac-a-mole': As soon as one criticizes one aspect of an adaptive design or group sequential design, the proponents produce other optimality criteria. We think it highly useful to have a paper that clearly identifies *a priori* important optimality criteria, finds the best adaptive design and the best GST design that satisfy those optimality criteria, and then discusses the additional esthetic aspects of each type of design. The substance of this commentary, then, is to provide the greater detail that we would have wanted and to allow readers to examine the extent to which we are correct in our contention that Dr Mehta and Pocock's manuscript is best viewed as an inadequate exploration of a possible 'lead' in the area of adaptive RCT design. There is, of course, some overlap between this commentary and our previously published views on adaptive design [3, 4], and we suspect that a proper evaluation would reveal that even when implemented optimally, the use of an adaptive rule in this setting would constitute, to our minds, a 'dead end.'

Because a major part of our criticism is that the evaluation of the methods is inadequate, we organize this commentary in a manner analogous to that used in evaluation of a new treatment. We first consider the context for the proposed approach by describing the 'unmet need,' consisting of the optimality criteria that one might hope to satisfy with a particular RCT design, and the 'current standard of care', which would include fixed sample and group sequential designs. We then consider the theoretical background and implementation of the proposed adaptive design. In the context of the two examples used by Drs Mehta and Pocock, we then present what we feel is a more complete comparison of the proposed adaptive procedure to the current standard of care. We conclude with some comments regarding the factors that might have led to what we believe to be an unfortunate selection of suboptimal RCT designs.

1. Unmet need: Potential optimality criteria for RCT design

The human experimentation inherent in clinical trials demands careful consideration of science, statistics, ethics, and logistics. The ideal would be to design a clinical trial that addresses an important scientific question in a statistically credible manner so as to ensure that inactive treatments are rarely adopted, active treatments are usually adopted, and adopted treatments are overwhelmingly truly active. Furthermore, all such RCTs would ideally meet these goals while exposing a minimal number of patients to inferior treatments, while not delaying the rapid introduction of new beneficial therapies, and while not engendering excessive costs. It is impossible to meet all of these criteria. Thus, clinical trial design necessarily involves the collaboration of a great many disciplines having often competing goals. The ultimate design selected for a particular RCT will likely not perfectly satisfy any single optimality criterion, instead representing the best balance among several criteria. Common criteria that might be considered in sequential RCT include statistical operating characteristics, the sample size distribution, the precision of inference, study cost, and flexibility of implementation. We briefly consider the aspects of each of these criteria that are relevant to this commentary below. A more complete discussion of the evaluation of clinical trial designs can be found in our previously published tutorials [5, 6].

1.1. Statistical operating characteristics

Control of the type I error rate, the probability of falsely judging an ineffective treatment as effective, is typically viewed as a crucial criterion in confirmatory phase III clinical trials. Even when an acceptable Bayesian approach is used in the design of a trial, regulatory authorities generally want some evaluation of the type I error rate. In a sequential trial, this is generally interpreted to mean control/assessment of the experiment-wise type I error across the repeated analyses. Similarly, we are generally also interested in assessing the study design's power, the probability of accurately judging an effective treatment as effective. As this probability is dependent on the magnitude of the treatment effect, we typically quantify the power of the study under some design alternatives. Ideally, that design alternative would represent

the minimal clinically important difference, but as Drs Mehta and Pocock note, a larger alternative is sometimes chosen owing to logistics (most often) or strong prior evidence of the plausibility of the larger effect (less often).

1.2. Sample size distribution

The sample size treated in an RCT is important in assessing the ethics (how many patients are potentially receiving an inferior treatment), the feasibility (how many patients can we accrue), and, hence, the potential cost (see below) of the clinical trial. When sequential sampling is to be considered, the sample size is a random variable with a distribution depending on the magnitude of the true treatment effect. In that setting, various summary measures of the sample size distribution should be considered.

(1) The average sample size (ASN) is often used as a measure of the average efficiency, with a goal of minimizing the ASN when an experimental therapy is not sufficiently better than the current standard or when one treatment is markedly better than the other. When only considering the primary endpoint, such a measure is arguably the most important summary measure for cooperative groups in which each RCT is followed by another RCT studying the same disease/patient population: minimizing the ASN is tantamount to maximizing the number of new therapies studied (and, hopefully, adopted) in a specified calendar time.
(2) For industrial sponsors who perhaps have interest in only a single proposed therapy, the ASN is still of interest, but the maximal sample size might be of equal concern: the feasibility of mounting the RCT must consider the worst-case scenario. The sponsor cannot as much rely on the averages.
(3) Owing to the stochastic nature of the sequential sampling, the RCT will not always need to accrue the maximal sample size, and hence it is of interest to also consider the probability of exceeding various sample sizes.
(4) Of course, the primary endpoint is not the sole issue in any RCT. In order to be able to assess safety and important secondary outcomes, there is often a minimal number needed to treat to satisfy regulatory requirements. Clearly, this is of greatest concern when a new treatment will ultimately be adopted, and this is the reason for the popularity of 'conservative-early' sequential designs such as the O'Brien–Fleming stopping rule [7]. It is often the case, however, that clinical trialists worry about 'learning curves' with a new treatment and therefore also have a minimal sample size requirement before they would abandon some experimental treatment as insufficiently effective to warrant further study (i.e. before declaring an RCT as 'futile').

Additional issues arise in longitudinal studies having a primary endpoint that can only be measured after some delay. In those settings, clinical trialists might be interested in quantifying the sample size accrued as well as quantifying the statistical information accrued. For instance, when the primary endpoint is based on potentially censored measurements of time to event, we need to consider both the number of subjects accrued as well as the number of observed events required to provide the desired precision. This introduces tradeoffs between subject accrual and calendar time, which may drive decisions about optimality [8]. Similarly, in studies involving repeated measurements made over time on accrued subjects, we can consider the number of accrued subjects and the number of measurements available on each subject over the course of the study. In the latter situation, we note that partial follow-up on some subjects due to too recent accrual (e.g. having only 3 months follow-up on the most recently accrued subjects) still provides information about a primary endpoint that is defined as patient outcomes at a fixed point in time (e.g. 6 months) owing to the missing at random missing data mechanism that obtains in the absence of time trends in accrual patterns [9].

1.3. Accuracy and precision of inference

The results of RCT are not only important for the regulatory decisions for or against approval of the new therapy. Evidence-based medicine dictates that clinicians consider the relative effectiveness of all approved therapies as they choose particular treatment strategies for a specific patient. But to be able to do this, the clinician must have access to estimates of treatment effect measured without bias, as well as quantification of the precision of those estimates and our confidence in the magnitude of effect. Hence, precision of inference that can be achieved with a particular RCT design is an important criterion. When considering a sequential sampling strategy, this translates to understanding the precision of inference available at each possible stopping time. For instance, when concluding a trial with a smaller sample size (and hence perhaps less information about secondary endpoints), it is important to be confident

that the new treatment is not just better than the standard, but that it is markedly better. Hence, we would want, say, a 95 per cent confidence interval to be well above a null hypothesis. Such a criterion argues in favor of early conservatism. We note that many argue for early conservatism on the grounds of extreme confidence (say 99 per cent) that the new treatment is better (perhaps just barely excluding the null), but we believe the true clinical imperative is that we be certain of a major effect.

1.4. Study cost

RCT design can affect study costs in at least four ways.

(1) The obvious expense is the per patient cost associated with patient care, data acquisition, and data management.
(2) To the sponsor, the major costs of a clinical trial are often more dependent upon the elapsed calendar time, which engenders infrastructure costs irrespective of the per patient costs, as well as financing costs (e.g. interest, stockholder return, etc.). Variations in study design, particularly in studies with delayed measurement of treatment outcome, can lead to major differences in calendar time. For instance, in cases where the primary endpoint is based on potentially censored observations of time to event, the statistical precision is more related to the number of observed events than it is to the number of accrued subjects. There are thus tradeoffs possible between accruing fewer subjects who are followed for longer periods of time and accruing more subjects who are followed for shorter periods of time [8].
(3) There are typically costs related to the conduct of formal interim analyses that exceed the costs associated with routine monitoring of the conduct and safety of the RCT. Such costs arise primarily from the understandable desire to have markedly cleaner data on which to base decisions regarding the completion of the study. Hence, the number and timing of interim analyses may represent important optimality criteria. If all other things were equal, a study design that entailed fewer formal interim analyses would likely be judged superior.
(4) Lastly (and certainly with the least ethical justification), sponsors are often concerned with the effect that termination of a study with 'negative' results might have on their stock valuation. As a rule, such is of less concern to the largest pharmaceutical companies that have many ongoing development programs. But to the smaller biotech companies with only a single product in development, the press release associated with the report of disappointing results is often associated with a large drop in the company's book value and/or a decrease in their ability to attract venture capital. (Our mention of this concern is not meant to condone its influence on continuation of studies long after the ultimate decision of an insufficiently beneficial effect has been determined. It is instead just to acknowledge that such motivations can influence a sponsor's choice of RCT design.)

1.5. Flexibility of implementation

RCTs are expensive, complex studies whose conduct is completed long after the planning stage. As a result, the conditions in place at the time of study design often no longer obtain mid-study. Thus, RCT design approaches that can easily accommodate necessary changes without impacting the scientific credibility and statistical precision of the study are greatly to be preferred. Approaches that allow flexibility in the number and timing of analyses [10, 11] or that use blinded data to maintain study power in the face of inaccurate preliminary estimates of measurement variability are readily applied [11–13].

Adaptive clinical trial designs have also been proposed that allow use of unblinded estimates of treatment effect to re-power the RCT, modify the study structure, modify the statistical hypotheses, and/or change the scientific hypotheses while controlling the experiment-wise type I error. We note that while these methods can often accommodate such modifications (at least in terms of control of the type I error) even when they were not specified at the time of study design, both the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMEA) have issued draft guidances that would require that all methods used in 'adequate and well-controlled' confirmatory trials be prespecified.

2. Current standards: fixed and group sequential designs

Drs Mehta and Pocock consider the setting of a phase III confirmatory trial, where the current standard approach would be a fixed sample or group sequential design. We adopt the same notation in which

totally independent observations $X_i \sim \mathcal{N}(\mu_c + \delta T_i, \sigma^2)$, where indicator variable $T_i = 0$ or 1 according to whether the ith subject is in the control or treatment groups, respectively. In a one-sided level α fixed sample design, data on n subjects are gathered, and a standardized difference in means Z_n is compared with the upper critical value $z_{1-\alpha}$ from a standard normal distribution. In order to satisfy any regulatory requirements for prespecification of design, it is sufficient to either specify n exactly, or to specify how n might depend on interim blinded estimates of measurement variability. In a one-sided group sequential design, analyses of the data accrued to date would be potentially performed at sample sizes $n_1, n_2, \ldots, n_J$. At the jth analysis, a standardized difference in means Z_{n_j} based on the first n_j observations is compared with the critical values a_j and d_j: If $Z_j \leqslant a_j$, the trial is stopped with a decision to not reject the null hypothesis $H_0 : \delta = 0$, if $Z_j \geqslant d_j$, the trial is stopped with a decision to reject the null hypothesis in favor of alternative $H_1 : \delta > 0$, and if $a_j < Z_j < d_j$, the trial is continued to observe the next $\tilde{n}_{j+1} = n_{j+1} - n_j$ subjects. A choice of $a_J = d_J$ ensures termination at the Jth analysis. A group sequential sufficient statistic is (M, Z_M) where M is the smallest value of $j \in \{1, 2, \ldots, J\}$ satisfying $Z_j \notin (a_j, d_j)$. The critical values a_j and d_j are typically chosen for $j = 1, \ldots, J$ in such a way as to have $Pr(Z_M \geqslant d_M | H_0) = \alpha$. We note that fixed sample designs correspond to a group sequential designs with $J = 1$ and $d_J = z_{1-\alpha}$. Again, regulatory requirements for prespecification of design can be met by either prespecifying the number and timing of analyses exactly or prespecifying the rule by which the schedule of analyses might depend on calendar time and/or interim blinded estimates of measurement variability independent of the estimate of treatment effect.

There are of course an infinite number of fixed sample and group sequential designs that can be considered as the sample size(s), schedule of analyses, and group sequential critical values (stopping boundary relationships) are varied. Depending on optimality criteria, a clinical trialist might want to restrict attention to only those group sequential designs meeting one or more of specified criteria, such as a specified type I error α, a specified power $Pwr(\delta_1) = 1 - \beta$ at a specific 'design alternative' $\delta = \delta_1$, a maximal number of interim analyses J, a maximal sample size n_J, a minimal sample size n_1 at the first interim analysis, only futility boundaries (so $d_j = \infty$ for $j < J$), only efficacy boundaries (so $a_j = -\infty$ for $j < J$), both futility and efficacy boundaries, or critical values at the first analysis that meet some predetermined need for early conservatism.

We note that the dimensionality of design decisions in sequential sampling plans is extremely high and the effect of particular design parameters on optimality criteria is highly nonlinear. It is easy to choose a sampling plan that does not provide the gains in efficiency and ethical treatment of patients one was hoping for. Hence iterative evaluation of candidate clinical trial designs is extremely important [5, 6], and a natural comparison is between a group sequential design and alternative fixed sample designs. When comparing fixed sample designs with a particular sequential sampling plan, useful comparisons would include

- A fixed design with the same maximal sample size (so $n = n_J$). This design will have the greatest statistical power of any group sequential or fixed sample designs having that same maximal sample size, so such a comparison allows consideration of the maximal loss of power from the use of a sequential sampling plan.
- A fixed design having $n = n_{\text{Pwr}}$ such that it provides the same type I error and the same power under the design alternative $\delta = \delta_1$ as the sequential sampling plan. Comparisons can then be made regarding
  - the inflation in maximal sample size n_J / n_{Pwr} required to accommodate the sequential sampling,
  - the probability $Pr(n_M \geqslant p n_{\text{Pwr}} | \delta)$ that the sequential sampling plan would continue past some specified inflation factor p of n_{Pwr} as a function of the true treatment effect δ, and
  - the relative average efficiency $n_J / ASN(\delta)$ of the sequential sampling plan relative to the fixed sample design having the same power. We note that if the sequential sampling plan is providing any advantage in ethics or efficiency, we would typically hope that the relative average efficiency would be greater than 1 for δ close to 0 when futility bounds are being used and for δ close to δ_1 when efficacy bounds are being used.

3. Drs Mehta and Pocock's new approach

In their paper, Drs Mehta and Pocock advocate the use of a particular adaptive clinical trial design to implement preliminary decisions about the likelihood of futility or efficacy in a phase III confirmatory

clinical trial. As with reports of investigations of new treatments, we find it useful to first discuss some of the background for the proposed method, and then to discuss the details specific to the proposal.

3.1. Background on the relevant adaptive clinical trial design methods

Previous authors [14–18] described methods in which the maximal sample size of a study might be modified at the penultimate of J interim analyses. Suppose, we have a group sequential clinical trial in which up to J analyses will be performed. Using the authors' setting and notation in which n_j and Z_j refer to the cumulative data, $\tilde{n}_j$ and $\tilde{Z}_j$ refer to incremental data accrued between the $(j-1)$th and jth analyses, and adaptive revisions of the sample size and the test statistics are denoted with asterisks (so n_J^*, $\tilde{n}_J^*$, Z_J^*, and $\tilde{Z}_J^*$), we can consider an adaptive modification $n_J^*(Z_{j-1})$ of some prespecified n_J. Regardless of the way $\tilde{n}_J$ is chosen, the (sub)density for Z_{J-1} will be based on the critical values (a_j, d_j), $j=1,\ldots,J-1$ as described for group sequential densities by Armitage *et al.* [19] and the conditional distribution for the incremental statistic at the last analysis will satisfy $\tilde{Z}_J|\tilde{n}_J \sim \mathcal{N}(\delta\sqrt{\tilde{n}_J/V}, 1)$, where $V=4\sigma^2$ based on 1:1 randomization.

When the sample size at the Jth stage depends on $\tilde{Z}_{J-1}$ (so $\tilde{n}_J^*=\tilde{n}_J^*(\tilde{Z}_{J-1})$), it is easily shown that the distribution for Z_J^* is actually based on a mixture of (possibly truncated) normals. In this setting, Z_J^* will not satisfy a location shift model. Intuitively, this can be expected due to the variability of n_J^* as a function of δ. This result has important ramifications related to the authors' goal of allowing valid use of fixed sample methods that ignore the adaptive rule. Even if the type I error is not inflated when using the naive fixed sample analysis, in the presence of an adaptive stopping rule, coverage of a confidence interval can still be anti-conservative.

Furthermore, regardless of the way n_J^* is chosen, it is easily shown that (n_J^*, Z_J^*) is a sufficient statistic for δ and, if $J>1$ with (a_j, d_j) a proper subset of $(-\infty, \infty)$ for some $j<J$, that it is minimal sufficient. Hence, the principle of sufficiency would argue for the use of inference that depends on the data only through (n_J^*, Z_J^*).

Much of the work with adaptive designs has been focussed on their use when rules for adaptation are not prespecified. In that setting, it can be shown that an experiment-wise type I error is preserved if one either appropriately reweights the observations from the last stage or, equivalently, appropriately modifies the critical value d_J^* at the Jth analysis [14–17]. In either case, the resulting inference violates the sufficiency principle. We note that the last sentence in the Introduction of Gao *et al.* [2] (which paper provides the foundation for the current manuscript) is incorrect in that regard. (For the sake of completeness, we note that if $n_J^*(\tilde{Z}_{J-1})$ is an invertible function, there will not be a violation of the sufficiency principle; however, we have seen no authors propose such an adaptation even theoretically. Owing to the discrete nature of subject accrual, it is in fact impossible to have an invertible function, though with a large enough sample size, the discreteness may have less of an impact.)

The violation of the sufficiency principle argues that inference will likely not be as efficient as it could have been, though the degree of inefficiency might be slight depending on the adaptation used (see below). However, even if one were to base inference on the exact distribution of a function of the sufficient statistic, the sampling rule may be markedly inefficient. In order to gain intuition about the inefficiency of many of the proposed methods, we find it extremely useful to carefully examine the results of Proschan and Hunsberger [15].

As the authors note, almost all of the voluminous literature on adaptive methods focuses on control of the experiment-wise type I error in the presence of adaptation using an interim estimate of treatment effect. There are of course an infinite number of ways that type I error can be controlled, including the use of only the first-stage data, using only the second-stage data, using R.A. Fisher's method for combining independent P values [14], using L. Fisher's variance spending functions to compute weighted averages across stages [16], using weighted averages of the incremental statistics [17] or (in the absurd case) ignoring the data from the trial altogether and basing inference on an independently drawn standard uniform random variable. The true key issue is how to address the type I error while attaining high power to detect relevant alternatives.

As the authors further note, however, the preponderance of the literature on adaptive methods focuses on conditional power. This was, to our mind, best laid out by Proschan and Hunsberger [15], who considered the case $J=2$ and described a general approach for controlling the experiment-wise type I

error based on a conditional error function $A(z) \in [0, 1]$ that is increasing on $(-\infty, \infty)$ and that satisfies

$$\int_{-\infty}^{\infty} A(z)\phi(z)\,dz = \alpha$$

Those authors noted that for a prespecified first-stage sample size $\tilde{n}_1$ and conditional error function $A(z)$, an experiment-wise level α test can be obtained by appropriately choosing second-stage sample size $\tilde{n}_2^*(\tilde{z}_1)$ and critical value $c(\tilde{n}_2^*, \tilde{z}_1)$ such that

$$Pr_{\delta=0}(Z_2^* \geqslant c(\tilde{n}_2^*, \tilde{z}_1) | \tilde{n}_2^*(\tilde{z}_1)) = A(\tilde{z}_1)$$

(Note that Proschan and Hunsberger did not restrict attention to inference that satisfies the Sufficiency Principle. We note also that this approach of Proschan and Hunsberger is sufficient for the control of the type I error, but not necessary. In particular, we can achieve control of the type I error with a non-increasing conditional error function.)

Putting the above approach into practice is difficult owing to the dimensionality of specifying the conditional error function $A(\tilde{z}_1)$. An approach adopted by many researchers is to base the conditional error function and critical values on a fixed sample test that has a prespecified value of $\tilde{n}_2$. That is, if we fix the values of $n_1 < n_2$, we can use the uniformly most powerful test based on Z_2 in a non-adaptive setting to define conditional error function

$$A(z) = Pr_{\delta=0}(Z_2 \geqslant \Phi^{-1}(1-\alpha) | \tilde{Z}_1 = z, \tilde{n}_2 = n_2 - n_1)$$

and then find the critical value $c(\tilde{n}_2^*, \tilde{Z}_1)$ for a completely arbitrary value of $\tilde{n}_2^*$ such that

$$Pr_{\delta=0}(Z_2 \geqslant c(\tilde{n}_2^*, \tilde{z}_1) | \tilde{Z}_1 = z_1, \tilde{n}_2 = \tilde{n}_2^*) = A(z_1)$$

Again, it should be noted that this strategy is sufficient, but not necessary, to control the type I error. And though it is not based on a sufficient statistic when the function $\tilde{n}_2^*(\tilde{Z}_1)$ is known, it has a perceived advantage of flexibility: A user need not specify how $\tilde{n}_2^*$ is chosen. As noted above, however, this perceived advantage is of diminished importance in a regulatory setting. Both the FDA and EMEA have draft guidances requiring that any adaptive strategy be totally prespecified.

Now the most interesting result (at least to us) of Proschan and Hunsberger [15] was their quantification of the worst-case type I error one can achieve with a two-stage adaptive design. That is, (in the current manuscript's notation) they describe the function $\tilde{n}_2^*(\tilde{Z}_1)$ that will maximize $Pr_{\delta=0}(Z_2 \geqslant c)$ for a specified value of c. Interestingly, using that worst-case adaptation in a two-stage design, one can achieve $Pr_{\delta=0}(Z_2 \geqslant 1.96) = 0.0616$, which is more than double the 'nominal' type I error of 0.025 (We note that this more than doubling of the type I error clearly cannot occur for every choice of α, but it does for this very important case).

At a first glance, this is quite paradoxical: The user only performed two analyses of the data, yet a Bonferroni correction based on two analyses does not protect the type I error. The paradox can perhaps be resolved; however, by considering the stochastic nature of the two analyses that were actually performed. In fact, the user considered an extremely large number of analyses (one at each possible sample size), but was able to avoid performing some of those analyses based on a reasonably accurate prediction that significant results would probably not be attained. This 'imputation' of results allowed the user to only actually conduct the analysis that stood the greatest chance of being significant. However, the ultimate type I error is seemingly affected by many more analyses than were actually performed. The end result is an inefficient design owing to the imprecision of the 'imputation.' A user will not get the efficiency of a fully sequential design.

Based on the above, the inefficiency of a non-prespecified adaptation using an interim estimate of the treatment effect is apparent: A user would have to protect himself against the worst-case scenario. Hence, most authors consider restrictions on the adaptation similar to those used by the current manuscript: The sample size is only modified if the stage 1 results are in some promising range, though as we will discuss below, there are many nuances to what should be considered 'promising.' And the above results about the most general approach should serve as warning that adaptive designs need to be evaluated carefully lest they be too compromised by the worst-case scenarios.

A critical value of an adaptive test can be found to protect against inflation of the type I error according to the prespecified bounds on adaptation. The methods for ensuring control of the type I error ultimately depend on a formulation of the power curve for an adaptive test by integrating out the dependence on

the first-stage statistic. This can be evaluated for a prespecified function $\tilde{n}_2^*(\tilde{Z}_1)$, because we can derive the power curve for a test based on the sufficient statistic $(n_2^*, Z_2^* = (\sqrt{\tilde{n}_1}\tilde{Z}_1 + \sqrt{\tilde{n}_2^*}\tilde{Z}_2^*)/\sqrt{\tilde{n}_1 + \tilde{n}_2^*})$ using

$$Pr_\delta(Z_2^* \geqslant c(n_2^*)) = \int Pr_\delta(Z_2^* \geqslant c(n_2^*) | \tilde{Z}_1 = \tilde{z}_1, \tilde{n}_2^*(\tilde{z}_1)) \phi\left(z_1 - \sqrt{\frac{\tilde{n}_1}{V}}\delta\right) \mathrm{d}z_1$$

for arbitrary prespecified critical value function $c(n_2^*)$. Such an approach can be implemented using standard group sequential software through a straightforward generalization of the methods described in the Appendix of Emerson *et al*. [8].

Most authors have focused on the imprecise methods based on reweighted statistics, whose statistics requires information beyond a minimal sufficient statistic. Depending on the restrictions on the adaptation and the conditional error function used to determine critical values, the resulting inference might be negligibly or markedly different from what would be obtained when using the sufficient statistic. Using the exact sampling distribution for the sufficient statistic, we have investigated the authors' comment that 'It has yet to be demonstrated that there is any appreciable loss of efficiency due to the use of the weighted statistic in these settings.' We have found that if one uses the conditional error function from a relatively efficient one-sided symmetric group sequential test with Pocock [20] boundary relationships to control the type I error in an efficient adaptation, there is little difference between the power curve for inference based on the sufficient statistic and that for inference based on the reweighted statistic. However, if the relatively inefficient (on average) O'Brien–Fleming boundary relationships are used to adapt to a more efficient design, the power using the sufficient statistic of 0.975 is reduced to 0.898 using the weighted statistic.

3.2. The recommended approach

Drs Mehta and Pocock advocate an approach to sequential sampling that entails three separate aspects, though they are somewhat interrelated in their implementation. Below, we therefore separately discuss their general advocacy of an adaptive approach, their definition of the 'promising region,' and their rule for sample size modification.

3.2.1. Use of adaptive designs. Drs Mehta and Pocock propose the use of a rather straightforward adaptive approach based on repowering a study at a penultimate analysis, and we might therefore expect that we could obtain any benefit that such designs afford. However, the true costs and benefits of that adaptive approach in a general setting remain unclear, in part due to the ever changing optimality criteria invoked by statistical researchers in the literature. The authors mention the previous literature that demonstrates that there exists a group sequential design that is more efficient than any adaptive design [21, 22], but they criticize that prior work as being irrelevant to most RCT settings. While we do not agree that it is irrelevant, we do acknowledge that there may be some optimality criteria that are not held constant in those comparisons. It is thus useful to examine some general behavior of the adaptive versus group sequential designs.

First, the authors' contention that a major advantage of the adaptive design in the phase III confirmatory trial is that it avoids an upfront commitment of large sample sizes is clearly false, at least in the current regulatory environment. Given the need to prespecify adaptive designs, a sponsor using one of the authors' proposed designs would need to commit to increasing the sample size twofold in some settings. The authors could try to put a spin on this requirement, by claiming that: 'The sample size is n_2, unless interim results observed at sample size n_1 correspond to an estimated treatment effect that is between 61 per cent and 92 per cent of the anticipated treatment effect, in which case a higher sample size will be used. The sample size will go as high as $2n_2$ when the interim estimate of the treatment effect is between 61 per cent and 67 per cent of the design alternative. If the true treatment effect is 20 per cent smaller than planned, the statistical power is 65.8 per cent, and we would expect the sample size at study conclusion to be $1.12n_2$.' This is no different than what can be said in the setting of group sequential trials, when endpoints are immediate. In that case, we can consider using a similarly powered symmetric group sequential design with first analysis at n_2. When O'Brien–Fleming stopping boundaries are used, we would be able to say: 'The sample size is n_2, unless interim results observed at sample size n_2 correspond to an estimated treatment effect that is between 56 per cent and 77 per cent of the anticipated treatment effect, in which case the sample size will be $1.16n_2$. If the true treatment effect is 20 per cent smaller than planned, the statistical power is 65.8 per cent, and we

would expect the sample size at study conclusion to be $1.03n_2$.' Using a Pocock boundary relationship would change the sentences to read: 'The sample size is n_2, unless interim results observed at sample size n_2 correspond to an estimated treatment effect that is between 59 per cent and 73 per cent of the anticipated treatment effect, in which case the sample size will be $1.23n_2$. If the true treatment effect is 20 per cent smaller than planned, the statistical power is 65.8 per cent, and we would expect the sample size at study conclusion to be $1.03n_2$.'

From the above, it seems clear that the authors are incorrect in their statements that the efficiency of group sequential alternatives to adaptive designs 'produce appreciable efficiency gains only if there are no over-runs, a large number of interim analyses, a large up-front sample size commitment, and aggressive early-stopping boundaries'. In the above examples, we used only two interim analyses, we had a lower up-front sample size commitment, and, owing to the spacing of the analyses, we achieved virtually the same results with either the O'Brien–Fleming boundary relationships typically chosen for their conservatism, or the Pocock boundary relationships that tend to exhibit greater average efficiency. (We address the impact of overruns later in our discussion of the authors' results for Example 1.)

Of course, 'one swallow does not a summer make,' and the above comparisons are specific to the adaptive rules espoused by the authors (see below). We have pursued some further investigations into comparisons that we find enlightening, though not truly definitive. For instance, if 1000 subjects would provide 97.5 per cent power to detect a treatment effect in a level 0.025 one-sided test, then the optimal (in terms of ASN under the design alternative or null) symmetric group sequential test with two analyses would have analyses performed with 500 and 1180 observations and would use a boundary relationship close to that of the Pocock relationships (unified family [23] parameter $P=0.542$, while a Pocock design has $P=0.5$ and O'Brien–Fleming relationship has $P=1.0$). The trial would continue to the second analysis if the conditional power based on the MLE were between 4.9 and 95.1 per cent, or, equivalently, if the conditional power based on the design alternative were between 81.8 and 99.0 per cent. Under either the null or alternative hypothesis, the ASN is 685.4. If we consider an adaptive modification of the sample size at the final analysis, we find that the optimal (again, by ASN under the design alternative or null) adaptation that only considers two alternative final sample sizes would have a maximal sample size of 1240 and an ASN of 683.1. Additional, negligible improvements in ASN were observed up to designs that allowed adaptation to five alternative samples sizes, in which case the maximal sample size was 1260 and the ASN was 682.4. Increasing the number of alternative sample sizes at the final analysis beyond 5 did not affect the ASN in the fourth significant digit (i.e. the ASN continued to be 682.4).

Adaptation in this setting provided only very minor improvements in average efficiency and degradation in the maximal sample size required. Similarly, dubious benefits from adaptive designs were seen in a setting in which constraints were placed on the minimal acceptable sample size at study termination [24]. By way of comparison, if we consider a group sequential test that has three analyses, one at each of the sample sizes used in the optimal adaptive design that allowed two alternative final sample sizes, the ASN under either the null or alternative hypothesis is 666.6. Hence, it would appear that the most efficient symmetric adaptive design having two analyses (with no restriction on the number of possible sample sizes used at the final analysis) is on average less efficient than a group sequential test having just three analyses). If we considered five analyses at sample sizes corresponding to an approximately optimal adaptive design, that ASN is 657.6. We see that the introduction of additional analyses has far greater impact than does allowing alternative, adaptively chosen sample sizes. We attribute this behavior to the double-edged sword of trying to 'impute' likely results at future analyses: On the one hand, the estimates are not precise enough to be able to state definitively whether the future results will meet criteria for a particular decision, but apparently the estimates are sufficiently correlated with the future outcomes to cause inflation of the type I error, unless the adaptation rule is chosen carefully.

Lastly, we do acknowledge that the above results are based on symmetric designs (to reduce the dimensionality of the search space) and based on starting with efficient group sequential designs. If we start with an inefficient design (e.g. with O'Brien–Fleming relationships), we can of course adapt toward a markedly more efficient design. But that more efficient adaptive design will mimic a more efficient group sequential design.

3.2.2. Definition of the promising region. The concept of a promising region for interim estimates of treatment effect is sound. In group sequential designs, the promising region would correspond to the continuation region. And in group sequential designs, there is the advantage that 'promise' only need be defined as meriting continuing to the next analysis. In adaptive designs of the type considered by

Drs Mehta and Pocock, however, the burden for the 'promising region' is much greater: the data at the final analysis might be three times larger than the data available at the adaptive interim analysis.

The criterion used by Drs Mehta and Pocock for defining the promising region seems quite strange to us. They have essentially defined as 'promising' any result that would allow adaptation without inflating the type I error. In fact, they define 'promising' as those results that will lead them to have conservative inference, albeit only slightly conservative in the examples they examine. *In fact, the results presented above suggest that at least part of the (to our minds) poor performance of their adaptive strategy is due to the poor criterion for the promising region*. The optimal group sequential tests had continuation regions that corresponded to between 4.9 and 95.1 per cent conditional power computed using the MLE—a much wider range than used in their promising region. We have observed similar results in the optimal adaptive strategies when not allowing early termination of a study [24].

We do not see any advantage in adapting the sample size according to how the adaptation might affect the sampling density. Not all commonly used test statistics have a standard normal distribution, so this does not really pose a new problem. For instance, when providing exact unconditional inference in 2×2 contingency tables, we do not worry because our critical value does not agree with 1.96. Nor are we bothered using the standard uniform distribution for Fisher's exact test. Furthermore, even when our inference is ultimately based on an approximately normally distributed statistic, we often transform that inference into a scale on which standard errors are not easily reported. For instance, in logistic or proportional hazards regression, we often report estimates and CI for the odds ratio and the hazard ratio. The key point is that on those scales, the CI is not the estimate ±1.96 times the standard error, yet communication of the results to applied scientists does not suffer.

The true goal of sequential sampling is to protect the safety of the patients and the economic interests of the sponsor. It is exactly the differences in the sampling distribution that provide that protection.

3.2.3. Sample size modification strategy. The authors recommend the use of a conditional error function with an adaptation that is based on increasing the conditional power in some promising regions in such a way as to have a sampling density that (at least under the null) has an upper α quantile that differs little from the standard normal. We now turn to the advisability of using revised conditional power as the criterion upon which future sample sizes will be based.

Conditional power is a statistical measure with difficult inferential properties. It is essentially a prediction of the sample one would obtain at the end of the study, rather than a measure of the statistical inference one would make using the data at hand. Conditional power is entirely a statement about the sample space, rather than a statement about the parameter space. Nonetheless, it is a measure that is frequently used to parameterize early stopping of studies and adaptive modification of sample size. In doing this, practitioners should be aware of the following difficulties [25]:

- Conditional power depends very much on the assumption of the treatment effect. The authors are choosing to use the current estimate of the treatment effect, while other authors will use the hypothesized effect used at the time of RCT design. There is, of course, a 1:1 correspondence between the conditional power computed under the current estimate and that computed under the design hypothesis. The chief difference, therefore, will be the ease with which thresholds can be described on the two scales. It seems clear to us that the authors were led astray by their misinterpretation of what might have been a 'low' conditional power. We see above that in a group sequential test with two analyses, the efficient design consider a conditional power as low as 4.9 per cent as worth of further study, when that conditional power is computed using the MLE. On the other hand, if conditional power is computed using the design alternative, a conditional power as high as 81.8 per cent is judged so low as to suggest the futility of further study.
- From a Bayesian standpoint, either one of these is placing prior mass 1 on a single treatment effect. The advantage of using the current estimate relative to the use of the original design alternative is that with the latter the user might be conditioning on a treatment effect that the available data have already eliminated. But conditioning on the estimated treatment effect as known assumes there is no variability. Use of predictive power with a noninformative prior would likely provide an improvement over the authors' approach without substantially complicating their formulas.
- Predicting the effect on unconditional power of adaptations expressed on the scale of conditional or predictive power is difficult. For instance, as the statistical information varies, futility rules based on a lower threshold of conditional power might have greater loss of unconditional power than a futility rule based on a higher threshold of conditional power with a larger sample size. Similar

difficulties are demonstrated in the authors' manuscript: The adaptation in their first example shows little gain in unconditional power despite their planning for a large gain in conditional power.

So while the authors' conditional power approach is certainly one way to parameterize sample size modifications, we know no statistical theory that suggests that increasing the sample size to achieve a constant conditional power over some range would lead to any appreciable efficiency. Instead, the comparisons we make in the next section a seem to suggest that it is not advantageous.

4. Results of comparisons

Below we present the results of our comparison of the authors' designs to additional designs that we feel should have also been considered. In order to examine the operating characteristics of the adaptive design, simulations of size 10 000 000 (under the null and design alternative) and 1 000 000 (under other alternatives) was performed (code available on request). With 10 000 000 simulations, a 95 per cent prediction interval for a type I error of 0.025 is ± 0.000097, and a 95 per cent prediction interval for coverage of a 95 per cent CI is ± 0.00014. With 1 000 000 simulations, a 95 per cent prediction interval for a power of 0.50 (the worst-case precision) is 0.00098. S+SeqTrial was used in S-Plus and R for computing power and ASN of the fixed sample and group sequential designs. The results of our simulations and computations are in good agreement with those of the authors.

4.1. Example 1: Schizophrenia

In their first example, the authors consider RCT designs

- *Fxd442*: A fixed sample trial having 442 subjects that was designed to provide 80 per cent power to detect $\delta_1 = 2.0$ (and coincidentally provides 61 per cent power to detect $\delta_1 = 1.6$).
- *Fxd690*: A fixed sample trial having 690 subjects that was designed to provide 80 per cent power to detect $\delta_1 = 1.6$ (and coincidentally provides 94 per cent power to detect $\delta_1 = 2.0$).
- *OBF694*: A GST having a maximum of 694 subjects that was designed to provide 80 per cent power to detect $\delta_1 = 1.6$ (and coincidentally provides 94 per cent power to detect $\delta_1 = 2.0$). A single interim analysis is conducted after observing complete data on 208 subjects, at which time an additional 208 subjects will have been accrued, but not yet have reached the 26 weeks of observation required for ascertaining the primary endpoint on the trial. Early stopping at the interim analysis would be guided by an O'Brien–Fleming-type stopping boundary for efficacy. No early stopping is planned for futility at the interim analysis.
- *Adapt*: An adaptive design having a maximum of 884 subjects. An interim analysis is conducted after observing complete data on 208 subjects, at which time the final sample size will be determined according to the value of an unblinded estimate of the treatment effect: for results corresponding to conditional powers (computed using the current MLE of the treatment effect) between 0.365 and 0.8, the final sample size will provide a conditional power of 0.8 to a maximum of 884 subjects. Such a design provides 65 per cent power to detect $\delta_1 = 1.6$ and provides 83 per cent power to detect $\delta_1 = 2.0$).

Key aspects of the RCT constraints that we gleaned from the authors' chronology were that (1) the investigators did consider the possibility of early stopping for efficacy and (2) though they gave some brief consideration to attaining 80 per cent power to detect an alternative of $\delta_1 = 1.6$, the investigators ultimately regarded that 65 per cent power to detect such an alternative was adequate. Relative to the adaptive design eventually advocated by the authors, the $Fxd690$ and $GST694$ designs take on the appearance of straw men: The adaptive design does not come anywhere close to achieving the 80 per cent power specified for them, and we do not further consider these designs.

Given the above considerations and the authors' obvious regard that their adaptive design is in some sense acceptable, we consider that designs that should also be included in any comparison

- *Fxd492*: A fixed sample trial having 492 subjects that was designed to provide 65.8 per cent power to detect $\delta_1 = 1.6$ (and coincidentally provides 84.1 per cent power to detect $\delta_1 = 2.0$). (We matched the power that we simulated for the authors' adaptive design.)

Table I. Comparison of RCT designs for Example 1.

	Hypothesized treatment effect						
Design	$\delta=0$	$\delta=1.5$	$\delta=1.6$	$\delta=1.7$	$\delta=1.8$	$\delta=1.9$	$\delta=2.0$
Power							
*Fxd*442	2.5%	55.6%	61.1%	66.3%	71.3%	75.9%	80.0%
Adapt	2.5%	60.4%	65.8%	70.8%	75.4%	79.6%	83.4%
*Fxd*492	2.5%	60.2%	65.8%	71.0%	75.9%	80.2%	84.1%
*Fut*492	2.5%	59.8%	65.4%	70.6%	75.4%	79.8%	83.7%
*OBF*492	2.5%	59.6%	65.2%	70.4%	75.3%	79.6%	83.5%
Expected number accrued							
*Fxd*442	442	442	442	442	442	442	442
Adapt	464	496	495	494	492	490	488
*Fxd*492	492	492	492	492	492	492	492
*Fut*492	468	488	489	490	490	490	491
*OBF*492	467	485	485	485	485	484	484
Expected number completed							
*Fxd*442	442	442	442	442	442	442	442
Adapt	464	496	495	494	492	490	488
*Fxd*492	492	492	492	492	492	492	492
*Fut*492	353	472	475	478	481	483	485
*OBF*492	352	455	455	454	452	449	445
Expected calendar time (months)							
*Fxd*442	18.8	18.8	18.8	18.8	18.8	18.8	18.8
Adapt	19.4	20.3	20.3	20.3	20.2	20.1	20.1
*Fxd*492	20.2	20.2	20.2	20.2	20.2	20.2	20.2
*Fut*492	16.2	19.6	19.7	19.8	19.9	19.9	20.0
*OBF*492	16.1	19.1	19.1	19.1	19.0	19.0	18.8

- *Fut492*: A GST having a maximum of 492 subjects and incorporates a futility stopping boundary at a single interim analysis conducted after observing complete data on 208 subjects. Though actually derived on a boundary scale that we find more appealing, when expressed in terms of conditional power, the study would terminate for futility only if the conditional power computed under $\delta=\hat{\delta}_1$ were less than 0.44 per cent, or, equivalently, if the predictive power computed under a flat prior were less than 4.43 per cent. This futility boundary is one that we find typically satisfies the tradeoffs between smaller ASN and loss of power that is acceptable to a sponsor and DSMB when the sponsor does not want to increase maximal sample size.
- *OBF492*: A GST having a maximum of 492 subjects and that at a single interim analysis conducted after observing complete data on 208 subjects incorporates both the futility stopping boundary described above for the *Fut492* and an O'Brien–Fleming efficacy boundary similar to that considered by the authors in their *OBF694* design.

Table I provides statistical power, ASN accrued (corresponding to the authors' numbers with overruns), ASN completed (corresponding to the authors' numbers without overruns), and expected calendar time of completion of data collection for the study. In computing the number of subjects accrued, we assume that 442 subjects will have been accrued at the time that a decision can be made using complete data on the first 208 subjects. Such an assumption allows only three weeks for preparation of the DSMB report and evaluation of that report by the DSMB, but this assumption is at least partially in keeping with that made by the authors (the authors use 416, hence ignoring the time for data analysis and review). We did not add in that three weeks for the calendar time of completion, because the delay associated with completing necessary data entry and reporting the ultimate decision about significant treatment effects will tend to vary according to whether the study was terminated at an interim analysis or at the final analysis.

Not presented in Table I is the worst-case behavior of the designs with respect to the maximal sample size. When using the authors' adaptive design in the current regulatory setting, the sponsor must at the start of the study prespecify a willingness to commit the resources necessary to increase the sample size by as much as 100 per cent. With the *Adapt* design, the probability of increasing the sample size by 25 per cent or more is 0.142 when $\delta=2.0$ and 0.162 when $\delta=1.6$. Even under a true null hypothesis, the sample size will be increased by at least 25 per cent with probability 0.064, and owing to the

nature of the adaptive rule, the preponderance of such large inflations of the sample size will be for a full doubling of the sample size. In contrast, none of the three additional designs we considered ever require more than an 11 per cent increase in the sample size beyond the originally planned 442.

In constructing the $Fxd492$, $Fut492$, and $OBF492$ designs, we made no attempt to find optimally efficient designs, instead just matching the power curves for the *Adapt* design that was chosen by the authors. By adding an extremely conservative futility rule (and without increasing the maximal sample size), we were able to also match the operating characteristics of the *Adapt* design in the presence of an ineffective treatment. Obviously, less conservative futility rules could be used to great advantage in terms of average efficiency when the treatment is ineffective. By adding an efficacy boundary similar to that considered by the authors' $GST694$ rule (and without increasing the maximal sample size), we could also improve on the operating characteristics of the *Adapt* design in the presence of an effective treatment.

In the presence of an effective treatment, the $OBF492$ design shortens the duration of the RCT by just over 5 per cent (a month) on average. We have found that sponsors are generally far more concerned with calendar time costs than with per patient costs. We present our more detailed consideration of these tradeoffs in a survival setting, and the role that adaptive designs might play in addressing those concerns, in a separate manuscript [8].

It should again be noted that the matching of typical operating characteristics (ASN and power) and the improvements in the maximal sample size and study duration were achieved (1) in the presence of delayed ascertainment of outcomes and potential for over-runs (the average numbers of accrued subjects are the same for *Adapt* and $OBF492$), (2) without a large number of interim analyses (only two analyses were performed in either case), (3) without a large up-front sample size commitment (the increase in power was obtained with a worst-case 11 per cent increase in the sample size, compared to a worst-case doubling of the sample size for *Adapt*), and (4) without aggressive early stopping boundaries (the O'Brien–Fleming stopping boundaries are generally well accepted).

We further note that better statistical performance is readily obtained by more careful consideration of the use of partial data on all accrued subjects: The incomplete data represent missingness-at-random (modulo no large time trends in types of subjects recruited), and it can therefore be handled using methods such as those described by Kittelson *et al.* [9], who demonstrated that even when only interested in the outcome measured at end of follow-up, using the incomplete data on some subjects increases the effective sample size by as much as 10 per cent. Hence, in a carefully planned study that considered the partial data, the 'over-runs' do not present a problem, as they can be incorporated into the analyses. Planning for the use of that data is thus very similar to the way survival data is typically handled. (And it may sometimes be the case that the issues driving the minimal acceptable sample size are short-term safety, in which case the incomplete cases might actually provide a cost-efficient mechanism to address the important questions.) In the setting of fully considering the partial information at interim analyses, the sponsor might find acceptable a group sequential approach that would provide 80 per cent power for $\delta = 1.6$, by considering additional analyses performed with more efficient stopping boundaries. We would not necessarily advocate a less conservative efficacy rule at analyses occurring before the minimal sample size of 442. However, given that stopping for efficacy was judged acceptable at 442, we should be free to consider more efficient stopping boundaries at analyses occurring with larger sample sizes.

4.2. *Example 2*: *Acute coronary syndrome*

As noted by the authors, there is no issue with delayed ascertainment of outcome in this example, and hence the comparison of alternative designs is more straightforward.

The authors consider RCT designs:

- *Fxd8000*: A fixed sample trial having 8000 subjects that provides 83 per cent power to detect a relative risk $\rho = 0.80$ (and coincidentally provides 58 per cent power to detect $\rho = 0.85$).
- *OBF8000*: A GST having a maximum of 8000 subjects that was designed to provide 80 per cent power to detect $\rho = 0.80$ (and coincidentally provides 57 per cent power to detect $\rho = 0.85$). Interim analyses are conducted after accruing 4000 and 5600 subjects. Early stopping at the interim analyses would be guided by an O'Brien–Fleming-type stopping boundary for efficacy. No early stopping is planned for futility at the interim analyses.
- *OBF13853*: A GST having a maximum of 13 853 subjects that was designed to provide 80 per cent power to detect $\rho = 0.85$ (and coincidentally provides 97 per cent power to detect $\rho = 0.80$). Interim analyses are conducted after accruing 6926 and 9697 subjects. Early stopping at the

interim analyses would be guided by an O'Brien–Fleming-type stopping boundary for efficacy. No early stopping is planned for futility at the interim analyses.

- *Adapt2*: An adaptive design having a maximum of 16 000 subjects. Interim analyses are conducted after accruing 4000 and 5600 subjects. Early stopping at the interim analyses would be guided by an O'Brien–Fleming-type stopping boundary for efficacy. No early stopping is planned for futility at the interim analyses. (This early stopping at the first and second analyses is the same as for $OBF8000$.) At the time of the second analysis, the sample size for the third and final analyses may be increased beyond 8000 according to the value of an unblinded estimate of the treatment effect: for results corresponding to conditional powers (computed using the current MLE of the treatment effect) between 0.33 and 0.8, the final sample size will provide a conditional power of 0.8 to a maximum of 16 000. subjects. Such a design provides 86 per cent power to detect $\rho=0.8$ and provides 62 per cent power to detect $\rho=0.85$).
- *Adapt2fut*: An adaptive design having efficacy boundaries as described for $Adapt2$, and with futility boundaries corresponding to an MLE for the difference in event rates of 0.01 or more (so in the wrong direction) at the first analysis and corresponding to a conditional power (computed using the MLE) less than 0.20 at the second interim analysis.

We again make comparisons to alternative GST designs. For this example, we consider GST that agree with $OBF8000$ at the first two analyses, and either perform a single additional analysis or then perform analyses at 8000 and a final sample size N_4. In the latter case, at the third analysis we include both an efficacy and futility boundary: because a final sample size of 8000 was considered an alternative, there must be no constraint on regarding some results as futile or convincingly efficacious at that sample size. The exact stopping boundaries and final sample size were found in a trial-and-error search.

- *MatchPwr*: A GST having analyses at 4000 and 5600 subjects using boundaries like $OBF8000$. If the trial did not stop at the second analysis, it continued until a total of 9100 subjects had been accrued, and the final critical value was chosen to ensure a level 0.025 test. (We tried to match the power that the authors reported for their adaptive design $Adapt2$.)
- *MatchPwr4*: A GST having four, rather than three interim analyses. At 4000 and 5600 subjects the boundaries match $OBF8000$. At 8000 subjects, the GST would recommend stopping for futility if $Z>-1.5$ (or equivalently the conditional power computed under the MLE were less than 0.1307) and stopping for efficacy if $Z<-2.35$ (or equivalently the conditional power computed under the MLE were greater than 0.9194. If the trial did not stop at the third analysis, it continued until a total of 9200 subjects had been accrued, and the final critical value was chosen to ensure a level 0.025 test. (We tried to match the power that the authors reported for their adaptive design $Adapt2$.)
- *MatchPwrFut*: A GST having efficacy boundaries the same as $MatchPwr$, except for the inclusion of an O'Brien–Fleming futility boundary at the first and second interim analyses. On the MLE event rate difference scale, these futility boundaries are 0.0034 and −0.0040. On the conditional power scale (relative to statistical significance at an analysis with 9200 subjects), the futility boundary is 0.0003 and 0.0195.
- *MatchPwr4Fut*: A GST having efficacy boundaries the same as $MatchPwr4$, except for the inclusion of an O'Brien–Fleming futility boundary at the first and second interim analyses. On the MLE event rate difference scale, these futility boundaries are 0.0034 and −0.0040. On the conditional power scale (relative to statistical significance at an analysis with 9200 subjects), the futility boundary is 0.0003 and 0.0195.

Table II provides statistical power and ASN accrued for selected alternative hypotheses for all of the above defined RCT designs, except $OBF13853$, which appears to be a straw man irrelevant to the power properties that were actually adopted for the trial. (For $Adapt2$ and $Adapt2fut$, we use the authors' reported operating characteristics.)

The results in Table II show that, as with Example 1, the adaptive approach espoused by the authors does not really do a very good job at improving the power of the RCT design in an efficient manner. It was relatively trivial to improve on either the ASN or the power of $Adapt2$, while committing a far smaller 'up-front' sample size than did $Adapt2$, which had to commit to the possibility of 16 000 subjects.

Furthermore, the futility boundary chosen by the authors resulted in nearly a complete loss of the power gained in the adaptive design: Comparing the right-hand column of Drs Mehta and Pocock's Table X with the results in their Table VI shows very little difference between $Adapt2fut$ and

Table II. Comparison of RCT designs for Example 2.

	Hypothesized treatment effect					
Design	$\rho=1.00$	$\rho=0.85$	$\rho=0.83$	$\rho=0.80$	$\rho=0.77$	$\rho=0.75$
Power						
$Fxd8000$	2.5%	57.4%	68.6%	82.6%	91.9%	95.6%
$OBF8000$	2.5%	56.6%	67.8%	82.0%	91.6%	95.4%
$Adapt2$	2.5%	62%	72%	86%	93%	97%
$MatchPwr$	2.5%	61.7%	73.1%	86.3%	94.4%	97.2%
$MatchPwr4$	2.5%	61.6%	72.9%	86.2%	94.3%	97.2%
$Adapt2fut$	2.1%	58%		82%		95%
$MatchPwrFut$	2.5%	61.4%	72.7%	86.0%	94.1%	97.1%
$MatchPwr4Fut$	2.5%	61.3%	72.6%	86.0%	94.1%	97.1%
Expected number accrued						
$Fxd8000$	8000	8000	8000	8000	8000	8000
$OBF8000$	7980	7263	7001	6533	6015	5669
$Adapt2$	8242	8288	7957	7313	6580	6052
$MatchPwr$	9071	8080	7723	7094	6410	5963
$MatchPwr4$	8045	7618	7320	6764	6151	5754
$Adapt2fut$	5346	7482		6983		5959
$MatchPwrFut$	5938	7614	7424	6957	6355	5936
$MatchPwr4Fut$	5775	7294	7111	6666	6111	5734

$OBF8000$ in terms of statistical power (though there are improvements in ASN). We suspect that this again arises from poor guesses by the authors as to appropriate thresholds for conditional power computed under the MLE, accompanied by a failure to fully evaluate other choices. To us this is another example of why conditional and predictive power are difficult scales to use: Even experienced clinical trialists do not know the threshold to use in order to achieve good operating characteristics. Statistical theory of the Neyman–Pearson lemma says that we should consider tradeoffs between type I and type II error. Conditional power is considering only a single hypothesis, rather than considering the relative likelihood of the outcomes under the null and various alternatives. Of course, the easy way to do this is just to consider the impact that introduction of a futility rule has on the unconditional power, relative to any desired impact its introduction would have on the ASN curve.

5. Discussion

From the preceding discussion, even the most casual reader will be able to ascertain that we are not impressed with the operating characteristics of the adaptive designs proposed by Drs Mehta and Pocock. In the results we presented above, we were able to easily find fixed sample and group sequential tests that matched the power and ASN of the adaptive designs. Those alternative fixed sample and group sequential designs were clearly superior with respect to maximal sample size required. We note that in a real-life situation, we would have explored other designs that might have markedly improved on the power and ASN of the adaptive designs, but that would require a more detailed understanding of all of the sponsor's optimality criteria, including any regulatory concerns about treatment safety and the results of other studies that would be included in a regulatory submission.

We will acknowledge that we did not give any weight to what seemed to be the authors main optimality criteria: (1) that an adaptive design be used, (2) that conditional power be increased, and (3) that naive use of fixed sample inferential methods be conservative.

In confirmatory phase III studies, we see no advantage of adaptive designs, unless they appear advantageous using standard statistical criteria. However, the authors repeatedly made the statement that the adaptive designs would obviate the need for large up-front commitment of resources (countered, of course, by our repeated statement in this commentary that this was not true), and this caused us to reflect further on why that erroneous statement might seem so attractive to a sponsor. What is the 'secondary gain' of the sponsor in choosing an adaptive design even when there is no need for added conservatism and the adaptive design does not improve the power or ASN and makes worse the maximal sample size demands? One possible area is the sponsors' desire to suppress bad news. That is, if a GST is planned, then stopping a trial at the first analysis is typically reported relatively

promptly in a press release. However, if an adaptive design is not prolonged, the sponsors might be able to take considerably longer to release the unfavorable results and do so in as muted a fashion as they can manage. Personally, we do not think that this is appropriate and are thus not sympathetic to the wasteful use of a precious resource (RCT patients) just to support the stock price of a sponsor and/or the continued salary of the study team. There are many reasons that a trial with 'negative' results on its primary endpoint might be ethically continued to gain additional information that would benefit the patient population. But concerns about stock prices should not be the only motivation for choosing inappropriate designs and continuing studies that are not benefiting the patient population. As we noted previously, it is just as easy to put the same sort of spin on a group sequential design as an adaptive design. A sequential sampling plan is just that: when the sampling plan suggests the study is over, we only need say the study is over. It is not over early, unless decisions are made counter to the sampling plan (as they sometimes must be for safety).

With regards to conditional power, we have already discussed our feeling that it was the authors' reliance on conditional power estimates that led them astray. Here we just add that their arguments that we should be more concerned with the conditional power of their adaptation than the unconditional power of the study design seem erroneous to us. It is the unconditional power that we care about at the time of study design. For instance, with a carefully chosen adaptation rule, we could increase conditional power from 20 to 99 per cent with minimal impact on either the ASN or the unconditional power (we would perform no other adaptation except in a small neighborhood of one possible interim outcome). This would be a pretty silly adaptation, as it serves no real purpose. But by the authors' criteria, this would be impressive.

Lastly, the authors' desire that naive use of fixed sample inference not be anti-conservative seems to us to be misplaced. Instead, our goal should be accuracy and precision. For reasons stated above, we place more emphasis on estimation than testing. In an earlier version of their paper, the authors seemed to agree when they said that 'clinical trial reports nowadays focus much more on estimation and confidence intervals rather than just on p-values' and that 'there is an attractive straightforward need to see conventional estimates, confidence intervals and p-values'. Their belief at that time was that conservative p values would necessarily imply conservative confidence intervals. These references to estimation and confidence intervals were removed, however, after our first referee's report demonstrated that the naive confidence intervals had lower than nominal coverage. Such coverage was only negligibly below the nominal level in the example we provided, but it nevertheless indicates that we cannot blindly use the authors' naive approach for estimation or, by natural extension, in non-inferiority studies. While we were very much for the removal of their earlier statement that naive CI would be conservative, we do not believe the inability of their methods to provide adequate protection in this regard negates the validity of their statements about the importance of CI.

We note that in the setting of prespecified adaptive design, we have found that standard group sequential software routines can be used to compute the sampling density of the sufficient statistic, and in a separate manuscript we describe the ways that standard methods for inference in the group sequential setting might be applied to the sufficient statistics from adaptive clinical trials [26]. We are not yet certain, however, whether the previously described relative behavior of the alternative estimation procedures [27, 28] will translate to adaptive designs.

The authors also discuss the use of their adaptive designs in the setting of regulated confirmatory phase III studies. While acknowledging the current requirement for 'well-understood' prespecified designs, they remark that the only concern with non-prespecified designs is operational bias. We do not believe this to be the case. Regulatory agencies must always consider the 'intent-to-cheat' analysis: What investigator biases might intentionally or unintentionally lead to adaptations that lead to anti-conservative inference? Borrowing from *Anna Karenina*, 'All unbiased clinical trials are alike, every biased clinical trial is biased in its own way.' There are many subtle changes that can be made to a trial that the investigators do not recognize are introducing bias. For instance, in a time-to-event analysis using the proportional hazards model, the null hypothesis of greatest scientific interest might be the weak null hypothesis of an average hazard ratio (defined somehow) that is 1, rather than the strong null hypothesis of exact equality of survival curves. The operating characteristics of the hazard ratio estimate (and logrank test) in such a setting might be unduly influenced by changes in the censoring distribution—changes that are easily effected in an adaptive design, but having impact that it is difficult to explore. The FDA has taken the stance that until more experience has been gained regarding such possible mechanisms of bias, we should not rely on the adaptive designs for confirmatory studies, unless the adaptive procedure is completely prespecified.

In summary then, we do not adhere to the above criteria espoused by the authors, and when using more standard optimality criteria, we find that the adaptive designs proposed in the paper by Drs Mehta and Pocock are not ones that we would recommend using.

Personally, we are heading back to the entrance of the cave in order to explore new leads.

Acknowledgements

This work was supported by the following grants: National Institutes of Health U01HL077893 and T32NS048005 and also by NIH T32NS048005.

References

1. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine* 2011.
2. Gao P, Ware J, Mehta C. Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics* 2008; **18**(6):1184–1196.
3. Emerson SS. Issues in the use of adaptive clinical trial designs. *Statistics in Medicine* 2006; **25**(19):3270–3296.
4. Emerson SS, Fleming TR. Adaptive methods: telling 'the rest of the story'. *Journal of Biopharmaceutical Statistics* 2010; **20**:1150–1165.
5. Emerson SS, Kittelson JM, Gillen DL. Frequentist evaluation of group sequential designs. *Statistics in Medicine* 2007; **26**(28):5047–5080.
6. Emerson SS, Kittelson JM, Gillen DL. Bayesian evaluation of group sequential designs. *Statistics in Medicine* 2007; **26**(7):1431–1449.
7. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
8. Emerson SC, Rudser KD, Emerson SS. Exploring the benefits of adaptive sequential designs in time-to-event endpoint settings. *Statistics in Medicine* 2011.
9. Kittelson JM, Sharples K, Emerson SS. Group sequential clinical trials for longitudinal data with analyses using summary statistics. *Statistics in Medicine* 2005; **24**(16):2457–2475.
10. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**(3):659–663.
11. Burington BE, Emerson SS. Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics* 2003; **59**:770–777.
12. Mehta CR, Tsiatis AA. Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal* 2001; **35**(4):1095–1112.
13. Pampallona S, Tsiatis AA, Kim K. Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. *Drug Information Journal* 2001; **35**(4):1113–1121.
14. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041.
15. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**: 1315–1324.
16. Fisher LD. Self-designing clinical trials. *Statistics in Medicine* 1998; **17**(14):1551–1562.
17. Cui L, Hung HMJ, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55**:853–857.
18. Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001; **57**(3):886–891.
19. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A* 1969; **132**(2):235–244.
20. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–200.
21. Tsiatis AA, Mehta CR. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**:367–378.
22. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; **22**(6):971–993.
23. Kittelson JM, Emerson SS. A unifying family of group sequential test designs. *Biometrics* 1999; **55**:874–882.
24. Levin GP, Emerson SC, Emerson SS. Adaptive clinical trial designs with pre-specified rules for modifying the sample size: Understanding efficient types of adaptation. UW Biostatistics Working Paper Series. *Working Paper 377*. Available from: http://www.bepress.com/uwbiostat/paper377.
25. Emerson SS, Kittelson JM, Gillen DL. On the use of stochastic curtailment in group sequential clinical trials. *UW Biostatistics Working Paper Series*, 2005.
26. Emerson SC. Application of group sequential estimation methods to adaptive designs. *Technical Report*, Oregon State University, Department of Statistics, 2011.
27. Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. *Biometrics* 1984; **40**:797–803.
28. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika* 1990; **77**:875–892.