Statistica Sinica **7**(1997), 1085-1102

ADAPTIVELY CHANGING SUBGROUP PROPORTIONS IN CLINICAL TRIALS

Dean Follmann

National Heart, Lung, and Blood Institute

Abstract: During the course of a fixed sample size clinical trial, differences in the size of the treatment effect between strata may become pronounced. If rejecting the null hypothesis of no treatment effect is of paramount concern, it may make sense to increase representation of the more responsive stratum to increase power of the trial. Additionally, strata in which treatment is harmful may need to be dropped for ethical reasons. This paper provides conditions under which such adaptation does not affect the type I error rate. The change in power resulting from adaptation under various strategies is investigated. Frequentist and Bayesian approaches to decision making are explored and a simulation is used to provide guidelines as to whether stratum proportions should be altered. A clinical trial with an early and substantial between stratum difference is reanalyzed under adaptation.

Key words and phrases: Bayesian statistics, decision theory, interim monitoring, interaction, strata.

1. Introduction

In clinical trials, there will sometimes be differential response to the treatment for subgroups or strata of patients. Sicker patients may respond more than healthier patients or certain prognostic variables may identify responders. Such heterogeneity may be desirable in that the results of the trial will be generalizable to a wide class of patients. At times, however, interest may focus on finding *any* group for which treatment works. Trials that "screen" treatments or trials for regulatory purposes, where the emphasis is on rejecting the null hypothesis for some group of patients, are settings where wide generalizability of the study may be of secondary interest.

If during the course of a fixed sample size trial substantial differences in response between strata become apparent, it may be tempting to quit randomizing patients from unresponsive strata and only randomize in the responsive strata. Intuitively, if we can accurately identify responsive and unresponsive strata midway through the trial, and subsequently randomize only from the "responsive" subgroups, then we should have a more powerful trial. Another reason for dropping a stratum occurs when the treatment shows substantial harm (see e.g. Coronary Drug Project Group (1981)).

Our interest in this area was spawned by the Prevention and Treatment of Hypertension Study (PATHS) study (Cushman et al. (1994)). Briefly, moderate drinkers were randomized to a control group or were counseled to reduce alcohol consumption. Two strata, 80-89 mm Hg and 90-99 mm Hg diastolic blood pressure, were specified during the design with a goal of at least 45% of the total sample size coming from the upper stratum. Early on in the study, the goal for the upper stratum was not being met and the screening criterion was modified with a net effect of reducing representation from the lower stratum. The screen modification was made after stratum specific estimates of treatment effect were known to some members of the executive committee. We were worried that changing the stratum proportions based on trial data held the potential for inflating the type I error rate.

This paper explores adaptively changing stratum proportions during the course of a clinical trial. We give two different conditions under which the type I error rate is not affected by adaptation. One condition requires that, on the null hypothesis, all the observations are i.i.d. (so that the stratum labels are statistically meaningless) and that we use an unstratified test. Another sufficient condition holds if we use a stratified test statistic and assume that observations are i.i.d. within each stratum on the null hypothesis. If we base our adaptation solely on the between stratum difference in treatment effect, and this estimated difference is independent of the final test statistic, the type I error is unaffected. Such independence occurs when the difference and stratified test statistic have normal distributions.

We also consider the power of the adaptive approach as a function of the stratum specific treatment effects and the proportion allocated to the first stratum and investigate when adaptation is more powerful than a static trial. Since the true treatment effects will be unknown in practice, we also provide a Bayesian formulation of the problem. We evaluate the distribution of the expected increase in the test statistic following adaptation relative to no adaptation, using a conjugate prior for the difference in stratum specific treatment effects. This approach, under a noninformative prior, corresponds to the frequentist approach of monitoring the test of interaction using Pocock's (Pocock (1977)) boundary. We evaluate the power of several decision rules based on this criterion, provide an example and discuss practical issues concerning implementation.

2. Protection of the Type I Error

2.1. Homogeneity null hypothesis

Suppose that we have a two-armed clinical trial with two strata, and a fixed total sample of size n. For each individual we obtain a normally distributed endpoint X, a treatment indicator Z which is 1 (0) for the "active" ("control") arm, and a stratum indicator S which is 1 or 0. At the conclusion of the study, our data consists of $\underline{X} = (X_1, \ldots, X_n)$, $\underline{Z} = (Z_1, \ldots, Z_n)$, and $\underline{S} = (S_1, \ldots, S_n)$.

We want to look at our data midway through the trial, after a proportion $p = m/n$ patients have been randomized (phase 1), with an eye toward changing the allocation proportions amongst strata for the remainder of the trial (phase 2). We define $\bar{S}_1 = \sum_{i=1}^{m} S_i/m$ and $\bar{S}_2 = \sum_{i=m+1}^{n} S_i/(n-m)$, $\underline{S}_1 = (S_1, \ldots, S_m)$, and $\underline{S}_2 = (S_{m+1}, \ldots, S_n)$. Under this setup, $\bar{S}_1$ should be close to the population proportion of people from stratum 1, with $\bar{S}_2$ being quite different from $\bar{S}_1$ if an adaptation is made. At times an experimenter may select $\bar{S}_2$ to be 0 or 1 for reasons of ethics or power. Throughout we will assume equal randomization between arms at the end of each phase and will treat the Zs as fixed constants.

Under a homogeneity null hypothesis we assume that the stratum label is statistically meaningless and that treatment has no effect. Thus the X_is are i.i.d. normal and we have

$$H_0^o : (\delta_0, \delta_1) = (0, 0) \;\; \sigma_0^2 = \sigma_1^2 = \sigma^2,$$

where $\delta_s = E(X_i|Z = 1, S_i = s) - E(X_i|Z = 0, S_i = s)$ is the treatment effect in stratum s, and $\sigma_s^2 = \mathrm{Var}(X_i|S_i = s)$. Without loss, we assume that $E(X_i|Z_i = 0, S_i = s) = 0$, that σ_s^2 is known and define $\Delta = \delta_1 - \delta_0$.

The usual test for this setting is the difference in means:

$$T_u(\underline{X}, \underline{Z}) = \frac{\sum[X_i Z_i - X_i(1 - Z_i)]}{\sqrt{n\sigma^2}}. \tag{1}$$

Note that $T_u(\underline{X}, \underline{Z})$ is free of the stratum labels. Thus we can set $\bar{S}_2$ equal to what we want for whatever reason we want based on the first phase data without affect on the null distribution. Further note that this argument does not make distributional assumptions about $\underline{X}$ so that any test statistic that is only a function of $T_u(\underline{X}, \underline{Z})$ will be unaffected by choosing $\bar{S}_2$ based on any function of the first phase data. Also note that the same argument applies to multiple strata.

2.2. Heterogeneity null hypothesis

Suppose that the distributional assumptions of the previous section hold except that the variance of the response depends on the stratum. This defines a heterogeneity null hypothesis:

$$H_0^e : (\delta_0, \delta_1) = (0, 0) \;\; \sigma_0^2 \neq \sigma_1^2.$$

Let D_s denote the estimate of treatment effect in stratum s at the end of the study. In other words, D_s is the treatment less control difference in the sample means of X for stratum s. Under H_0^e, D_s is normally distributed with mean 0 and variance $V_s = \mathrm{Var}(D_s)$.

The stratified test statistic for this setting (see e.g. Fleiss (1986), Chap. 6) is the weighted sum of the within stratum estimates of treatment effect:

$$T_s(\underline{X}, \underline{Z}, \underline{S}) = \frac{D_0/V_0 + D_1/V_1}{\sqrt{1/V_0 + 1/V_1}}. \tag{2}$$

Note that T_s depends on $\underline{S}$, unlike T_u. If $\underline{S}$ can be treated as a vector of constants, as is typically done, the distribution of $T_s|\underline{S}$ is standard normal. If $\underline{S}$ is determined by some random mechanism which is independent of T_s then $T_s|\underline{S}$ is still standard normal under H_0^e. Thus we can change the distribution of $\underline{S}_2$ based on first phase data and use the usual null distribution as long as T_s and $\underline{S}$ remain independent.

A natural function of the first phase data to consider is the difference in the stratum specific estimates of the treatment effect or $\hat{\Delta} = D_1(m) - D_0(m)$, where $D_s(m)$ is the estimate for the sth stratum after m total subjects have been evaluated. It can be shown that $\hat{\Delta}$ and $T_s(\underline{X}, \underline{Z}, \underline{S})$ have covariance 0, under H_0^e, and hence are independent since they are jointly normal. Note that $\hat{\Delta}/\sigma(\hat{\Delta})$ is also independent of $T_s(\underline{X}, \underline{Z}, \underline{S})$, where $\sigma(\hat{\Delta})$ is the standard deviation of $\hat{\Delta}$.

The following argument shows how using a function of the first phase data not independent of the final test statistic can inflate the type I error. Suppose that $m << n$, $\sigma_0^2 \approx 0$ and $\sigma_1^2/n \approx \infty$. If the test based on the first m patients is "significant" (using a standard normal reference distribution) at level α, we then choose $\bar{S}_2 = 1$ so that the remaining $n - m$ patients are from stratum 1. Because these patients have such a large variance, the test statistic at the end of the study is essentially the same as the test statistic at the end of the first phase and thus significant. If the first phase test statistic is not significant which happens with probabilitiy $1 - \alpha$, we choose $\bar{S}_2 = 0$, so that we obtain $n - m$ additional patients, each with a small variance. Since $m << n$ the final test statistic is virtually independent of the first phase test statistic and our chance of rejection here is close to α. Thus our overall chance of rejection is approximately $\alpha + \alpha(1 - \alpha)$.

If there are more than two strata, we can change strata proportions without affecting the type I error if our decision to modify strata is based on data that is independent of the final test statistic. As an example, suppose the final stratified test statistic involves K multiple strata,

$$T_s = \frac{D_1/V_1 + \cdots + D_K/V_K}{\sqrt{1/V_1 + \cdots + 1/V_K}},$$

where D_j, V_j are the estimated treatment effect, and its variance in stratum j based on all the data. It can be shown that, under a heterogeneity null hypothesis, T_s has 0 covariance with $D_k - \bar{D}$ and hence T_s is independent of $D_1 - \bar{D}, \ldots, D_K - \bar{D}$. It follows that if we choose the stratum with the largest mean based on $m < n$ patients, the type I error is not affected. Similarly, we could drop strata with the lowest means.

The argument for non-inflation of the type I error is extended to a wide class of test statistics in the appendix.

3. Adaptive Power by Monitoring the Test of Interaction

While one can adaptively modify the allocation proportions between strata during the course of a clinical trial without inflating the type I error rate, static trials also do not inflate the type I error rate. The results of the previous section provide reassurance concerning the potential for subtle bias as in the PATHS example. However they can also be used to justify adaptation chosen explicitly to increase power. Intuitively, adaptation should result in increased power if there is a large difference in response between strata, and if we are likely to correctly identify the stratum with the larger response. In this section we evaluate power under a few simple adaptive strategies where we allow ourselves a single adaptation. Our strategies are based on monitoring the test of treatment by strata interaction.

For simplicity, we evaluate a clinical trial under the distributional assumptions of the previous section with homogeneous variance σ^2. At the end of the study, we use the unstratified test statistic (1). The first phase between stratum difference $\hat{\Delta}$ has a normal distribution with mean Δ and variance $\sigma^2(\hat{\Delta}) = 4\sigma^2/[np\bar{s}_1(1-\bar{s}_1)]$. The treatment by stratum test of interaction has test statistic $\hat{\Delta}/\sigma(\hat{\Delta})$.

For illustration, we consider the adaptive procedure where we stop randomizing to a stratum if the first phase interaction test is large enough. We will call this extreme adaptation *adaptive exclusion.* If an adaptation is not made we choose $\bar{S}_2 = \bar{s}_1$, a fixed constant. Our decision rule can thus be written as

$$\bar{S}_2 = \begin{cases} 0, & \text{if } \hat{\Delta} < -c\,\sigma(\hat{\Delta}), \\ \bar{s}_1, & \text{if } |\hat{\Delta}| \leq c\,\sigma(\hat{\Delta}), \\ 1, & \text{if } \hat{\Delta} > c\,\sigma(\hat{\Delta}). \end{cases}$$

Given $\bar{S}_2$, (and the constant $\bar{s}_1$), the unstratified test statistic based on all the data has a normal distribution with variance 1 and mean given below:

$$2\sqrt{n}E(T_u(\underline{X},\underline{Z})|\bar{S}_2) = \begin{cases} m[\delta_1\bar{s}_1 + \delta_0(1-\bar{s}_1)] + (n-m)\delta_0, & \text{if } \bar{S}_2 = 0, \\ n[\delta_1\bar{s}_1 + \delta_0(1-\bar{s}_1)], & \text{if } \bar{S}_2 = \bar{s}_1, \\ m[\delta_1\bar{s}_1 + \delta_0(1-\bar{s}_1)] + (n-m)\delta_1, & \text{if } \bar{S}_2 = 1. \end{cases}$$

Thus the overall power of a procedure that allows adaptation is

$$Pow(\delta_0, \delta_1, m, c) = \sum_{\bar{S}_2} P(T_u(\underline{X},\underline{Z}) > 1.96|\bar{S}_2)P(\bar{S}_2).$$

To evaluate power, we imagined planning a trial under the assumption that $\delta_0 = \delta_1 = \delta$ and n was chosen to yield 80 percent power. We wanted to evaluate the power if in fact $(\delta_0, \delta_1) = (0, \delta)$. While this is a fairly large departure from $\delta_0 = \delta_1 = \delta$, it is useful for illustration. Less dramatic departures should have proportionally smaller effects. We consider three different first phase allocation proportions of $\bar{s}_1 = .2$, .5, and .8, and let $c = 0$ or 2. Note that if $\delta_0 = \delta_1$ then no matter when we look or what decision we make, power is unaffected, so changing allocation proportions cannot compromise power here.

Figure 1 graphs the power for the three allocation proportions for $c = 0$ and 2 as a function of $p = m/n$. Thus the power at the extreme right of each curve corresponds to an unadaptive design. First note that when the allocation proportion is either .2 or .5 we substantially increase power with adaptation using a $c = 0$ decision rule, with the best time to make a decision fairly early in the study. A large gain is possible when $\bar{s}_1 = .2$ since with no adaptation, most patients will come from the stratum with the smaller treatment effect.

With an allocation proportion of .8, however, the $c = 0$ decision rule loses power. The reason is that with a static trial, 80% of the patients will end up in the stratum with the largest treatment effect. Furthermore, if we make a decision early, there is a larger chance that we will choose the stratum with no effect and substantially reduce power. Thus there is relatively little improvement and moderate harm possible with adaptive exclusion in this setting. For all $\bar{s}_1$, the $c = 2$ decision rule is similar to no adaptation.

Figure 2 graphs the power for a $c = 0$ decision rule where we show the effect of four (δ_0, δ_1)s subject to $\delta_0 + \delta_1 = .5$ with $\bar{s}_1 = .5$. For these parameters, it is apparent that looking too early may compromise power and that the best time to look is about midway through the trial. With only moderate departures from $\delta_0 = \delta_1$, there is relatively little benefit if we solely base our decision rule on within trial data.

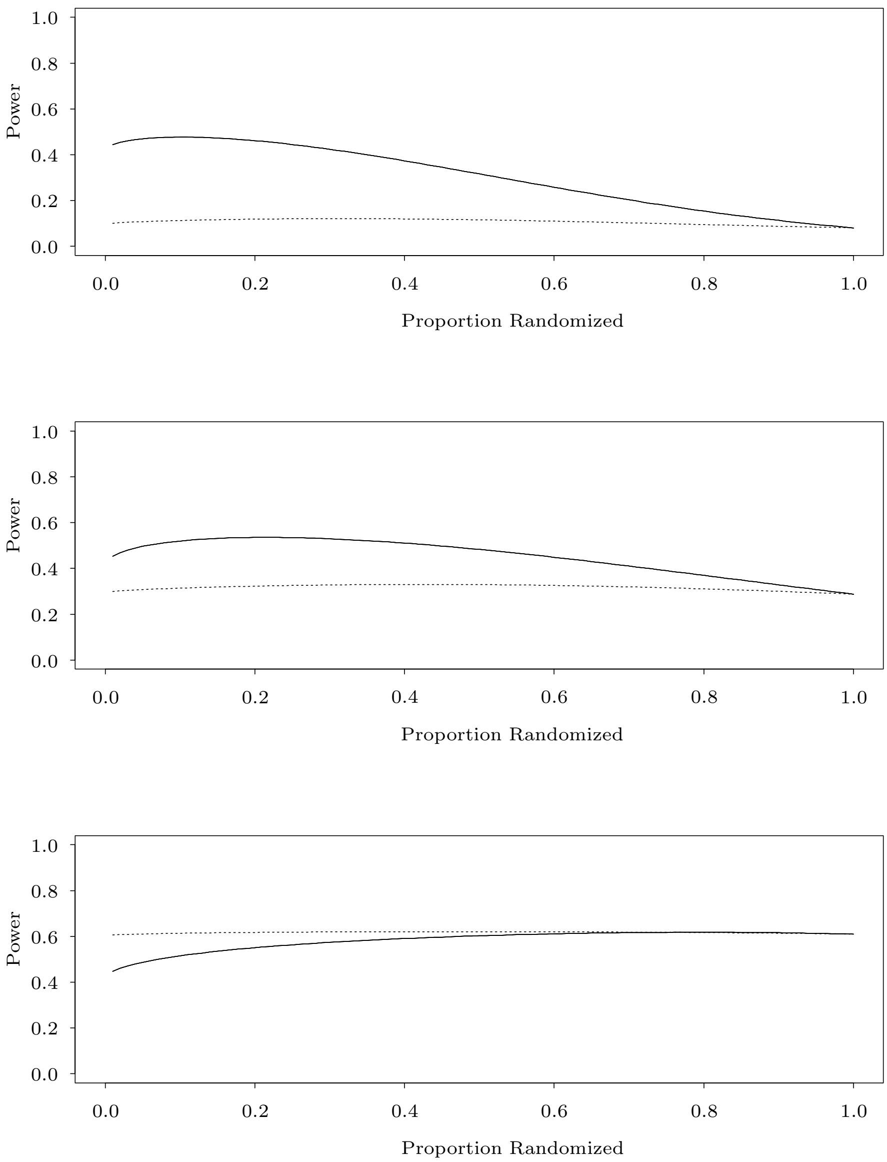

Figure 1. Power of a single look adaptive procedure as a function of when we look. We randomize only from the stratum with the larger observed mean if the between stratum difference in treatment effects $(\hat{\Delta})$ exceeds $c\sigma(\hat{\Delta})$ in absolute value, where $c = 0$ (solid curve) or $c = 2$ (dashed curve). The standardized treatment effects in the two strata are (0, .28) for all three figures with first phase allocation proportions $\bar{s}_1$ for the $\delta_1 = .28$ stratum going from .2 to .5 to .8 (top to bottom).

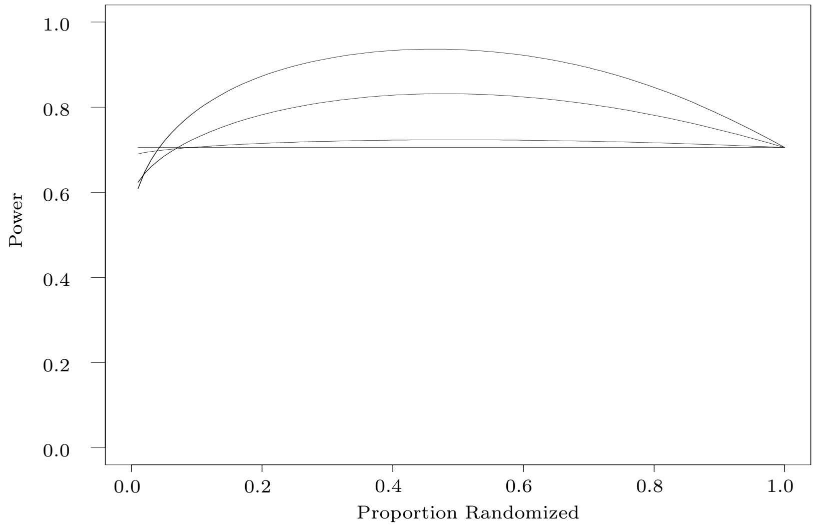

Figure 2. Power of a single look adaptive procedure as a function of when we drop the stratum with the smaller observed treatment effect. Standardized treatment effects in the two strata are (.50,.0), (.4,.1), (.3,.2), and (.25,.25), going from the top curve to the bottom curve.

4. Bayesian Formulation

The approach of the previous section makes the decision to drop a stratum solely on the trial data. If information concerning δ_0, δ_1 is available, incorporating it should give us more powerful decisions earlier. In particular, gains in power larger than those presented in the figures are possible. Such considerations lead naturally to a Bayesian approach to the problem. We choose a simple formulation—we assume that we are conducting a trial, that a proportion p of the patients have been randomized, and we are trying to decide whether or not to adapt at this single point. We are interested in testing H_0^o and will use the unstratified test at the end of the study. We make the distributional assumptions of the previous section and assume that σ^2 is known and Δ is unknown.

To describe the uncertainty about Δ, we use a conjugate prior. At the start of the trial, we assume that knowledge about Δ can be described by a normal distribution with mean d and variance τ^2. At the end of the first phase, $\hat{\Delta}$ summarizes the trial information about Δ. The posterior distribution of Δ is normal with mean and variance

$$\frac{d/\tau^2 + \hat{\Delta}/\sigma^2(\hat{\Delta})}{1/\tau^2 + 1/\sigma^2(\hat{\Delta})}, \quad \frac{1}{1/\tau^2 + 1/\sigma^2(\hat{\Delta})},$$

respectively (DeGroot (1970), Chap. 9).

A natural way to quantify the potential gain following adaptation is to calculate the difference in the expected values of the test statistics for adaptation and a static trial. If we fail to adapt, we assume that the second phase stratum proportions will be about the same as the first phase, thus the expected gain is

$$U_\Delta(\bar{s}_1, S) = E(T_u(\underline{X}, \underline{Z})|\bar{S}_1 = \bar{s}_1, \bar{S}_2 = S) - E(T_u(\underline{X}, \underline{Z})|\bar{S}_1 = \bar{s}_1, \bar{S}_2 = \bar{s}_1)$$

which reduces to

$$U_\Delta(\bar{s}_1, S) = \frac{n[(1-p)(S-\bar{s}_1)]}{2\sigma}\Delta. \tag{3}$$

To aid in deciding on whether or not to adapt, we suggest calculating the posterior probability that choosing stratum S for the remainder of the study results in a larger test statistic (in expectation) than a static trial:

$$P(U_\Delta(\theta, S) > 0) = P(a(S)\Delta > 0) = \bar{\Phi}(-\text{sgn}[a(S)]E(\Delta)/\sigma(\Delta)), \tag{4}$$

where $a(S) = (n[(1-p)(S-\bar{s}_1)])/(2\sigma)$ and $E(\Delta), \sigma(\Delta)$ are the posterior mean and standard deviation of Δ. If (4) is large, one would be inclined to drop the inferior stratum.

Note that if a noninformative prior is used, monitoring based on (4) is equivalent to monitoring with the test of interaction, because in this case $E(\Delta)/\sigma(\Delta) = \hat{\Delta}/\sigma(\hat{\Delta})$, and $\bar{\Phi}(z)$ is a monotone function of z. Specifically, the inferior stratum is dropped if $|\hat{\Delta}/\sigma(\hat{\Delta})|$ exceeds $c = \bar{\Phi}^{-1}(\eta)$. This can be viewed as applying Pocock's method (Pocock (1977)) for monitoring the test of efficacy to monitoring the test of interaction. Of course, the purpose here is much different.

While a prior on Δ may be difficult to specify, a noninformative prior on Δ will tend to encourage early dropping of a stratum relative to a proper prior with $d = 0$ and τ^2 finite. Noninformative priors should be evaluated with this in mind. In general, we recommend using a few priors that bracket the range of plausibility for Δ and only dropping a stratum if this decision is robust to prior specification.

If we use a stratified test statistic to test H_0^e with normal data (3) can still be used, say with σ being some combination of σ_0, σ_1. Since (3) depends only on the unknown parameter Δ, we can update our prior for Δ using the first phase data only through $\hat{\Delta}$. Thus we need not worry about type I error inflation here either. Unfortunately, (3) does not equal $E(T_s(\underline{X}, \underline{Z}, \underline{S})|\bar{s}_1, S) - E(T_s(\underline{X}, \underline{Z}, \underline{S})|\bar{s}_1, \bar{s}_1)$ and cannot be interpreted as an expected increase in the test statistic. The difference $E(T_s(\underline{X}, \underline{Z}, \underline{S})|\bar{s}_1, S) - E(T_s(\underline{X}, \underline{Z}, \underline{S})|\bar{s}_1, \bar{s}_1)$ involves both δ_0 and δ_1. If priors for both δ_0 and δ_1 are updated using first phase data from the study, there is the potential for type I error inflation.

While, strictly speaking, our formulation has been for a single examination of the data, it can be readily generalized to multiple examinations. At any point, no matter how many times we have updated before, we can calculate an updated posterior and apply (4) to decide whether to adapt. The timing of the multiple examinations, however, needs to be considered. Most large clinical trials are formally monitored, primarily for safety and efficacy, on a few discrete occasions roughly evenly spaced during followup. We would recommend that adaptation decisions be made at these times as well. In particular, we would generally recommend that a decision to drop a stratum be made by the time the study is halfway through, that we should not modify stratum accrual until a decision to drop is made, and that only a few looks should be made. In most applications, examining the data on a few occasions (say after $n/4$ and $n/2$ subjects) should be adequate, and adaptation after more than half the patients have been randomized would not be worth the bother. This recommendation also ensures that losses in power based on very early adaptation, as illustrated in the figures, will be avoided.

5. Simulation of Adaptive Power

To evaluate the performance of a clinical trial that allows for adaptation, we conducted a small simulation. We imagined a clincal trial where analysis occurs after each quarter of the sample has been evaluated. We make the homogeneity distributional assumptions of the previous section, and only record the treatment less control difference in means for each strata over the four phases of the study. We have two strata, 0 and 1, which occur with probability $(1-\theta)$ and θ, respectively. We imagine a large study and take $8\sigma^2/N = 1$ for simplicity. We thus generate $Y_{s\ell}$ from a normal distribution with mean δ_s and variance $2/[s\theta + (1-s)(1-\theta)]$ where $s = 0, 1$, and $\ell = 1, 2, 3, 4$. These correspond to the means over the four phases for a study with no adaptation where $\bar{s}_1 = \bar{s}_2 = \theta$.

We allow for adaptation after 1/4 or 1/2 of the data are evaluated so we also generate $Y^*_{s\ell}$, from a normal distribution with mean δ_s and variance 2 for $s = 0, 1$ and $\ell = 2, 3, 4$. Thus if we adapt to stratum s at look 1, the data for our clinical trial consists of $Y_{01}, Y_{11}, Y^*_{s2}, Y^*_{s3}, Y^*_{s4}$.

Our decision rule is to choose stratum S for the remainder of the study if

$$P(U_\Delta(\theta, S) > 0) = \bar{\Phi}(-\text{sgn}[a(S)]E(\Delta)/\sigma(\Delta)) > \eta$$

for $\eta \geq .5$, where $E(\Delta), \sigma(\Delta)$ are the current posterior mean and standard deviation. This allows for a rich class of possible decision rules. For example, if $\eta = .5$, we adapt to the stratum with the larger mean at look 1. Increasing values of η specify greater assurance that the adaptation will be beneficial.

We considered various combinations of δ_0, δ_1 and θ and evaluated different decision rules as specified by η. For each scenario 10,000 clinical trials were evaluated. We specified a true Δ and chose δ_1, δ_0 so that a static trial with the stratum 1 proportion fixed at θ would have power = .50. We specified the prior variance as a fraction f of the sampling variance of $\hat{\Delta}$ we would obtain in a study without adaptation, thus $\tau^2 = f\sigma^2(\hat{\Delta})$. We assumed the prior mean to be 0 and took f large so that the prior is essentially noninformative. A noninformative prior is interesting because the performance of adaptation would likely be better with an informative prior not centered at 0. Additionally, a noninformative prior corresponds to the frequentist strategy of monitoring the test of interaction.

Table 1. Evaluation of a Bayesian strategy for adaptive exclusion under a noninformative prior. Strategy is equivalent to frequentist monitoring of the test of interaction with a Pocock-type boundary. The inferior stratum is dropped if the probability that adaptation will result in a larger test statistic is larger than η. Adaptation is allowed to occur after 1/4 or 1/2 of the patients have been evaluated. A static trial has power .50 for all settings.

True			Power		
Δ	θ	η	Overall	Given Adaptation	P(adapt)
1	.2	.5	.65	.65	1.00
		.8	.61	.67	.60
		.9	.50	.71	.04
1	.5	.5	.56	.56	1.00
		.8	.55	.59	.61
		.9	.50	.66	.05
1	.8	.5	.43	.43	1.00
		.8	.48	.46	.59
		.9	.51	.55	.04
3	.2	.5	.85	.85	1.00
		.8	.82	.93	.76
		.9	.57	.96	.15
3	.5	.5	.82	.82	1.00
		.8	.81	.89	.82
		.9	.58	.86	.24
3	.8	.5	.60	.60	1.00
		.8	.63	.67	.77
		.9	.53	.66	.15

Table 1 presents the results. For each pair (Δ, θ), three adaptation strategies ($\eta = .5, .8$, and .9) are presented. Recall that for a static trial, power would be

.50. Another benchmark corresponds to $\eta = .5$ where we always adapt to the stratum with the larger mean at the first look. The power decreases as θ goes from .2 to .5 to .8, i.e. the stratum with the larger mean goes from rare to common. The probability of adaptation is larger for larger Δ and smaller for larger η.

The use of $\eta = .5$, which makes it easy to adapt, results in larger gains in overall power when $\theta = .2$ than other strategies, but larger losses in overall power when $\theta = .8$. In particular, for $\Delta = 1$, use of $\eta = .5$ gives power of .43. Since no adaptation gives power of .50, this strategy can result in a nontrivial loss of power. For $\eta = .8$ or .9, the overall power is essentially unchanged. The use of $\eta = .9$ results in power which is closest to .5 for all scenarios, thus $\eta = .9$ is best if one is risk averse.

We imagine that adaptation will often not be planned at the start of a clinical trial, but might be considered if there were strong evidence that there was a difference between the two strata. It seems that typically one would only want to adapt if one were quite certain that adaptation would result in substantially greater power. If so, the power given that an adaptation occured is particularly interesting. The conditional power increases with η while the frequency of adaptation decreases. If one wanted to adapt rarely and only if one were quite sure it would be beneficial, use of $\eta = .9$ is best.

In passing, we note that if Δ is zero, power is .50 for all strategies. The probability of switching was simulated at 1.00, .17, and .03 for $\eta = .5, .8, .9$ respectively. Thus $\eta = .9$ would result in quite infrequent adaptations if in fact Δ were 0.

We also evaluated power based on an informative prior centered at 0 with $f = 1$. For $\eta = .5$ and $\eta = .9$ the results was, respectively, identical and very similar. For $\eta = .8$ the informative prior resulted in adaptation about half as often, similar conditional power, and thus overall powers closer to .50 than with the uninformative prior. Presumably, for informative priors centered away from 0, the power under an adaptive strategy should be larger than that of Table 1.

6. Example

A recent 2 period crossover clinical trial was undertaken to test the efficacy of a new drug in reducing epileptic seizures. Patients needed to have a history of simple partial or complex partial seizures to be enrolled. An initial baseline period of 8 weeks was followed by active and control periods, each lasting 11 weeks. To allow for the drug to reach its full effect, in our analysis seizures are counted only during the last 8 weeks of each period. Randomization was stratified by whether ($s = 1$) or not ($s = 0$) patients had a history of secondarily generalized tonic-clinic (GTC) seizures. There was felt to be a possibility of differential treatment effect between strata. A total of 56 patients were randomized and finished the study, 22 of whom were from stratum 1.

As primary endpoint we take truncated relative change:

$$R = \frac{\#\text{ Seizures during control phase} - \#\text{ Seizures during active phase}}{\#\text{ Seizures during baseline phase}}.$$

We truncate R at $+1$ or -1 if $|R|$ exceeds 1 and denote the truncated endpoint by X. We treat the X's as being i.i.d. from some distribution and use the standardized unstratified mean of X as our test statistic: $T_u(\underline{X}) = \bar{X}/\sigma(\bar{X})$. $T_u(\underline{X})$ is a one sample paired difference test with known variance. For simplicity, we treat the final sample variance .1895 as the known value for $\sigma^2 = \text{Var}(X)$. We treat $\bar{X}$ as being approximately normal throughout our illustration. We are interested in testing the homogeneity null hypothesis:

$$H_0^0 : (\delta_0, \delta_1) = (0, 0) \;\; \sigma_0^2 = \sigma_1^2 = \sigma^2,$$

where $\delta_s = E(X|S = s)$, and $\sigma_s^2 = \text{Var}(X|S = s)$.

After 1/4 of the patients had been evaluated (7 from each stratum), the estimates of treatment effect were .205 for patients with a history of GTC seizures and -.236 for patients without a history of GTC seizures. Thus there was a substantial difference between the two strata. The test of interaction gives a value of 1.90. Based on frequentist criteria, adaptation to $S = 1$ seems worth considering.

To apply the Bayesian method, we specify a prior distribution for Δ that is normal with mean zero and consider three prior variances equal to .1,1, and 10 times the variance of the estimate of Δ that will be obtained at the end of the study. That is, $\tau^2 = .1895(1/34 + 1/22)f$, with $f = .1, 1, 10$. Thus $f = 10$ can be loosely interpreted that what we know about Δ before the study is 1/10 of what we will know about Δ following the experiment. To aid in deciding on whether or not to adapt, we form $U_\Delta(\bar{s}_1, S) = E(T_u(\underline{X})|\bar{s}_1, S) - E(T_u(\underline{X})|\bar{s}_1, \bar{s}_1)$, and calculate $P(U_\Delta(\bar{s}_1, S) > 0)$ using the current posterior distribution for Δ. This reduces to the RHS of (4).

Table 2 reports the expected increase in the test statistic under the three different priors. For $f = .1$, the probablity of a larger mean under adaptation is only slightly larger than .5. Note that the probability of a larger mean after choosing the inferior stratum is .48. Thus if we are fairly certain that the true Δ is zero it does not matter which stratum we choose and thus we would want to continue randomizing to both arms. If we are less certain that the true Δ is zero (e.g. $f = 1$ or 10) adapation seems attractive. In passing, we note that for a noninformative prior, the probability of a larger mean after adaptation is .84.

Table 2. The posterior probability that the final test statistic will have a larger mean under adaptation relative to a static trial, The posterior distribution is based on 1/4 of the data. The prior variance is f times the variance of the estimate of Δ based solely on trial data at the end of the study.

Prior Moments		Posterior Moments		Posterior
d	f	$E(\Delta)$	$\sigma^2(\Delta)$	$P(U_\Delta(\bar{s}_1, S = 1) > 0)$
0	.1	.011	.037	.52
0	1	.092	.106	.61
0	10	.320	.193	.77

Based on these probabilities, one might have decided to only randomize patients without a history of GTCs after 1/4 of the trial patients had been evaluated. How would the study have turned out with this adaptation? Because no adaptation was undertaken, we pretend that the second phase estimate of the treatment effect in the GTC stratum of .286 was based on 49 patients rather than 15 patients. Our guess as to the test statistic in an adaptive trial is thus $((.205 - .236) \times 7 + .286 \times 49)/\sqrt{56(.1859)} = 4.27$. The actual test statistic based on no adaptation was 3.22. Thus, in this trial adaptation could have resulted in a substantially stronger conclusion being drawn, albeit for a different population of patients.

7. Discussion

At the simplest level, this paper has given conditions under which we need not worry about inflating the type I error when we adaptively change stratum proportions. The PATHS study discussed in the introduction could vigorously attempt to achieve a stratum goal without fear of type I error inflation even though the extent of vigor might have been influenced by early knowledge of stratum-specific treatment effects. Hypothetically, one might discover that a treatment was harmful to patients in one stratum and helpful to patients in the other stratum. While ethically bound to discontinue the trial in the inferior stratum, one could finish the trial in the other stratum without afffecting the type I error rate.

More ambitiously, this paper has provided a framework to evaluate where an underpowered study may be modified to increase power. Such adaptation should be appealing in the early stages of treatment evaluation when relatively little is known about which groups of patients would benefit. The adavantages should be most pronouced if one observes a substantial difference between strata and the trial has poor power unless the trial is changed. At times, it may offer a way to salvage trials that would otherwise be compromised due to heterogenous response across strata.

While our discussion has focused on the setting with two strata, the basic idea remains the same for multiple strata. Of course, looking at the treatment effects from many strata midway through the trial makes it more likely that one stratum will seem to be substantially better than the others. This could affect the power of the procedure.

Our Bayesian approach to dynamically changing allocation probablities to increase power is quite similar to the Bayesian approach to adaptively changing randomization probabilities to the treatment and control arms so that most patients get randomized to the superior arm. Papers that have explored changing randomization probabilties include Anscombe (1963), Colton (1963), Zelen (1969), and Cornfield, Halperin and Greenhouse (1969). More generally, adaptive exclusion can be viewed as a type of two-armed bandit problem (see e.g. Berry and Fristedt (1985)). The more elaborate approach of Cornfield et al can be readily applied to our setting. Under their formulation one is allowed a single look at any time during the trial. The expectation of the final test statistic is then maximized as a function of p, and $\bar{s}_1, \bar{s}_2$.

In practice, changing the stratum proportions may cause logisitical or administrative difficulties. In such settings the most that may be possible is to aggressively pursue patients from the apparently superior stratum to enroll in the study. Patients from the apparently inferior stratum would not be so vigorously recruited. Such an adaptation should improve power, though not as substantially as when one drops the stratum with the poorer observed response.

The issue of generalizability of the study is of some concern. The stratum proportions have been chosen adaptively and some may find it unappealing to condition on the stratum proportions and pretend that no adaptation took place. However, in any clinical trial with a significant result, it would be hard to argue that there is no evidence of effect in the subgroup with the largest observed effect. Thus at the worst, the results of the study should be generalizable to the adaptively chosen subgroup. Under this view, our formulation can be viewed as a simple way to correct for the type I error inflation that would result if we only included the chosen subgroup.

Adaptive exclusion trials are similar in spirit to enrichment designs, where prior to randomization, patients who are likely to show a response are identified and then randomzied. Temple (1994) argues that enrichment designs can be useful in the initial stages of drug development when the "first task is to find *any* group in which the drug can be shown to work". He also feels that concern over generalizability of such "early" studies may be overstated.

Additionally, adaptive exclusion trials can be viewed as consistent with the "uncertainty principle" approach to randomization (Byar et al. (1990)). Under the uncertainty principle, if a clinician or patient feels that one of the arms would be definitely inferior, the patient is not randomized. In slightly different terms, one can imagine that each clinician is adaptively modifying inclusion criteria to exclude strata for which randomization is felt to be unethical. Adaptive exclusion dynamically modifies inclusion criteria, but the decision is made for the entire trial and on the basis of within trial data. Proponents of the uncertainty principle can ensure that the type I error rate will not be affected by judiciously choosing what interim data to release to clinicians.

It is important to recognize that dropping a stratum need not imply that there is convincing evidence of a true difference between strata. An analogy should make the point. Suppose that one had to make bets on whether a coin would turn up heads based on the sequence HHHTHTHTHHT and no other information. A test of whether the coin was biased would not reject. Nonetheless, a good strategy is to bet on the historical winner. One does not need to infer at the end whether there is significant evidence that the coin is biased. One merely collects one's winnings. In our setting we are not trying to say subgroups differ. We are trying to determine if treatment is beneficial to some subgroup of patients.

8. Appendix:Protection of the Type I Error in General

In general, we can write a global homogeneity hypothesis as

$$H_0^o : F(x|Z=1, S=s) = F(x|Z=0, S=s) = F(x) \text{ for all } s,$$

where an individual's response X_i, given Z, S has distribution $F(x|Z,S)$ and may be vector valued. As long as we use a test statistic of the form $T_u(\underline{X}, \underline{Z})$ that is constant as a function of $\underline{S}$, then under H_0^o the marginal distribution of T_u is unaffected by choice of $\underline{S}$ for the reasons given previously.

In general, we can write a heterogeneity null hypothesis as

$$H_0^e : F(x|Z=1, S=s) = F(x|Z=0, S=s) = F(x|S=s).$$

As before, if our stratified test statistic at the end of the study (T_s) is independent of $\underline{S}$, then the null distribution for $T_s|\underline{S}$ which treats the stratum labels as fixed constants can be used. As before, we can choose a distribution for $\underline{S}_2$ on the basis of functions of the first phase data that are independent of T_s.

The first phase between stratum difference is independent of the final test statistic for a wide variety of settings, at least asymptotically. Let t denote the calendar time of the study with $t = 0$ being the time of first randomization and τ being the end of followup. Denote the estimate of treatment effect within the sth stratum at calendar time t by $\mathcal{D}_s(t)$. For a wide variety of endpoints, $\mathcal{D}_s(t)$ converges to a Gaussian Process $\mathcal{D}_s^*()$ with mean δ_s for all t and

$\mathrm{Cov}[\mathcal{D}_s^*(t), \mathcal{D}_s^*(t')] = V_s(t')$, with $t \le t'$. Examples here include the one and two sample t-tests, the log-rank statistic, and tests based on a linear random effects model for each subject in a trial with repeated measurements (see Lan and Zucker (1993)). In general, however, tests based on repeated measurements from each subject do not have the above covariance structure. Other cases for survival endpoints include weighted log-rank statistics (Tsiatis (1982)), the Pepe-Fleming statistics (Murray and Tsiatis (1996)), and the two-sample difference in Kaplan-Meier proportion (Lin, Shen, Ying and Breslow (1996)). In the parlance of Lan and Zucker (1993), $\mathcal{D}_s^*(\ )$ is known as an $E - process$.

The stratified test at the end of the study can be written as

$$\mathcal{T}_s(\tau) = \frac{\mathcal{D}_0(\tau)/V_0(\tau) + \mathcal{D}_1(\tau)/V_1(\tau)}{\sqrt{1/V_0(\tau) + 1/V_1(\tau)}}.$$

Using the fact that $\mathcal{D}_0^*(), \mathcal{D}_1^*()$ are independent Gaussian processes with covariances as above, one can show that $\mathcal{T}_s(\tau)$ is asymptotically independent of $\mathcal{D}_1(t) - \mathcal{D}_0(t)$ under the null hypothesis H_0^e. It follows that the type I error rate is unaffected even if $\bar{S}_2$ depends on $\mathcal{D}_1(t) - \mathcal{D}_0(t)$.

Acknowledgements

I would like to thank Michael Proschan for several provocative discussions on this topic and Todd Sahlroot for discussion about the data.

References

Anscombe, F. J. (1963). Sequential medical trials. *J. Amer. Statit. Assoc.* **58**, 365-383.

Berry, D. A. and Fristedt, B. (1985). *Bandit Problems.* Chapman and Hall, New York.

Byar, D. et al. (1990). Design considerations for AIDS trials. *New England Journal of Medicine* **323**, 1343.

Coronary Drug Project Research Group (1981). Practical aspects of decision making in clinical trials: the coronary drug project as a case study. *Controlled Clinical Trials* **1**, 363-376.

Colton, T. (1963). A model for selecting one of two medical treatments. *J. Amer. Statit. Assoc.* **58**, 388-400.

Cornfield, J., Halperin, M. and Greenhouse, S. (1969). An adaptive procedure for sequential clinical trials. *J. Amer. Statit. Assoc.* **64**, 759-770.

Cushman, W. C., Cutler, J. A., Bingham, S., Harford, T., Hanna, E., Dubbert, P., Collins, J., Dufour, M., Follmann, D., Allender, P. (1994). Prevention and treatment of hypertension study (PATHS): Rationale and design. *Amer. J. Hypertension*, in press.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments.* John Wiley, New York.

Lan, K. K. G. and Zucker D. M. (1993). Sequential monitoring of clinical trials: The role of information and brownian motion. *Statist. Medicine* **12**, 753-766.

Lin, D. Y., Shen, L., Ying, Z. and Breslow, N. E. (1996). Group sequential designs for monitoring survival probabilities. *Biometrics* **52**, 1033-1041.

Murray, S. and Tsiatis, A. (1996). Sequential methods for the two sample censored data problem using the weighted difference of integrated survival curves. *Biometrics* **52**, 137-151.

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

Temple, R. J. (1994). Special study designs: Early escape, enrichment, studies in non-responders. *Comm. Statist. Theory Methods* **23**, 499-531.

Tsiatis, A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *J. Amer. Statist. Assoc.* **77**, 855-861.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *J. Amer. Statist. Assoc.* **64**, 131-146.

Office of Biostatistics Research, National Heart, Lung, and Blood Institute, 2 Rockledge Center,
Bethesda, MD 20892-7938, U.S.A.

E-mail: follmand@gwgate.nhlbi.nih.gov

(Received March 1995; accepted March 1997)