STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1593–1611
Published online 15 February 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3201

Efficient adaptive designs with mid-course sample size adjustment in clinical trials

Jay Bartroff[1,*,†] and Tze Leung Lai[2]

[1] *Department of Mathematics, University of Southern California, Los Angeles, CA 90089, U.S.A.*
[2] *Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305, U.S.A.*

SUMMARY

Adaptive designs have been proposed for clinical trials in which the nuisance parameters or alternative of interest are unknown or likely to be misspecified before the trial. Although most previous works on adaptive designs and mid-course sample size re-estimation have focused on two-stage or group-sequential designs in the normal case, we consider here a new approach that involves at most three stages and is developed in the general framework of multiparameter exponential families. This approach not only maintains the prescribed type I error probability but also provides a simple but asymptotically efficient sequential test whose finite-sample performance, measured in terms of the expected sample size and power functions, is shown to be comparable to the optimal sequential design, determined by dynamic programming, in the simplified normal mean case with known variance and prespecified alternative, and superior to the existing two-stage designs and also to adaptive group-sequential designs when the alternative or nuisance parameters are unknown or misspecified. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: adaptive design; conditional power; futility; Kullback–Leibler information; sample size re-estimation

1. INTRODUCTION

In standard clinical trial designs, the sample size is determined by the power at a given alternative (e.g. treatment effect). In practice, especially for new treatments about which there is little information on the magnitude and sampling variability of the treatment effect, it is often difficult

[*] Correspondence to: Jay Bartroff, Department of Mathematics, University of Southern California, 3620 South Vermont Ave., KAP 108, Los Angeles, CA 90089-2532, U.S.A.
[†] E-mail: bartroff@usc.edu

Contract/grant sponsor: NIH; contract/grant number: RO1-CA08890
Contract/grant sponsor: NSF; contract/grant numbers: DMS-0403105, DMS-0305709

Received 26 November 2006
Accepted 26 November 2007

for investigators to specify a realistic alternative at which sample size determination can be based. Therefore, the problem of sample size re-estimation based on an observed treatment difference at some time before the prescheduled end of the trial has attracted considerable attention during the past decade, see, e.g. Jennison and Turnbull [1, Section 14.2], Shih [2], and Whitehead *et al.* [3]. Moreover, there are concerns from the regulatory perspective regarding possible inflation of the type I error probability when such sample size adjustments are used in pharmaceutical trials. For normally distributed outcome variables, Proschan and Hunsberger [4], Fisher [5], Posch and Bauer [6], and Shen and Fisher [7] 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. Jennison and Turnbull [8] gave a general form of these methods and showed that they performed considerably worse than group-sequential tests. Tsiatis and Mehta [9] independently came to the same conclusion, pointing out their inefficiency because the adjusted test statistics are not sufficient statistics. It is possible to adhere to efficient generalized likelihood ratio statistics in a mid-course adaptive design if one uses the non-normal sampling distribution (due to the mid-course adaptation) of the test statistic, instead of ignoring the non-normality and thereby resulting in type I error inflation. A way to do this was proposed by Li *et al.* [10], but it was shown by Turnbull [11] to be relatively inefficient compared with group-sequential tests. Jennison and Turnbull [12] 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 also showed how the corresponding optimization problem can be solved numerically by using the backward induction algorithms for 'optimal sequentially planned' designs developed by Schmitz [13]. Jennison and Turnbull [14] found that standard (non-adaptive) group-sequential tests with the first stage chosen optimally are nearly as efficient as their optimal adaptive counterparts that are considerably more complicated, and we use these as a benchmark for our comparisons in Section 3.

With the goal of achieving similar efficiency in more complicated situations where the alternative of interest and/or nuisance parameters are not known, we give in Section 2 a simple adaptive test that updates the sample size after the initial stage by using estimates of the unknown parameters and adjustments for the uncertainty of these estimates. This is done first for the one-parameter case in Section 2.1 and extended to the multiparameter setting in Section 2.3. These tests usually terminate at the first or second stage, but allow the possibility of a third stage to account for uncertainties in the second-stage sample size estimate. The tests control the type I error probability and have power close to the uniformly most powerful fixed sample test. Section 3 gives a comprehensive simulation study, which is the first of its kind, of the adaptive tests in the aforementioned references and compares them with the adaptive tests developed in Section 2 and with fixed sample size (FSS) and standard group-sequential tests having the same minimum and maximum sample sizes. A thorough evaluation of the performance of these tests is presented, involving the power, mean number of stages, and the mean, 25th, 50th, and 75th percentiles of the sample size distribution under a wide range of alternatives, subject to the prescribed constraints on type I error probability and first-stage and maximum sample sizes. Section 3.1 also compares the proposed adaptive test with the benchmark optimal adaptive test of Jennison and Turnbull [12, 14], and the variance unknown case is considered in Section 3.2. An example from the National Heart, Lung and Blood Institute (NHLBI) Coronary Intervention Study is given in Section 3.3. Section 4 gives some concluding remarks.

2. EFFICIENT ADAPTIVE TESTS WITH THREE OR FEWER STAGES

In this section, we consider one-sided tests of the null hypothesis $H_0 : \theta \leqslant \theta_0$ on the natural parameter θ in a one-parameter exponential family $f_\theta(x) = \mathrm{e}^{\theta x - \psi(\theta)}$ of densities with respect to some measure on the real line. Let $X_1, X_2, \ldots$ denote the successive observations, and let $S_n = X_1 + \cdots + X_n$. A sufficient statistic based on $(X_1, \ldots, X_n)$ is $\bar{X}_n = S_n/n$, and the maximum likelihood estimate of θ is $\widehat{\theta}_n = (\psi')^{-1}(\bar{X}_n)$. The special case of normal X_i with mean θ and known variance 1 is widely used in the literature on sample size re-estimation as a prototype, which can be used to approximate more complicated situations via the central limit theorem, as in the references in Section 1.

In practice, there is an upper bound M on the allowable sample size for a clinical trial because of funding and duration constraints and because there are other trials that compete for patients, investigators, and resources. The re-estimated sample size in two-stage designs has to be restricted within this bound; see, e.g. Li *et al.* [10, p. 283]. Lai and Shih [15, p. 511] have pointed out that M implies constraints on the alternatives that can be considered in power calculations to determine the sample size. Specifically, by the Neyman–Pearson lemma, the FSS test that rejects H_0 if $S_M \geqslant c_{\alpha,M}$ has maximal power at any alternative $\theta > \theta_0$ and, in particular, at the alternative θ_1 at which the FSS test has prescribed power $1-\widetilde{\alpha}$. Here, $c_{\alpha,n}$ denotes the critical value of the level-α FSS test based on a sample of size n, i.e. $\mathrm{pr}_{\theta_0}\{S_n \geqslant c_{\alpha,n}\} = \alpha$. Typical sample size re-estimation procedures in the literature (see, e.g. the references in Section 1) first use the initial sample of size m, which is some fraction of M, to provide an estimate $\widehat{\theta}_m$ of θ and then evaluate the sample size of the FSS test that has conditional power $1-\widetilde{\alpha}$ given the alternative $\widehat{\theta}_m$, assuming that $\widehat{\theta}_m > \theta_0$. This results in a two-stage procedure, which does not incorporate the sampling variability of the estimate $\widehat{\theta}_m$. A simple way to make 'uncertainty adjustments' in the above procedure that attempts to 'self-tune' itself to the actual θ value is to allow the possibility of not stopping at the second stage when H_0 is not rejected, by including a third (and final) stage with total sample size M.

2.1. An efficient test of H_0 with at most three stages

To test $H_0 : \theta \leqslant \theta_0$ at significance level α, suppose that no fewer than m but no more than M observations are to be taken. Let θ_1 be the alternative 'implied' by M, in the sense that M can be determined as the sample size of the level-α Neyman–Pearson test with power $1-\widetilde{\alpha}$ at θ_1. Alternatively, θ_1 can be specified separately from M as a clinically relevant or realistic anticipated effect size based on prior experimental, observational, or theoretical evidence, if such information is available. A fundamental result in sequential testing theory is that Wald's sequential probability ratio test (SPRT) of the simple hypotheses $\theta = \theta'$ *vs* $\theta = \theta''$ has the smallest expected sample size at $\theta = \theta'$ and θ'' among all tests with the same or smaller type I and II error probabilities; see Reference [16]. Moreover, letting α and $\widetilde{\alpha}$ denote the type I and II error probabilities and $T(\theta', \theta'')$ be the sample size of the SPRT, Chernoff [16, p. 66] has derived the approximations:

$$E_{\theta''}(T(\theta', \theta'')) \approx |\log \alpha| / I(\theta'', \theta'), \quad E_{\theta'}(T(\theta', \theta'')) \approx |\log \widetilde{\alpha}| / I(\theta', \theta'') \tag{1}$$

where

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

is the Kullback–Leibler information number. To test the one-sided hypothesis $H_0 : \theta \leqslant \theta_0$, suppose that we use the maximum likelihood estimator $\widehat{\theta}_m$ from the first stage of the study in place of the

DOI: 10.1002/sim

alternative θ'' in (1) with $\theta'=\theta_0$, in the event $\widehat{\theta}_m>\theta_0$. Then the first relation in (1) suggests that an efficient second-stage sample size would be around $|\log\alpha|/I(\widehat{\theta}_m,\theta_0)$. On the other hand, if $\widehat{\theta}_m\leqslant\theta_0$, then we can consider the possibility of stopping due to futility by choosing $\theta'=\widehat{\theta}_m$ and $\theta''=\theta_1$ in the SPRT; hence, the second relation in (1) suggests $|\log\widetilde{\alpha}|/I(\widehat{\theta}_m,\theta_1)$ as an efficient second-stage sample size. Adjusting for the sampling variability in $\widehat{\theta}_m$ by inflating by the factor $1+\rho_m$, we therefore define the second-stage sample size:

$$n_2=m\vee\{M\wedge\lceil(1+\rho_m)n(\widehat{\theta}_m)\rceil\} \tag{2}$$

where $\rho_m>0$, $\vee$ and $\wedge$ denote maximum and minimum, respectively, $\lceil x\rceil$ denotes the smallest integer $\geqslant x$ (and $\lfloor x\rfloor$ denotes the largest integer $\leqslant x$), and

$$n(\theta)=\frac{|\log\alpha|}{I(\theta,\theta_0)}\wedge\frac{|\log\widetilde{\alpha}|}{I(\theta,\theta_1)} \tag{3}$$

which is an approximation to Hoeffding's [17] lower bound for the expected sample size $E_\theta(T)$ of a test that has type I error probability α at θ_0 and type II error probability $\widetilde{\alpha}$ at θ_1. Note that (2) includes the cases $n_2=m$ and $n_2=M$ associated with using just one or two stages. Moreover, the stopping rule defined below by (4)–(6) allows the possibility of stopping after the first or second stage. Therefore, the actual number of stages used by the 'three-stage' test is in fact a random variable taking the values 1, 2, 3.

The three-stage test uses rejection and futility boundaries similar to those of the efficient group-sequential tests introduced by Lai and Shih [15]. Letting n_i denote the total sample size at the ith stage, the test stops at stage $i\leqslant 2$ and rejects H_0 if

$$n_i<M,\quad \widehat{\theta}_{n_i}>\theta_0 \quad\text{and}\quad n_iI(\widehat{\theta}_{n_i},\theta_0)\geqslant b \tag{4}$$

where $n_1=m$ and n_2 is given by (2). The test stops at stage $i\leqslant 2$ and accepts H_0 if

$$n_i<M,\quad \widehat{\theta}_{n_i}<\theta_1 \quad\text{and}\quad n_iI(\widehat{\theta}_{n_i},\theta_1)\geqslant\widetilde{b} \tag{5}$$

It rejects H_0 at stage $i=2$ or 3 if

$$n_i=M,\quad \widehat{\theta}_M>\theta_0 \quad\text{and}\quad MI(\widehat{\theta}_M,\theta_0)\geqslant c \tag{6}$$

accepting H_0 otherwise. Letting $0<\varepsilon,\widetilde{\varepsilon}<1$, define the thresholds $b,\widetilde{b}$, and c by the following equations:

$$\text{pr}_{\theta_1}\{(5)\text{ occurs for } i=1 \text{ or } 2\}=\widetilde{\varepsilon}\,\widetilde{\alpha} \tag{7}$$

$$\text{pr}_{\theta_0}\{(5)\text{ does not occur for } i\leqslant 2,\text{ and (4) occurs for } i=1\text{ or }2\}=\varepsilon\alpha \tag{8}$$

$$\text{pr}_{\theta_0}\{(4)\text{ and (5) do not occur for } i\leqslant 2,\text{ and (6) occurs}\}=(1-\varepsilon)\alpha \tag{9}$$

Note that (8) and (9) imply that the type I error probability is exactly α, and we have found in our simulations (see Section 3) that the power at θ_1 is generally close to, but slightly less than, $1-\widetilde{\alpha}$. The values $\varepsilon,\widetilde{\varepsilon}$ are the fractions of type I and II error probabilities 'spent' at the first two stages, and in theory any values $0<\varepsilon,\widetilde{\varepsilon}<1$ may be used. In practice, we recommend using $0.2\leqslant\varepsilon,\widetilde{\varepsilon}\leqslant 0.8$, and we have found that the power and expected sample size of the above adaptive test vary very little with changes in $\varepsilon,\widetilde{\varepsilon}$. In particular, the three examples in Section 3 use $\varepsilon=\widetilde{\varepsilon}=\frac{1}{3}$, $(\varepsilon,\widetilde{\varepsilon})=(\frac{1}{2},\frac{3}{4})$, and $\varepsilon=\widetilde{\varepsilon}=\frac{1}{2}$. The factor ρ_m in (2) is a small inflation of $n(\widehat{\theta}_m)$ to adjust for the uncertainty in $\widehat{\theta}_m$.

Lorden [18] gives an asymptotic upper bound for ρ_m as a function of θ_0, θ_1, α, and $\widetilde{\alpha}$. We advocate simply fixing ρ_m to a small maximum inflation that the practitioner is comfortable with and have found that $\rho_m = 0.05$ or 0.1 works well in practice, which we use in the examples in Section 3. As with M, the choice of m is often determined by practical considerations such as funding and duration. To aid such considerations or in the absence of them, if the practitioner has bounds $\underline{\theta} < \theta_0$ and $\overline{\theta} > \theta_1$ in mind (e.g. $\overline{\theta}$ might be the largest realistic treatment effect likely to be seen), then m could be chosen to be $n(\underline{\theta}) \wedge n(\overline{\theta})$, an approximation to Hoeffding's [17] lower bound for the smallest expected sample size of a test with error probabilities $\alpha, \widetilde{\alpha}$ at θ_0, θ_1 when $\theta = \underline{\theta}$ or $\overline{\theta}$.

The probabilities in (7)–(9) can be computed by Monte Carlo or recursive numerical integration, using normal approximations to signed-root likelihood ratio statistics. Further details are given in Section 2.2. The original idea to use (2) as the second-stage sample size and to allow the possibility of a third stage to account for uncertainty in the estimate $\widehat{\theta}_m$ (and hence n_2) is due to Lorden [18], although his test uses very conservative upper bounds on the error probabilities. Here, we have modified Lorden's test to control the type I error α exactly and provided algorithms to implement the modified test. It can be shown that our three-stage test is asymptotically optimal: If N is the sample size of our three-stage test above, then

$$E_\theta(N) \sim m \vee \left\{ M \wedge \frac{|\log \alpha|}{I(\theta, \theta_0) \vee I(\theta, \theta_1)} \right\} \tag{10}$$

as $\alpha + \widetilde{\alpha} \to 0$, $\log \alpha \sim \log \widetilde{\alpha}$, $\rho_m \to 0$ and $\rho_m \sqrt{m/\log m} \to \infty$; if T is the sample size of any test of $H_0 : \theta \leqslant \theta_0$ whose error probabilities at θ_0 and θ_1 do not exceed α and $\widetilde{\alpha}$, respectively, then

$$E_\theta(T) \geqslant (1 + o(1)) E_\theta(N) \tag{11}$$

simultaneously for all θ. The proof uses Hoeffding's [17] lower bound for $E_\theta(T)$ as in [18] and can be found in [19].

Since $\log \alpha \sim \log(\varepsilon\alpha)$ as $\alpha \to 0$ for any fixed $0 < \varepsilon < 1$, the asymptotic formula for $E_\theta(N)$ in (10) is unchanged if one replaces the type I error probability α by a fraction of it, and this is why Lorden [18] can use crude bounds of the type above for the type I error probability. For values of the type I error probability α (e.g. 0.05 or 0.01) commonly used in practice, replacing α by $\alpha/10$, say, can substantially increase $E_\theta(N)$. Note that our adaptive test keeps the error probability at θ_0 to be α (instead of less than α) by using Monte Carlo or recursive numerical integration to evaluate it, discussed in the next section.

2.2. *The normal case and recursive numerical integration*

The thresholds b, $\widetilde{b}$, and c can be computed by solving in succession (7), (8), and (9). Univariate grid search or Brent's method [20] can be used to solve each equation. Suppose X_i are $\mathrm{N}(\theta, 1)$. Without loss of generality, we shall assume that $\theta_0 = 0$. Since $I(\theta, \lambda) = (\theta - \lambda)^2/2$, we can rewrite (7) as

$$\mathrm{pr}_{\theta_1}\{S_m - m\theta_1 \leqslant -(2\widetilde{b}m)^{1/2}\} + \mathrm{pr}_{\theta_1}\{S_m - m\theta_1 > -(2\widetilde{b}m)^{1/2}, S_{n_2} - n_2\theta_1 \leqslant -(2\widetilde{b}n_2)^{1/2}\} = \widetilde{\varepsilon}\,\widetilde{\alpha} \tag{12}$$

and (8) and (9) as

$$\mathrm{pr}_0\{S_m/\sqrt{2m} \geqslant b^{1/2}\} + \mathrm{pr}_0\{\widetilde{b}^{1/2} < S_m/\sqrt{2m} < b^{1/2}, S_{n_2}/\sqrt{2n_2} \geqslant b^{1/2}\} = \varepsilon\alpha \tag{13}$$

$$\mathrm{pr}_0\{\widetilde{b}^{1/2} < S_m/\sqrt{2m} < b^{1/2}, \widetilde{b}^{1/2} < S_{n_2}/\sqrt{2n_2} < b^{1/2}, S_M/\sqrt{M} \geqslant c^{1/2}\} = (1-\varepsilon)\alpha \tag{14}$$

DOI: 10.1002/sim

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

$$\mathrm{pr}_0\{S_{n_2}\geqslant(2bn_2)^{1/2}|S_m=mx\}=\Phi\left(\frac{mx-[2bk(x)]^{1/2}}{[k(x)-m]^{1/2}}\right) \tag{15}$$

$$\mathrm{pr}_0\{S_{n_2}\in \mathrm{d}y, S_M\in \mathrm{d}z|S_m=mx\}=\varphi_{k(x)-m}(y-mx)\varphi_{M-k(x)}(z-y)\,\mathrm{d}y\,\mathrm{d}z \tag{16}$$

where Φ is the standard normal cumulative distribution function and φ_v is the N$(0,v)$ density function, i.e. $\varphi_v(w)=(2\pi v)^{-1/2}\exp(-w^2/2v)$. The probabilities under θ_1 can be computed similarly. Hence, standard recursive numerical integration algorithms can be used to compute the probabilities in (7)–(9).

As an example, we compute the thresholds $b,\widetilde{b}$, and c for the following adaptive test whose performance is studied in Section 3.1. Here $M=120$, $\alpha=0.025$, and we wish the power to be close to $1-\widetilde{\alpha}=0.9$ at $\theta=\theta_1=0.3$. Setting $\varepsilon=\widetilde{\varepsilon}=\frac{1}{3}$ and $\rho_m=0.1$, we first find $\widetilde{b}$ by solving (12), which can be expressed as

$$\Phi(-[2\widetilde{b}]^{1/2})+\int_{[2\widetilde{b}/m]^{1/2}}^{\infty}\Phi\left(\frac{-m(x-\theta_1)-[2\widetilde{b}k(x)]^{1/2}}{[k(x)-m]^{1/2}}\right)\varphi_m(mx)m\,\mathrm{d}x=\widetilde{\varepsilon}\,\widetilde{\alpha}=\frac{0.1}{3}$$

by the analog of (15) for $\theta=\theta_1$, where φ_v is as in (16). The integral is computed by numerical integration, and a few iterations of the bisection method give $\widetilde{b}=1.99$. This value is next used to find b similarly by solving (13), which can be expressed as

$$\Phi(-[2b]^{1/2})+\int_{[2\widetilde{b}/m]^{1/2}}^{[2b/m]^{1/2}}\Phi\left(\frac{mx-[2bk(x)]^{1/2}}{[k(x)-m]^{1/2}}\right)\varphi_m(mx)m\,\mathrm{d}x=\varepsilon\alpha=\frac{0.025}{3}$$

by (15). The bisection method gives $b=3.26$, which we, in turn, use to find c by solving (14), which is

$$\int_{[2\widetilde{b}/m]^{1/2}}^{[2b/m]^{1/2}}\int_{[2\widetilde{b}k(x)]^{1/2}}^{[2bk(x)]^{1/2}}\Phi\left(\frac{[cM]^{1/2}-y}{[M-k(x)]^{1/2}}\right)\varphi_{k(x)-m}(y-mx)\varphi_m(mx)m\,\mathrm{d}y\,\mathrm{d}x=(1-\varepsilon)\alpha=\frac{0.05}{3}$$

by (16), giving $c=2.05$.

2.3. Multiparameter extension

Suppose $X_1,X_2,\ldots$ are independent d-dimensional random vectors from a multiparameter exponential family $f_\theta(x)=\exp\{\theta^{\mathrm{T}}x-\psi(\theta)\}$ of densities. The three-stage test in Section 2.1 can be readily extended to test $H_0:u(\theta)\leqslant u_0$, where u is any smooth real-valued function. As in Section 2.1, $n_1=m$ and $n_3=M$. The stopping rule of the three-stage test of $H_0:u(\theta)\leqslant u_0$ is the same as (4)–(6) but with $nI(\widehat{\theta}_n,\theta_j)$ replaced by

$$\inf_{\theta:u(\theta)=u_j} nI(\widehat{\theta}_n,\theta) \tag{17}$$

$j=0,1$, where $u_1>u_0$ is the alternative implied by the maximum sample size M and the desired type II error probability $\widetilde{\alpha}$; see [15, Section 3.4]. In particular, the test stops and rejects H_0 at stage $i\leqslant 2$ if

$$n_i<M, \quad u(\widehat{\theta}_{n_i})>u_0 \quad \text{and} \quad \inf_{\theta:u(\theta)=u_0} n_i I(\widehat{\theta}_{n_i},\theta)\geqslant b \tag{18}$$

which is analogous to (4). Early stopping for futility (accepting H_0) can also occur at stage $i\leqslant 2$ if

$$n_i<M, \quad u(\widehat{\theta}_{n_i})<u_1 \quad \text{and} \quad \inf_{\theta:u(\theta)=u_1} n_i I(\widehat{\theta}_{n_i},\theta)\geqslant \widetilde{b} \tag{19}$$

which is analogous to (5). The test rejects H_0 at stage $i=2$ or 3 if

$$n_i=M, \quad u(\widehat{\theta}_M)>u_0 \quad \text{and} \quad \inf_{\theta:u(\theta)=u_0} MI(\widehat{\theta}_M,\theta)\geqslant c \tag{20}$$

accepting H_0 otherwise. The thresholds b, $\widetilde{b}$, and c are chosen to ensure certain type I and type II error probability constraints that are similar to (7)–(9) and are computed by using the normal approximation to the signed-root likelihood ratio statistic

$$\ell_n(\delta)=n\{\mathrm{sign}(u(\widehat{\theta}_n)-\delta)\}\left\{2\inf_{\theta:u(\theta)=\delta} I(\widehat{\theta}_n,\theta)\right\}^{1/2}$$

under the hypothesis $u(\theta)=\delta$; see [15, p. 513]. Note that this normal approximation can be used for the choice of u_1 implied by the maximum sample size M and the type II error probability $\widetilde{\alpha}$. The sample size n_2 of the three-stage test is given by (2) with

$$n(\theta)=\min\left\{|\log\alpha| \Big/ \inf_{\lambda:u(\lambda)=u_0} I(\theta,\lambda), |\log\widetilde{\alpha}| \Big/ \inf_{\lambda:u(\lambda)=u_1} I(\theta,\lambda)\right\} \tag{21}$$

which is a generalization of (3). Examples of the multiparameter case are given in Section 3.2 for normally distributed data with unknown variance and in Section 3.3 for two binomial populations.

3. COMPARISON WITH OTHER TESTS

3.1. Normal mean with known variance

We consider the special case of normal X_i with unknown mean θ and known variance 1 and compare a variety of adaptive tests of $H_0:\theta\leqslant 0$ in the literature with the tests proposed in Section 2.1. In this normal setting, $\widehat{\theta}_n=\overline{X}_n$ and $I(\theta,\lambda)=(\theta-\lambda)^2/2$. It is widely recognized that the performance of adaptive tests is difficult to evaluate and compare because it depends heavily on the choice of first-stage and maximum sample sizes, the number of groups (stages) allowed, and the parameter values at which the tests are evaluated. For this reason, the tests evaluated here use the same first-stage and maximum sample sizes, except for a few illustrative examples discussed below. In addition, we report a variety of operating characteristics for each test—power, mean number of stages, and the 25th, 50th, and 75th percentiles in addition to the mean of the sample size distribution—over a wide range of θ values. A comprehensive evaluation of adaptive and group-sequential tests similar to this has not appeared previously in the literature. We also include the uniformly most powerful FSS test

DOI: 10.1002/sim

with the same maximum sample size and type I error probability α, which provides the appropriate benchmark for the power of any test of H_0. Another relevant comparison—especially given their widespread use in clinical trials—made here is with standard (non-adaptive) group-sequential tests having a similar number of stages as the adaptive test.

To test $H_0: \theta \leqslant 0$, Proschan and Hunsberger [4] proposed a two-stage test, based on the conditional power criterion, which uses the usual z-statistic but with a data-dependent critical value to maintain the type I error at a prescribed level α. The test allows early stopping to accept (or reject) the null hypothesis if the test statistic is below a user-specified upper normal quantile z_{p^*} (or above some level k) at the end of the first stage. Choosing a data-dependent critical value is tantamount to multiplying the z-statistic by a data-dependent factor and using a fixed critical value. Li *et al.* [10] proposed to use the z-statistic with a fixed critical value c while still determining the second-stage sample size by conditional power and maintaining the type I error at α. Their test stops after the first stage if the test statistic falls below h or above k. For each h and conditional power level, their test has a maximum allowable k, which they denote by $k_1^*(h)$. Fisher [5] proposed a 'variance spending' method for weighting the observations so that the type I error of his test does not exceed α, despite its data-dependent second-stage sample size that is given by the conditional power criterion. To avoid a very large second-stage sample size if the first-stage estimate of θ lies near the null hypothesis, Shen and Fisher [7] proposed early stopping due to futility whenever the upper $100(1-\alpha_0)$ per cent confidence bound for θ falls below some specified alternative $\theta_1 > 0$.

Table I compares these tests, a FSS test, and two standard group-sequential tests with the adaptive test described in Section 2.1. The values of the user-specified parameters of the tests are summarized in the list below. The user-specified parameters are chosen so that they have the same first-stage sample size $m=40$ (except for the FSS test), maximum sample size $M=120$ (except for SF$'$; see the last paragraph of this section), type I error not exceeding $\alpha=0.025$, and nominal power (or conditional power level in the case of conditional power tests) equal to 0.9.

- *ADAPT*: The adaptive test described in Section 2.1 that uses $b=3.26$, $\widetilde{b}=1.99$, and $c=2.05$ corresponding to $\varepsilon=\widetilde{\varepsilon}=\frac{1}{3}$ in (7)–(9), and $\rho_m=0.1$ (see Section 2.2 for details).
- FSS_{120}: The FSS test having sample size 120.
- OBF_{PF}, OBF_{SC}: O'Brien and Fleming's [21] one-sided group-sequential tests having three groups of size 40. $\mathrm{OBF}_{\mathrm{PF}}$ uses power family futility stopping ($\Delta=1$ in [1, Section 4.2]) and $\mathrm{OBF}_{\mathrm{SC}}$ uses stochastic curtailment futility stopping ($\gamma=0.9$ in [1, Section 10.2]). Both $\mathrm{OBF}_{\mathrm{PF}}$ and $\mathrm{OBF}_{\mathrm{SC}}$ use reference alternative $\theta_1=0.3$; see below.
- *PH*: Proschan and Hunsberger's [4] test that uses $p^*=0.0436$ and $k=2.05$.
- *L*: Li *et al.*'s [10] test that uses $h=1.63$ and $k=k_1^*(h)=2.83$.
- SF, SF': Two versions of Shen and Fisher's [7] test; SF uses $\alpha_0=0.425$ and SF$'$ uses $\alpha_0=0.154$.

The tests are evaluated at the θ values where FSS_{120} has power $0.01, 0.025, 0.6, 0.8, 0.9, 0.95$, and at $\theta=0.15$, the midpoint of $\theta=0$ and $\theta=\theta_1=0.3$, the alternative implied by $M=120$ since FSS_{120} has power $1-\widetilde{\alpha}=0.9$ there. This is also the alternative used by the OBF tests for futility stopping. Each entry in Table I is computed by Monte Carlo simulation with 100 000 replications. To compare tests T, T' with type I error probability α but with different type II error probabilities $\widetilde{\alpha}_T(\theta)$, $\widetilde{\alpha}_{T'}(\theta)$ and expected sample sizes $E_\theta T$, $E_\theta T'$ at $\theta>0$, Jennison and Turnbull [12] defined

Table I. Power (bold), expected sample size (bold), sample size quantiles T_q, expected number of stages (bold), and efficiency ratio (at $\theta>0$) with respect to ADAPT, of FSS, adaptive, and group-sequential tests with maximum sample size $M=120$ except for SF$'$ that uses $5M$.

Test	ADAPT	FSS$_{120}$	OBF$_{PF}$	OBF$_{SC}$	PH	L	SF	SF$'$
$\theta=-0.03$	**1.1** per cent	**1.0** per cent	**0.9** per cent	**1.0** per cent	**1.5** per cent	**1.3** per cent	**0.6** per cent	**0.9** per cent
	68.5	**120.0**	**72.3**	**91.7**	**40.8**	**41.4**	**41.3**	**72.3**
$T_{0.25}$	40	120	40	80	40	40	40	40
$T_{0.5}$	40	120	80	80	40	40	40	40
$T_{0.75}$	120	120	80	120	40	40	40	40
#	**1.53**	**1.00**	**1.81**	**2.29**	**1.01**	**1.03**	**1.03**	**1.14**
$\theta=0$	**2.5** per cent	**2.5** per cent	**2.3** per cent	**2.5** per cent	**2.4** per cent	**2.5** per cent	**1.2** per cent	**2.2** per cent
	75.1	**120.0**	**77.8**	**96.4**	**41.1**	**42.2**	**41.2**	**82.3**
$T_{0.25}$	40	120	80	80	40	40	40	40
$T_{0.5}$	60	120	80	80	40	40	40	40
$T_{0.75}$	120	120	80	120	40	40	40	40
#	**1.64**	**1.00**	**1.94**	**2.41**	**1.02**	**1.05**	**1.05**	**1.20**
$\theta=0.15$	**35.6** per cent	**37.6** per cent	**35.7** per cent	**37.1** per cent	**18.7** per cent	**20.9** per cent	**13.8** per cent	**36.1** per cent
	98.6	**120.0**	**98.9**	**110.2**	**44.5**	**48.3**	**47.2**	**115.3**
$T_{0.25}$	71	120	80	120	40	40	40	40
$T_{0.5}$	120	120	120	120	40	40	40	47
$T_{0.75}$	120	120	120	120	40	40	40	146
#	**2.05**	**1.00**	**2.47**	**2.76**	**1.09**	**1.22**	**1.22**	**1.53**
$R_\theta(T,N)$	100	78.5	99.5	86.4	332	289	358	84.5
$\theta=0.20$	**57.2** per cent	**60.0** per cent	**57.9** per cent	**59.5** per cent	**30.2** per cent	**33.2** per cent	**24.8** per cent	**53.5** per cent
	99.4	**120.0**	**101.4**	**108.0**	**45.9**	**50.8**	**50.6**	**124.3**
$T_{0.25}$	76	120	80	80	40	40	40	40
$T_{0.5}$	120	120	120	120	40	40	40	65
$T_{0.75}$	120	120	120	120	40	51	52	157
#	**2.07**	**1.00**	**2.54**	**2.70**	**1.11**	**1.30**	**1.86**	**1.66**
$R_\theta(T,N)$	100	88.5	99.7	97.2	98.1	99.3	70.1	73.1

Table I. *Continued.*

Test	ADAPT	FSS_{120}	$\mathrm{OBF}_{\mathrm{PF}}$	$\mathrm{OBF}_{\mathrm{SC}}$	PH	L	SF	SF′
$\theta=0.26$	**77.4** per cent	**80.0** per cent	**78.0** per cent	**79.6** per cent	**44.2** per cent	**47.5** per cent	**38.2** per cent	**67.5** per cent
	95.2	**120.0**	**99.8**	**102.0**	**46.6**	**52.7**	**52.9**	**120.2**
$T_{0.25}$	59	120	80	80	40	40	40	41
$T_{0.5}$	120	120	80	120	40	40	40	68
$T_{0.75}$	120	120	120	120	40	60	60	145
#	**2.00**	**1.00**	**2.47**	**2.55**	**1.13**	**1.38**	**1.47**	**1.78**
$R_\theta(T,N)$	100	84.7	96.8	98.6	91.4	88.4	67.4	62.7
$\theta=\theta_1=0.3$	**88.8** per cent	**90.0** per cent	**88.6** per cent	**89.5** per cent	**55.2** per cent	**58.0** per cent	**49.1** per cent	**75.5** per cent
	89.2	**120.0**	**94.5**	**96.4**	**46.8**	**53.3**	**54.0**	**111.8**
$T_{0.25}$	40	120	80	80	40	40	40	43
$T_{0.5}$	118	120	80	80	40	40	42	65
$T_{0.75}$	120	120	120	120	40	62	62	128
#	**1.91**	**1.00**	**2.36**	**2.41**	**1.14**	**1.42**	**1.58**	**1.85**
$R_\theta(T,N)$	100	77.5	93.8	94.7	82.6	77.5	61.5	55.6
$\theta=0.33$	**94.0** per cent	**95.0** per cent	**94.1** per cent	**94.7** per cent	**63.5** per cent	**66.5** per cent	**57.3** per cent	**80.8** per cent
	83.0	**120.0**	**90.1**	**91.1**	**46.7**	**53.3**	**54.3**	**103.5**
$T_{0.25}$	40	120	80	80	40	40	40	43
$T_{0.5}$	89	120	80	80	40	40	44	61
$T_{0.75}$	120	120	120	120	40	62	62	114
#	**1.81**	**1.00**	**2.25**	**2.28**	**1.13**	**1.44**	**1.65**	**1.89**
$R_\theta(T,N)$	100	72.8	92.6	94.3	76.4	71.8	56.9	52.0

DOI: 10.1002/sim

the efficiency ratio of T to T':

$$R_\theta(T,T') = \frac{(z_\alpha + z_{\widetilde{\alpha}_T(\theta)})^2/E_\theta T}{(z_\alpha + z_{\widetilde{\alpha}_{T'}(\theta)})^2/E_\theta T'} \times 100 \tag{22}$$

noting that $(z_\alpha + z_{\widetilde{\alpha}_T(\theta)})^2/\theta^2$ is the sample size of the FSS test with the same type I error probability and power as T. Table I contains $R_\theta(T,N)$ for all tests T and $\theta>0$, where N is the sample size of ADAPT.

ADAPT has power comparable to FSS_{120} at all values of θ while achieving substantial savings in sample size, as shown by the percentiles and mean of the sample size. The three-stage OBF tests have power comparable to ADAPT and FSS_{120}, but ADAPT has sample size savings over the OBF tests, especially for larger $\theta>0$, reflected by the efficiency ratio. The mean number of stages (denoted by #) reveals that although ADAPT allows for the possibility of three stages, most frequently it uses only one or two stages.

The conditional power tests PH, L, SF, and SF$'$ are underpowered at values of $\theta>0$ in Table I. In particular, PH, L, and SF all have power less than 0.6 at $\theta_1 = 0.3$, where ADAPT, FSS_{120}, and the OBF tests have power around 0.9. The lack of power of PH, L, and SF shown by Table I is caused by stopping too early for futility. For example, the PH test stops for futility after the first stage if $S_m/\sqrt{m}$ falls below $z_{p^*} = 1.71$. However, $\text{pr}_{\theta_1}\{S_m/\sqrt{m}<1.71\} = 0.44$, well exceeding the nominal type II error of 0.1. On the other hand, such stringent futility stopping is necessary to control the sample size of conditional power tests. For example, the 0.025-level PH test that stops for futility only when $\widehat{\theta}_m \leqslant 0$ (i.e. with $p^* = 0.5$) has expected sample size greater than 10^7 at all values of θ in Table I, yet power less than 0.9 at θ_1. SF and SF$'$ provide another example of this behavior. Since these tests stop for futility at the first stage when $S_m/m \leqslant \theta_1 - z_{\alpha_0}/\sqrt{m}$, the choice of α_0 determines the maximum sample size. For maximum sample size $M = 120$, SF uses $\alpha_0 = 0.425$, a high rate of first-stage futility stopping that results in small expected sample sizes, low power, and a reduced type I error of 0.012, which is $\alpha = 0.025$ in the absence of futility stopping. In contrast, SF$'$ uses less stringent futility stopping with $\alpha_0 = 0.154$ that corresponds to maximum sample size $5M = 600$, which results in a type I error closer to 0.025 and better power, although it is still underpowered and its expected sample size exceeds 120 at $0.2 \leqslant \theta \leqslant 0.26$. The smallest α_0 that does not perturb the type I error of 0.025 of Shen and Fisher's test is $\alpha_0 = 0.039$, but the resultant test has expected sample size 1856 at $\theta = 0$ and maximum sample size 52 341.

The efficiency ratios relative to ADAPT in Table I are all less than 100 with the exception of PH, L, and SF at $\theta = 0.15$, but it is not clear that the efficiency ratio has much meaning in this case where the power of these tests is so low. For the other cases, it is natural to ask if much more improvement is possible. A benchmark for answering this question is provided by the optimal adaptive tests of Jennison and Turnbull [12, 14] that minimize the expected sample size averaged over a collection of θ values, subject to a given type I error probability and power level at a prespecified alternative θ'. Table II contains the expected sample size of T_k^*, the k-stage test minimizing

$$[E_0(T) + E_{\theta'}(T) + E_{2\theta'}(T)]/3 \tag{23}$$

among all k-stage tests with maximum sample size $M = 120$, type I error probability $\alpha = 0.025$ and power 0.8 at θ', the alternative where FSS_{100} has power 0.8, from [14, Table III]. To this benchmark, we compare ADAPT with the same first group size $m = 29$ as T_3^*, $M = 120$, θ_1 fixed at θ', and $b = 2.94$, $\widetilde{b} = 0.7$, and $c = 2.05$ corresponding to $\varepsilon = \frac{1}{2}$, $\widetilde{\varepsilon} = \frac{3}{4}$. Also included in Table II is the optimal

DOI: 10.1002/sim

Table II. Expected sample size of ADAPT, optimal adaptive and group-sequential tests, with $\alpha=0.025$, power 0.8 at θ', first group size m and maximum sample size $M=120$ for normal data with known variance.

Test	m	E_0N	$E_{\theta'}N$	$E_{2\theta'}N$
ADAPT	29	58.1	81.2	41.5
T_3^*	29	54.9	78.9	39.5
OGS(3)	34	58.2	78.1	43.0
T_2^*	43	64.0	85.3	49.0
OGS(2)	43	64.6	86.2	48.9
T_4^*	24	50.9	75.2	36.0
OGS(4)	29	55.1	74.8	39.8

k-stage 'ρ-family' group-sequential test (denoted by OGS(k)) with $M=120$, groups $2,\dots,k$ of size $(M-m)/(k-1)$, and with m and ρ chosen to minimize (23). Jennison and Turnbull [14] concluded that OGS(k) is a computationally easier alternative to T_k^*, and Table II shows that their expected sample sizes are close at $\theta=0,\theta',2\theta$. Note that ADAPT has expected sample size close to OGS(3) and T_3^* even though the probability that ADAPT uses only one or two stage is 96.4, 83.1, and 98.4 per cent for $\theta=0,\theta'$, and $2\theta'$, respectively, showing that ADAPT very often behaves like a two-stage test. ADAPT has substantially smaller expected sample size than T_2^* and OGS(2), however. On the other hand, T_4^* is more efficient than ADAPT, but this is due in part to its smaller first group of $m=24$, afforded by its additional stage. Here, we have matched the first group $m=29$ of ADAPT to the that of T_3^* for the purpose of comparison, but in practice there is flexibility in its choice of m. The T_k^* and OGS(k) tests, on the other hand, are rigid in their choice of m that is determined by dynamic programming from the prespecified alternative θ', about which there may be some uncertainty before the trial.

Lokhnygina [22], who considers somewhat different objective functions than (23), has computed and plotted the data-dependent total sample size of the optimal two-stage design as a function of the first-stage sample mean $\overline{X}_m$. Her results show the total sample size to be a unimodal function of $\overline{X}_m$, peaking between 0 and θ'. For comparison, Figure 1 plots the function $n(\theta)$ (3) in the sample size updating rule (2) of ADAPT for the setting of Table II. A similar shape is exhibited by Figure 2.2 of [22] on the total sample size function (which is m plus the second-stage sample size) of the optimal two-stage test. This is not surprising because to be optimal, the expected sample size cannot differ much from Hoeffding's [17] lower bound, of which $n(\theta)$ is a close approximation. Figure 1 differs dramatically from the total sample size function of any untruncated two-stage conditional power rule which increases to infinity as $\widehat{\theta}_m$ approaches 0. Jennison and Turnbull [23, p. 672] have also pointed out this and suggested that this is a source of inefficiency of two-stage conditional power tests.

3.2. Case of unknown variance

The optimal adaptive test T_k^* and the optimal group-sequential tests OGS(k) in Table II require the variance of the observations to be known. As pointed out above, in practice there is often little information about the sampling variability before the trial. Dynamic programming is difficult to carry out for the optimal adaptive test when the $X_1, X_2, \dots$ are i.i.d. $\mathrm{N}(\mu,\sigma^2)$, and both μ and σ are unknown, and no analog of T_k^* has been developed in this setting. However, the optimal

DOI: 10.1002/sim

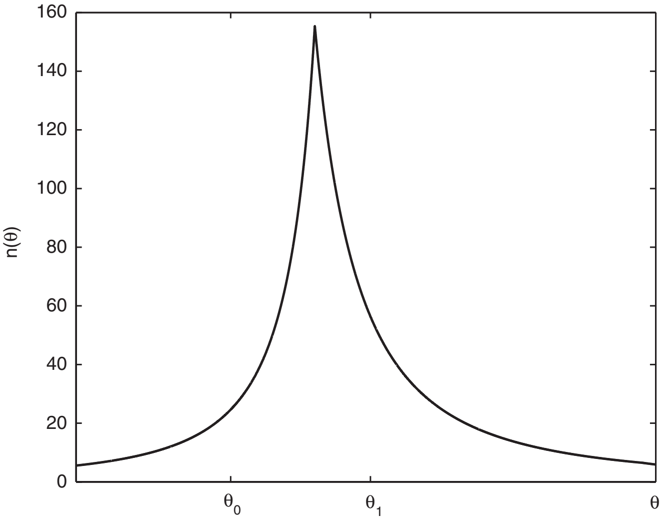

Figure 1. The function $n(\theta)$ when the X_i are $\mathrm{N}(\theta,1)$, $\alpha=0.025$, and $\widetilde{\alpha}=0.2$.

group-sequential tests OGS(k) in Table II can be modified for the present setting by applying their error-spending functions to the sequential t-statistics, as described in [1, Section 11.5], which are denoted by OGS*(k). In this section, we compare ADAPT with OGS*(k) and other tests for a normal mean when the variance is unknown. In the notation of Section 2.3, $\theta=(\mu,\sigma)^{\mathrm{T}}$, $u(\theta)=\mu$, $u_0=0$, and the generalized likelihood ratio statistic (17) is

$$(n/2)\log\left[1+\left(\frac{\overline{X}_n-u_j}{\widehat{\sigma}_n}\right)^2\right]$$

where $\widehat{\sigma}_n$ is the MLE of σ. Denne and Jennison [24] proposed an adaptive group-sequential extension of Stein's [25] two-stage t-test in which the total sample size and stopping boundaries are updated at each stage as a function of the current estimate of σ^2. Lai and Shih [15] introduced tests of composite hypotheses for a multiparameter exponential family which use the same stopping rules (18)–(20) as ADAPT but with prespecified group sizes. The expected number of stages, power, and expected sample size of these tests are given in Table III at various (μ,σ) values. All tests use $m=34$ (with the exception of OGS*(4)) and $M=120$ (with the exception of DJ that has unbounded maximum sample size), the first-stage and maximum sample sizes of OGS(3) in Section 3.1, and nominal power levels $\alpha=0.025$ and $1-\widetilde{\alpha}=0.8$ at $(\mu,\sigma)=(0,1)$ and $(\theta',1)$, respectively, where θ' is as in Section 3.1. Other values of the user-specified parameters of the tests are listed as follows:

- *ADAPT*: The adaptive test described in Section 2.3 with u_1 fixed at θ', $b=2.49$, $\widetilde{b}=0.59$ and $c=2.7$ corresponding to $\varepsilon=\frac{1}{2}$, $\widetilde{\varepsilon}=\frac{3}{4}$.
- *OGS**(k): Jennison and Turnbull's [1, Section 11.5] group-sequential t-test with k groups and the same m and error-spending function as that of OGS(k) in Table II: OGS*(3) uses $\rho=0.99$ and group sizes 34, 43, 43; OGS*(4) uses $\rho=1.13$ and group sizes 29, 30, 30, 31.

DOI: 10.1002/sim

- *DJ*: The adaptive three-stage t-test of Denne and Jennison [24] with $\rho = 0.99$, to match OGS*(3).
- *LS*: Lai and Shih's [15, Section 3.4] group-sequential test with group sizes 34, 43, 43, so that $m = 34$ and $M = 120$.

When $\sigma = 1$, ADAPT, OGS*, and DJ have similar power and expected sample size properties, with ADAPT having the smallest expected number of stages and smaller expected sample size than OGS*(3). LS has the highest power of the five tests, but highest expected sample size too. When $\sigma < 1$ and $\mu = 0$, ADAPT has substantially smaller expected sample size than OGS*(3) and even OGS*(4). DJ has similar operating characteristics to ADAPT and LS when $\sigma < 1$. However, when $\sigma > 1$, the expected sample size of DJ becomes much larger than those of other tests because its total sample size is chosen to be proportional to the estimate of σ^2 at the end of the previous stage. In all cases evaluated, ADAPT has the smallest expected number of stages, less than 2 in each case, showing that it most often behaves like a FSS or two-stage test, as in the variance known setting of Table I.

3.3. Coronary intervention study

The NHLBI Type II Coronary Intervention Study [26] was designed to investigate the cholesterol-lowering affects of cholesytyramine on patients with type II hyperlipoproteinemia and coronary artery disease. Patients were randomized into cholesytyramine and placebo groups, and coronary angiography was performed before and after five years of treatment. It was found that the disease had progressed in 20 of 57 in the placebo group and 15 of 59 in the cholesytyramine group. Proschan and Hunsberger [4] and Li *et al.* [10] have considered how this study could have been extended by using their two-stage tests for the difference in two normal means with common unknown variance. To apply these tests to the NHLBI study, they assumed the first-stage sample size to be $58 = (57+59)/2$ for the normal problem and used the arcsine transformation so that the difference between the transformed binomial frequencies, p_1 for the placebo group and p_2 for the treatment group, is approximately normally distributed; details are given in the following paragraph. As an alternative, we apply the three-stage test in Section 2.3 to two binomial populations. In the notation of Section 2.3, to test $H_0 : p_2 \leqslant p_1$ we have $\theta = (p_1, p_2)^{\mathrm{T}}$, $u(\theta) = p_2 - p_1$, $u_0 = 0$, and the test statistic $\inf_{\theta : u(\theta) = \delta} nI(\widehat{\theta}_n, \theta)$ in (18)–(20) takes the following form:

$$n\left\{\widehat{p}_{1,n}\log\left(\frac{\widehat{p}_{1,n}}{p_{\delta,n}}\right)+\widehat{q}_{1,n}\log\left(\frac{\widehat{q}_{1,n}}{1-p_{\delta,n}}\right)+\widehat{p}_{2,n}\log\left(\frac{\widehat{p}_{2,n}}{p_{\delta,n}+\delta}\right)+\widehat{q}_{2,n}\log\left(\frac{\widehat{q}_{2,n}}{1-p_{\delta,n}-\delta}\right)\right\}$$

where $\widehat{p}_{i,n}$ is the maximum likelihood estimator of p_i based on n observations, $\widehat{q}_{i,n} = 1 - \widehat{p}_{i,n}$, and $p_{\delta,n}$ is the maximum likelihood estimator of p_1 under the assumption $p_2 - p_1 = \delta$. The treatment and placebo groups are assumed to have the same per-group sample size during interim analyses, following Proschan and Hunsberger [4] and Li *et al.* [10].

Letting S_n denote the sum of independent normal random variables with mean μ and variance 1, and following a pilot study of size m resulting in $S_m = s_m$, Proschan and Hunsberger's [4] test chooses n_2 and critical value c to satisfy the conditional power criterion:

$$\mathrm{pr}\{S_{n_2}/n_2^{1/2} > c \mid S_m = s_m, \mu = s_m/m^{1/2}\} \geqslant 1 - \widetilde{\alpha} \tag{24}$$

and type I error constraint

$$\mathrm{pr}_0\{S_{n_2}/n_2^{1/2} > c \mid S_m = s_m\} = \alpha \tag{25}$$

DOI: 10.1002/sim

Table III. Expected number of stages, power and expected sample size of ADAPT and other tests for normal data with unknown variance.

(μ, σ)	ADAPT	OGS*(3)	OGS*(4)	DJ	LS
$\sigma=1$					
$(0, 1)$	1.3	1.6	1.8	2.2	1.8
	2.5 per cent	2.5 per cent	2.5 per cent	2.6 per cent	2.5 per cent
	56.5	58.3	53.5	59.7	67.0
$(0.25, 1)$	1.6	2.1	2.6	2.5	2.4
	62.1 per cent	69.9 per cent	68.8 per cent	57.2 per cent	73.0 per cent
	78.6	82.2	77.8	71.2	93.6
$(\theta', 1)$	1.6	2.1	2.5	2.5	2.3
	72.4 per cent	79.3 per cent	78.5 per cent	68.1 per cent	82.3 per cent
	76.7	79.7	75.0	70.0	90.2
$(0.3, 1)$	1.6	2.0	2.5	2.4	2.2
	78.1 per cent	84.5 per cent	83.7 per cent	74.3 per cent	87.1 per cent
	74.9	77.3	72.7	69.2	87.3
$(2\theta', 1)$	1.1	1.3	1.4	2.1	1.3
	99.6 per cent	99.9 per cent	99.9 per cent	99.7 per cent	99.9 per cent
	42.7	45.0	41.0	52.5	47.8
$\sigma<1$					
$(0, \frac{3}{4})$	1.1	1.6	1.8	1.9	1.4
	2.1 per cent	2.5 per cent	2.5 per cent	2.4 per cent	1.8 per cent
	42.8	58.4	53.5	47.8	51.2
$(0, \frac{1}{2})$	1.0	1.6	1.8	1.0	1.1
	1.4 per cent	2.6 per cent	2.5 per cent	2.4 per cent	0.7 per cent
	34.2	58.1	53.7	34.1	37.0
$(\theta', \frac{1}{2})$	1.0	1.3	1.4	1.0	1.3
	88.3 per cent	99.9 per cent	99.9 per cent	88.7 per cent	95.3 per cent
	35.8	45.0	41.0	34.1	45.6
$(0, \frac{1}{3})$	1.0	1.6	1.8	1.0	1.0
	1.4 per cent	2.5 per cent	2.5 per cent	2.3 per cent	0.5 per cent
	34.0	58.3	53.4	34.0	34.0
$(0, \frac{1}{4})$	1.0	1.6	1.8	1.0	1.0
	1.4 per cent	2.4 per cent	2.4 per cent	2.5 per cent	0.5 per cent
	34.0	58.2	53.4	34.0	34.0
$\sigma>1$					
$(0, 2)$	1.6	1.6	1.8	2.3	2.5
	2.5 per cent	2.5 per cent	2.5 per cent	2.4 per cent	2.5 per cent
	84.2	58.3	53.6	225.5	97.2
$(2\theta', 2)$	1.8	2.1	2.5	2.2	2.4
	78.2 per cent	79.4 per cent	78.4 per cent	99.6 per cent	84.5 per cent
	86.9	79.6	75.3	195.5	94.2
$(3\theta', 2)$	1.5	1.7	1.9	2.0	1.8
	97.8 per cent	97.4 per cent	98.4 per cent	99.9 per cent	99.4 per cent
	64.9	63.0	56.7	178.1	66.6
$(4\theta', 2)$	1.2	1.3	1.4	2.0	1.3
	99.9 per cent	99.9 per cent	99.9 per cent	99.9 per cent	99.9 per cent
	42.9	45.0	42.0	177.6	47.8
$(4\theta', 3)$	1.3	1.8	2.1	2.0	2.0
	96.2 per cent	95.7 per cent	95.3 per cent	99.9 per cent	97.8 per cent
	69.0	67.0	62.2	395.5	75.5

In order to solve for n_2 and c, a parametric form for the probability in (25) is assumed, which contains a user-specified futility boundary h and critical value k for the internal pilot. Li *et al.* [10] introduce a modification of Proschan and Hunsberger's [4] test in which the critical value c is specified before the internal pilot study but h, k, and n_2 are chosen to satisfy (24) and (25) after the internal pilot study. This modification allows approximations to the probabilities in (24) and (25) to be used *in lieu* of a specific parametric form. For the coronary intervention study, Proschan and Hunsberger [4] and Li *et al.* [10] propose using these tests with the variance-stabilizing transformation $S_n=(2n)^{1/2}\{\arcsin(\widehat{p}_{1,n}^{1/2})-\arcsin(\widehat{p}_{2,n}^{1/2})\}$.

Table IV gives the power, per-group expected sample size, and efficiency ratio (22), using the normal approximation, relative to ADAPT (for alternatives $p_2>p_1$) of the following tests for various values of p_1, p_2 near $\frac{15}{59}=0.254$ and $\frac{20}{57}=0.351$, the values observed in the NHLBI study [26].

- *L*: Li *et al.*'s [10] test with $h=1.036$, $k=1.82$, $c=1.7$, $\alpha=0.05$, conditional power level 0.8, and first-stage size $m=58$.
- *PH*: Proschan and Hunsberger's [4] test with $h=1.036$, $k=1.82$, $\alpha=0.05$, conditional power level 0.8, and first-stage size $m=58$.

Table IV. Power, expected sample size, and efficiency ratio (in parentheses and at $p_2>p_1$) of the tests of $H_0: p_2 \leqslant p_1$.

p_1	p_2	L	PH	ADAPT
0.20	0.15	0.7 per cent 63.4	0.7 per cent 63.0	0.3 per cent 98.6
	0.20	5.2 per cent 75.8	5.2 per cent 74.5	5.0 per cent 158.2
	0.30	53.0 per cent 102.0 (89.7)	51.8 per cent 97.2 (90.8)	81.8 per cent 206.1 (100)
	0.35	77.1 per cent 95.3 (73.3)	76.2 per cent 90.7 (75.1)	97.4 per cent 160.5 (100)
0.25	0.20	0.8 per cent 64.7	1.0 per cent 64.5	0.4 per cent 111.2
	0.25	5.2 per cent 77.3	5.1 per cent 75.8	5.0 per cent 171.2
	0.35	**48.3** per cent **97.7 (90.5)**	**47.0** per cent **93.3 (91.9)**	**79.2** per cent **213.1 (100)**
	0.40	72.7 per cent 94.1 (74.1)	71.7 per cent 89.7 (76.3)	96.7 per cent 170.3 (100)
0.30	0.25	0.9 per cent 65.5	0.9 per cent 64.7	0.4 per cent 122.2
	0.30	5.1 per cent 75.1	5.0 per cent 73.7	5.0 per cent 177.0
	0.40	45.3 per cent 96.4 (92.7)	44.3 per cent 92.0 (95.1)	76.6 per cent 218.3 (100)
	0.45	70.9 per cent 96.1 (75.2)	69.9 per cent 91.4 (77.6)	96.2 per cent 176.9 (100)

Note: The boldface numbers represent those at the NHLBI parameter values, where L and PH are markedly under-powered.

- *ADAPT*: The adaptive test described in a previous paragraph with $m=58$, $M=302$ (the maximum sample size of L), and thresholds $b=2.36$, $\widetilde{b}=1.1$, and $c=1.55$ corresponding to $\alpha=0.05$, $\tilde{\alpha}=0.2$, and $\varepsilon=\widetilde{\varepsilon}=\frac{1}{2}$.

All three tests use the same first-stage size $m=58$. ADAPT matches the maximum sample size $M=302$ of L, and the parameters of PH determine its maximum sample size to be slightly larger at 354. The actual power of L and PH is around 50 per cent for the values of p_1 and p_2 in Table IV with $p_2-p_1=0.1$ and is less than 50 per cent when $p_1=0.254$ and $p_2=0.351$ where they were designed to have conditional power 80 per cent. This is caused in part by premature stopping for futility at the end of the first stage. Indeed, L and PH use the same futility boundary and their probability of stopping at the end of the first stage when $p_1=0.254$ and $p_2=0.351$ is 0.47, well exceeding the nominal type II error probability 0.2. One might ask if a conditional power test can avoid this phenomenon by using a larger first-stage sample size so that the estimate $\widehat{p}_2-\widehat{p}_1$ is near 0 less often after the first stage when the true difference p_2-p_1 is substantially greater than 0. If the first-stage sample size of L is raised to 162 (raising the maximum sample size to 1331), the resultant test has power 79 per cent when $p_1=0.254$ and $p_2=0.351$, approximately equal to the power of ADAPT. However, the expected sample size of this version of L is 264 at this alternative, compared with the expected sample size 213.1 of ADAPT. Similar oversampling also occurs for the values of p_1 and p_2 in Table IV with $p_2-p_1>0.1$, where the power of L and PH is closer to the nominal conditional power level of 80 per cent, but the efficiency ratio drops to around 75 per cent.

4. DISCUSSION

Most previous works in the literature on adaptive design of clinical trials and mid-course sample size adjustments have focused on two-stage designs whose second-stage sample size is determined by the results from the first stage using conditional power. 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. This can result in substantial power loss, as shown in Section 3.1. Although Jennison and Turnbull [8] and Tsiatis and Mehta [9] have pointed out the inefficiency of this approach and advocate instead using group-sequential designs, their critique focuses on the use of non-sufficient ‘weighted’ test statistics and variability in the interim estimate. Through our extensive simulation studies, we have shown that another problem with conditional power methods in practice is potential lack of power, which results from the difficulty in bridging conditional power with actual power and in choosing a futility stopping rule.

In their recent survey of adaptive designs, Burman and Sonesson [27] pointed out that previous criticisms of the statistical principles and properties of these 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, whereas most efficient group-sequential designs require the prespecification of the relevant alternative and variance, as in the case of the optimal adaptive tests of Jennison and Turnbull [12, 14]. Moreover, conditional power tests are easy to implement, while optimal adaptive tests require substantial dynamic programming computations. The adaptive tests of Section 2 combine the attractive features of both the conditional power and group-sequential tests. Rather than achieving exact optimality at a specified collection of alternatives through

DOI: 10.1002/sim

dynamic programming, they achieve asymptotic optimality over the entire range of alternatives, resulting in near optimality in practice; see Section 3.1. These tests are based on efficient generalized likelihood ratio statistics 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 (through the implied alternative) that conditional power tests enjoy. As shown in Section 2.3, these generalized likelihood ratio statistics and the associated adaptive tests can be readily extended to multiparameter settings with nuisance parameters and they enjoy near optimality in these more complicated and realistic settings as well, see Sections 3.2 and 3.3.

The possibility of adding a third stage to improve two-stage designs dated back to Lorden [18], who used upper bounds for the type I error probability that are overly conservative for applications to clinical trials, which need to maintain the type I error probability of the test at a prescribed level because of regulatory and publication requirements, see the references in Section 1. We have modified Lorden’s three-stage test by combining its basic features to preserve its asymptotic optimality with those of Lai and Shih [15] for efficient group-sequential designs. The adaptive test in Section 2 makes use of the maximum sample size M to come up with an implied alternative that 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) FSS test of the null hypothesis *vs* the implied alternative. This idea has led to the superior power properties of ADAPT in Table I, comparable to those of the FSS test. Moreover, the expected number of stages of ADAPT in Table I ranges from 1.5 to 2.07 and is less than 2 for all cases in Table III. Therefore, ADAPT is not much less convenient to run than the FSS test (with only 1 stage), in contrast with group-sequential tests with 3, 4, 5 or more interim analyses of the accumulated data. In practical terms, this can provide substantial savings in the operational costs of the trial by eliminating the need for data monitoring at interim analyses since the updated sample size and stopping rule are completely determined at the end of the pilot stage.

On the other hand, there are situations where adding an additional stage or increasing the maximum sample size may be desired, as pointed out by Cui *et al.* [28] and Lehmacher and Wassmer [29]. For example, Cui *et al.* [28] cite 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 $N_{\max}$. The basic idea underlying the proposed test in Section 2 can be easily modified to allow the increase in the maximum sample size from M to no more than $N_{\max}$ after the second stage, resulting in a test with at most four stages. The type I error probability of the modified test can be computed numerically by recursive integration or by Monte Carlo simulations, as described in Section 2.

REFERENCES

1. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC: New York, 2000.
2. Shih WJ. Sample size re-estimation—journey for a decade. *Statistics in Medicine* 2001; **20**:515–518.
3. Whitehead J, Whitehead A, Todd S, Bollard K, Sooriyarachi MR. Mid-trial design reviews for sequential clinical trials. *Statistics is Medicine* 2001; **20**:165–176.

4. Proschan M, Hunsberger S. Designed extension studies based on conditional power. *Biometrics* 1995; **51**: 1315–1324.
5. Fisher L. Self-designing clinical trials. *Statistics in Medicine* 1998; **17**:1551–1562.
6. Posch M, Bauer P. Adaptive two stage designs and the conditional error function. *Biometrical Journal* 1999; **41**:689–696.
7. Shen Y, Fisher L. Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* 1999; **41**:190–197.
8. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; **22**:971–993.
9. Tsiatis AA, Mehta C. On the efficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**:367–378.
10. Li G, Shih WJ, Xie T, Lu J. A sample size adjustment procedure for clinical trials. *Biostatistics* 2002; **3**:277–287.
11. Turnbull BW. Discussion on 'Standard versus adaptive monitoring procedures: a commentary' by Thomas R. Fleming. *Statistics in Medicine* 2006; **25**:3320–3325.
12. Jennison C, Turnbull BW. Adaptive and nonadaptive group sequential tests. *Biometrika* 2006; **93**:1–21.
13. Schmitz N. *Optimal Sequentially Planned Decision Procedures*. Lecture Notes in Statistics, vol. 79. Springer: New York, 1993.
14. Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine* 2006; **25**:917–932.
15. Lai TL, Shih MC. Power, sample size and adaptation considerations in the design of group sequential clinical trials. *Biometrika* 2004; **91**:507–528.
16. Chernoff H. *Sequential Analysis and Optimal Design*. Society for Industrial and Applied Mathematics: Philadelphia, PA, 1972.
17. Hoeffding W. Lower bounds for the expected sample size and the average risk of a sequential procedure. *Annals of Mathematical Statistics* 1960; **31**:352–368.
18. Lorden G. Asymptotic efficiency of three-stage hypothesis tests. *Annals of Statistics* 1983; **11**:129–140.
19. Bartroff J, Lai TL. Supplement to 'Efficient adaptive designs with mid-course sample size adjustment in clinical trials.' 2007. Available from: http://www-rcf.usc.edu/˜bartroff/research/adaptive_supp.pdf.
20. Press NH, Flannery BP, Teukolsky SA, Vitterling WT. *Numerical Recipes in C*: *The Art of Scientific Computing* (2nd edn). Cambridge University Press: Cambridge, 1992.
21. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
22. Lokhnygina Y. Topics in design and analysis of clinical trials. *Ph.D. Thesis*, North Carolina State University, 2004.
23. Jennison C, Turnbull BW. Discussion on 'Are flexible designs sound?' *Biometrics* 2006; **62**:670–673.
24. Denne JS, Jennison C. A group sequential t-test with updating of sample size. *Biometrika* 2000; **87**:125–134.
25. Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* 1945; **16**:243–258.
26. Brensike JF, Kelsey SF, Passamani ER, Fisher MR, Richardson JM, Loh IK, Stone NJ, Aldrich RF, Battaglini JW, Moriarty DJ, Marianthopoulos MB, Detre KM, Epstein SE, Levi RI. NHLBI type II coronary intervention study: design, methods and baseline characteristics. *Controlled Clinical Trials* 1982; **3**:91–111.
27. Burman CF, Sonesson C. Are flexible designs sound? (with Discussion). *Biometrics* 2006; **62**:664–683.
28. Cui L, Hung HM, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55**:835–857.
29. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**:1286–1290.

DOI: 10.1002/sim