Sequential Analysis, 27: 254–276, 2008

ISSN: 0747-4946 print/1532-4176 online
DOI: 10.1080/07474940802241009

Taylor & Francis
Taylor & Francis Group

Generalized Likelihood Ratio Statistics and Uncertainty Adjustments in Efficient Adaptive Design of Clinical Trials

Jay Bartroff[1] and Tze Leung Lai[2]
[1]Department of Mathematics, University of Southern California, Los Angeles, California, USA
[2]Department of Statistics, Stanford University, Stanford, California, USA

Abstract: A new approach to adaptive design of clinical trials is proposed in a general multiparameter exponential family setting, based on generalized likelihood ratio statistics and optimal sequential testing theory. These designs are easy to implement, maintain the prescribed Type I error probability, and are asymptotically efficient. Practical issues involved in clinical trials allowing mid-course adaptation and the large literature on this subject are discussed, and comparisons between the proposed and existing designs are presented in extensive simulation studies of their finite-sample performance, measured in terms of the expected sample size and power functions.

Keywords: Hoeffding's information bound; Internal pilot; Kullback–Leibler information; Modified Haybittle–Peto test; Multiparameter exponential family; Sample size re-estimation.

Subject Classifications: 62L10; 62F03; 62P10.

1. INTRODUCTION

Because of the ethical and economic considerations in the design of clinical trials to test the efficacy of new treatments and because of lack of information on the magnitude and sampling variability of the treatment effect at the design stage, there has been increasing interest from the biopharmaceutical industry in sequential methods that can adapt to information acquired during the course of the trial. Beginning with Bauer (1989), who introduced sequential adaptive test strategies over a planned series of separate trials, and Wittes and Brittain (1990), who discussed

Received December 17, 2007, Revised February 25, 2008, and June 1, 2008, Accepted June 2, 2008
Recommended by Nitis Mukhopadhyay
Address correspondence to Tze Leung Lai, Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA; E-mail: lait@stanford.edu

internal pilot studies, a large literature has grown on adaptive design of clinical trials. Depending on the topics covered, the term "adaptive design" in this literature is sometimes replaced by "sample size re-estimation," "trial extension" or "internal pilot studies." In standard clinical trial designs, the sample size is determined by the power at a given alternative, but in practice, it is often difficult for investigators to specify a realistic alternative at which sample size determination can be based. Although a standard method to address this difficulty is to carry out a preliminary pilot study, the results from a small pilot study may be difficult to interpret and apply, as pointed out by Wittes and Brittain (1990), who proposed to treat the first stage of a two-stage clinical trial as an internal pilot from which the overall sample size can be re-estimated. The problem of sample size re-estimation based on observed treatment difference at some time before the prescheduled end of a clinical trial has attracted considerable attention during the past decade; see, e.g., Gould and Shih (1992), Herson and Wittes (1993), Birkett and Day (1994), Denne and Jennison (1999, 2000), Jennison and Turnbull (2000, Section 14.2), Shih (2001), and Whitehead et al. (2001). For normally distributed outcome variables with known variances, Proschan and Hunsberger (1995), Fisher (1998), Posch and Bauer (1999), and Shen and Fisher (1999) have proposed ways to adjust the test statistics after mid-course sample size modification so that the Type I error probability is maintained at the prescribed level. By making use of a generalization of the Neyman–Pearson lemma, Tsiatis and Mehta (2003) showed that these adaptive tests of a simple null versus a simple alternative hypothesis are inefficient because they are not based on likelihood ratio statistics. Jennison and Turnbull (2003) gave a general weighted form of these test statistics and demonstrated in simulation studies that the adaptive tests performed considerably worse than group sequential tests. Jennison and Turnbull (2006a) recently introduced adaptive group sequential tests that choose the jth group size and stopping boundary on the basis of the cumulative sample size n_{j-1} and the sample sum $S_{n_{j-1}}$ over the first $j-1$ groups and that are optimal, in the sense of minimizing a weighted average of the expected sample sizes over a collection of parameter values, subject to prescribed error probabilities at the null and a given alternative hypothesis. They showed how the corresponding optimization problem can be solved numerically by using the backward induction algorithms for "optimal sequentially planned" designs developed by Schmitz (1993). Jennison and Turnbull (2006b) found that standard (nonadaptive) group sequential tests with the first stage chosen optimally are nearly as efficient as their optimal adaptive tests.

Except for Jennison and Turnbull's optimal adaptive group sequential tests and the extensions of the sample size re-estimation approach to group sequential testing considered by Cui et al. (1999), Lehmacher and Wassmer (1999), and Denne and Jennison (2000), previous works in the literature on mid-course sample size re-estimation have focused on two-stage designs whose second-stage sample size is determined by the results from the first stage (internal pilot), following the seminal work of Stein (1945) in this area. Although this approach is intuitively appealing, it does not adjust for the uncertainty in the first-stage parameter estimates that are used to determine the second-stage sample size. Moreover, it considers primarily the special problem of comparing the means of the two normal populations, using the central limit theorem for extensions to more general situations. The case of unknown common variance at a prespecified alternative for the mean difference was considered first, as in Stein (1945) and the first set of references in the preceding paragraph. Then

the case of known variances in the absence of a prespecified alternative for the mean difference was studied, as in the second set of references above.

Bartroff and Lai (2008) recently gave a unified treatment of both cases in the general framework of multiparameter exponential families. It uses efficient generalized likelihood ratio (GLR) statistics in this framework and adds a third stage to adjust for the sampling variability of the first-stage parameter estimates that determine the second-stage sample size. Specifically, let $X_1, X_2, \ldots$ be independent d-dimensional random vectors from a multiparameter exponential family $f_\theta(x) = \exp\{\theta' x - \psi(\theta)\}$ of densities with respect to some measure ν on $\mathbf{R}^d$. Let $S_n = X_1 + \cdots + X_n$. A sufficient statistic based on $(X_1, \ldots, X_n)$ is the sample mean $\overline{X}_n = S_n/n$, which is the maximum likelihood estimate of the mean $\nabla\psi(\theta)$. Consider the hypothesis $u(\theta) = u_j$, where $j = 0$ or 1, $u : \Theta \to \mathbf{R}$ is continuously differentiable, and $\Theta = \{\theta : \int e^{\theta' x} d\nu(x) < \infty\}$ is the natural parameter space. As noted by Lai and Shih (2004, p. 513), the GLR statistic $\Lambda_{i,j}$ for total sample size n_i at stage i has the form

$$\Lambda_{i,j} = n_i\big[\theta'_{n_i}\overline{X}_{n_i} - \psi(\hat{\theta}_{n_i})\big] - \sup_{\theta:u(\theta)=u_j} n_i\big[\theta'\overline{X}_{n_i} - \psi(\theta)\big] = \inf_{\theta:u(\theta)=u_j} n_i I(\hat{\theta}_{n_i}, \theta), \tag{1.1}$$

in which $\hat{\theta}_n = (\nabla\psi)^{-1}(\overline{X}_n)$ and $I(\theta, \lambda)$ is the Kullback–Leibler information number given by

$$I(\theta, \lambda) = E_\theta\big\{\log\big[f_\theta(X_i)/f_\lambda(X_i)\big]\big\} = (\theta - \lambda)'\nabla\psi(\theta) - [\psi(\theta) - \psi(\lambda)]. \tag{1.2}$$

The possibility of adding a third stage to improve two-stage designs dated back to Lorden (1983). Whereas Lorden used crude upper bounds for the Type I error probability that are too conservative for practical applications, Bartroff and Lai (2008) overcome this difficulty by developing numerical methods to compute the Type I error probability and also extended the three-stage test to multiparameter and multi-armed settings, thus greatly broadening the scope of these efficient adaptive designs. A review of their method is given in Section 2.1. In Section 4 we prove the asymptotic optimality of these adaptive tests in the multiparameter case, extending Lorden's (1983) result for the special case $d = 1$.

Another new addition to this asymptotic optimality theory of adaptive designs is related to the problem of trial extension considered in Section 2.2. As pointed out by Cui et al. (1999), the issue of increasing the maximum sample size sometimes arises after interim analysis in group sequential trials. They cited a study protocol, which was reviewed by the Food and Drug Administration, involving a Phase III group sequential trial for evaluating the efficacy of a new drug to prevent myocardial infarction in patients undergoing coronary artery bypass graft surgery. During interim analysis, the observed incidence for the drug achieved a reduction that was only half of the target reduction assumed in the calculation of the maximum sample size M, resulting in a proposal to increase the maximum sample size to $\widetilde{M}$ ($N_{\max}$ in their notation). Cui et al. (1999) and Lehmacher and Wassmer (1999) extended the sample size re-estimation approach to adaptive group sequential trials by adjusting the test statistics as in Proschan and Hunsberger (1995), and allowing the future group sizes to be increased or decreased during interim analyses, so that the overall sample size does not exceed $\widetilde{M}$ ($>M$) and the Type I error probability is maintained at the prescribed level. In Section 2.2 we propose an

alternative approach that is shown to be asymptotically efficient in Section 4. Whereas the adaptive designs in Section 2.1 assume a given maximum sample size M and require at most three stages, Section 2.2 extends them to allow mid-course increase of M to $\widetilde{M}$. These adaptive designs involve no more than four stages and an adjustment in the maximum sample size is made at the third stage. Computational algorithms to implement them are provided in Section 2.3 and simulation results on their performance are given in Section 3.4.

Bartroff and Lai (2008) carried out comprehensive simulation studies of the performance, measured in terms of the expected sample size and power functions, of their adaptive test and compared it with other adaptive tests in the literature. In the case of normal mean with known variance and Type I and II error constraints under the null and a given alternative hypothesis, they showed that their adaptive test is comparable to the benchmark optimal adaptive test of Jennison and Turnbull (2006a,b), which is superior to the existing two-stage adaptive designs. On the other hand, whereas the benchmark optimal adaptive test needs to assume a specified alternative, these adaptive two-stage tests and the adaptive tests of Bartroff and Lai (2008) do not require such assumptions as they consider the estimated alternative at the end of the first stage. In their recent survey of adaptive designs, Burman and Sonesson (2006) pointed out that previous criticisms of the statistical properties of two-stage adaptive designs may be unconvincing in some situations when flexibility and not having to specify parameters that are unknown at the beginning of a trial (like the relevant treatment effect or variance) are more imperative than efficiency or being powerful. The adaptive designs in Bartroff and Lai (2008) and this article can fulfill the seemingly disparate requirements of flexibility and efficiency on a design. Rather than achieving exact optimality at a specified collection of alternatives through dynamic programming, they achieve asymptotic optimality over the entire range of alternatives, resulting in near optimality in practice. They are based on efficient test statistics of the GLR type, which have an intuitively "adaptive" appeal via estimation of unknown parameters by maximum likelihood, ease of implementation, and freedom from having to specify the relevant alternative; see Section 2.

Bauer and Einfalt (2006, p. 504) have found from a search of the medical literature that adaptive designs have not been widely used in practice and that "adaptations in practice are rather limited to sample size reassessment." Perhaps one reason why these two-stage adaptive designs have not gained wide acceptance is their use of seemingly unnatural and convoluted test statistics (e.g., the inefficient test statistics mentioned in the first paragraph). This can be circumvented by the use of efficient GLR statistics in our adaptive tests with no more than three (or four) stages. Another reason may be the lack of routine medical studies to which adaptive designs can lead to substantial improvements over current practice. In Section 3.1 we consider one such potential application in Phase II cancer studies. We show how our adaptive designs offer improvements over Simon's (1989) optimal two-stage designs, which are commonly used in single-arm cancer trials, and over their analogs, due to Thall et al. (1988), for randomized trials in Section 3.2. Section 3.3 considers another issue that often arises in the design of clinical trials, namely, nuisance parameters. "Very often, statistical information also depends on nuisance parameters (e.g., the standard deviation of the response variable). Extension of statistical information is a design adaptation that occurs most frequently" (Hung et al., 2006, p. 566). The simulation results in Section 3.3 show how our adaptive

test resolves the difficulties with conventional two-stage designs to treat nuisance parameters, which are estimated at the end of the first stage and then used to estimate the second stage sample size. Further discussion of our proposed approach to adaptive designs and some concluding remarks are given in Section 5.

2. EFFICIENT ADAPTIVE DESIGN AND GLR TESTS

2.1. An Adaptive Three-Stage GLR Test

Whereas Tsiatis and Mehta (2003) consider the case of simple null and alternative hypotheses $\theta = \theta_j$ $(j = 0, 1)$ for which likelihood ratio tests are most powerful even in their group sequential designs, Bartroff and Lai (2008) use the GLR statistics (1.1) in an adaptive three-stage test of the composite null hypothesis $H_0 : u(\theta) \le u_0$, where u is a smooth real-valued function such that

$$I(\theta, \lambda) \text{ is increasing in } u(\lambda) \text{ for every fixed } \theta. \tag{2.1}$$

Let $n_1 = m$ be the sample size of the first stage (or internal pilot study) and $n_3 = M$ be the maximum total sample size, both specified before the trial. Let $u_1 > u_0$ be the alternative implied by the maximum sample size M and the reference Type II error probability $\tilde{\alpha}$. That is, u_1 $(>u_0)$ is the alternative where the fixed sample size (FSS) GLR test with Type I error probability α and sample size M has power $\inf_{\theta:u(\theta)=u_1} P_\theta\{\text{Reject } H_0\}$ equal to $1 - \tilde{\alpha}$, as in Lai and Shih (2004, Section 3.4). The three-stage test of $H_0 : u(\theta) \le u_0$ stops and rejects H_0 at stage $i < 2$ if

$$n_i < M, \quad u(\hat{\theta}_{n_i}) > u_0 \quad \text{and} \quad \Lambda_{i,0} \ge b. \tag{2.2}$$

Early stopping for futility (accepting H_0) can also occur at stage $i < 2$ if

$$n_i < M, \quad u(\hat{\theta}_{n_i}) < u_1 \quad \text{and} \quad \Lambda_{i,1} \ge \tilde{b}. \tag{2.3}$$

The test rejects H_0 at stage $i = 2$ or 3 if

$$n_i = M, \quad u(\hat{\theta}_M) > u_0 \quad \text{and} \quad \Lambda_{i,0} \ge c, \tag{2.4}$$

accepting H_0 otherwise. The sample size n_2 of the three-stage test is given by

$$n_2 = m \vee \left[M \wedge \lceil (1 + \rho_m) n(\hat{\theta}_m) \rceil\right] \tag{2.5}$$

with

$$n(\theta) = \min\left[|\log \alpha| / \inf_{\lambda:u(\lambda)=u_0} I(\theta, \lambda), |\log \tilde{\alpha}| / \inf_{\lambda:u(\lambda)=u_1} I(\theta, \lambda) \right], \tag{2.6}$$

where $\rho_m > 0$ is an inflation factor to adjust for uncertainty in $\tilde{\theta}_m$; see the examples in Section 3. Letting $0 < \varepsilon, \tilde{\varepsilon} < 1$, define the thresholds $b, \tilde{b}$ and c to satisfy the equations

$$\sup_{\theta:u(\theta)=u_1} P_\theta[(2.3) \text{ occurs for } i = 1 \text{ or } 2] = \tilde{\varepsilon}\tilde{\alpha}, \tag{2.7}$$

$$\sup_{\theta:u(\theta)=u_0} P_\theta[(2.3) \text{ does not occur for } i \le 2, (2.2) \text{ occurs for } i = 1 \text{ or } 2] = \varepsilon\alpha, \tag{2.8}$$

$$\sup_{\theta:u(\theta)=u_0} P_\theta[(2.2) \text{ and } (2.3) \text{ do not occur for } i \le 2, (2.4) \text{ occurs}] = (1 - \varepsilon)\alpha. \tag{2.9}$$

The probabilities in (2.7)–(2.9) can be computed by using the normal approximation to the signed-root likelihood ratio statistic

$$\ell_{i,j} = \{\text{sign}[u(\hat{\theta}_{n_i}) - u_j]\}(2n_i\Lambda_{i,j})^{1/2},$$

($1 \le i \le 3$ and $j = 0, 1$) under $u(\theta) = u_j$, as in Lai and Shih (2004, p. 513). When $u(\theta) = u_j$, $\ell_{i,j}$ is approximately normal with mean 0, variance n_i, and the increments $\ell_{i,j} - \ell_{i-1,j}$ are asymptotically independent. We can therefore approximate $\ell_{i,j}$ by a sum of independent standard normal random variables under $u(\theta) = u_j$ and thereby determine b, $\tilde{b}$ and c. Note that this normal approximation can also be used for the choice of u_1 implied by M and $\tilde{\alpha}$. Computational details are given in Section 2.3, as well as an alternate method for computing the thresholds by Monte Carlo.

A special multiparameter case of particular interest in clinical trials involves K independent populations having density functions $\exp[\theta_k x - \tilde{\psi}_k(\theta_k)]$ so that $\theta' x - \psi(\theta) = \sum_{k=1}^K [\theta_k x_k - \tilde{\psi}(\theta_k)]$. In multi-armed trials, for which different numbers of patients are assigned to different treatments, the GLR statistic $\Lambda_{i,j}$ for testing the hypothesis $u(\theta_1, \ldots, \theta_K) = u_j$ ($j = 0$ or 1) at stage i has the form

$$\Lambda_{i,j} = \sum_{k=1}^K n_{ki}[\tilde{\theta}_{k,n_{ki}}\overline{X}_{k,n_{ki}} - \tilde{\psi}(\hat{\theta}_{k,n_{ki}})] - \sup_{\theta:u(\theta_1,\ldots,\theta_K)=u_j} \sum_{k=1}^K n_{ki}[\theta_k \overline{X}_{k,n_{ki}} - \tilde{\psi}(\theta_k)],$$

in which n_{ki} is the total number of observations from the kth population up to stage i. Let $n_i = \sum_{k=1}^K n_{ki}$. As pointed out in Section 3.4 of Lai and Shih (2004), the normal approximation to the signed root likelihood ratio statistic is still applicable when $n_{ki} = p_k n_i + O_p(n_i^{\frac{1}{2}})$, where $p_1, \ldots, p_K$ are nonnegative constants that sum up to 1, as in random allocation of patients to the K treatments (for which $p_k = 1/K$).

2.2. Mid-Course Modification of Maximum Sample Size

We now modify the adaptive designs in the preceding section to accommodate the possibility of mid-course increase of the maximum sample size from M to $\widetilde{M}$. Let u_2 be the alternative implied by $\widetilde{M}$ so that the level-α GLR test with sample size $\widetilde{M}$ has power $1 - \tilde{\alpha}$. Note that $u_1 > u_2 > u_0$. Whereas the sample size n_3 is chosen to be M in Section 2.1, we now define

$$\tilde{n}(\theta) = \min\Big[|\log\alpha| / \inf_{\lambda:u(\lambda)=u_0} I(\theta,\lambda), |\log\tilde{\alpha}| / \inf_{\lambda:u(\lambda)=u_2} I(\theta,\lambda)\Big],$$

$$n_3 = n_2 \vee \big[M' \wedge \lceil(1+\rho_m)\tilde{n}(\hat{\theta}_{n_2})\rceil\big],$$

where $M < M' \le \widetilde{M}$ and $n_2 = m \vee [M \wedge (1+\rho_m)\tilde{n}(\hat{\theta}_m)]$. We can regard the test as a group sequential test with four groups and $n_1 = m$, $n_4 = \widetilde{M}$, but with adaptively chosen n_2 and n_3. If the test does not end at the third stage, continue to the fourth and final stage with sample size $n_4 = \widetilde{M}$. Its rejection and futility boundaries are similar to those in Section 2.1. Extending our notation $\Lambda_{i,j}$ in (1.1) to $1 < i < 4$ and $0 \le j < 2$, the test stops at stage $i \le 3$ and rejects H_0 if

$$n_i < \widetilde{M}, \quad u(\hat{\theta}_{n_i}) > u_0, \quad \text{and} \quad \Lambda_{i,0} \ge b, \tag{2.10}$$

stops and accepts H_0 if

$$n_i < \widetilde{M}, \quad u(\hat{\theta}_{n_i}) < u_2, \quad \text{and} \quad \Lambda_{i,2} \geq \tilde{b}, \tag{2.11}$$

and rejects H_0 at stage $i = 3$ or 4 if

$$n_i = \widetilde{M}, \quad u(\hat{\theta}_{\widetilde{M}}) > u_0, \quad \text{and} \quad \Lambda_{i,0} \geq c, \tag{2.12}$$

accepting H_0 otherwise. The thresholds $b, \tilde{b}$ and c can be defined by equations similar to (2.7)–(2.9) to insure the overall Type I error probability to be α. For example, in place of (2.7),

$$\sup_{\theta: u(\theta) = u_2} P_\theta[(2.11) \text{ occurs for some } i \leq 3] = \tilde{\varepsilon}\tilde{\alpha}. \tag{2.13}$$

The basic idea underlying (2.13) is to control the Type II error probability at u_2 so that the test does not lose much power there in comparison with the GLR test that has sample size $\widetilde{M}$ (and therefore power $1 - \tilde{\alpha}$ at u_2).

2.3. Implementation via Normal Approximation or Monte Carlo

To begin with, suppose the X_i are $N(\theta, 1)$ and $u(\theta) = \theta$. We write θ_j instead of u_j and, without loss of generality, we shall assume that $\theta_0 = 0$. The thresholds $b, \tilde{b}$ and c of the three-stage test in Section 2.1 can be computed by solving in succession (2.7), (2.8), and (2.9). Univariate grid search or Brent's method (Press et al., 1992) can be used to solve each equation. Since $I(\theta, \lambda) = (\theta - \lambda)^2/2$, we can rewrite (2.7) as

$$P_{\theta_1}\big[S_m - m\theta_1 > -(2\tilde{b}m)^{\frac{1}{2}}, S_{n_2} - n_2\theta_1 \leq -(2\tilde{b}n_2)^{\frac{1}{2}}\big] + P_{\theta_1}\big[S_m - m\theta_1 \leq -(2\tilde{b}m)^{\frac{1}{2}}\big] = \tilde{\varepsilon}\tilde{\alpha},$$

and (2.8) and (2.9) as

$$P_0\big[S_m/(2m)^{\frac{1}{2}} \geq b^{\frac{1}{2}}\big] + P_0\big[\tilde{b}^{\frac{1}{2}} < S_m/(2m)^{\frac{1}{2}} < b^{\frac{1}{2}}, S_{n_2}/(2n_2)^{\frac{1}{2}} \geq b^{\frac{1}{2}}, n_2 < M\big] = \varepsilon\alpha,$$

$$P_0\big[\tilde{b}^{\frac{1}{2}} < S_m/(2m)^{\frac{1}{2}} < b^{\frac{1}{2}}, n_2 < M, \tilde{b}^{\frac{1}{2}} < S_{n_2}/(2n_2)^{\frac{1}{2}} < b^{\frac{1}{2}}, S_M/(2M)^{\frac{1}{2}} \geq c^{\frac{1}{2}}\big]$$
$$+ P_0\big[b^{\frac{1}{2}} < S_m/(2m)^{\frac{1}{2}} < b^{\frac{1}{2}}, n_2 = M, S_M/(2M)^{\frac{1}{2}} \geq c^{\frac{1}{2}}\big] = (1 - \varepsilon)\alpha.$$

The probabilities involving n_2 can be computed by conditioning on the value of $S_m/m = x$, which completely determines the value of n_2, denoted by $k(x)$. For example, the probabilities under $\theta = 0$ can be computed via

$$P_0\big[S_{n_2} \geq (2bn_2)^{\frac{1}{2}} \mid S_m = mx\big] = P\big\{N(0,1) \geq \big[2bk(x)n_2^{\frac{1}{2}} - mx\big]/[k(x) - m]^{\frac{1}{2}}\big\}, \tag{2.14}$$

$$P_0(S_{n_2} \in dy, S_M \in dz \mid S_m = mx) = \varphi_{k(x)-m}(y - mx)\varphi_{M-k(x)}(z - y)dy\,dz, \tag{2.15}$$

where φ_v is the $N(0, v)$ density function; i.e., $\varphi_v(w) = (2\pi v)^{-\frac{1}{2}} \exp(-w^2/2v)$. The probabilities under θ_1 can be computed similarly. Hence, standard recursive numerical integration algorithms can be used to compute the probabilities in (2.7)–(2.9); see Jennison and Turnbull (2000, Section 19.2). For the general

multiparameter exponential family, this method can be used to compute the thresholds $b, \tilde{b}$ and c for (2.2)–(2.4) since the problem can be approximated by that of testing a normal mean, as discussed in Section 2.1.

For mid-course modification of the maximum sample size in Section 2.2, the above recursive numerical algorithm can be modified to handle the randomness of n_2 and n_3. The basic idea is that conditional on $S_m/m = x$, the value of n_2 is completely determined as $k(x)$, and conditional on $S_m/m = x$ and $S_{n_2}/n_2 = y$, the value of n_3 is completely determined as $h(x, y)$. Therefore, analogous to (2.15), we now have

$$\begin{aligned} &P(S_{n_3} \in du, S_{\tilde{M}} \in dw \mid S_m/m = x, S_{n_2}/n_2 = y) \\ &\quad = \varphi_{h(x,y)-k(x)}[u - yk(x)]\varphi_{\tilde{M}-h(x,y)}(w-u)\,du\,dw, \end{aligned} \tag{2.16}$$

and can use bivariate recursive numerical integration. For the general exponential family, normal approximation to the signed-root likelihood ratio statistic can again be used.

An alternative to normal approximation is to use Monte Carlo similar to that used in bootstrap tests. While using Monte Carlo simulations to compute error probabilities is an obvious idea, it is far from being clear which distribution from a composite hypothesis should be chosen to simulate from. Bootstrap theory suggests that we can simulate from the estimated distribution under the assumed hypothesis as the GLR statistic is an approximate pivot under that hypothesis. Since the "estimated distribution" needs data to arrive at the estimate, we make use of the first-stage data to determine b and $\tilde{b}$ and then use the second-stage data to determine c for the three-stage test in Section 2.1. Specifically, the Monte Carlo method to determine $b, \tilde{b}$ and c proceeds as follows. At the end of the first stage, compute the maximum likelihood estimate $\hat{\theta}_{m,j}$ under the constraint $u(\theta) = u_j$, $j = 0, 1$. Determine $\tilde{b}$, b and c successively by solving

$$P_{\hat{\theta}_{m,1}}[(2.3) \text{ occurs for } i = 1 \text{ or } 2] = \tilde{\varepsilon}\tilde{\alpha}, \tag{2.17}$$

$$P_{\hat{\theta}_{m,0}}[(2.3) \text{ does not occur for } i \le 2, \text{ and } (2.2) \text{ occurs for } i = 1 \text{ or } 2] = \varepsilon\alpha, \tag{2.18}$$

$$P_{\hat{\theta}_{n_2},0}[(2.2) \text{ and } (2.3) \text{ do not occur for } i \le 2, \text{ and } (2.4) \text{ occurs}] = (1-\varepsilon)\alpha, \tag{2.19}$$

noting that c does not have to be determined until after the second stage when n_2 observations become available for the updated estimate $\hat{\theta}_{n_2,0}$. The probabilities in (2.17)–(2.19) can be computed by Monte Carlo simulations. Similarly, to determine thresholds $b, \tilde{b}$ and c of the adaptive test in Section 2.2, we can use Monte Carlo simulations instead of normal approximation and numerical integration to compute the corresponding probabilities.

3. APPLICATIONS AND NUMERICAL EXAMPLES

3.1. Application to Single-Arm Phase II Cancer Trials

As pointed out by Vickers et al. (2007, p. 972), in a typical phase II study of a novel cancer treatment, "a cohort of patients is treated, and the outcomes are related to

the prespecified target or bar. If the results meet or exceed the target, the treatment is declared worthy of further study; otherwise, further development is stopped. This has been referred to as the go/no go decision. Most often, the outcome specified is a measure of tumor response; e.g., complete or partial response using Response Evaluation Criteria in Solid Tumors, expressed as a proportion of the total number of patients. Response can also be defined in terms of the proportion who have not progressed or who are alive at a predetermined time (e.g., one year) after treatment is started." The most widely used designs for these single-arm phase II trials are Simon's (1989) optimal two-stage designs, which allow early stopping of the trial if the treatment has not shown beneficial effect that is measured by a Bernoulli proportion. These designs are optimal in the sense of minimizing the expected sample size under the null hypothesis of no viable treatment effect, subject to Type I and II error probability bounds. Given a maximum sample sample size M, Simon considered the design that stops for futility after $m < M$ patients if the number of patients exhibiting positive treatment effect is r_1 ($\leq m$) or fewer, and otherwise treats an additional $M - m$ patients and rejects the treatment if and only if the number of patients exhibiting positive treatment effect is r_2 ($\leq M$) or fewer. Simon's designs require that a null proportion p_0, representing some "uninteresting" level of positive treatment effect, and an alternative $p_1 > p_0$ be specified. The null hypothesis is $H_0 : p \leq p_0$, where p denotes the probability of positive treatment effect. The Type I and II error probabilities $P_{p_0}\{\text{Reject } H_0\}$, $P_{p_1}\{\text{Accept } H_0\}$ and the expected sample size $E_{p_0}N$ can be computed for any design of this form, which can be represented by the parameter vector (m, M, r_1, r_2). Using computer search over these integer-valued parameters, Simon (1989) tabulated the optimal designs in his Tables 1 and 2 for different values of (p_0, p_1).

Whether the new treatment is declared promising in a phase II trial depends strongly on the prescribed p_0 and p_1. In their systematic review of 134 papers reporting phase II trials in *J. Clin. Cancer Res.*, Vickers et al. (2007, p. 972) found seventy papers referring to historical data for their choice of the null or alternative response rate, and that nearly half (i.e., 32) of these papers did not cite the source of the historical data used, while only nine gave clearly a single historical estimate of their choice of p_0. Moreover, no study "incorporated any statistical method to account for the possibility of sampling error or for differences in case mix between the phase II sample and the historical cohort." The adaptive designs in Section 2.1 applied to this setting do not require the specification of the alternative p_1, a desirable property to prevent well-intentioned but misguided practitioners from choosing p_1 artificially small to inflate the appearance of a positive treatment effect, if one exists; uncertainty in the choice of p_0 is also an important issue and is addressed in Section 3.2. For now, assume p_0 to be given along with initial and maximum sample sizes m and M. The adaptive test takes p_1 to be the alternative where the FSS test, with Type I error probability α at p_0, has power $1 - \beta$; i.e., the solution of $F_{M,p_1}(F^{-1}_{M,p_0}(1-\alpha)) = \beta$, where $F_{M,p}$ is the distribution function of the Bin(M, p) distribution. The GLR statistic (1.1) at the ith stage is

$$n_i\left[\hat{p}_{n_i}\log\left(\frac{\hat{p}_{n_i}}{p_j}\right) + (1-\hat{p}_{n_i})\log\left(\frac{1-\hat{p}_{n_i}}{1-p_j}\right)\right]$$

for $j = 0$ or 1, and the critical values $b, \tilde{b}$ and c are chosen to satisfy (2.7)–(2.9). Because of discreteness of the binomial distribution it may be impossible

Table 1. Description of ADAPT and Sim2 for two cases

S_m	ADAPT	Sim2 ($p_1 = .3$ or .44)
	(a) $m = 10, M = 29, p_0 = .1, \alpha = .05, \beta = .2$	
≤ 1	Accept H_0.	Accept H_0.
2	$n_2 = M$; reject H_0 if $S_{n_2} \geq 6$.	$n_2 = M$ and
3	$n_2 = 20$	reject H_0 if $S_M \geq 6$.
	(i) If $S_{n_2} \leq 3$, accept H_0.	
	(ii) If $S_{n_2} \geq 6$, reject H_0.	
	(iii) If $4 \leq S_{n_2} \leq 5$ and $S_M \geq 6$, reject H_0.	
≥ 4	Reject H_0.	$n_2 = M$; rej. H_0 if $S_M \geq 6$.
	(b) $m = 30, M = 82, p_0 = .1, \alpha = \beta = .1$	
≤ 8	Accept H_0.	Accept H_0.
9	$n_2 = 57$	Accept H_0.
	(i) If $S_{n_2} \leq 19$, accept H_0.	
	(ii) If $S_{n_2} \geq 24$, reject H_0.	
	(iii) If $20 \leq S_{n_2} \leq 23$ and $S_M \geq 32$, reject H_0.	
10–13	$n_2 = M$, reject H_0 if $S_{n_2} \geq 31$.	$n_2 = M$ and
≥ 14	Reject H_0.	reject H_0 if $S_M \geq 30$.

to satisfy (2.7)–(2.9) exactly, in which case (2.7) is satisfied approximately and (2.8)–(2.9) are satisfied conservatively. The stopping rule defined by (2.2)–(2.4) may alternatively be stated in terms of the number of cumulative successes S_{n_i} at the ith stage. Table 1 describes the adaptive design (denoted by ADAPT) and Simon's (1989) optimal two-stage design (denoted by Sim2) for two choices of m, M, α, β, p_0, and Table 2 contains their operating characteristics, computed exactly using the Bin(n, p) distribution. ADAPT has expected sample size close to Sim2 for p near p_0 and smaller sample size when p is roughly midway between p_0 and p_1 or is larger;

Table 2. Expected sample size, power (in parentheses), and expected number of stages (in brackets) of phase II designs

p	ADAPT			Sim2		
	(a) $m = 10, M = 29, p_0 = .1, \alpha = .05, \beta = .2$					
.05	11.6	(.3%)	[1.1]	11.6	(.2%)	[1.1]
$p_0 = .1$	14.5	(5%)	[1.3]	15.0	(4.7%)	[1.3]
.2	18.8	(43.3%)	[1.6]	21.9	(43.1%)	[1.6]
$p_1 = .3$	18.1	(79.4%)	[1.6]	26.1	(79.6%)	[1.8]
.4	14.8	(94.9%)	[1.4]	28.1	(95.0%)	[2.0]
.5	12.1	(98.9%)	[1.2]	28.8	(98.9%)	[2.0]
.6	10.1	(99.9%)	[1.0]	29.0	(99.9%)	[2.0]
	(b) $m = 30, M = 82, p_0 = .3, \alpha = \beta = .1$					
.2	34.9	(.3%)	[1.1]	33.2	(.03%)	[1.1]
$p_0 = .3$	51.8	(10.0%)	[1.5]	51.4	(10.0%)	[1.4]
.35	60.4	(35.0%)	[1.7]	63.4	(36.2%)	[1.6]
$p_1 = .44$	52.9	(88.7%)	[1.5]	77.7	(87.8%)	[1.9]
.5	42.4	(98.4%)	[1.3]	80.9	(97.5%)	[2.0]
.6	31.9	(99.9%)	[1.0]	82.0	(99.9%)	[2.0]

$p_1 = .3$ in the top panel of Table 2 and $p_1 = .44$ in the bottom panel. It is not surprising that the expected sample size of Sim2 increases with p since Sim2 only stops early for futility. The expected number of stages shows a similar pattern, while their power functions are nearly identical. Note that even though ADAPT has a maximum of three stages, its expected number of stages is less than two for all p and usually close to one.

3.2. Extension to Randomized Phase II Cancer Trials

As noted by Vickers et al. (2007), uncertainty in the choice of p_0 and p_1 can increase the likelihood that (a) a treatment with no viable positive treatment effect proceeds to phase III, or (b) a treatment with positive treatment effect is abandoned at phase II. To circumvent the problem of choosing an artificially small p_0, either intentionally by a practitioner wanting to give the treatment the "best chance" of showing a positive effect if one exists, or unintentionally because of inaccurate information about the control, Ellenberg and Eisenberger (1985) proposed to perform a controlled two-arm phase II trial in which patients are randomized into both treatment and control groups. After randomizing $2n_1$ patients into treatment and control arms, the trial is stopped for futility when the number in the treatment group showing positive effect is not greater than in the control group. Otherwise, the trial continues until a total of $2n_2$ patients have been randomized into the study and then a standard fixed sample binomial test is performed. Letting p and q denote the probability of positive effect in the treatment and control groups, respectively, Thall et al. (1988) subsequently chose n_1, n_2 and two other design parameters y_1 and y_2, described below, to minimize the "average" expected sample size

$$\text{AvSS} = \frac{1}{2}[E(N \mid p = q) + E(N \mid p = q + \delta)] \tag{3.1}$$

subject to Type I and II error probability constraints α and β. The two-stage test of $H_0 : p \le q$ stops for futility after the first stage if an approximately normally distributed test statistic Z_1, based on the two-population binomial data, is no greater than y_1, otherwise continuing with a second stage and rejecting H_0 if $Z_2 > y_2$. Because the expectations in (3.1) depend on p and q, both must be specified to calculate the trial design.

The adaptive designs of Section 2 apply naturally to this two-arm setting. The GLR statistic (1.1) at the ith stage is

$$n_i\Bigg[\hat{p}_i \log\left(\frac{\hat{p}_i}{\hat{p}_{\delta_j}}\right) - (1-\hat{p}_i)\log\left(\frac{1-\hat{p}_i}{1-\hat{p}_{\delta_j}}\right) + \hat{q}_i \log\left(\frac{\hat{q}_i}{\hat{p}_{\delta_j} - \delta_j}\right) - (1-\hat{q}_i)\log\left(\frac{1-\hat{q}_i}{1-\hat{p}_{\delta_j} + \delta_j}\right)\Bigg],$$

where $\hat{p}_i, \hat{q}_i$ are the fraction of successes in the treatment and control arms and $\hat{p}_{\delta_j}$ is the MLE of p under $p - q = \delta_j$, $j = 0, 1$, where $\delta_0 = 0$ and δ_1 is the implied alternative (see the example below). The boundaries $b, \tilde{b}$ and c can be computed using a normal approximation or Monte Carlo simulations as described

in Section 2.3; the latter, with 1 million simulations for each probability calculation, is used in the following comparative study.

Thall and Simon (1994) describe a phase II trial of fludarabine + ara-C + granulocyte colony stimulating factor (G-CSF) for treatment of acute myelogenous leukemia (AML). Although this trial was designed by using other methods described in their paper, we use it here as a real setting to compare the adaptive design with Thall et al.'s (1988) design. The standard therapy for AML at the time was fludarabine + ara-C alone, and in a preceding study, 22 out of 45 patients achieved complete remission of the leukemia, the clinical endpoint of interest, suggesting an initial estimate of q of .5. The study was conducted to detect an increase in remission rate of $\delta = .2$. For $\alpha = .05$ and $\beta = .2$, Thall et al.'s (1988) optimal two-stage design (denoted by Opt2) for detecting a 20% improvement when the control remission rate q is .5 has first stage of 33 per arm, followed by a second stage of 45 per arm if $Z_1 > y_1 = .356$, and rejecting the null hypothesis after the second stage if $Z_2 > y_2 = 1.584$. The adaptive design (denoted by ADAPT) with first stage $m = 25$ and maximum sample 78 per arm, the same as Opt2, uses boundaries $b = 2.12$, $\tilde{b} = 1.03$ and $c = 1.56$ for $\alpha = .05$, $\tilde{\alpha} = .2$ and $\varepsilon = \tilde{\varepsilon} = 1/2$. Table 3 contains the operating characteristics of ADAPT and Opt2 for a variety of treatment and control response rates (p, q) around (.7, .5). Each result is based on 100,000 simulations. ADAPT has substantially smaller expected sample size than Opt2. This is in part because Opt2 only stops early for futility, although the parameters of Opt2 in this case are chosen to minimize (3.1), yet there is substantial savings both when $p = q$ and $p = q + \delta$. ADAPT and Opt2 have similar expected number of stages near the null hypothesis, with ADAPT decreasing as $p - q$ increases while Opt2 steadily increases to 2, again due to the latter's early stopping only for futility. The power

Table 3. Expected sample size, power (in parentheses), expected number of stages (in brackets) and average expected sample size (3.1) (denoted by AvSS) of 2-arm phase II designs

q	p	ADAPT			Opt2		
.4	.3	33.3	(0.4%)	[1.1]	37.8	(0.2%)	[1.1]
	.4	46.1	(5.3%)	[1.5]	48.9	(5.3%)	[1.4]
	.5	57.5	(32.3%)	[1.8]	63.3	(35.6%)	[1.7]
	.6	56.4	(76.0%)	[1.8]	73.5	(78.9%)	[1.9]
	.7	43.8	(97.0%)	[1.5]	77.3	(97.7%)	[2.0]
AvSS		51.3			61.2		
.5	.4	34.7	(0.4%)	[1.2]	38.2	(0.2%)	[1.1]
	.5	47.3	(5.0%)	[1.5]	49.0	(5.6%)	[1.4]
	.6	57.5	(32.2%)	[1.8]	63.3	(35.5%)	[1.7]
	.7	55.1	(77.8%)	[1.8]	73.7	(80.4%)	[1.9]
	.8	41.0	(97.6%)	[1.4]	77.5	(98.2%)	[2.0]
AvSS		51.2			61.4		
.6	.5	34.7	(0.4%)	[1.2]	38.2	(0.2%)	[1.1]
	.6	46.0	(5.2%)	[1.5]	48.9	(5.3%)	[1.4]
	.7	55.8	(33.2%)	[1.7]	63.3	(35.6%)	[1.7]
	.8	52.3	(81.1%)	[1.7]	74.4	(84.2%)	[1.9]
	.9	35.9	(98.5%)	[1.3]	77.8	(99.4%)	[2.0]
AvSS		49.2			61.7		

functions of the tests are similar, with Opt2 having slightly higher power. Note that the Type I error probability of Opt2 is inflated above $\alpha = .05$ at $p = q = .5$ due to the normal approximations to Z_i used to compute the design parameters.

3.3. Comparison with Adaptive Tests for Difference of Means with Unknown Variances

Let $X_i, X_2, \ldots$ and $Y_i, Y_2, \ldots$ be independent normal observations with unknown means μ_X, μ_Y and variances σ_X^2, σ_Y^2, respectively. Even if the variances σ_X^2, σ_Y^2 are assumed equal, no fixed sample size test of the hypothesis $\mu_X < \mu_Y$ can achieve specified error probabilities $\alpha, \tilde{\alpha}$ at $\mu_X - \mu_Y$ equal to 0 and some specified $\delta > 0$ without knowing the true value of the variance, as demonstrated by Dantzig (1940). To overcome this difficulty in the case of a single normal mean, Stein (1945) proposed a two-stage procedure that uses the first-stage sample to estimate the variance. The total sample size is then determined as a function of this estimate so that the test statistic, which uses the data from both stages to estimate the mean but only the first stage to estimate the variance, is exactly t-distributed under $\mu_X - \mu_Y = 0$ at which the test has Type I error α; the power of the test at $\mu_X - \mu_Y = \delta$ is strictly larger than $1 - \tilde{\alpha}$. Stein's procedure can be easily extended to the mean difference problem if $\sigma_X^2 = \sigma_Y^2$, and has been modified for clinical trials by Wittes and Brittain (1990), Birkett and Day (1994), and Denne and Jennison (1999). These modifications of Stein's procedure use the overall sample to estimate both the mean difference and the common variance in modifying Stein's test statistic and therefore may not maintain the Type I error probability at the prescribed level. Wittes and Brittain (1990) also assume a prior estimate σ_0^2 (e.g., from previous studies in the literature) of the common variance to modify Stein's formula for the total sample size.

In Table 4 we compare the performance of the adaptive test in Section 2.1 (denoted by ADAPT) with Stein's test, denoted by S, and the modified versions of Wittes and Brittain (1990, denoted by WB), Birkett and Day (1994, denoted by BD), and Denne and Jennison (1999, denoted by DJ) in the context of a phase II hypercholesterolemia treatment efficacy trial described by Facey (1992). In this trial, patients were randomized into treatment and placebo groups and serum cholesterol level reductions, X_i and Y_i, assumed to be normally distributed, were measured after four weeks of treatment. A difference in reductions of serum cholesterol levels, in mmol/liter, between the treatment and placebo groups of 1.2 was of clinical interest. Based on previous studies, it was anticipated that the standard deviation of the reductions would be about 0.7 for both groups. If the standard deviation were known to be $\sigma_0 = 0.7$, the size of the fixed sample t-test with error probabilities $\alpha = \tilde{\alpha} = .05$ at mean difference 0 and $\delta = 1.2$ is 9 per group. Following Denne and Jennison (1999), we assume a first-stage per group sample size of $m = 5$, which is approximately half of 9. If the standard deviation were in fact $2\sigma_0 = 1.4$, the per group sample size of the same t-test is 31, which we take as a reasonable maximum sample size M for our three-stage test with $\rho_m = .1$ and $\varepsilon = \tilde{\varepsilon} = 1/3$. Table 4 contains the power and per group expected sample size of ADAPT and the aforementioned procedures in the literature, evaluated by 100,000 simulations at various values of $\mu_X - \mu_Y \in [0, \delta]$ and $\sigma = \sigma_X = \sigma_Y$. Whereas the Stein-type tests require this assumption of equal variances, the three-stage tests defined in Section 2.1 do not, so for comparison

Table 4. Power and per group expected sample size of tests of $H_0 : \mu_X \leq \mu_Y$

$(\mu_X - \mu_Y, \sigma)$ I	S	WB	BD	DJ	ADAPT	ADAPT$_{\neq}$
$(0, \sigma_0/2)$	5.0%	4.9%	5.0%	5.0%	1.6%	1.8%
$I = 0$	5.0	9.0	5.0	5.0	5.0	5.0
$(0, \sigma_0)$	5.0%	5.4%	6.0%	5.6%	4.0%	4.1%
$I = 0$	10.1	10.1	10.1	10.3	9.4	10.2
$(\delta/2, \sigma_0)$	53.0%	59.5%	55.4%	57.3%	65.0%	68.1%
$I = .169$	10.1	10.1	10.1	10.3	15.5	13.7
(δ, σ_0)	96.6%	98.0%	95.8%	96.4%	97.8%	98.5%
$I = .551$	10.1	10.1	10.1	10.3	9.4	8.0
$(0, 2\sigma_0)$	5.0%	5.5%	5.5%	4.6%	5.0%	5.3%
$I = 0$	38.2	38.2	38.2	30.7	22.1	22.7
$(\delta/2, 2\sigma_0)$	50.3%	50.0%	49.7%	53.8%	44.0%	44.5%
$I = .045$	38.1	38.2	38.2	30.7	25.5	26.0
$(\delta, 2\sigma_0)$	95.2%	89.1%	91.3%	92.9%	91.9%	91.4%
$I = .169$	38.1	38.2	38.2	30.7	22.1	20.2
$(0, 3\sigma_0)$	5.0%	5.3%	5.3%	4.6%	5.1%	5.2%
$I = 0$	85.2	85.2	85.3	67.7	26.3	26.7
$(0, 5\sigma_0)$	5.0%	5.4%	5.4%	4.7%	5.2%	5.1%
$I = 0$	236	235	236	186	27.6	28.6
$(0, 10\sigma_0)$	5.0%	5.4%	3.8%	4.9%	5.1%	5.1%
$I = 0$	942	940	943	754	28.7	29.3

we also include in Table 4 the three-stage test that does not assume $\sigma_X = \sigma_Y$, which we denote by ADAPT$_{\neq}$. The Kullback–Leibler information number $I = \min\{I[(\mu_X, \mu_Y, \sigma^2), (\tilde{\mu}_X, \tilde{\mu}_Y, \tilde{\sigma}^2)] : \tilde{\mu}_X - \tilde{\mu}_Y = 0\}$, where $I(\theta, \tilde{\theta})$ is defined in (1.2), is also reported in the first column. When the true standard deviations σ_X and σ_Y are equal to the specified value σ_0, ADAPT and ADAPT$_{\neq}$ have similar power but smaller expected sample size than the other tests for values of $\mu_X - \mu_Y$ near 0 and δ. When the standard deviations σ_X and σ_Y are larger than the specified value σ_0, the adaptive tests have much smaller expected sample sizes than the Stein-type tests, whose second-stage sample size increases without bound as a function of the first-stage sample variance; in particular, see the last three rows of Table 4.

An alternative approach to Stein-type designs has been used by Proschan and Hunsberger (1995) and Li et al. (2002), who simply replace the σ^2 in their two-stage tests that assume known variance with its current estimate at each stage. To compare ADAPT with these tests, which rely on stable variance estimates, we allow a larger first-stage sample size of $m = 20$. Table 5 contains the power and per group expected sample size of Proschan & Hunsberger's test (denoted by PH), two choices of the early stopping boundaries (h, k) in Table 1 of Li et al. (2002) for their test, which we denote by L1 and L2, and our three-stage test ADAPT, for various values of $\mu_X - \mu_Y$ and σ, each entry being the result of 100,000 replications. To compare these tests on equal footing we have chosen the maximum sample size $M = 121$ for ADAPT because this is the maximum sample size of L1 and is quite close to the maximum sample sizes of PH and L2, which are 122 and 104, respectively. The PH, L1, and L2 tests are designed to achieve Type I

Table 5. Maximum sample size M, power, and per group expected sample size for the tests L1 and L2 of Li et al. (2002), Proschan and Hunsberger (1995, PH), and ADAPT

$(\mu_X - \mu_Y, \sigma)$ I	L1	L2	PH	ADAPT
(0, 1)	5.5%	5.3%	5.5%	4.8%
$I = 0$	26.3	25.5	25.9	56.5
(0, 2)	5.3%	5.3%	5.4%	5.4%
$I = 0$	26.2	26.3	25.9	93.5
(1/4, 1)	29.9%	29.3%	29.0%	48.3%
$I = .016$	32.8	31.0	31.6	76.1
(3/8, 1)	49.5%	48.7%	48.3%	77.5%
$I = .035$	34.5	32.7	33.1	73.5
(1/2, 1)	67.8%	66.4%	66.4%	92.8%
$I = .061$	34.3	32.8	32.7	63.3
(1/2, 2)	12.0%	29.9%	28.9%	56.1%
$I = .016$	29.1	32.9	31.7	98.7
(3/4, 2)	49.8%	48.5%	48.1%	85.6%
$I = .035$	34.5	32.7	33.2	87.0

error probability .05 and they choose the sample size of their second stage based on a conditional power level of 80%. The threshold values $b = 2.68$, $\tilde{b} = 1.75$, $c = 1.75$ used by ADAPT are thus computed using $\alpha = .05$, $\tilde{\alpha} = .20$. The results in Table 5 show that the true power of L1, L2, and PH falls well below their nominal conditional power level of 80%. When $\sigma = 2$, the L1, L2, and PH tests have power less than 50% for all values of $\mu_X - \mu_Y$ considered, which is caused by stopping prematurely for futility at the end of the first stage; see in particular the rows in Table 5 that correspond to $\mu_X - \mu_Y = 0$. Since the conditional power criterion is not valid when the estimated difference of means is near zero, the L and PH tests must stop for futility when this occurs even though the true difference of means may be substantially greater than zero.

3.4. Comparison of Tests Allowing Mid-Course Modification of Maximum Sample Size

As pointed out in Section 1, Cui et al. (1999) have proposed a method to modify the group size of a given group sequential test of $H_0 : \theta \leq 0$ in response to protocol amendments during interim analyses. In the example considered by Cui et al. (1999, p. 854), the maximum sample size is initially $M = 125$ for detecting $\theta_1 = .29$ with power .9 and $\alpha = .025$ but can be subsequently increased up to $\widetilde{M} = 500$; their sample sizes are twice as large because they consider variance 2. They consider modifying the group size at the end of a given stage L if the ratio of conditional power at the observed alternative $\hat{\theta}_{n_L}$ to the conditional power at θ_1 is greater than 1 or less than .8, in which case the group size is then modified so that the new maximum sample size is

$$\widetilde{M} \wedge M(\theta_1/\hat{\theta}_{n_L})^2. \tag{3.2}$$

If (3.2) is less than the already sampled n_L, error spending can be used to end the trial. The crux of this method is that the original critical values can be used for the weighted test statistic without changing the Type I error probability regardless of how the sample size is changed. Table 6 compares their proposed adaptive group sequential tests with FSS tests, standard (nonadaptive) group sequential tests, and the adaptive test described in Section 2.2. Each result is based on 100,000 simulations. All adaptive tests in Table 6 use the first-stage sample size $m = 25$, maximum sample size initially $M = 125$ with the possibility of extension up to

Table 6. Power (bold), expected sample size (bold), sample size quantiles T_q, and expected number of stages (bold) of FSS, group sequential, and adaptive tests that allow mid-course modification of the maximum sample size

Test	ADAPT	FSS_{125}	FSS_{500}	OBF^5_{SC}	C^4	C^5	C^5_{SC}	C^5_{PF}
$\theta = -.03$	**0.7%**	**1.0%**	**0.3%**	**0.4%**	**1.0%**	**.8%**	**0.8%**	**0.8%**
	263.1	**125.0**	**500.0**	**305.4**	**338.5**	**361.6**	**269.1**	**187.6**
$T_{.25}$	125	125	500	200	154	185	178	117
$T_{.5}$	134	125	500	300	445	500	263	144
$T_{.75}$	500	125	500	400	500	500	381	263
#	**2.99**	**1.00**	**1.00**	**3.06**	**3.98**	**4.99**	**4.19**	**3.09**
$\theta = 0$	**2.5%**	**2.5%**	**2.5%**	**2.4%**	**2.6%**	**2.5%**	**2.5%**	**2.4%**
	316.12	**125.0**	**500.0**	**371.0**	**341.5**	**362.5**	**294.8**	**215.9**
$T_{.25}$	125	125	500	300	160	189	190	125
$T_{.5}$	500	125	500	400	457	500	309	235
$T_{.75}$	500	125	500	500	500	500	381	263
#	**3.21**	**1.00**	**1.00**	**3.71**	**3.98**	**4.98**	**4.38**	**3.42**
$\theta = .13$	**72.6%**	**28.8%**	**80.0%**	**78.8%**	**51.4%**	**54.8%**	**55.2%**	**54.1%**
	388.2	**125.0**	**500.0**	**400.2**	**279.2**	**288.6**	**285.0**	**268.5**
$T_{.25}$	250	125	500	300	100	125	125	121
$T_{.5}$	500	125	500	400	286	285	281	263
$T_{.75}$	500	125	500	500	500	500	457	381
#	**3.50**	**1.00**	**1.00**	**3.99**	**3.73**	**4.55**	**4.52**	**4.34**
$\theta = \theta_2 = .15$	**84.0%**	**36.7%**	**90.0%**	**88.8%**	**61.1%**	**64.6%**	**64.6%**	**63.4%**
	368.5	**125.0**	**500.0**	**371.3**	**285.1**	**263.6**	**262.3**	**253.1**
$T_{.25}$	208	125	500	300	94	109	109	105
$T_{.5}$	500	125	500	400	247	263	263	263
$T_{.75}$	500	125	500	500	412	381	381	381
#	**3.42**	**1.00**	**1.00**	**3.71**	**3.64**	**4.40**	**3.39**	**4.28**
$\theta = .22$	**98.6%**	**70.0%**	**99.9%**	**99.8%**	**78.4%**	**81.3%**	**81.5%**	**81.4%**
	245.0	**125.0**	**500.0**	**272.0**	**175.2**	**176.6**	**175.9**	**175.7**
$T_{.25}$	125	125	500	200	61	71	71	71
$T_{.5}$	192	125	500	300	119	125	125	125
$T_{.75}$	500	125	500	300	342	263	263	263
#	**2.88**	**1.00**	**1.00**	**2.72**	**3.26**	**3.87**	**3.86**	**3.85**
$\theta = \theta_1 = .29$	**99.6%**	**90.0%**	**100.0%**	**100.0%**	**83.8%**	**86.3%**	**86.4%**	**86.2%**
	153.8	**125.0**	**500.0**	**221.6**	**122.5**	**125.0**	**125.3**	**124.5**
$T_{.25}$	74	125	500	200	43	49	49	49
$T_{.5}$	125	125	500	200	75	88	88	86
$T_{.75}$	183	125	500	200	183	199	203	199
#	**2.43**	**1.00**	**1.00**	**2.22**	**3.00**	**3.57**	**3.57**	**3.57**

$\widetilde{M} = 500$, and Type I error probability not exceeding $\alpha = .025$, matching the setting considered in Section 2 of Cui et al. (1999). Since the maximum sample size can vary between $M = 125$ and $\widetilde{M} = 500$, the two relevant implied alternatives are $\theta_1 = .29$, where FSS_{125} has power $1 - \tilde{\alpha} = .9$, and $\theta_2 = .15$, where FSS_{500} has power .9. The values of the user-specified parameters of the tests in Table 6 are summarized below.

- ADAPT: The adaptive test, described in Section 2.2, that uses $b = 3.48$, $\tilde{b} = 2.1$, and $c = 2.31$ corresponding to $\varepsilon = \tilde{\varepsilon} = 1/2$, $\rho_m = .1$ and $M' = 250$.
- FSS_{125}, FSS_{500}: The FSS tests having sample sizes $M = 125$ and $\widetilde{M} = 500$, respectively.
- OBF^5_{SC}: A one-sided O'Brien–Fleming group sequential test having five groups of size 100 and that uses stochastic curtailment futility stopping ($\gamma = .9$ in Section 10.2 of Jennison and Turnbull, 2000) with reference alternative $\theta_2 = .15$; see below.
- C^4, C^5: Two versions the adaptive group sequential test of Cui et al. (1999) that adjusts the group size at the end of the first stage; C^4 uses four stages and C^5 uses five stages.
- C^5_{SC}, C^5_{PF}: Two modifications of C^5 to allow for futility stopping; C^5_{SC} uses stochastic curtailment futility stopping ($\gamma = .9$ in Section 10.2 of Jennison and Turnbull, 2000) and C^5_{PF} uses power family futility stopping ($\Delta = 1$ in Section 4.2 of Jennison and Turnbull, 2000). Both C^5_{SC} and C^5_{PF} use reference alternative $\theta_2 = .15$.

Since OBF^5_{SC}, C^5_{SC}, and C^5_{PF} have maximum sample size $\widetilde{M} = 500$, the futility stopping boundaries of these tests are designed to have power .9 at θ_2. We have also included C^4 because our adaptive test uses no more than four stages. The tests are evaluated at the θ values where FSS_{125} has power .01, .025, .7, .8, and .9, and where FSS_{500} has power .7, .8, and .9.

Even though the C tests have maximum sample size $\widetilde{M} = 500$, they are underpowered at $0 < \theta \le \theta_2$, the alternative implied by M, when compared with ADAPT, FSS_{500}, and OBF^5_{SC}. In particular, the C tests have power less than .65 at θ_2. Since C^4 and C^5 use no futility stopping, this suggests that their updated maximum sample size (3.2) (with $L = 1$) has contributed to the power loss. The large expected sample sizes of C^4 and C^5 at $\theta \le 0$ reveal another problem with this sample size updating rule: It does not consider the sign of $\hat{\theta}_m$; a negative value of $\hat{\theta}_m$ could result in the same sample size modification as a positive one, causing a large increase in the group size when it should be decreased toward futility stopping. ADAPT has only a slight loss of power in comparison with FSS_{500} and the five-stage OBF^5_{SC} at $\theta > 0$, with substantially smaller expected sample size. The mean number of stages of ADAPT at $\theta_1 = .29$ shows that it behaves like a two- or three-stage test there. OBF^5_{SC}, on the other hand, has the largest expected sample size at $\theta \ge 0$ of the tests in Table 6 other than FSS_{125}.

4. ASYMPTOTIC THEORY AND A MODIFIED CONDITIONAL POWER TEST

In this section we prove the asymptotic optimality of the adaptive tests in Sections 2.1 and 2.2. The proof also sheds light on how the two-stage conditional

power tests, which are shown to be severely underpowered in Section 3, can substantially increase their power by adding a third stage. These modified conditional power tests are still not asymptotically efficient because they try to mimic the optimal FSS test when the alternative is given, whereas the adaptive test of Section 2.1 tries to mimic the SPRT instead, assuming H_0 to be simple.

Theorem 4.1. *Let N denote the sample size of the three-stage GLR test in Section* 2.1, *with m, M and $m \vee [M \wedge \lceil (1+\rho_m) n(\hat{\theta}_m) \rceil]$ being the possible values of N. Let T be the sample size of any test of $H_0 : u(\theta) \le u_0$ versus $H_1 : u(\theta) \ge u_1$, sequential or otherwise, which takes at least m and at most M observations and whose Type I and Type II error probabilities do not exceed α and $\tilde{\alpha}$, respectively. Assume that* $\log \alpha \sim \log \tilde{\alpha}$,

$$m/|\log \alpha| \to a, \quad M/|\log \alpha| \to A, \quad \rho_m \to 0 \ \ \textit{but}\ m^{\frac{1}{2}} \rho_m / (\log m)^{\frac{1}{2}} \to \infty \tag{4.1}$$

as $\alpha + \tilde{\alpha} \to 0$, with $0 < a < A$. Then for every fixed θ, as $\alpha + \tilde{\alpha} \to 0$,

$$E_\theta(N) \sim m \vee \left\{ M \wedge |\log \alpha| / \left[\inf_{\lambda : u(\lambda) = u_0} I(\theta, \lambda) \vee \inf_{\lambda : u(\lambda) = u_1} I(\theta, \lambda) \right] \right\}, \tag{4.2}$$

$$E_\theta(T) \ge [1 + o(1)] E_0(N). \tag{4.3}$$

Proof. Let $\Theta_0 = [\theta : u(\theta) \le u_0]$, $\Theta_1 = [\theta : u(\theta) \ge u_1]$. By (2.1), for $i = 0, 1$,

$$\inf_{\lambda \in \Theta_i} I(\theta, \lambda) = I_i(\theta), \quad \text{where } I_i(\theta) = \inf_{\lambda : u(\lambda) = u_i} I(\theta, \lambda). \tag{4.4}$$

Take any $\lambda \in \Theta_0$ and $\tilde{\lambda} \in \Theta_1$. In view of (4.4) and Hoeffding's (1960) lower bound for the expected sample size $E_\theta(T)$ of a test that has error probabilities α and $\tilde{\alpha}$ at λ and $\tilde{\lambda}$ and takes at least m and at most M observations,

$$E_\theta(T) \ge m \vee \left\{ M \wedge \frac{[1 + o(1)] |\log \alpha|}{I_0(\theta) \vee I_1(\theta)} \right\} \tag{4.5}$$

as $\alpha + \tilde{\alpha} \to 0$ such that $\log \alpha \sim \log \tilde{\alpha}$. We next show that the asymptotic lower bound in (4.5) is attained by N. Since $m \sim a|\log \alpha|$ and $M \sim A|\log \alpha|$ and since the thresholds $b, \tilde{b}$ and c are defined by (2.7)–(2.9), we can use an argument similar to the proof of Theorem 2(ii) of Lai and Shih (2004, p. 525) to show that (4.2) holds. In fact, the second-stage sample size of the adaptive test is a slight inflation of the Hoeffding-type lower bound (4.5) with θ replaced by the maximum likelihood estimate $\hat{\theta}_m$ at the end of the first stage. The assumption $\rho_m \to 0$ but $\rho_m \succ m^{-1/2} (\log m)^{1/2}$ is used to accommodate the difference between θ and its substitute $\hat{\theta}_m$, noting that as $m \to \infty$,

$$P_\theta \{ \sqrt{m} |\hat{\theta}_m - \theta| \ge r (\log m)^{1/2} \} = o(m^{-1})$$

is r if sufficiently large, by standard exponential bounds involving moment generating functions.

We can now extend the Hoeffding-type lower bound (4.5) to establish the asymptotic optimality of the adaptive test in Section 2.2 that allows mid-course modification of the maximum sample size. This adaptive test can be regarded as a mid-course amendment of an adaptive test of $H_0 : u(\theta) \le u_0$ versus $H_1 : u(\theta) \ge u_1$,

with a maximum sample size of M, to that of H_0 versus $H_2 : u(\theta) \geq u_2$, with a maximum sample size of $\widetilde{M}$. Whereas (4.5) provides an asymptotic lower bound for tests of H_0 versus H_1, any test of H_0 versus H_2 with error probabilities not exceeding α and $\tilde{\alpha}$ and taking at least m and at most $\widetilde{M}$ observations likewise satisfies

$$E_\theta(T) \geq m \vee \left\{ \widetilde{M} \wedge \frac{[1 + o(1)]|\log \alpha|}{I_0(\theta) \vee I_2(\theta)} \right\} \tag{4.6}$$

as $\alpha + \tilde{\alpha} \to 0$ such that $\log \alpha \sim \log \tilde{\alpha}$. Note that $\Theta_1 = [\theta : u(\theta) \geq u_1] \subset \Theta_2 = [\theta : u(\theta) \geq u_2]$ and therefore $I_2(\theta) \leq I_1(\theta)$. The four-stage test in Section 2.2, with $M' = \widetilde{M}$, attempts to attain the asymptotic lower bound in (4.5) prior to the third stage and the asymptotic lower bound in (4.6) afterwards. It replaces $I_1(\theta)$ in (4.5), which corresponds to early stopping for futility, by $I_2(\theta)$ that corresponds to rejection of H_2 (instead of H_1) in favor of H_0. Thus, the second-stage sample size n_2 corresponds to the lower bound in (4.5) with θ replaced by $\hat{\theta}_m$ and I_1 replaced by I_2, while the third-stage sample size corresponds to that in (4.6) with θ replaced by $\hat{\theta}_{n_2}$. The arguments used to prove the asymptotic optimality of the three-stage test in Theorem 4.1 can be readily modified to prove the following.

Theorem 4.2. *Let N^* denote the sample size of the four-stage GLR test in Section 2.2, with $M' = \widetilde{M}$. Assume that $\log \alpha \sim \log \tilde{\alpha}$ as $\alpha + \tilde{\alpha} \to 0$, that (4.1) holds and $\widetilde{M}/|\log \alpha| \to \widetilde{A}$ with $0 < a < A < \widetilde{A}$. Then*

$$E_\theta(N^*) \sim \begin{cases} m \vee [1 + o(1)]|\log \alpha|/I_0(\theta) & \text{if } I_0(\theta) > A^{-1} \\ m \vee \{\widetilde{M} \wedge [1 + o(1)]|\log \alpha|/[I_0(\theta) \vee I_2(\theta)]\} & \text{if } I_0(\theta) < A^{-1}. \end{cases}$$

The simulation results in Table 5 show that the two-stage tests of Proschan and Hunsberger (1995) and Li et al. (2002), which use conditional power to determine the second-stage sample size, have actual power much lower than the adaptive test of Section 2.1. Without assuming a prespecified alternative θ_1, the usual approach in the literature on sample size re-estimation considers the case $d = 1$ and $u(\theta) = \theta$ and determines the second-stage sample size via the conditional power criterion

$$\tilde{n}(\theta) = \min[n \geq m : P_\theta(S_n \geq c_{\alpha,n} \mid S_m) \geq 1 - \tilde{\alpha}], \tag{4.7}$$

choosing $n_2 = \tilde{n}(\hat{\theta}_m)$ if $\hat{\theta}_m \geq \theta_0 + \delta$ and stopping at the first stage due to futility otherwise, where $P_{\theta_0}(S_n \geq c_{\alpha,n}) = \alpha$ and δ is chosen "to set an upper bound to limit the sample size of the second stage"; see Li et al. (2002, p. 283). Although the conditional power given $\hat{\theta}_m \geq \theta_0 + \delta$ is at least $1 - \tilde{\alpha}$ by choosing the second-stage sample size to be $\tilde{n}(\hat{\theta}_m)$, the actual (unconditional) Type II error probability of the test at θ $(> \theta_0)$ may substantially exceed $\tilde{\alpha}$ if m is not large enough since $P_\theta(\hat{\theta}_m < \theta_0 + \delta)$ may well exceed $\hat{\alpha}$. Stopping due to futility at the end of the first stage when $\hat{\theta}_m < \theta_0 + \delta$ can lead to serious loss of power of the two-stage test. By allowing the test to have a possible third stage, we do not have to stop prematurely when $\hat{\theta}_m$ falls below θ_0, for which the conditional power criterion (4.7) is not applicable. Thus, a three-stage test that uses conditional power to determine the second-stage sample size chooses $n_1 = m, n_3 = M$ and $n_2 = \min[\tilde{n}(\hat{\theta}_m), M\}$, where $\tilde{n}(\theta)$ is defined by (4.7) if $\theta > \theta_0$, and by

$$\tilde{n}(\theta) = \max[m, \lceil |\log \tilde{\alpha}|/I(\theta, \theta_1) \rceil] \quad \text{if } \theta \leq \theta_0. \tag{4.8}$$

The rejection and futility boundaries are given by (2.2)–(2.4) as in the three-stage test of Section 2.1. The asymptotic properties of the test, whose sample size is denoted by $\widetilde{N}$, are given by the following.

Theorem 4.3. *Define* η_θ *for* $\theta > \theta_0$ *by*

$$\theta_0 < \eta_\theta < \theta \quad \text{and} \quad I(\eta_\theta, \theta_0) = I(\eta_\theta, \theta). \tag{4.9}$$

Then as $\alpha + \tilde{\alpha} \to 0$, $E_\theta(\widetilde{N}) \sim \tilde{n}(\theta)$ *if* $\theta < \theta_0$ *and*

$$E_\theta(\widetilde{N}) \sim m \vee [M \wedge |\log \alpha| / I(\eta_\theta, \theta_0)] \quad \text{if } \theta > \theta_0. \tag{4.10}$$

Proof. Suppose $\theta > \theta_0$. From (4.1) and the law of large numbers, it follows that $P_\theta[\hat{\theta}_m > \theta_0] \to 1$ as $\alpha + \tilde{\alpha} \to 0$, and therefore

$$P_\theta\big[\hat{\theta}_m > \theta_0 \text{ and } \widetilde{N} = \tilde{n}(\hat{\theta}_m) \wedge M\big] \to 1. \tag{4.11}$$

Since $\tilde{n}(\theta)$ is given by (4.7) in this case, application of Theorem 2(i) of Lai and Shih (2004) then yields

$$\tilde{n}(\theta) \sim m \vee [|\log \alpha| / I(\eta_\theta, \theta)]. \tag{4.12}$$

In view of (4.1), we can apply Lebesgue's dominated convergence theorem to obtain (4.10) from (4.11) and (4.12) in this case. Note in this connection that $I(\eta_\theta, \theta)$ is continuous in θ and that $\hat{\theta}_m$ converges to θ in probability.

We next consider the case $\theta < \theta_0$. Then $\tilde{n}(\theta)$ is given by (4.8), which is less than $M \sim |\log \alpha| / I(\theta^*, \theta_1)$ as $\alpha + \tilde{\alpha} \to 0$ such that $\log \alpha \sim \log \tilde{\alpha}$. By the law of large numbers, $P_\theta\{\hat{\theta}_m < \theta_0 \text{ and } \widetilde{N} = \tilde{n}(\hat{\theta}_m)\} \to 1$. Continuity of $I(\theta, \theta_1)$ and dominated convergence can then be used to show that $E_\theta(\widetilde{N}) \sim \tilde{n}(\theta)$.

Since $I(\eta_\theta, \theta_0) < I(\theta, \theta_0)$ by (4.9), it follows from (4.2) and (4.10) that the three-stage test using the conditional power criterion (4.7) is not asymptotically efficient. This is not surprising since (4.7) is the sample size for the level-α FSS test to have at least $1 - \tilde{\alpha}$ power at the alternative θ ($>\theta_0$). However, the optimal test with error probabilities α at θ_0 and $\tilde{\alpha}$ at θ is Wald's sequential probability ratio test whose expected sample size is of the smaller order $|\log \alpha| / I(\theta, \theta_0)$ under the assumptions of Theorem 4.1.

5. DISCUSSION

A major drawback of the commonly used conditional power approach to two-stage designs is pointed out in Section 3.3. The actual power can be much lower than the conditional power since the estimated alternative at the end of the first stage can be quite different from the actual alternative. In particular, if the estimated alternative falls in the region of the null hypothesis and misleads one to stop for futility, there can be substantial loss of power. On the other hand, early stopping for futility is critical for keeping the sample size of a conditional power test within a manageable bound M. Our three-stage test makes use of M to come up with an implied alternative which is used to choose the rejection and futility boundaries

appropriately so that the test does not lose much power in comparison with the (most powerful) fixed sample size test of the null hypothesis versus the implied alternative. This idea underlying (2.7)–(2.9) that define the stopping boundaries of three-stage tests has been used earlier by Lai and Shih (2004) to develop efficient group sequential tests.

Our approach estimates the second-stage sample size by using an approximation to Hoeffding's lower bound for the expected sample size of sequential tests satisfying a prescribed Type I error constraint and a Type II error constraint at the alternative that is estimated at the end of the first stage. As shown in the simulation studies in Bartroff and Lai (2008) and in the asymptotic theory in Section 4, this approach yields adaptive tests that are comparable to the benchmark optimal adaptive test of Jennison and Turnbull (2006a,b) for a normal mean, which assumes known variance and a specified alternative. Our approach can serve to bridge the gap between the two "camps" in the adaptive design literature: One camp focuses on efficient designs, under restrictive assumptions, that involve sufficient statistics and optimal stopping rules, while the other camp emphasizes flexibility to address the difficulty of coming up with realistic alternatives at the design stage. As pointed out in Section 1, our approach that is built on the foundations of sequential testing theory is able to resolve the dilemma between efficiency and flexibility. Like the "efficiency camp," it adheres to the GLR test statistics whose efficiency is well established in the theory of FSS tests. An important innovation is that it uses the Markov property to compute error probabilities when the fixed sample size is replaced by a data-dependent sample size that is based on the estimated alternative at the end of the first stage, like the "flexibility camp."

ACKNOWLEDGMENTS

Bartroff's work was supported by grant DMS-0403105 from the National Science Foundation. Lai's work was supported by grants RO1-CA088890 and DMS-0305749 from the National Institutes of Health and the National Science Foundation.

REFERENCES

Bartroff, J. and Lai, T. L. (2008). Efficient Adaptive Designs with Mid-Course Sample Size Adjustment in Clinical Trials, *Statistics in Medicine* 27: 1593–1611.

Bauer, P. (1989). Multistage Testing with Adaptive Designs (with Discussion), *Biometrie und Informatik in Medizin und Biologie* 20: 130–148.

Bauer, P. and Einfalt, J. (2006). Application of Adaptive Design—A Review, *Biometrical Journal* 48: 493–506.

Birkett, M. and Day, S. (1994). Internal Pilot Studies for Estimating Sample Size, *Statistics in Medicine* 13: 2455–2463.

Burman, C. F. and Sonesson, C. (2006). Are Flexible Designs Sound? (with Discussion), *Biometrics* 62: 664–683.

Cui, L., Hung, H. M., and Wang, S. J. (1999). Modification of Sample Size in Group Sequential Clinical Trials, *Biometrics* 55: 835–857.

Dantzig, G. B. (1940). On the Non-Existence of Tests of Student's Hypothesis Having Power Functions Independent of σ, *Annals of Mathematical Statistics* 11: 186–192.

Denne, J. S. and Jennison, C. (1999). Estimating the Sample Size for a *t*-Test Using an Internal Pilot, *Statistics in Medicine* 18: 1575–1585.

Denne, J. S. and Jennison, C. (2000). A Group Sequential *t*-Test with Updating of Sample Size, *Biometrika* 87: 125–134.

Ellenberg, S. S. and Eisenberger, M. A. (1985). An Efficient Design for Phase III Studies of Combination Chemotherapies, *Cancer Treatment Reports* 69: 1147–1154.

Facey, K. M. (1992). A Sequential Procedure for a Phas II Efficacy Trial in Hypercholesterolemia, *Controlled Clinical Trials* 13: 122–133.

Fisher, L. (1998). Self-Designing Clinical Trials, *Statistics in Medicine* 17: 1551–1562.

Gould, A. L. and Shih, W. J. (1992). Sample Size Re-estimation without Unblinding for Normally Distributed Outcomes with Unknown Variance, *Communications in Statistics, Series A* 21: 2833–2853.

Herson, J. and Wittes, J. (1993). The Use of Interim Analysis in Sample Size Adjustment, *Drug Information Journal* 27: 753–760.

Hoeffding, W. (1960). Lower Bounds for the Expected Sample Size and the Average Risk of a Sequential Procedure, *Annals of Mathematical Statistics* 31: 352–368.

Hung, H. M. J., O'Neill, R. T., Wang, S.-J., and Lawrence, J. (2006). A Regulatory View on Adaptive/Flexible Clinical Trial Design, *Biometrical Journal* 48: 565–573.

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

Lai, T. L. and Shih, M. C. (2004). Power, Sample Size and Adaptation Considerations in the Design of Group Sequential Clinical Trials, *Biometrika* 91: 507–528.

Lehmacher, W. and Wassmer, G. (1999). Adaptive Sample Size Calculations in Group Sequential Trials, *Biometrics* 55: 1286–1290.

Li, G., Shih, W. J., Xie, T., and Lu, J. (2002). A Sample Size Adjustment Procedure for Clinical Trials, *Biostatistics* 3: 277–287.

Lorden, G. (1983). Asymptotic Efficiency of Three-Stage Hypothesis Tests, *Annals of Statistics* 11: 129–140.

Posch, M. and Bauer, P. (1999). Adaptive Two Stage Designs and the Conditional Error Function, *Biometrical Journal* 41: 689–696.

Press, N. H., Flannery, B. P., Teukolsky, S. A., and Vitterling, W. T. (1992). *Numerical Recipes in C: The Art of Scientific Computing*, Second edition, Cambridge University Press.

Proschan, M. and Hunsberger, S. (1995). Designed Extension Studies Based on Conditional Power, *Biometrics* 51: 1315–1324.

Schmitz, N. (1993). *Optimal Sequentially Planned Decision Procedures*, Volume 79 of Lecture Notes in Statistics, New York: Springer-Verlag.

Shen, Y. and Fisher, L. (1999). Statistical Inference for Self-Designing Clinical Trials with a One-Sided Hypothesis, *Biometrics* 41: 190–197.

Shih, W. J. (2001). Sample Size Re-Estimation—Journey for a Decade, *Statistics in Medicine* 20: 515–518.

Simon, R. (1989). Optimal Two-Stage Designs for Phase II Clinical Trials, *Controlled Clinical Trials* 10: 1–10.

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.

Thall, P. F. and Simon, R. (1994). Practical Bayesian Guidelines for Phase IIB Clinical Trials, *Biometrics* 50: 337–349.

Thall, P. F., Simon, R., Ellenberg, S. S., and Shrager, R. (1988). Optimal Two-Stage Designs for Clinical Trials with Binary Response, *Statistics in Medicine* 7: 571–579.

Tsiatis, A. A. and Mehta, C. (2003). On the Efficiency of the Adaptive Design for Monitoring Clinical Trials, *Biometrika* 90: 367–378.

Vickers, A. J., Ballen, V., and Scher, H. I. (2007). Setting the Bar in Phase III Trials: The Use of Historical Data for Determining "Go/No Go" Decision for Definitive Phase II Trials, *Clinical Cancer Research* 13: 972–976.

Whitehead, J., Whitehead, A., Todd, S., Bollard, K., and Sooriyarachi, M. R. (2001). Mid-Trial Design Reviews for Sequential Clinical Trials, *Statistics in Medicine* 20: 165–176.

Wittes, J. and Brittain, E. (1990). The Role of Internal Pilots in Increasing the Efficiency of Clinical Trials, *Statistics in Medicine* 9: 65–72.