Biometrics 62, 664–683
September 2006

Are Flexible Designs Sound?

Carl-Fredrik Burman* **and Christian Sonesson**

AstraZeneca R & D, SE-431 83 Mölndal, Sweden
**email:* carl-fredrik.burman@astrazeneca.com

Summary. Flexible designs allow large modifications of a design during an experiment. In particular, the sample size can be modified in response to interim data or external information. A standard flexible methodology combines such design modifications with a weighted test, which guarantees the type I error level. However, this inference violates basic inference principles. In an example with independent $N(\mu, 1)$ observations, the test rejects the null hypothesis of $\mu \leq 0$ while the average of the observations is negative. We conclude that flexible design in its most general form with the corresponding weighted test is not valid. Several possible modifications of the flexible design methodology are discussed with a focus on alternative hypothesis tests.

Key words: Adaptive design; Inference principles; Sample size re-estimation; Sufficiency.

1. Introduction

Flexible designs have attracted great interest since the work of Bauer and Köhne (1994) and Proschan and Hunsberger (1995). The methodology offers an option to change the design during an experiment. This flexibility makes it possible to adapt the design of later stages to the findings of earlier ones. It is also possible to allow external information to influence the redesign of the experiment. Many features can be modified. In a clinical trial, for example, the experimenter may drop doses, change the null hypothesis, modify the sample size, etc., possibly after evaluating unblinded data. Fisher (1998) discussed a number of different modifications and Posch, Bauer, and Brannath (2003) cited several papers that have investigated such possibilities. Regardless of how the idea of flexible designs is utilized, it is possible to protect the overall type I error rate by applying a predefined weighting of the data obtained in the different stages.

Most of the interest in flexible designs has focused on methods for sample size modification based on observed interim data. The usual analysis of such designs, treated in Section 2.2, uses a statistic with weights that are not proportional to the information obtained in the different stages. Thus, equally informative observations will be weighted unequally. In this article we discuss whether this analysis is valid. Several alternative analyses, using either weighted or unweighted statistics, will also be considered.

The literature on flexible designs has almost solely taken examples from clinical trials. This is understandable, to some extent, as clinical trials constitute an important and very expensive class of experiments. Regulatory authorities have also shown an interest in this type of design. A reflection paper on flexible designs has been issued by the European Medicines Agency (2006) and several statisticians employed by the FDA have written about the subject. One clinical trial application is found in Zeymer et al. (2001). However, flexible designs can potentially be applied to virtually any experiment in which data are collected sequentially. One example in the field of genetics is given in Scherag et al. (2003). Owing to the wide range of potential applications of flexible designs, we believe that questions concerning the validity of the design and analysis of such experiments are of interest to a wide statistical community.

In Section 2, we discuss flexible designs, including sample size modifications (SSM) based on interim estimates of effect size, and compare them with traditional group sequential designs. Sample size modifications based only on variance estimates (Stein, 1945; Wittes and Brittain, 1990; Gould, 2001) will not be discussed. Section 3 considers more specifically the case of SSM. We describe some of the criticism previously raised against SSM (Jennison and Turnbull, 2003; Tsiatis and Mehta, 2003). Then we examine SSM in the light of basic inference principles and discuss the validity of the inference based on the weighted analysis. Our position is that unrestricted use of the weighted analysis is not sound. Section 4 discusses alternative analyses based on an unweighted statistic, using a design with a prespecified SSM rule. The combination of a weighted and an unweighted statistic is also considered. The resulting test does not require that the SSM rule is fully specified and therefore allows full flexibility. Conclusions are given in Section 5.

2. Different Types of Sequential Designs

2.1 *Group Sequential Designs*

Group sequential designs (Pocock, 1977; O'Brien and Fleming, 1979; Jennison and Turnbull, 2000) are well accepted. A group sequential design consists of a number of stages where interim analyses are made after each stage. A fundamental property of the group sequential design is the possibility of stopping the experiment at an interim analysis

to either accept or reject the null hypothesis. This decision is typically based on a sufficient statistic. Critical values for stopping early are adjusted to achieve a certain type I error. Group sequential designs require some prespecifications (e.g., of the error spending function). The sample sizes of the stages are not allowed to depend on the interim effect estimates.

2.2 *Flexible Designs*

A flexible design has an arbitrary number of stages that are carried out in sequence, just as in a group sequential design. However, flexible designs require fewer prespecifications. The sample sizes of the stages can be chosen freely, even depending on unblinded interim estimates of the treatment effect. A conventional, sequential or nonsequential, design may be transformed into a flexible one at any time and unplanned interim analyses may be added (Müller and Schäfer, 2001). All that is required when applying the weighted test (Bauer and Köhne, 1994) is that the experimenter specifies the weight, v_k, for the data from stage k independently of the data from the kth stage and from later stages. This usually means that the weight is determined before the data in the stage are collected. In a fully blinded trial, this determination can be made later but must be done before breaking the blind. The weights have to be nonnegative, $v_k \geq 0$, and sum up to unity, $\sum_k v_k = 1$. The number of stages, m, is determined simply by assigning in the last stage the remaining weight $v_m = 1 - \sum_{k=1}^{m-1} v_k > 0$. We will in the following assume that the observations from the different stages are independent given the design of the stages.

One way of viewing the weighted analysis of a flexible design is to temporarily consider the stages as separate trials and calculate a p-value, p_k, for each stage k. These p-values are weighted together in an overall test. The joint null hypothesis $\mathrm{H} = \mathrm{H}_1 \cap \cdots \cap \mathrm{H}_m$ is tested by calculating the overall p-value as $p = \Phi(\sum_{k=1}^{m} \sqrt{v_k}\Phi^{-1}(p_k))$, where Φ denotes the cumulative distribution function of the standard normal distribution (Lemacher and Wassmer, 1999). There are other possible ways of weighting the p-values, but these will not be discussed here because they do not change the essence of the problem (see, e.g., Bauer and Köhne, 1994).

We will assume that under H the p-values from the different stages are independent and uniformly distributed on [0, 1]. (This assumption is easily relaxed to the general situation, with p-values stochastically larger than the uniform distribution, without altering the validity of the argument given here; Brannath, Posch, and Bauer, 2002.) Thus $Z_k = -\Phi^{-1}(p_k)$ follows a standard normal distribution. Due to the independence of the p-values, this holds also for the weighted statistic $Z^w = \sum_{k=1}^{m} (v_k)^{1/2} Z_k$. Consequently, with $p = \Phi(-Z^w)$ as given above, we have $P_{\mathrm{H}}(p \leq \alpha) = \alpha$. This proves that the weighted test protects the type I error rate.

Point estimates and confidence intervals are discussed by Brannath et al. (2002) and Lawrence and Hung (2003). Wang et al. (2001) describe procedures for showing noninferiority or superiority.

Flexible designs may be combined with the possibility of stopping the experiment at an interim analysis (Bauer and Köhne, 1994; Cui, Hung, and Wang, 1999; Brannath et al., 2002). This includes the possibility of accepting the null hypothesis (stopping for futility) or declaring statistical significance. As this is done in essentially the same way as for group sequential designs we will not focus on this possibility here. Müller and Schäfer (2001) elaborate on the idea of turning a group sequential design into a flexible design.

Several authors have restricted the flexibility of flexible designs. Early on, Bauer and Köhne (1994) required some prespecifications of how to update the sample size in the study plan. Some examples of prespecified flexible designs will be discussed in Section 4.1.

2.3 *Flexible Designs versus Group Sequential Designs*

The main argument for using flexible designs is the flexibility they provide. Group sequential designs offer only a limited degree of adaptation to the unblinded results of early stage data and no adaptation to external factors. The criticism raised against flexible designs has been relatively mild and rare and mainly focused on SSM. Jennison and Turnbull (2003) and Tsiatis and Mehta (2003) criticized the weighted analysis, arguing that it is inefficient, and advocated instead using group sequential designs. Although this criticism is important (and will be further treated in Section 3.1) we do not regard it as a fully convincing reason for not using SSM. There are certainly situations, taking into account the cost of a trial and possible external information, etc., in which a flexible design may be preferable. The most fundamental question is, however, not whether flexible designs are efficient but rather what inference following a flexible design is valid. This will be discussed in Section 3.2.

3. Sample Size Modifications

Many authors have searched for optimal implementations of SSM. To update the sample size, one alternative is to use the effect size assumed at the start of the trial as a basis for the sample size determination of the coming parts (Denne, 2001). Another way is to use the effect size estimated during the trial (Proschan and Hunsberger, 1995). A decision analytic approach to choosing the sample sizes for the different stages is suggested by Thach and Fisher (2002). Proschan, Liu, and Hunsberger (2003) base an SSM on a Bayesian prior for the effect, which is later updated with interim data. It should be noted that Bayesian ideas are only utilized for internal decision making; the results communicated externally are analyzed in a frequentist way.

3.1 *Criticisms Raised against SSM*

Tsiatis and Mehta (2003) show that there is always a group sequential design that is in a certain sense more efficient than a proposed prespecified flexible design with SSM. One drawback of the proposed group sequential design is that an interim analysis is required at every occasion when the flexible design might have a decision point. This makes the logistics of the trial difficult.

Example 1. Consider a two-stage flexible design including one single interim analysis performed after $N_1 = 100$ observations. Assume that the sample size N_2 for the second stage is a deterministic function of the interim results. The optimal group sequential design, according to Tsiatis and Mehta's definition, has an interim analysis at N_1 and at every possible value of $N_1 + N_2$. For example, if the sample space for N_2 is $\{1, 2, \ldots, 200\}$, the optimal design would then have 200 planned interim analyses (if no early stopping occurs) and a final analysis after 300 observations.

The problem with many interim analyses is obvious, and Tsiatis and Mehta state that less frequent monitoring in the group sequential design, say five to ten times, will typically give a design with similar properties as the optimal one. A decision analytic optimality criterion is useful as it takes the costs of interim analyses into account. With this perspective, even a group sequential design with five interim analyses may be inferior to a flexible design.

Jennison and Turnbull (2003) also proposed group sequential designs as an alternative to flexible designs. The sample size is often based on a power calculation given a plausible effect value, δ. A flexible design will typically increase the sample size if the interim estimated effect is somewhat lower than anticipated. This indicates that the experimenter is interested in demonstrating a positive effect even if the true effect is lower than δ. A group sequential design with very high power given an effect of δ would, according to Jennison and Turnbull, be a better alternative than flexible designs. Such a group sequential design could have a good chance of early stopping if the effect is δ while the power could also be reasonable for smaller effect values, although at the cost of a relatively large sample size. Thus, a group sequential design could achieve the same objectives as a prespecified flexible design.

Flexible designs are more flexible than the alternatives, however. Group sequential designs may stop early in response to interim data but will not react to information from outside the trial. Results of other trials may, for example, be useful in assessing the plausible effect and the conditional power, that is, the power given the observations from the already carried out stages in the ongoing trial. Issues regarding funding and the anticipated consequences of different possible trial results may also change over time. The scientific question that the study has set out to answer may gain in importance during the study, and this could call for higher sample sizes resulting in better precision and higher power.

Liu, Proschan, and Pledger (2002, Example 10), while promoting SSM, criticize SSM rules that are not prespecified, pointing at potential measurability problems. Technically, in Liu et al.'s formulation, N_2 has to be a measurable function of p_1. They argue that a fully flexible rule, where the experimenter is free to choose any sample size after seeing the first stage data, is not necessarily measurable. Rules that are not specified in advance may also lead to other difficulties, as mentioned later in Section 4.

3.2 *Inference Principles*

The key points about the validity of the weighted inference can be made under the simplifying assumption that observations $X_1, X_2, \cdots$ are independent and normally distributed with mean μ and known variance equal to 1. It is usually straightforward to generalize the results to other situations such as, for example, unknown variance, other distributions, and/or two-sample comparisons.

It is convenient to use the formulation suggested by Fisher (1998) of flexible designs in terms of individual observations with individual weights. Before observing X_k, the corresponding weight, $w_k \geq 0$, is specified. In general a weight w_k is random and can depend on all previous observations, $\mathbf{X_{k-1}} = \{X_1, \ldots, X_{k-1}\}$, all previous weights, $\mathbf{w_{k-1}} = \{w_1, \ldots, w_{k-1}\}$, and external factors that we will model by including a nuisance parameter, λ. Thus, $w_k = w_k(\mathbf{X_{k-1}}, \mathbf{w_{k-1}}; \lambda)$. We require that there exists, with a probability of 1, an integer $N = N(w_1, w_2, \cdots)$, interpreted as the total sample size, such that $\sum_{k=1}^{N} w_k = 1$ and $w_N > 0$. The test statistic $Z^w = \sum_{k=1}^{N} (w_k)^{\frac{1}{2}} X_k$ then follows a standard normal distribution under the null hypothesis $\mu = 0$. Let the probability (density) function for the kth weight given the previous weights and observations be $f(w_k \,|\, \mathbf{w_{k-1}}, \mathbf{X_{k-1}})$, with $f(w_1 \,|\, \mathbf{w_0}, \mathbf{X_0})$ interpreted as the unconditional probability function for w_1. With φ denoting the standard normal density, the likelihood for the data generated by the experiment is

$$\begin{aligned}
&\prod_{k=1}^{N} f(w_k \mid \mathbf{w_{k-1}}, \mathbf{X_{k-1}}) \varphi(X_k - \mu) \\
&\quad = (2\pi)^{-N/2} \exp\left(-N\mu^2/2 + \mu \sum_{k=1}^{N} X_k - \sum_{k=1}^{N} X_k^2/2 \right) \\
&\qquad \cdot \prod_{k=1}^{N} f(w_k \mid \mathbf{w_{k-1}}, \mathbf{X_{k-1}}).
\end{aligned}$$

If the weights are completely determined by previous weights and observations, then $f(w_k \,|\, \mathbf{w_{k-1}}, \mathbf{X_{k-1}}) = 1$ for all k. In general, w_k is random given $\mathbf{w_{k-1}}$ and $\mathbf{X_{k-1}}$ and its conditional distribution depends on λ but not on μ. The likelihood can therefore be divided into one part $\exp(-N\mu^2/2 + \mu \sum_{k=1}^{N} X_k)$ that depends on the parameter of interest, μ, and the statistic $S = \{N, \sum_{k=1}^{N} X_k\}$, and a remaining part that does not depend on μ but possibly on the observed variables and weights and the nuisance parameter, λ. The statistic S is therefore minimal sufficient or, in the presence of λ, S-sufficient (Barndorff-Nielsen, 1978). In many statistical models with random sample size, the sample size is independent of the parameter and therefore ancillary. In our situation, however, the total sample size, N, is generally dependent on μ.

We will now focus on the inference procedure in the type of flexible trial described in Section 2. According to the sufficiency principle (Cox and Hinkley, 1974; Barndorff-Nielsen, 1978), the inference should be based on the minimal sufficient statistic alone. Because the weighted test statistic Z^w weights different X_k's differently, the weighted test is not consistent with the sufficiency principle. This was brought up by Jennison and Turnbull (2003) and Posch et al. (2003). Thach and Fisher (2002, p. 436) highlighted the problem of very different weights. The invariance and conditionality principles (Cox and Hinkley, 1974) are also violated, as the weighted test depends on the order of exchangeable observations.

The violation of these principles is problematic, at least in a strictly theoretical sense. However, one might raise the question of their practical importance. Are there any consequences? The following example clearly illustrates that the weighted test may lead to questionable conclusions.

Example 2. Assume that the interest lies in testing $\mu \leq 0$ versus the alternative, $\mu > 0$. With 1000 experimental units, at a level of $\alpha = 5\%$, and thus a critical limit $C_\alpha = \Phi^{-1}(1 - \alpha) = 1.645$ for Z, this sample size gives a power of 81% if $\mu = 0.08$ and $\sigma = 1$. After $N_1 = 100$ observations, it

is decided to take an interim look at the data. Disappointingly, $\sum_{k=1}^{N_1} X_k = -3$, that is, the observed average effect is slightly negative, -0.03. The total weight for stage 1, v_1, is 0.1, because the sample size was originally planned to be 1000. Assuming that only one more stage will be carried out, the second stage will therefore have the weight $v_2 = 1 - v_1 = 0.9$. If the experiment is continued according to plan with 900 observations in the second stage, the conditional power, given the observations from stage 1, is now only 71% under the assumption that the true mean μ is 0.08, as compared to the original 81%. Several authors (e.g., Proschan et al., 2003) have suggested that the sample size modification could be based on a combination of the originally anticipated effect and the observed average. The conditional power is only 37% assuming that $\mu = 0.05$. It might be that the experimenter does not find it worthwhile to continue the experiment as planned with 900 observations. Consider instead the alternative of taking only one single observation in stage 2. For $\mu = 0$, 0.05, and 0.08, for example, the conditional power will then be 3.3%, 3.7%, and 4.0%, respectively. If the experimenter is keen on finding a significant result, this chance may be worth taking. If the observed value of X_{101} happens to be 2.5, then $Z^w = (v_1)^{\frac{1}{2}} Z_1 + (v_2)^{\frac{1}{2}} Z_2 = 0.1^{\frac{1}{2}} \cdot (-0.3) + 0.9^{\frac{1}{2}} \cdot 2.5 \approx 2.28$. Thus, the hypothesis test is clearly significant and it is concluded that $\mu > 0$. However, the average of the observations is $(\sum_{k=1}^{101} X_k)/101 \approx -0.005$. We have the counterintuitive situation of concluding that μ is positive although the average of the observations is negative. This is due to the different weighting of the observations taken in the study and illustrates the danger of violating the inference principles.

4. Alternative Testing Procedures

Having concluded that the previously proposed weighted analysis of a trial involving SSM is questionable, we will investigate whether there is a reasonable frequentist analysis of such a trial.

There are some attempts in the literature to avoid the problems of unequal weighting. Several papers have restricted the way in which the sample size is changed by giving lower and/or upper bounds for the sample size. Often, only increases of the sample size are allowed; see Proschan et al. (2003). As seen by reversing the stages in the example of Section 3.3, this rule is not fully convincing. Furthermore, the restricted SSM using the weighted analysis still violates the sufficiency principle. However, restricted rules that limit the difference between weights will prevent the most extreme consequences of unrestricted SSM.

4.1 *Tests Based on Unweighted Observations*

Because unequal weighting of observations is a problem, a natural attempt could be to base the test on $Z = (\sum_{k=1}^{N} X_k)/N^{\frac{1}{2}}$. There may be considerable inflation of the type I error rate if a naive test is applied (Proschan and Hunsberger, 1995; Shun et al., 2001), rejecting the null hypothesis at a nominal level α if $Z > \Phi(1-\alpha)$. However, provided that the SSM rule is known, the critical level may be adjusted so that the correct type I error rate is achieved. The resulting unweighted test is intuitively reasonable but not necessarily optimal. In order to investigate other alternatives we will further explore what can be said about sufficiency and ancillarity in this model.

One version of the minimal sufficient statistic is $\{N, Z\}$. The next step of the analysis could be to search for an ancillary statistic that is a function of the minimal sufficient statistic. There is no such ancillary statistic here. Note that, in an SSM design, the sample size, N, depends on the observations and thus on the unknown parameter of interest, μ. Even though it is clear that N is not ancillary, one might guess that it is nearly ancillary and that conditioning with respect to N would give a reasonable analysis. However, as we will see, N is highly informative in some situations and then conditioning is not sound.

Example 3. Consider a two-stage design with Bernoulli distributed responses. Assume that the SSM rule $N_2 = N_2(\sum_{k=1}^{N_1} X_k)$ is deterministic and one-to-one. This means that the different outcomes of the first stage correspond to different values of the total sample size, N. Conditioning on N therefore implies conditioning on the results of the first stage. Consequently, the conditional analysis would completely ignore the data from the first stage, regardless of the choice of weights v_1 and v_2.

The assumption of Bernoulli distributed observations makes the point in Example 3 most obvious. However, similar examples may also be constructed for continuous distributions.

As there is no obvious other conditioning, a reasonable analysis is a likelihood ratio (LR) test. Such inference requires that the way in which N depends on the observations is known. A completely flexible SSM rule, without any prespecification, will not satisfy this.

Here we will consider the LR test for a two-stage design as an example. Denote by Y_k the sum of the N_k observations in stage k. Assume a deterministic SSM rule, with $N_1 = n_1$, and $A_n = \{Y_1 : N(Y_1) = n\}$. Let $\varphi(\cdot)$ denote the standard normal probability density. The likelihood of parameter μ for fixed $N = n$ and $\sum_{k=1}^{N} X_k = s$ is

$$\begin{aligned} L(\mu, n, s) &= \frac{d}{ds} P(Y_1 \in A_n, Y_1 + Y_2 \le s) \\ &= \int_{y \in A_n} \varphi\left(\frac{y - n_1\mu}{\sqrt{n_1}}\right) \varphi\left(\frac{(s-y) - (n-n_1)\mu}{\sqrt{n-n_1}}\right) dy \\ &= \varphi\left(\frac{s - n\mu}{\sqrt{n}}\right) \int_{y \in A_n} \varphi\left(\frac{y - sn_1/n}{\sqrt{n_1(n-n_1)/n}}\right) dy. \end{aligned}$$

Note that the last integral is independent of the parameter μ. Given the null hypothesis $\mu = 0$ and alternative $\mu = \mu'$, the LR test statistic is therefore

$$\mathrm{LR}_{\mu'}(n, s) = \frac{L(\mu', n, s)}{L(0, n, s)} = \frac{\varphi((s - n\mu')/\sqrt{n})}{\varphi(s/\sqrt{n})}.$$

Thus, using the transformation to the Z-value, $z = s/n^{\frac{1}{2}}$, we have $\mathrm{LR}_{\mu'}(n, s) = \exp(z\mu' n^{\frac{1}{2}} - \mu'^2 n/2)$, independent of the SSM rule. A one-sided LR test rejects the null hypothesis if and only if $\mathrm{LR}_{\mu'}(n, s) > c$, where c is chosen to achieve a certain type I error. Although the likelihood ratio does not depend on the SSM rule, it is clear that the test does, as the critical value depends on the distribution of the sample size.

Example 4. Assume that $N_1 = 100$ and that the total sample size is 200 or 300 depending on whether $Z_1 > 2.0$ or ≤ 2.0. We optimize the LR test for the alternative $\mu' = 0.2$. Note that the rejection region depends on the value of μ'. From numerical calculations, the rejection region for the LR test, using a significance level of 2.5%, is $\{N = 200,\ Z > 1.81\} \cup \{N = 300,\ Z > 2.05\}$. The uncorrected naive test, with rejection region $Z > 1.96$, would have an inflated level of $\alpha = 2.75\%$. A test based only on Z, ignoring N, which has a correct level $\alpha = 2.5\%$, rejects the null hypothesis if $Z > 2.00$. The LR test gives slightly higher power, 91.7%, compared to 91.4% for this test.

4.2 *The Dual Test*

A major advantage of flexible designs is that there is no requirement for prespecification of how and under which conditions the design is to be changed. In particular, the experimenter can base design modifications on external information. If the flexibility of flexible design is to be preserved, it is therefore typically impossible to characterize the distribution of N. In this case, we cannot construct a test with a correct type I error rate that is based solely on the minimal sufficient statistic.

One idea is to combine the weighted test of Section 2 with a test based on Z (Denne, 2001; Posch et al., 2003; Chen, DeMets, and Lan, 2004). Recall that the weighted statistic is $Z^w = \sum_{k=1}^{N} (w_k)^{\frac{1}{2}} X_k$ and the unweighted statistic is $Z = \sum_{k=1}^{N} (\frac{1}{N})^{\frac{1}{2}} X_k$. The dual test rejects the null hypothesis at one-sided level α if and only if $\min(Z^w, Z) > \Phi(1 - \alpha)$, that is, both the weighted and the naive tests reject at level α. Because the weighted test controls the type I error, the dual test has at most size α.

5. Conclusions

This article is mainly concerned with frequentist inference following a design that uses sample size modifications based on interim effect estimates. It is clear that there is no problem in the Bayesian paradigm. The prior is updated with the information given by the variables actually observed, and the way in which the sample size was determined is irrelevant. Also from a Bayesian point of view, unequal weighting of equally informative observations is not acceptable.

We have identified four possible hypothesis tests from the literature (the naive, the unweighted, the weighted, and the dual), added the LR test as the fifth, and also considered a conditional test. The weighted test (Bauer and Köhne, 1994) is historically closely connected to the ideas of flexible designs and SSM. With this test, great design flexibility is allowed while the type I error is controlled. However, the weighted test violates inference principles, as it weights equally informative observations unequally, and may lead to unreasonable results. Section 3.2 gave an example in which statistically significant evidence for a positive mean was declared while the average of the observations was in fact negative.

There are alternative tests based on the unweighted statistic, Z. The naive test, ignoring that SSM was used, typically results in an inflated type I error level. This inflation can be large. If the SSM rule is known, then the naive test is inferior to the unweighted test where the critical value c is chosen such that the rejection region $\{Z > c\}$ has the correct type I error. The unweighted test seems to be a viable and rather simple option. The critical value is easily calculated numerically for a two-stage design and can also be simulated without difficulty for any SSM design.

For a fixed alternative, the LR test is, by Neyman–Pearson's lemma, the most powerful of all possible tests. As the minimal sufficient statistic is $\{N, Z\}$, the critical value of Z depends on the observed total sample size, N. It is worth noting that the likelihood ratio does not depend on the SSM rule. Still, the rejection region of the LR test depends on the null distribution generated by this rule. One objection to the LR test for the SSM situation is related to well-known issues of ancillarity and mixture of experiments (e.g., Lehmann, 1986, Chapter 10.1). It is usually accepted that, if one of several possible experiments was chosen solely on the basis of an independent random mechanism, such as coin flipping (in our case choosing the sample size), then the analysis should be made conditional on the experiment chosen. Consider an SSM rule where $N_1 = 1$ and where N varies greatly depending on X_1. The situation is then similar to the coin flipping example. In both cases, the LR test has the highest power, but we regard a conditional test as being more sensible. The conditional test is not always attractive, however, as demonstrated by an example in Section 4.1. The reason is that the total sample size is sometimes highly informative.

The unweighted, LR and conditional tests all require a known SSM rule and thus restrict the design flexibility considerably. An interesting alternative analysis is the dual test proposed by Denne (2001) and further studied, for example, in Posch et al. (2003) and Chen et al. (2004). This test requires both the weighted test and the naive test to be significant. The dual test clearly protects the type I error as the weighted test has correct size. One might think that the dual test would be severely conservative. Depending on how the sample sizes are chosen, however, the dual test may be as powerful as the weighted test for a fixed significance level. For a two-stage design, it is clear at the interim analysis for which values of N_2 the conditional power will be the same as that of the weighted test. If such a sample size is always chosen, there is no loss of power. The critical observation is that, given Z_1 and for each potential value of N_2, it is easy to calculate which one of the critical levels of the weighted test and the naive test will be largest. In the former case, the naive test is automatically significant when the weighted test is.

It is clear that the dual test does not obey the sufficiency principle. The problem with this seems connected with power rather than with the strength of evidence in the case of a significant result. It is often instructive to view a statistical procedure from different angles. The weighted test shows a very strange behavior from a perspective of likelihood-based inference or Bayesianism. However, a significant result in the dual test implies that the value of Z is large. For example, a Bayesian analysis with uninformative prior would then agree that there is good evidence for a positive effect. On the other hand, it may happen that Z is large also when the dual test fails to reject the null hypothesis. This is reflected in the suboptimal power of the dual test. Only some SSMs can be done

without losing power compared to the weighted test. Furthermore, the weighted test is less powerful than the LR test.

Much more work is needed to explore flexible designs and the inference following such designs. First we would like to see continued discussion on the validity of the tests. The dual test deserves further attention as it preserves Bauer and Köhne's original flexibility, with no need to prespecify the SSM rule. Second, the efficiency of the tests should be compared. Third, given better answers to how the analysis should be done, the relative merits of flexible designs, as compared, for example, with group sequential methods, are open to additional study.

Acknowledgements

We have benefited from general discussions of SSM with Professors David Cox and Marianne Frisén. However, the responsibility for the viewpoints expressed in this article is solely our own. We thank the co-editors, an associate editor, and a referee for helpful comments.

References

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. New York: Wiley.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50,** 1029–1041.

Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* **97,** 236–244.

Chen, Y. H. J., DeMets, D. L., and Lan, K. K. G. (2004). Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine* **23,** 1023–1038.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cui, L., Hung, H. M. J., and Wang, S.-J. (1999). Modifications of sample size in group sequential clinical trials. *Biometrics* **55,** 853–857.

Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine* **20,** 2645–2660.

European Medicines Agency. (2006). Draft reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan. Available from `http://www.emea.eu.int/pdfs/human/ewp/245902en.pdf` (accessed March 2006).

Fisher, L. D. (1998). Self-designing clinical trials. *Statistics in Medicine* **17,** 1551–1562.

Gould, A. L. (2001). Sample size re-estimation: Recent developments and practical considerations. *Statistics in Medicine* **20,** 2625–2643.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. London: Chapman and Hall.

Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on observed treatment effect. *Statistics in Medicine* **22,** 971–993.

Lawrence, J. and Hung, H. M. J. (2003). Estimation and confidence intervals after adjusting the maximum information. *Biometrical Journal* **45,** 143–152.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. New York: Wiley.

Lemacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55,** 1286–1290.

Liu, Q., Proschan, M. A., and Pledger, G. W. (2002). A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association* **97,** 1034–1041.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential trials. *Biometrics* **57,** 886–891.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35,** 549–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64,** 191–199.

Posch, M., Bauer, P., and Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine* **22,** 953–969.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51,** 1315–1324.

Proschan, M. A., Liu, Q., and Hunsberger, S. (2003). Practical midcourse sample size modification in clinical trials. *Controlled Clinical Trials* **24,** 4–15.

Scherag, A., Müller, H.-H., Dempfle, A., Hebebrand, J., and Schäfer, H. (2003). Data adaptive interim modification of sample sizes for candidate-gene association studies. *Human Heredity* **56,** 56–62.

Shun, Z., Yuan, W., Brady, W. E., and Hsu, H. (2001). Type 1 error in sample size re-estimation based on observed treatment difference. *Statistics in Medicine* **20,** 497–513.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16**, 243–258.

Thach, C. T. and Fisher, L. D. (2002). Self-designing two-stage trials to minimize expected costs. *Biometrics* **58,** 432–438.

Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90,** 367–378.

Wang, S.-J., Hung, H. M. J., Tsong, Y., and Cui, L. (2001). Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* **20,** 1903–1912.

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9,** 65–72.

Zeymer, U., Suryapranata, H., Monassier, J. P., Opolski, G., Davies, J., Rasmanis, G., Linssen, G., Tebbe, U., Schroder, R., Tiemann, R., Machnig, T., and Neuhaus, K.-L. (2001). The Na^+/H^+ exchange inhibitor eniporide as an adjunct to early reperfusion therapy for acute myocardial infarction. *Journal of the American College of Cardiology* **38,** 1644–1650.

Received July 2004. *Revised September* 2005.
Accepted October 2005.

Discussions

Christopher Jennison

Department of Mathematical Sciences
University of Bath
Bath BA2 7AY, U.K.
email: *cj@maths.bath.ac.uk*

and

Bruce W. Turnbull

Department of Statistical Science
Cornell University
Ithaca, New York 14853-3801, U.S.A.
email: *bwt2@cornell.edu*

The authors raise important concerns about adaptive and flexible designs. They note that equally informative observations can be weighted unequally in these designs and such weighting is contrary to principles of statistical inference; their examples illustrate what many would agree to be inappropriate inferences. Traditional group sequential tests also offer a methodology for reaching a data-dependent choice of final sample size and the efficiency of new adaptive methods should be assessed against this alternative approach.

The emphasis in the article is on the "validity" of inference on conclusion of a study, but this is not a simple concept to define. It is clearly important that trial results should be credible to the intended audience and credibility may suffer if basic principles of inference are not followed. The same principles of inference are fundamental to issues of efficiency and admissibility; pursuing these questions further shows that efficiency depends crucially on other aspects of adaptive designs, in particular the rule for modifying sample size. We have found many published proposals for adaptive designs to be inefficient from this perspective but their proponents are usually keen to defend them. If a rule for modifying sample size were identified as inefficient, would the authors agree to label a design using this rule as "invalid"?

The authors mention other types of adaptive redesign besides modification of sample size, for example, dropping treatment arms for certain doses or modifying the null hypothesis being tested. These and other forms of adaptation have recently been surveyed by a PhRMA working group and conclusions are presented in Gallo et al. (2006). Our own discussion accompanying that paper complements our comments here, which focus particularly on efficiency.

1. Preplanned Adaptive Designs

We begin our discussion in the context of designs that allow adaptive sample size modification but do so in a preplanned, rather than flexible, manner. As in the article, we consider the case where there is no need to adjust sample size in response to updated estimates of a variance or other nuisance parameter. The authors note in Section 2.2 that "flexible designs may be combined . . . with stopping the experiment at an interim analysis" but they do not pursue this option. We shall consider the broader class of designs that permit early stopping. We start with an example to illustrate how a seemingly intuitive and appealing adaptive design can be quite inefficient when compared with a suitably chosen group sequential test.

1.1 *Example: A Variance Spending Design*

Suppose θ represents a treatment effect and we wish to test $\mathrm{H}_0 : \theta \leq 0$ against the alternative $\theta > 0$. The score statistic for data yielding information $\mathcal{I}$ for θ has distribution $S \sim N(\theta\mathcal{I}, \mathcal{I})$ and a fixed sample test of H_0 with power $1 - \beta$ at $\theta = \delta > 0$ requires information

$$\mathcal{I}_f = (z_\alpha + z_\beta)^2/\delta^2,$$

where z_p denotes the $1 - p$ quantile of the standard normal distribution. In many applications, "information" is directly related to sample size but in other cases the relation is less direct, for example, in a comparison of survival distributions information depends primarily on the number of observed failures. In Shen and Fisher's (1999) "variance spending" designs, data are collected in groups with successive groups providing information $r_1\mathcal{I}_f, r_2\mathcal{I}_f, \ldots$, for a prespecified sequence r_1, $r_2, \ldots$. The score statistic from group j is

$$S_j \sim N(\theta r_j \mathcal{I}_f, r_j \mathcal{I}_f).$$

Weights w_j, $j = 1, 2, \ldots$, are defined adaptively with w_j allowed to depend on $S_1, \ldots, S_{j-1}$ and $w_1, \ldots, w_{j-1}$. The design requires that a stage m is reached where

$$\sum_{j=1}^{m} w_j^2 = 1. \tag{1}$$

An overall test statistic is built up from contributions $w_j S_j/(r_j\mathcal{I}_f)^{\frac{1}{2}} \sim N(\theta w_j (r_j\mathcal{I}_f)^{\frac{1}{2}}, w_j^2)$. Then, under $\theta = 0$, it can be shown that

$$T_m = \sum_{j=1}^{m} \frac{w_j S_j}{\sqrt{(r_j\mathcal{I}_f)}} \sim N(0, 1).$$

With $\sum_j r_j = 1$ and $w_j = (r_j)^{\frac{1}{2}}$, this formula yields the fixed sample size test. However, the idea of the adaptive design is to allocate lower weights to future groups when the current estimate of θ is low, thereby extending the study and increasing power. Shen and Fisher (1999) propose a method for assigning weights adaptively. At the end of stage $j-1$, a target for further information $\mathcal{I}_j^*$ is calculated and the aim, at this point, is to gather groups of data until information $\mathcal{I}_j^*$ is reached, setting future group weights w_j proportional to $(r_j)^{\frac{1}{2}}$ while satisfying (1). If $r_j\mathcal{I}_f > \mathcal{I}_j^*$, the next group will exceed this information target and the study can then be brought to a close with $w_j = (1 - \sum_{i=1}^{j-1} w_i^2)^{1/2}$. If $r_j\mathcal{I}_f < \mathcal{I}_j^*$, more than one additional group is required and the weight for group j is set as

$$w_j = \left\{ \frac{r_j\mathcal{I}_f}{\mathcal{I}_j^*} \left(1 - \sum_{i=1}^{j-1} w_i^2 \right) \right\}^{1/2}.$$

If the overall target information level remained fixed, the future weights w_j would be proportional to the square root of the information provided by each group, but in reality the target will vary after each new group of observations.

Denote the maximum likelihood estimate of θ after stage $j-1$ by

$$\hat{\theta}_{j-1} = \frac{\sum_{i=1}^{j-1} S_i}{\sum_{i=1}^{j-1} r_i\mathcal{I}_f}.$$

In the design we have studied, $\mathcal{I}_j^*$ is chosen so that if just one additional group of observations were taken, yielding a score statistic $S_j \sim N(\theta\mathcal{I}_j^*, \mathcal{I}_j^*)$ weighted by $w_j = (1 - \sum_{i=1}^{j-1} w_i^2)^{1/2}$, the conditional power under $\theta = \hat{\theta}_{j-1}$ would be at least $1 - \beta$. This gives the condition, for positive values of $\hat{\theta}_{j-1}$,

$$\frac{T_{j-1} - z_\alpha}{\left(1 - \sum_{i=1}^{j-1} w_i^2 \right)^{1/2}} + \hat{\theta}_{j-1}\sqrt{\mathcal{I}_j^*} \geq z_\beta,$$

where $T_{j-1} = \sum_{i=1}^{j-1} w_i S_i / (r_i\mathcal{I}_f)^{\frac{1}{2}}$. If $T_{j-1} - z_\alpha \geq (1 - \sum_{i=1}^{j-1} w_i^2)^{1/2} z_\beta$, then $\mathcal{I}_j^* = 0$ would suffice, but it is still necessary to take one more group of observations for the variance spending design to terminate. In other cases with positive $\hat{\theta}_{j-1}$ we set

$$\mathcal{I}_j^* = \left\{ \frac{T_{j-1} - z_\alpha}{\left(1 - \sum_{i=1}^{j-1} w_i^2 \right)^{1/2}} - z_\beta \right\}^2 \frac{1}{\hat{\theta}_{j-1}^2}, \qquad (2)$$

truncating this when necessary to restrict the total information to an upper bound $\mathcal{I}_{\max}$. When $\hat{\theta}_{j-1}$ is negative but early termination for "futility" has not occurred, the same upper bound on total information is used to define $\mathcal{I}_j^*$ directly.

We have simulated this design with $\alpha = 0.025$ and $1 - \beta = 0.9$. Following Shen and Fisher's (1999) recommendations, we used $r_1 = 1/2$ and $r_2 = r_3 = \cdots = 1/6$ up to a maximum of 10 groups, so the test is terminated with $w_{10} = (1 - \sum_1^9 w_i^2)^{1/2}$ if it continues as far as the tenth stage and total information $2\mathcal{I}_f$. The value $2\mathcal{I}_f$ is also used for $\mathcal{I}_{\max}$ in the truncation described above. The design includes early termination for futility by stopping to accept H_0 at stage j if $\hat{\theta}_j < \tilde{\delta} - (\sum_{i=1}^{j} r_i\mathcal{I}_f)^{-1/2} z_{0.99}$ with $\tilde{\delta}$ set as $0.8\,\delta$, the value below which power starts to decline rapidly. This rule is of the form proposed at the end of Section 2 of Shen and Fisher (1999), but using $\tilde{\delta} = 0.8\,\delta$ rather than $\tilde{\delta} = \delta$ to avoid too much loss of power for θ values just below δ.

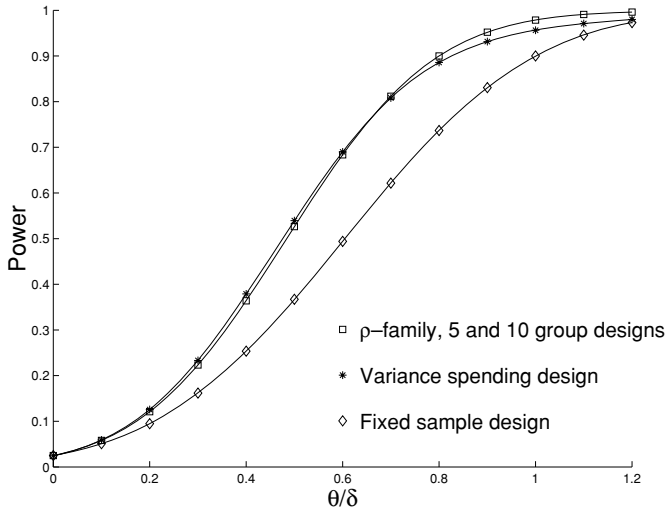

Figure 1. Power of the fixed sample test, variance spending design, and a 5- or 10-group ρ-family test ($\rho = 1$) with power 0.9 at $\theta = 0.8\ \delta$.

Figure 1 compares the power curve of this variance spending test with that of the underlying fixed sample design. Results are based on one million simulations and estimation error is negligible. The figure shows adaptation has been effective in increasing power above that of the fixed sample size test. Because the adaptive redesign is completely prespecified, it is perfectly possible to evaluate the design and compare it with other group sequential schemes before starting a study. One versatile class of group sequential tests is the ρ-family of error spending tests described by Jennison and Turnbull (2000, Section 7.3). A suitable choice here is the design with parameter $\rho = 1$ and group sizes set to attain power 0.9 at $0.8\ \delta$, which requires maximum information $1.95\mathcal{I}_f$ and $2.02\mathcal{I}_f$ for tests with 5 and 10 groups, respectively. The power curves for the ρ-family tests with 5 and 10 groups are indistinguishable and Figure 1 shows these are superior to the power curve of the variance spending test at the higher values of θ where power is most important. Figure 2 shows expected information on termination for the two group sequential tests and the variance spending test, expressed in units of $\mathcal{I}_f$. The curves for the group sequential ρ-family tests are lower by around 15–20% over the full range of θ values, indicating lower average sample size and a shorter expected time for the study to reach a conclusion. The group sequential test's superior power curve and significantly lower expected information curve show serious inefficiency in the variance spending design.

We have made similar comparisons to assess the efficiency of a variety of adaptive designs proposed in the literature and,

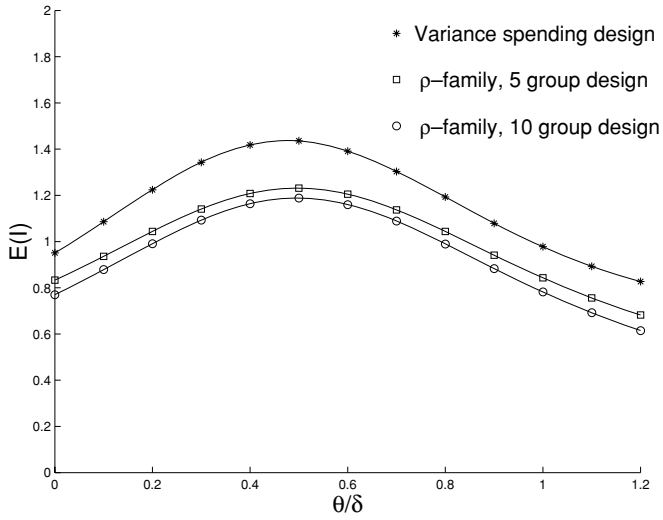

Figure 2. Expected information on termination of the variance spending design and 5- and 10-group ρ-family tests ($\rho = 1$) with power 0.9 at $\theta = 0.8\ \delta$, expressed in units of $\mathcal{I}_f$.

in all cases, found the ρ-family to provide efficient nonadaptive alternatives. In examples where adaptation is used to increase power more substantially, the degree of inefficiency can be much greater. For further examples, see Jennison and Turnbull (2003) and (2006a). It is, thus, appropriate to explore the theoretical foundations of efficiency and inefficiency in order to understand the principles behind sound trial design.

1.2 *Theory*

1.2.1 *Sufficient statistics.* We summarize here results derived in Jennison and Turnbull (2006a) for adaptive group sequential designs testing $\mathrm{H}_0 : \theta \leq 0$ against $\theta > 0$. In these designs, group sizes are chosen adaptively from a finite, but large, set of options, a maximum of K analyses is allowed, and early stopping is permitted at each analysis to accept or reject H_0. Designs are compared in terms of their power curves and expected information on termination at a set of θ values. A design is said to be inadmissible if another design has a lower expected information function and higher power curve (strictly speaking, the dominating design can be equal in some respects but there must be at least one instance of strict superiority). A design that is not inadmissible is said to be admissible.

The fundamental theoretical result is a "complete class theorem," which states that all admissible designs are solutions of Bayes sequential decision problems. Because Bayes designs are functions of sufficient statistics, any adaptive design defined through nonsufficient statistics is inadmissible and is dominated by a design based on sufficient statistics. This conclusion confirms that violation of the sufficiency principle has a negative impact on the efficiency of an adaptive design. Schmegner and Baron (2004) obtain similar conclusions on the inadmissibility of rules based on nonsufficient statistics in the special case where sampling and decision costs are combined in a single Bayes risk under a stated prior for θ.

In our example of a Shen and Fisher (1999) variance spending test, the weights w_j create a statistic T_m that is not sufficient for θ and, hence, this test can be outperformed by a test with the same sequence of information levels and a stopping rule based on sufficient statistics. Because the group sizes (and therefore information levels) of the variance spending test are predetermined, this dominating design is simply a nonadaptive group sequential test.

In many other adaptive designs, group sizes are chosen adaptively and a test dominating an adaptive design based on nonsufficient statistics may also have adaptively chosen group sizes. This raises the question as to when an adaptive design using nonsufficient statistics can be improved on by a *nonadaptive* group sequential design. It follows from our theoretical results that this is always possible, but with the same proviso required by Tsiatis and Mehta (2003), namely, the group sequential test has to be allowed an analysis at every cumulative information level that might arise in the adaptive design. Allowing so many analyses gives the group sequential test an unfair advantage and, as Burman and Sonesson note in Section 3.1, a trial design with a great many interim analyses could well be impractical.

The advantage held by adaptive designs in this discussion is that response-dependent choice of group sizes can itself be a source of improved efficiency, an idea first proposed by Schmitz (1993). We have explored the possible benefits of this feature by comparing optimal adaptive designs and optimal nonadaptive designs for specific criteria. Here, the optimal adaptive designs minimize expected information on termination averaged over a set of effect sizes; they are optimal among all possible stopping boundaries and sample size rules, a substantial advance on the "optimal implementations" Burman and Sonesson refer to at the beginning of Section 3. The benefits gained by adaptive choice of group sizes (or equivalently information levels) are quantified by Jennison (2003) and Jennison and Turnbull (2006a,b) and these results show that, when the maximum number of analyses is held fixed, adaptive choice of group sizes leads to only slight efficiency gains. These gains are of the order of 1% of $\mathcal{I}_f$, and it is unlikely they would justify the administrative complexity of an adaptive design. Combining the complete class theorem with these numerical results, we arrive at the following conclusions: any adaptive design based on nonsufficient statistics can be improved by an adaptive design using sufficient statistics; the performance of this improved design can be matched very closely by a group sequential test with the same number of analyses; if the adaptive design is inefficient in any respect, it is quite possible that a well-chosen group sequential test can outperform it completely.

1.2.2 *Sample size rule.* Although the origins of our discussion of efficiency lie in a concern over breaches of the sufficiency principle, it is another aspect of certain adaptive designs, namely, the rule for sample size modification (SSM), that we believe to be the major source of inefficiency. Our calculations of optimal adaptive designs show that optimized sample size rules are quite different from those proposed by many authors for their adaptive designs. When a design is

based on attaining fixed conditional power, either at a given effect size or under the current estimate of effect size, future sample size increases monotonically as the current test statistic decreases; in contrast, optimal adaptive designs have the largest future group sizes when the test statistic is in the center of the continuation region and group sizes are smallest near either stopping boundary.

In our example of a Shen and Fisher (1999) design, the current estimate of effect size was substituted into a sample size formula with no allowance for the high variance of such an interim estimate (remember that the fixed sample size test with information $\mathcal{I}_f$ is only just capable of distinguishing between $\theta = 0$ and $\theta = \delta$). Interim estimates of θ are used in a similar way in many proposed designs and their inherent variability leads to random variations in sample size that are themselves inefficient (see Jennison and Turnbull, 2003, Section 4.3, for further discussion of this point).

1.2.3 *Burman and Sonesson's likelihood ratio rule.* In Section 4.1, the authors propose a likelihood ratio (LR) test for data collected in a two-stage design. In their Example 4, an initial sample of 100 observations is taken in the first stage followed by a second group of 100 or 200 observations, depending on the first-stage data. Early stopping at the first stage is not permitted. It is easy to see that the final decision of the LR test is the Bayes rule for some prior and loss function with weights at effect sizes 0 and μ'. Thus, the LR procedure is admissible among designs with the same SSM rule. However, the SSM rule itself is arbitrary; moreover, the design does not take advantage of the opportunity to stop to accept or reject H_0 at the first stage. If we consider the LR procedure in the class of all possible two-stage designs with the option to stop at stage 1, it is not a Bayes design (because then it would stop at the first stage for some outcomes) and so it is inadmissible. Indeed, we have found nonadaptive group sequential tests with two groups of 150 observations that match the power curve of the authors' LR test and have expected sample size lower by 10–30% for effect sizes in the range 0–0.3.

The authors state in Section 2.3 that, to them, the "fundamental question is . . . not whether flexible designs are efficient but rather what inference following a flexible design is valid." We are concerned that this view can lead to the conclusion that the LR version of the study design in Example 4 is regarded as "acceptable" when other designs of comparable complexity can provide the same power curve for considerably smaller average sample sizes. We believe efficiency issues should be central to this discussion.

2. Adaptive Designs Used Flexibly

Much of the motivation for adaptive designs is the possibility of flexible usage, extending beyond the prespecified designs we have discussed thus far. The authors note in Section 3.1 that investigators may wish to respond to information from outside a trial, for example, results reported from other trials. The importance of the scientific question being addressed can change over time and the availability of new funding may increase the resources to pursue this question. Adapting a design in response to unanticipated external factors requires flexible methods. For many types of flexible redesign, the requirement to maintain the stated type I error probability makes it inevitable that the final decision will be based on a nonsufficient statistic. Thus, the combination of flexibility and strict control of the type I error rate means it is not possible to insist on some of the usual features of inference rules.

The authors appear to acknowledge this paradox in their conclusion that "unrestricted use of the weighted analysis is not sound" as this leaves open the option of using a weighted analysis when there is no other alternative. Combining this with an unweighted analysis in the "dual test" may provide reassurance when both tests agree. One should still worry though about the arguments that will arise in the (possibly rare) situations where the two tests lead to different conclusions.

The paper ends by encouraging "continued discussion on the validity of the tests." Various desiderata for tests have been aired in the paper and we would welcome a simple statement of what the authors see as the key criteria for a test to be "valid." As noted earlier, we would wish to see efficiency included with these criteria.

It is a pleasure to thank the authors for a stimulating paper.

References

Burman, C.-F. and Sonesson, C. (2006). Are flexible designs sound? *Biometrics* doi: 10.1111/j.1541-0420.2006.00626.x.

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., and Pinheiro, J. (2006). Adaptive designs in clinical drug development—An executive summary of the PhRMA working group, with discussion. *Journal of Biopharmaceutical Statistics* **16,** 275–312.

Jennison, C. (2003). Discussion of "Optimal dynamic treatment regimes" by S. A. Murphy. *Journal of the Royal Statistical Society, Series B* **65,** 356–357.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, Florida: Chapman & Hall/CRC.

Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* **22,** 971–993.

Jennison, C. and Turnbull, B. W. (2006a). Adaptive and nonadaptive group sequential tests. *Biometrika* **93,** 1–21.

Jennison, C. and Turnbull, B. W. (2006b). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine* **25,** 917–932.

Schmegner, C. and Baron, M. I. (2004). Principles of optimal sequential planning. *Sequential Analysis* **23,** 11–32.

Schmitz, N. (1993). *Optimal Sequentially Planned Decision Procedures*. New York: Springer-Verlag.

Shen, Y. and Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55,** 190–197.

Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90,** 367–378.

Michael A. Proschan

Biostatistics Research Branch
National Institute of Allergy and Infectious Diseases
Bethesda, Maryland 20892-7609, U.S.A.
email: *proscham@niaid.nih.gov*

I am thankful for the opportunity to respond to this important critical look at adaptive methods for clinical trials. Burman and Sonesson (2006) raise legitimate concerns while accurately characterizing the tradeoff between flexibility and optimality. Understanding this tradeoff is essential when deciding which adaptive methods, if any, should be used in a given trial. A similar tradeoff applies to the comparison of group sequential monitoring to no monitoring: monitoring allows us to stop early, but at the cost of loss of power compared to not monitoring. Adaptive sample size methods allow more leeway by permitting us to change the originally planned sample size after seeing data. Even within the class of adaptive methods, some are more flexible than others. Some rely on estimates of the treatment effect, while others use only nuisance parameter estimates. Burman and Sonesson focus on the former, but there are gradations of flexibility even within methods based on the treatment effect. Some require prespecification of a sample size rule dictating exactly how we will change the sample size for every conceivable first-stage outcome; others require us to decide the sample size only for the first-stage outcome actually observed, making the assumption that had we observed another value, our mind would have acted in a measurable way. The greater the flexibility we allow, the more care we must exercise to ensure a sensible test.

The authors offer, as proof of the unsoundness of adaptive designs, an example in which an adaptive design leads to the conclusion that $\mu > 0$ even though the average of the observations is negative. That this could happen with injudicious choice of adaptive method and/or sample size modification was pointed out at the bottom of page 1321 of Proschan and Hunsberger (1995). In fact, it was the prospect of this type of aberration that prompted Proschan and Hunsberger to consider, in the choice of sample size, whether the critical value for the usual z-score would differ substantially from that of a fixed-sample test (see Proschan, 2004 for a geometric perspective on this critical value). This informal sample size restriction prevents the type of nonsensical test that can result from extreme deviations from the originally planned sample size. The examples from Proschan and Hunsberger (1995) and Burman and Sonesson (2006) highlight the fact that greater care is needed the greater the flexibility one allows.

Burman and Sonesson's application of inference principles to adaptive methods is very useful. They consider the one-sample case of testing whether a mean exceeds 0 and focus on the sufficiency principle, though they also mention the invariance principle. I would like to expand on these and discuss another one—the likelihood principle. The authors emphasize that most adaptive methods violate the sufficiency principle because inference is not based solely on the sufficient statistic—the (random) sample size N and the sum of the N observations. The premise is that a specific rule $N(S_1)$ relating the final sample size to the first-stage sum has been prespecified and must be followed. If that is the case, then I agree that one should follow the sufficiency principle. But the premise eliminates much of the flexibility afforded by adaptive methods. In reality, the decision to change the sample size is a very complex one involving numerous factors, some of which may be foreseeable and others not. The idea that we could know how decision makers would react for every possible outcome is simply unrealistic. All we will ever really know is the sample size $N(s_{\text{obs}})$ they chose for the *observed* first-stage outcome s_{obs}. Suppose we prespecify a rule $N = N(S_1)$ and an α-level rejection region based on the sufficient statistic (N, S_N). That is, the rejection region is of the form $\cup_n (N = n, S_n \in R_n)$ for Borel sets R_n. Now suppose we do not follow the sample size rule. Will we still have an α-level procedure? In some situations we cannot even proceed if we do not follow the prespecified rule. For example, suppose the prespecified rule calls for a sample size of either 100 or 200, therefore the rejection region is of the form $(N = 100, S_{100} \in A_{100}) \cup (N = 200, S_{200} \in A_{200})$. If we decide to use a sample size of 150 instead, we cannot proceed. More can be said, however. In fact, there does not exist a sample size rule $N = N(S_1)$ and test *based on the sufficient statistic* that maintains level α even if one does not follow the sample size rule (one proof is given at the end of this discussion, though I think there must be a shorter one). The only tests that maintain level α even if we do not follow the prespecified sample size rule are not based on the sufficient statistic. For example, the test that rejects the null hypothesis when $(Z_1 + Z_2)/2^{1/2} > 1.96$, where $Z_1 = S_1/n_1^{1/2}$ and $Z_2 = S_2/(n - n_1)^{1/2}$ are the z-statistics from the first and second stage, has level α for any (nonzero) first- and second-stage sample sizes.

I must admit to some confusion about Burman and Sonesson's point that adaptive methods violate the invariance principle because inferences depend on the order of exchangeable observations. It was unclear to me which set of random variables they were asserting were exchangeable—the infinite set of all potential observations or the set of N observations actually observed, where N is random. The former set is exchangeable but the latter is not. To simplify the discussion, consider a trial with group sequential monitoring and no adaptive sample size modification. Assume the trial will have only two observations unless it is stopped after the first one, and the observations are independent and identically distributed (i.i.d.) standard normals under the null hypothesis. If we use the O'Brien–Fleming boundary, we will reject the null hypothesis after the first observation Z_1 if $Z_1 > 2.796$, and after the second observation if $(Z_1 + Z_2)/2^{1/2} > 1.977$. If we did not monitor, but rather always observed Z_1 and Z_2, then Z_1 and Z_2 would be i.i.d. and hence exchangeable. With monitoring, if we proceed to the second stage, the observations are no longer exchangeable; the likelihood of (2.9, 1.0) is 0—because we would have stopped at stage 1 without seeing

$z_2 = 1.0$—whereas the likelihood of (1.0, 2.9) is not. Thus, we have two ways of viewing things: (a) (Z_1, Z_2) are not exchangeable, in which case there is no violation of the invariance principle, or (b) (Z_1, Z_2) are exchangeable—because in a nonmonitoring setting, they are—in which case group sequential monitoring *without sample size modification* also violates the invariance principle.

The fact that Burman and Sonesson speak positively about the dual test, which they admit also violates the sufficiency principle, suggests that principles of inference only go so far. It is not uncommon for accepted statistical methods to violate at least one seemingly reasonable principle. For example, the likelihood principle states that our inference should depend only on the likelihood of the observed data, not on what we would have done if the outcome had been different. This seems like a reasonable principle, but following it means eliminating much of classical statistics, including hypothesis testing. In a sense, adaptive methods that are invariant to the choice of sample size function take a step toward appeasing proponents of the likelihood principle; at least our current decision does not depend on the sample size we would have chosen had the first-stage result been different.

I agree with Burman and Sonesson that previous criticisms of adaptive methods (Jennison and Turnbull, 2003; Tsiatis and Mehta, 2003) are not convincing. The proposed "improvements" either impose a rigid sample size rule that eliminates much of the appeal of adaptive methods or assume a maximum sample size N_{big}. That presupposes two things: (1) one is willing to use a sample size of N_{big} if necessary and (2) one is not willing to use a sample size of $N_{\text{big}} + 1$ under any circumstances. If one is willing to make these assumptions, then it is absolutely true that the group sequential design is preferable to the adaptive design. However, Lehmacher and Wassmer (1999) and Cui, Hung, and Wang (1999) showed how to improve any group sequential design with a given maximum sample size N_{big} by allowing a sample size increase such that if the original design is maintained, inference will be identical to the group sequential design. For example, suppose you specify a maximum sample size of $N_{\text{big}} = 200$, with an interim look after each group of 50 observations. Now suppose at the halfway point, you decide you want a little additional power by adding 30 observations. You are free to do so as long as you apply the group sequential boundaries to Z_1, $(Z_1 + Z_2)/2^{1/2}$, $(Z_1 + Z_2 + Z_3)/3^{1/2}$, and $(Z_1 + Z_2 + Z_3 + Z_4)/4^{1/2}$, where $Z_1, \ldots, Z_4$ are the z-scores from each of the four stages. If you decide not to increase the sample size, the boundary is the same as for the group sequential trial with no option to increase the sample size. Thus, if you feel that you can specify a number N_{big} such that you might use a sample size of N_{big} but under no circumstances would you use $N_{\text{big}} + 1$ or more, then group sequential methods are superior to adaptive methods. If you are not in that situation, adaptive methods are superior because they allow the possibility of a sample size increase while maintaining the original boundaries if the sample size is not changed.

Adaptive designs based on the treatment effect are not a panacea. They should be used only when very little information is known about the expected treatment effect and/or the minimally relevant effect. Burman and Sonesson raise important concerns that underscore the need for care when using very flexible methods. They are correct that such methods, if extremely abused, produce illogical conclusions, but that is no more a condemnation of adaptive methods than Jack the Ripper is a condemnation of cutlery.

Proof that no test based on the sufficient statistic can maintain level α irrespective of whether the prespecified sample size rule is followed.

For a given sample size function N, write the sufficient statistic as $(N, S_1 + S_2)$, where S_1 and S_2 are sums of the observations in stages 1 and 2, respectively. Any level-α test based on the sufficient statistic is of the form $\cup_n (N = n, S_1 + S_2 \in R_n)$ with

$$\begin{aligned}
\alpha &= \sum_n \Pr(N = n, S_1 + S_2 \in R_n) \\
&= \sum_n \Pr\left(N = n, \frac{S_2}{\sqrt{n - n_1}} \in \frac{R_n - S_1}{\sqrt{n - n_1}}\right) \\
&= \sum_n \Pr\left(N = n, Z_2 \in \frac{R_n - \sqrt{n_1} Z_1}{\sqrt{n - n_1}}\right) \\
&= \sum_n \int_{-\infty}^{\infty} \Pr\left(N(Z_1) = n, Z_2 \in \frac{R_n - \sqrt{n_1} Z_1}{\sqrt{n - n_1}} \,\middle|\, Z_1 = z_1\right) \phi(z_1)\, dz_1 \\
&= \sum_n \int_{-\infty}^{\infty} I(N(z_1) = n) \Pr\left(Z_2 \in \frac{R_n - \sqrt{n_1} z_1}{\sqrt{n - n_1}}\right) \phi(z_1)\, dz_1, \\
&= \sum_n \int_{-\infty}^{\infty} I(N(z_1) = n) A(z_1, n) \phi(z_1)\, dz_1, \qquad (1)
\end{aligned}$$

where Z_1 and Z_2 are the z-scores for the data of stages 1 and 2, respectively, $(R_n - n_1^{1/2} z_1)/(n - n_1)^{1/2}$ denotes the set of points $([x - n_1^{1/2} z_1]/[n - n_1]^{1/2} : x \in R_n)$, and

$$\begin{aligned}
A(z_1, n) &= \Pr\left(Z_2 \in \frac{R_n - \sqrt{n_1} z_1}{\sqrt{n - n_1}}\right) \\
&= \int_{-\infty}^{\infty} I\left(z_2 \in \frac{R_n - \sqrt{n_1} z_1}{\sqrt{n - n_1}}\right) \frac{\exp\left(-z_2^2/2\right)}{\sqrt{2\pi}}\, dz_2. \qquad (2)
\end{aligned}$$

It is important to keep in mind that in (2), n and z_1 are realized values of random variables, and are therefore not random (that is why we were able to drop the conditioning statement $Z_1 = z_1$ in the steps leading to (1)).

Now suppose we abandon the original sample size rule, but we continue to use the same rejection sets R_n. For the type I error rate to remain α irrespective of the sample size rule actually used, any two such rules N_1 and N_2 must yield the same type I error rate. For fixed z_1^*, define

$$N_1(z_1) = \begin{cases} k_1 & \text{if } z_1 \in \left[z_1^*, z_1^* + \epsilon\right] \\ m & \text{if } z_1 \notin \left[z_1^*, z_1^* + \epsilon\right] \end{cases}$$

$$N_2(z_1) = \begin{cases} k_2 & \text{if } z_1 \in \left[z_1^*, z_1^* + \epsilon\right] \\ m & \text{if } z_1 \notin \left[z_1^*, z_1^* + \epsilon\right]. \end{cases}$$

From (1), the difference in type I error rates for these two sample size rules is

$$\int_{z_1^*}^{z_1^* + \epsilon} \{A(z_1, k_1) - A(z_1, k_2)\} \phi(z_1)\, dz_1.$$

Because ϵ is arbitrary, for this integral to be 0, $A(z_1^*, k_1) = A(z_1^*, k_2)$. In other words, $A(z_1^*, n)$ does not depend on n.

Because z_1^* was also arbitrary, $A(z_1,\ n) = A(z_1)$ does not depend on n for any z_1.

I next show that $A(z_1,\ n)$ does not depend on z_1 either. Write $A(z_1,\ n)$ as $\Pr(Z_2 \in \frac{R_n}{(n-n_1)^{1/2}} - \epsilon_n(z_1))$, where $\epsilon_n(z_1) = n_1^{1/2} z_1/(n\ -\ n_1)^{1/2}$. Because $\epsilon_n(z_1) \to 0$ as $n \to \infty$, it is not difficult to show that $A(z_1, n) - \Pr(Z_2 \in \frac{R_n}{(n-n_1)^{1/2}}) \to 0$ as $n \to \infty$. Thus,

$$\begin{aligned} A(z_1) = A(z_1, n) &= \lim_{n\to\infty} A(z_1, n) \\ &= \lim_{n\to\infty} \left\{ \Pr\left(Z_2 \in \frac{R_n}{\sqrt{n-n_1}} \right) \right. \\ &\qquad \left. + A(z_1, n) - \Pr\left(Z_2 \in \frac{R_n}{\sqrt{n-n_1}} \right) \right\} \\ &= \lim_{n\to\infty} \Pr\left(Z_2 \in \frac{R_n}{\sqrt{n-n_1}} \right). \end{aligned} \tag{3}$$

In other words, $A(z_1)$ does not depend on z_1.

Recapping, $A(n,\ z_1)$ depends on neither n_1 nor z_1; it must be a constant. Fix n and write $A(n,\ z_1)$ as $\Pr(X \in R_n\}$, where $X = (n - n_1)^{1/2} Z_2 + n_1^{1/2} z_1$ is normally distributed with mean $\mu = \mu(z_1) = n_1^{1/2} z_1$ and fixed variance $\sigma^2 = n\ -\ n_1$. As z_1 ranges from $-\infty$ to ∞, so does μ. Thus, R_n is a set such that $\Pr(\mathrm{N}(\mu, \sigma^2) \in R_n)$ is the same for all $\mu \in (-\infty,\ \infty)$. This clearly implies that $\Pr(X_n \in R_n)$ is either 0 or 1, and because the value is the same for each n, the right-hand side of equation (1) is either 0 or 1 instead of α. As this is a contradiction, it cannot be that the original rejection region based on the sufficient statistic maintains level α irrespective of whether the original sample size rule is followed, completing the proof.

References

Burman, C.-F. and Sonesson, C. (2006). Are flexible designs sound? *Biometrics* doi: 10.1111/j.1541-0420.2006.00626.x.

Cui, L., Hung, H. M., and Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55,** 853–857.

Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on observed treatment effect. *Statistics in Medicine* **22,** 971–993.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55,** 1286–1290.

Proschan, M. A. (2004). The geometry of two-stage tests. *Statistica Sinica* **13,** 163–177.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51,** 1315–1324.

Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90,** 367–378.

P. Bauer

Section of Medical Statistics
Medical University of Vienna
A-1090 Vienna, Austria
email: *peter.bauer@meduniwien.ac.at*

The authors discuss some of the known problems related to flexible designs based on the combination test (Bauer, 1989; Bauer and Köhne, 1994) or conditional error function principle (Proschan and Hunsberger, 1995). For a single adaptive interim analysis, in both approaches the test statistic combines two statistics calculated from the sample units observed before and after an adaptive interim analysis (see, e.g., Posch and Bauer, 1999). How to combine the two statistics must be laid down in advance. Also some important features of the distributions of the test statistics under the null hypothesis have to be preserved by the adaptations. For example, when stagewise *p*-values are used they follow independent uniform distributions on [0, 1], or, if stagewise *z*-scores are used they follow independent standard normal distributions. The principle is recursive, because an additional adaptive interim analysis could be introduced for the remainder of the trial after the first interim analysis, again fixing in advance how the forthcoming two stagewise test statistics will be combined (Brannath, Posch, and Bauer, 2002). The method allows flexible changes in ongoing trials without compromising on the type I error rate.

Sample size modification has attracted most interest. Because the modified sample size may depend on the first-stage test statistic, it is not possible to use the actual sample size for combining the stagewise test statistics. Therefore the combination test will not be based on the sufficient statistic. Observations collected before and after sample size modification will be weighted differently, a (known) property criticized by the authors. They concede that to preserve the full flexibility of adaptive designs it is typically impossible to characterize the distribution of the resulting sample size: "In this case we cannot construct a test with a correct type I error rate that is based solely on the minimal sufficient statistics."

It may not be reasonable to force midtrial sample size modification as a rule. The conditional power calculated at the interim effect estimate generally will be a strongly biased estimate of the true conditional power (Bauer and König, 2006). The corresponding modified sample sizes can be highly variable with a large expectation (Jennison and Turnbull, 2003). Tsiatis and Mehta (2003) have shown that there is always a group sequential design in a sense more efficient than a design with a prespecified sample size reassessment rule. Burman and Sonesson point out a weakness of this result: costs and impact of performing many interim analyses are not accounted for. They do not believe that increasing the power in group-sequential trials (choosing relatively small a priori effect sizes) is the general answer. They mention situations where sample size modification may be a useful option.

In Section 4.2 the article briefly refers to proposals from the literature on how to avoid decisions of adaptive tests that are in conflict with decisions derived from the common test statistics ("dual test"): reject, if and only if the adaptive test and

the common test based on the overall sufficient statistics both reject at the level α. By the way, Example 2 refers to sample size modification after 100 experimental units, reducing the recruitment from 900 to 1 after having observed a negative effect! This is a misuse of the method. We could misuse also other types of inference, for example, by always choosing an unrealistically optimistic prior so that according to Bayesian inference the decision does not require any experiment. Possible directional conflicts and how to deal with it have been discussed from the very beginning (Bauer and Köhne, 1994). In reasonable adaptive designs the dual-test principle will not be associated with a prohibitive loss of power. The authors mention that there may be a subset of the sample space where sample size reassessment will never lead to inflation of the type I error rate of the conventional test based on the sufficient statistics. If for the test of a normal mean (variance known) we follow a rule that in the case of large observed interim effects the sample size is increased (or decreased in the case of a small effect) the conventional test never rejects if the adaptive test does not reject (see Posch, Bauer, and Brannath, 2003, Figure 5). However, the sample size rules that are likely to be applied in practice (increase it in the case of a small observed effect, or decrease it in the case of a large effect) may lead to anticonservative conventional tests. But with such rules the dual test does not lose power: a rejection of the adaptive test is always accompanied by the rejection of the conventional test. Here we will get large overall means that generally will be biased (Brannath, König, and Bauer, 2006).

A way to maintain the conventional test statistics in flexible designs is to adjust the level by considering the scenario that produces the maximum type I error rate. Such an adjusted likelihood ratio test with rejection region $(Z \geq \Phi^{-1}[1-\alpha_*])$, $\alpha_* < \alpha$, will be conservative. It can be improved upon by using an adaptive test that fully exploits the level α and rejects uniformly more often in any point of the sample space (Brannath et al., 2006, Section 4.2.3). This can be seen by considering the conditional error of the likelihood ratio test (Proschan and Hunsberger, 1995; Müller and Schäfer, 2001):

$$\mathrm{CE}_{\mathrm{lr},\alpha} = \Pr\mathrm{ob}(\sqrt{n_1/(n_1+n_2)}z_1 + \sqrt{n_2/(n_1+n_2)}Z_2 \geq z_{1-\alpha}),$$

where the probability is taken over the second-stage standardized mean Z_2 and $z_{1-\alpha}$ denotes the $(1-\alpha)$ quantile of the standard normal distribution. By always taking the second-stage sample size $n_2 = n_2(z_1)$ to maximize the conditional type I error rate $\mathrm{CE}_{\mathrm{lr},\alpha}$ (given n_1, α, z_1) we maximize its overall type I error rate. The resulting sample size reassessment rule is the worst-case scenario for which α_* has to be adjusted. It leads to the maximum conditional type I error rate of the likelihood ratio test $\mathrm{CE}_{\mathrm{max},\alpha}$, which is just a function of z_1. Taking the expectation of $\mathrm{CE}_{\mathrm{max},\alpha}$ over z_1 gives the maximum type I error rate, which exceeds the targeted level α. The adjusted likelihood ratio test applies a level $\alpha_* < \alpha$, such that the expectation of $\mathrm{CE}_{\mathrm{max},\alpha_*}$ over z_1 is equal to α. This $\mathrm{CE}_{\mathrm{max},\alpha_*}$ has been derived by Proschan and Hunsberger (1995) for an unconstrained second-stage sample size (assuming stopping for futility). In an interim analysis we can replace this "worst-case design" by another second-stage design with conditional type I error rate $\mathrm{CE}_{\mathrm{max},\alpha}$ leading to an adaptive test. Note that whenever we deviate from the worst-case sample size reassessment rule, the conditional error rate $\mathrm{CE}_{\mathrm{lr},\alpha_*}$ of the adjusted likelihood ratio test will be smaller than the maximum conditional type I error rate $\mathrm{CE}_{\mathrm{max},\alpha_*}$ used in the adaptive test. Hence the adaptive test with rejection region $(Z_2 \geq \Phi^{-1}[1-\mathrm{CE}_{\mathrm{max},\alpha_*}])$ rejects uniformly more often than the adjusted likelihood ratio test with rejection region $(Z_2 \geq \Phi^{-1}[1-\mathrm{CE}_{\mathrm{lr},\alpha_*}])$. We could take this as an indication that in flexible designs the use of likelihood ratio test statistics may not be most efficient in terms of power. Clearly, the unadjusted likelihood ratio test always rejects if the adjusted test rejects. Hence a dual adaptive test may also uniformly improve the adjusted test based on the likelihood ratio statistics. This seems to be a remarkable property.

A crucial methodological issue not addressed by the authors is estimation. Because the adaptation rules need not be specified a priori, there is no predefined sample space. Solutions have been discussed mainly for sample size modifications (e.g., Brannath et al., 2006) but further clarification is necessary. Note that the problem of bias does not arise from using unconventional test statistics but from the adaptation itself.

One of the motivations to deal with flexible designs has been the practice of performing design modifications by writing amendments to the study protocol without fully understanding the statistical impact. The real merits of flexible designs will be adaptations going beyond sample size reassessment. Examples of such adaptation are dropping treatments, changing doses, modifying the test statistics, shifting interest to subgroups, inserting or skipping interim analyses, or even such controversial options such as modifying the primary endpoint (see, e.g., Posch et al., 2003). It should be possible (although an ambitious exercise in real studies), for example, to predefine formally specific midtrial treatment selection rules based on variables quantifying efficacy, safety, costs, and ethical issues. However, flexibility may be needed because experimenters may not adhere to a formal rule in an environment of evolving information. In the original papers (Bauer, 1989; Bauer and Köhne, 1994; Bauer and Röhmel, 1995) changes in the null hypothesis were allowed. Hence general adaptive test procedures are testing an intersection null hypothesis and problems of interpretation arise following a rejection. Inference on the individual null hypotheses based on the closed testing principle can be performed (Bauer and Kieser, 1999; Hommel, 2001). Note that in adaptive designs only stagewise models have to apply. Complications in the interpretation of results in such complex designs are a trade-off for the variety of options offered by the design.

The concept of adaptive designs allows design modifications at any (unscheduled) time by replacing the remainder of the design by one which preserves the conditional type I error of the preplanned design (Müller and Schäfer, 2004). This corresponds to a "recursive" continuous application of combination tests (defined implicitly by the preplanned design). The conditional error principle will also lead to a deviation from the sufficient statistics, but it can be applied to introduce flexibility into conventional group sequential trials. We may plan to combine stagewise test statistics as in a group sequential trial. Then, if no design adaptation is performed the conventional group sequential analysis will apply. This corresponds to using the "inverse normal combination function" for stagewise p-values (Cui, Hung, and Wang, 1999; Lehmacher and Wassmer, 1999). Preservation of the conditional error rate is an ideal tool to deal with the unexpected. However, the exact

calculation of the conditional error function will be difficult if nuisance parameters are involved (for the *t*-test see Posch et al., 2004). Another precaution refers to situations with a delayed endpoint, if surrogate information for patients still waiting for their endpoint is used for adaptation (Bauer and Posch, 2004). Here strictly one has to work with the distribution of the forthcoming endpoints given the information used in the adaptation (Liu and Pledger, 2006), which may be difficult in practice.

To include "learning from experience" into a design may be an ethical or economic issue. The designs critiqued by the authors are a general tool to handle flexibility during an ongoing trial without sacrificing type I error control. Flexibility is needed in practice and the statistical price to be paid for the many options offered by the designs is known. The authors have summarized the ongoing discussion. Is wine sound? It is the way to drink it that matters. For adaptive designs it seems that it is the way they are used, presented, and interpreted that matters.

Acknowledgements

I thank W. Brannath, F. König, and M. Posch for their helpful comments.

References

Bauer, P. (1989). Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie* **20,** 130–148.

Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* **18,** 1833–1848.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50,** 1029–1041.

Bauer, P. and König, F. (2006). The reassessment of trial perspectives from interim data—A critical view. *Statistics in Medicine* **25,** 23–36.

Bauer, P. and Posch, M. (2004). Modification of the sample size and the schedule of interim analyses in survival trials based on data (letter to the editor). *Statistics in Medicine* **23,** 1333–1334.

Bauer, P. and Röhmel, J. (1995). An adaptive method for establishing a dose-response relationship. *Statistics in Medicine* **14,** 1595–1607.

Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* **97,** 1–9.

Brannath, W., König, F., and Bauer, P. (2006). Estimation in flexible two stage designs. *Statistics in Medicine*, in press.

Burman, C.-F. and Sonesson, C. (2006). Are flexible designs sound? *Biometrics* doi: 10.1111/j.1541-0420.2006.00626.x.

Cui, L., Hung, H. M. J., and Wang, S. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55,** 321–324.

Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* **43,** 581–589.

Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* **22,** 971–993.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55,** 1286–1290.

Liu, Q. and Pledger, G. (2006). On design and inference for two-stage adaptive clinical trials with dependent data. *Journal of Statistical Planning and Inference* **136,** 1962–1984.

Müller, H. H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57,** 886–891.

Müller, H. H. and Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* **23,** 2497–2508.

Posch, M. and Bauer, P. (1999). Adaptive two stage design and the conditional error function. *Biometrical Journal* **41,** 689–696.

Posch, M., Bauer, P., and Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine* **22,** 953–969.

Posch, M., Timmesfeld, N., König, F., and Müller, H. H. (2004). Conditional rejection probabilities of student's t-test and design adaptations. *Biometrical Journal* **46,** 389–403.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extensions of studies based on conditional power. *Biometrics* **51,** 1315–1324.

Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90,** 367–378.

Marianne Frisén

Statistical Research Unit
Göteborg University
SE 40530 Göteborg, Sweden
email: *marianne.frisen@statistics.gu.se*

1. Introduction

The article by Burman and Sonesson (B&S for short) is an important contribution because it takes up logical inference issues and gives good examples illustrating the problems.

To change a design might be well motivated. This has earlier caused great concern because it might hurt the trust of the study. Thus, it has been avoided as much as possible.

The paper by Bauer and Köhne (1994) on a test that keeps the significance level in spite of changes has many more than 100 citations. This indicates that many papers on this subject have followed. This also indicates how great the hope is for the possibility of flexible but valid studies.

The earlier criticism has been concentrated on inefficiency. This is important, but logical faults are of great scientific and practical concern. Like B&S, I will concentrate on the logical questions, but I will concentrate on questions related to the conditionality principle rather than the sufficiency principle. The main issues will be illustrated by their example of sample size modification by a test of the effect of a drug, and the same notation will be used.

2. Inference Principles

To have a fixed significance level is not enough. Statistical inference is a complicated kind of logic. The paper by B&S, and most other papers on flexible design, are set within the frequentist framework. General principles such as sufficiency and conditioning are important and have been much discussed. Even though no complete agreement has been reached on the exact formulation of the principles, a violation of commonly accepted principles should be a serious warning. It is important in the study of a new drug that the results are not manipulated or even suspected of being manipulated.

The sufficiency principle tells us that it is inefficient to use a statistic that is not sufficient for the problem. An example of a violation of the sufficiency principle is the use of the median instead of the mean when estimating the expected value of a normal distribution.

The conditionality principle concerns the danger of letting the conclusion depend on distributions that are ancillary for the problem or to disregard important ones. According to Fraser (2004) the conditionality principle is more fundamental than the sufficiency principle even though the latter is better known. To concentrate on the main issues I describe the case where the effect size μ is the only parameter. The sample size N is an ancillary variable for inference about the effect μ if the distribution of N does not depend on μ. As an example, consider a randomized test where the conclusion about the hypothesis depends on flipping a coin. This violates the conditionality principle and is seldom used in practice. The outcome of the coin flipping is an ancillary statistic and thus should not be allowed to influence the conclusion.

Sometimes a very minor change of the problem can produce an agreement with the inference principles. It is pointed out by Cox and Hinkley (1974) that it is of interest to see whether a statistic is approximately ancillary in some sense. A random variable (corresponding to the sample size) can be considered as an approximately ancillary statistic if it would have been exactly ancillary if some less important boundary values of the sample space were excluded. This was discussed by Frisén in connection with a clinical study using matched pairs. Reid (2003) gave many examples of approximate ancillaries based mainly on asymptotic considerations.

3. The Modification Does Not Depend on the Parameter of Interest

Here, the change of design is totally unrelated to anything depending on μ. Examples could be an administrative mistake or an unexpected cutting of funds by reasons unrelated to μ. In this case N is an ancillary statistic and one should, by the conditionality principle, condition on N and disregard the fact that N is stochastic. It is thus correct to use what is called the "naive" test if N is an ancillary statistic.

4. The Modification Depends on the Parameter of Interest

If the change of design is related to the effect, then one has to be very careful. If the distribution of N depends on μ, then N is not ancillary for conclusions on μ. N might be approximately ancillary but that is a separate issue.

The case where the modification depends on the observations achieved so far is the most obvious and challenging issue. However, the same kind of problem can arise if the change depends on external information related to the problem of interest. Some, but not all, suggested methods require that the design rule is known and the distribution of N can be used.

4.1 *Design Rule Not Reported*

Flexible designs have been advocated as being totally flexible for which one does not even have to know or report the change. Is there some method that is sound even if one withholds this information?

The likelihood principle allows a totally flexible sampling plan, for example, one can sample until significance (e.g., $Z > 1.96$) is achieved. In this case one will, with probability 1, get a significant result even when there is no effect.

The "weighted" test avoids this by forcing a certain error spending. This is done at the cost of violating the conditionality principle. The ordering of the observations is an ancillary statistic for a conclusion about the hypothesis. Thus, by the conditionality principle the test statistic should not depend on the ordering of the realized observations. B&S's Example 2 illustrates the possibilities for manipulation of significance, if one can give different weights to equally informative observations. In this example, the rule used was to increase the conditional power for small values of μ, when the first observations indicate that the drug is ineffective. The same can happen when the design rule is dependent on the parameter of interest through external information. Suppose that in the middle of a study, external information is obtained that makes it clear that the drug is worthless. By the weighted test, one now has the possibility of ending the study and, by only the small cost of one additional observation, still has a chance to announce that the drug is significantly better than the old one.

The "dual" method is a modification of the weighted test and cuts the edges so that very drastic examples cannot be constructed. However, it suffers from violation of the same inference principles. One can manipulate the significance by a combination of the likelihood principle for the unweighted test and by violating the conditionality principle for the weighted one. If one gets internal or external evidence that the drug has no effect, then one is allowed to sample with extremely low weight until $Z > 1.96$, and then take one extra observation with the large remaining weight and thus have the chance to announce that the drug is significantly superior.

Whether the conditional or unconditional power is more relevant has been discussed by Frisén (1980) and Brannath

and Bauer (2004). However, without a known rule for the design it is impossible to give the unconditional one, even if that is the more logical one.

4.2 *Design Rule Known*

If the design rule can be determined before the experiment is performed, then methods such as a group sequential test have several advantages. However, methods for the case where the design rule was not known until after the experiment was performed are of interest.

To use the full likelihood for the problem and consider the observed value of N and the distribution of N, as suggested by Burman and Sonesson, is a good approach. The likelihood ratio (LR) test can be used for different kinds of designs even if some design rules will result in a dubious method. If the rule was to do as in B&S's Example 2, namely, to increase the power for very low values of μ if one gets indications that the drug is very bad, then one will have a very peculiar test and a full report on the power function reveals that the design was not sound. The problem of deciding whether N is approximately ancillary for the problem (see Section 2) is not specific for the LR method.

The "unweighted test" utilizes the distribution of N to choose a critical level with the desired probability of rejecting the null hypothesis unconditional on N but disregards the observed value of N. Because N is part of the minimal sufficient statistic, it is not in accordance with the sufficiency principle and is inefficient. This is demonstrated by a slight difference in unconditional power in Example 4. However, the logical issue might be still more disturbing. To withhold that you know that you have performed a small experiment with little information, just because you might have performed a large-scale one, is not sound and is not in accordance with accepted inference principles. In my view the unweighted test is not viable.

To use a "conditional" test (see B&S, Example 3) when N is not an ancillary statistic violates the conditionality principle.

5. Conclusions

There is no easy way to handle an unplanned change in design. To gain control over the significance level is not good enough. There are very misleading procedures that have a controlled significance level. The concentration on a solution to get a fixed significance level takes the focus away from the important issue of evaluating information.

There is a correspondence between the conditionality and the sufficiency principles. However, sufficiency focuses on efficiency. Analysis of conditionality focuses on logical issues. Ancillary information should not be allowed to influence conclusions. However, this is the case when one gives equally informative observations on different weights. The danger of this is well demonstrated by B&S's Example 2, where the last observation gets a much larger weight than the others and this disturbs the correspondence between significance and a reasonable conclusion. The remarkable result is not that the efficiency is low but that manipulation can produce a higher power for very low values of μ and thus provide a chance for a poor drug to look good.

My conclusion is that one should use planned adaptive designs when one expects that it will be necessary to adapt the design. If one unexpectedly has strong reasons to change the plans, one should be very careful and give full information about that. It is not sound to be so flexible unnecessarily that methods which considerably violate accepted inference principles have to be used. Changes in designs should be avoided because there is no way to totally avoid that the trust in the results is damaged.

References

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50,** 1029–1041.

Brannath, W. and Bauer, P. (2004). Optimal conditional error functions for the control of conditional power. *Biometrics* **60,** 715–723.

Burman, C.-F. and Sonesson, C. (2006). Are flexible designs sound? *Biometrics* doi: 10.1111/j.1541-0420.2006.00626.x.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Fraser, D. A. S. (2004). Ancillaries and conditional inference. *Statistical Science* **19,** 333–369.

Frisén, M. (1980). Consequences of the use of conditional inference in the analysis of a correlated contingency table. *Biometrika* **67,** 23–30.

Reid, N. (2003). Asymptotics and the theory of inference. *Annals of Statistics* **31,** 1695–1731.

Rejoinder

Carl-Fredrik Burman
and Christian Sonesson

We are grateful for the stimulating comments and to the editors for finding such insightful and diverse discussants. Although quite different views are expressed, there is relative consensus on some central issues. All contributors seem to agree that flexibility is sometimes needed but that unrestricted use of the weighted test is not sound.

The inferential issues discussed in the article are of fundamental importance to the science of statistics. If the statistical community would find flexible designs to be statistically sound when applied to clinical trials, all experiments where data accumulate over time would be affected. For example, having counted half of the responses for an opinion poll, the investigator could decide to increase the sample size. Even if the sample size is not changed, it might have been, and an orthodox frequentist analysis can therefore not be performed, as the sample space is not defined.

1. Is the Weighted Test, Following Sample Size Modifications, Generally Valid?

It is well known that the weighted test violates the sufficiency principle. A main point in our article is that the weighted test may also lead to paradoxical conclusions. This is compellingly demonstrated by Proschan and Hunsberger (1995) and Example 2 in our article, where a significant positive effect is shown although the average response is negative. Consequently, all the discussants agree that the weighted test should not be applied in an unrestricted way.

2. Are Prespecified Sample Size Modifications and the Weighted Test Inadmissible?

We agree with Jennison and Turnbull (2006) that the weighted test, not being based on the sufficient statistic, is inadmissible and that adaptive choices of sample sizes lead to modest gains compared to a group sequential design with equal number of interim analyses. However, we would not agree to label a sample size modification (SSM) "invalid" if it is shown to be inefficient. Data from an inefficient design may be highly convincing to the scientific community. Consequently, we think that it is valuable to discuss the analysis of data generated by a preplanned SSM design even if we do not recommend this design. The examples in this discussion and the article are chosen to illustrate the inferential issues, not good designs.

3. Do We Have to Adhere to Inference Principles?

In the article, we have emphasized the sufficiency principle but we also briefly mentioned the invariance principle, which in our context implies that the *order* in which exchangeable observations are collected should not affect the analysis (Cox and Hinkley, 1974, p. 41–42). The conditionality principle implies the same thing. Frisén regards the conditionality principle as more fundamental than the sufficiency principle. We tend to agree. For a general distribution, not belonging to the exponential family, the weighted test will violate the conditionality principle but not the sufficiency principle. We focused on the sufficiency principle, rather than the invariance or conditionality principle, because we were thinking about the problem of exchangeability in a sequential trial, a topic also discussed by Proschan. For the examples in our article the invariance principle is clearly violated, because at least one observation is always taken after the first interim analysis. Proschan "agree(s) that one should follow the sufficiency principle" if the SSM "has been prespecified and must be followed." However, he argues that greater flexibility is needed and that we should always anticipate that a prespecified rule is not strictly followed.

Frisén stresses the importance of inference principles and states that "a violation of commonly accepted principles should be a serious warning." This is a useful formulation; we should not take the inference principles as absolute rules but rather use them as guidance when assessing whether proposed analyses are sensible. In many cases, the violation of inference principles will lead to unacceptable consequences. One such case is shown in Example 2.

4. How Should a Trial Be Analyzed after Preplanned Sample Size Modifications?

Preplanned SSM designs provide a stimulating framework for the discussion of statistical theory and highlight some of the problems therein. Some problems are not specific to SSM—as Frisén points out regarding the likelihood ratio (LR) test—but can be well illustrated in this framework. Both Frisén and our article focus on inferential logic, while Jennison and Turnbull emphasize efficiency. For the sufficiency principle, credibility and efficiency go hand in hand. However, the conditionality principle may sacrifice power for validity. In some cases, N is approximately ancillary, and the test should then be performed conditional on N. In other cases, however, N is highly informative about the effect and the conditional test is then not sound. When using SSM, N is typically partly informative and it is not obvious which test should be chosen.

Bauer discusses an "adjusted likelihood ratio test" and concludes that "the likelihood ratio test statistics may not be efficient." It should be pointed out that Bauer studies a test based on Z alone, and not on the LR statistic derived in our article. This statistic depends also on N. For each one-point alternative hypothesis, a true LR test is most powerful by Neymann–Pearson's lemma. Frisén is critical toward the conditional and the unweighted tests, saying that they violate the conditionality principle and the sufficiency principle, respectively. Formally, we disagree as the conditionality principle is only applicable when an ancillary statistic exists, and the unweighted test is a function of the minimal sufficient statistic. More importantly, however, we agree that the unweighted test is problematic. The critical level c may be lower than the nominal critical level $C = \Phi^{-1}(1-\alpha)$ for the naive test. To illustrate this, take $N_1 = 1$ and N_2 as either 1 or 1000 depending on whether Z_1 is negative or positive, respectively. Then under the null hypothesis, $P(Z > C \mid N = 2) \ll \alpha$ and $P(Z > C \mid N = 1001) \approx \alpha$. Consequently, $c < C$. If $\alpha = 5\%/2$, then $c = 1.69$ while $C = 1.96$. There is a continuum of LR tests, for different alternatives μ'. In the simple example above with only two possible sample sizes, both the conditional test and the weighted test belong to this class. The conditional type I error given that $N = 2$ is less than, greater than, or equal to α, depending on whether μ' is large, close to 0, or at a certain intermediate value. Intuitively, if both μ' and N are large, then the conditional type I error can be decreased while retaining the conditional power close to 1. It can be noted that all admissible tests have at least one of the critical limits smaller than C, as a result of the example's counterintuitive increase in sample size for large interim effects. Furthermore, and similarly with the weighted test, examples can be constructed where the conditional test or an LR test have a negative critical value for some values of N.

5. Is There Any Valid Frequentist Inference Following Nonprespecified Sample Size Modifications?

Flexibility is sometimes needed, at least to respond to important unforeseen external information. When the SSM rule is not predefined, however, the analysis is even more problematic. Proschan shows that no test based on the sufficient statistic can maintain the significance level in this situation.

Traditionally, frequentist statistics have required the sample space to be fully specified. The weighted test tries to overcome this problem by requiring predefined weights and by modeling the sample space of the p-values from different stages. While strongly promoting flexibility, Proschan states that "the greater the flexibility we allow, the more care we must exercise to ensure a sensible test." Having agreed that unrestricted use of the weighted test is not always sensible, we should therefore turn to consider whether a modified weighted test could be acceptable.

6. Is the Dual Test Sound?

The dual test, requiring that both Z^w and Z lie in the rejection region, joins frequentist α-level protection with Bayesian focus on the data at hand, ignoring sampling properties. For a fixed critical limit C, $Z^w > C$ implies $Z > C$ for a large class of SSMs. Bauer states that "in reasonable designs the dual test will not be associated with a prohibitive loss of power" compared to the weighted test. SSM followed by the dual test may be inefficient when compared to a nonflexible group sequential design for a fixed alternative. However, in the possibly rare occasions where flexibility is needed, we do not have a fixed alternative. External factors may radically change our mind about, for example, which magnitude of treatment effects is desirable to detect.

Jennison and Turnbull point out that the weighted and naive tests may disagree. The main problem is the case when Z is high while Z^w is smaller, because high observations have been downweighted. The data may then be convincing from a Bayesian perspective while the dual test is nonsignificant. This is a risk for the "producer," not the "consumer." We therefore view this problem as considerably smaller than the risk that significant results from a weighted test may be communicated while a low value of Z is ignored.

Frisén argues against the dual test. Although agreeing that it avoids very extreme examples, she presents an interesting example that combines a common criticism of Bayesianism with our Example 2. First, sample until $Z > 1.96$, which will occur almost surely under the null hypothesis. Then take one single observation with very high weight to have a chance of getting $Z^w > 1.96$. This example could be seen as artificial; under the null hypothesis the expected number of observations is infinite, yet the power is only α. However, one might interchange the two stages in the example. Start by taking one observation with weight close to 1. If the first-stage p-value p_1 is less than α, continue sampling until both $Z > C$ and N is at least some minimum number. The results of this trial would seem to be convincing if the only results communicated are the total sample size and the dual test p-value.

A large number of trials could be started and ignored whenever $p_1 < \alpha$. Of course, the trial could have been stopped after one observation and a significant effect could be declared. This problem is not new and extremely small trials will not be convincing. The new problem is that by continuing the experiment, the sample size seems to be large enough to make the results seem credible.

The core of the problem is that Z^w is essentially based on one observation and not, as it would appear, on N observations. It might be useful to define an "effective sample size" when interpreting the results from a weighted test. Remember that $Z^w = \Sigma\sqrt{w_k}X_k$. Note that the true conditional or unconditional variances of Z^w are undefined if the SSM rule is not prespecified and may be highly misleading even if the SSM rule is prespecified. If instead we naively ignore that the weights depend on the observations the obvious definition, based on the signal-to-noise ratio, of an effective sample size would be $N_{\text{eff}} = (\Sigma\sqrt{w_k})^2$. The same expression can be motivated by considering the length of the confidence interval (e.g., Brannath, König, and Bauer, 2006) based on the correspondence theorem. For Example 2, with 101 observations, we would have $N_{\text{eff}} = 16.9$, indicating that much information is wasted but still that the first stage contributes markedly to the weighted Z score. In the more likely practical example of equally weighted stages with 100 and 200 observations, $N_{\text{eff}} = 291.4$ is rather close to N. How to best define the effective sample size and whether the concept is a useful one is an open question.

7. Point Estimates and Confidence Intervals

Bauer points out that "a crucial methodological issue not addressed by the authors is estimation." In the article we wanted to focus on the fundamental question of whether the hypothesis test is valid. If that is not the case, the corresponding point and interval estimates are of limited interest. The correspondence theorem provides a rather general way to construct confidence intervals and median unbiased estimates (Brannath et al., 2006, Sections 3.2 and 4.1.1). In Example 2, this method gives the point estimate +0.554 and 95% symmetric confidence interval (+0.077, +1.031). Given that the average of the observations is in fact negative, –0.005, this analysis is not convincing. However, the length of the confidence interval indicates how much information has been lost due to the extremely unequal weighting.

8. Are Other Design Modifications of Value?

It is somewhat unfortunate that so much interest has been focused on sample size. Bauer makes the relevant comment that "the real merits of flexible designs will be adaptations going beyond sample size reassessment." Many of these adaptations imply that observations from different stages are not exchangeable. The intersection of null hypotheses, corresponding to different parameters, may need to be tested. However, if the same parameters are used for different stages, the sufficiency principle and much of our discussion are still relevant. If, for example, the residual variance is reduced by a certain amount during the trial, the sufficiency principle would indicate how observations should be weighted.

9. Practical Guidelines for Clinical Trials

Bauer says that Example 2 is a misuse of the method, and the possibility of such misuse was exactly the intent of the example. Choosing an unrealistically optimistic prior is, following Bauer, a similar misuse of Bayesian statistics. In the Bayesian example, the solution is transparency—the sufficient statistic should be given so that the consumer may substitute his or her own prior. Jennison and Turnbull conclude their comment by asking us what we see as key criteria for a "valid" test. Using Jennison and Turnbull's own words, we believe that validity is about being "credible to the intended audience." The risk with an "invalid" test is that it is misleading. As in Bauer's

Bayesian example, clear rules and transparency are needed when reporting SSM experiments, especially in the area of clinical trials. We therefore propose the following rules, based on an ideal of transparency, as minimum requirements for the communication of all clinical trials, adaptive or not, irrespective of sponsor and aims:

1. Register all clinical trials;
2. When reporting trial data, account for the preplanned design and analysis. State and motivate any departures from the planned design and analysis; and
3. Present the results of the preferred analysis in sufficient detail in a web-based register, including, for example, the sample size and a confidence interval. Give essential parts of the sufficient statistic. For adaptive designs, give the sufficient statistic for each stage.

By requiring preregistration of all clinical trials and publishing of the results in a web-based register, we can get more reliable meta-analyses. Careful preplanning, considering the possibility of low effects also, would eliminate much of the need for SSMs. The two last rules are partly inspired by regulatory guidelines and the recently issued draft reflection paper on flexible designs (European Medicines Agency, 2006) from the EU regulatory agency. However, we place more emphasis on the sufficient statistics. For a two-sample normal distribution trial with a small number of SSMs, the main analysis should contain the p-value, point estimate, and confidence interval for the preferred analysis. In addition, the sample size, group averages, and sample variance should be given for each stage. The overall sample size may be complemented with an effective sample size. In more complicated situations, it may be impractical to give the full sufficient statistics. For an analysis of covariance (ANCOVA) model, for example, components of the sufficient statistic related to nuisance variables may have to be omitted. When SSM is performed after each observation (Fisher, 1998), it may suffice that trends in responses are explored.

Adherence to the proposed rules would make it possible to assess whether the conclusions from the primary analysis are credible. If the investigator applies SSM and uses the weighted test, it is possible for other scientists to do alternative analyses, assess homogeneity of treatment effects between stages, and combine the data with other trial results. Additional information could be valuable, but we think that the minimal requirements are a good starting point.

10. Conclusions

We think that the statistical community should strive to reach a consensus on the requirements that should be posed on the use of flexible designs and the related inference. Our article and the concrete suggestions, in this reply, for reporting of clinical trials are attempts to provoke discussion about such requirements.

References

Brannath, W., König, F., and Bauer, P. (2006). Estimation in flexible two stage designs. *Statistics in Medicine*, in press.

Burman, C.-F. and Sonesson, C. (2006). Are flexible designs sound? *Biometrics* doi: 10.1111/j.1541-0420.2006.00626.x.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

European Medicines Agency. (2006). Draft reflection paper on methodological issues in confirmatory clinical trials after flexible design and analysis plan. Available from `http://www.emea.eu.int/pdfs/human/ewp/245902en.pdf` (accessed March 2006).

Fisher, L. D. (1998). Self-designing clinical trials. *Statistics in Medicine* **17,** 1551–1562.

Jennison, C. and Turnbull, B. W. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika* **93,** 1–21.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51,** 1315–1324.