Biostatistics (2013), **0**, 0, *pp.* 1–22
doi:10.1093/biostatistics/manuscript

Adaptive Enrichment Designs for Clinical Trials

Noah Simon*

Statistics Department, Stanford University

Richard Simon

Biometric Research Branch, National Cancer Institute

nsimon@stanford.edu

Summary

Modern medicine has graduated from broad spectrum treatments to targeted therapeutics. New drugs recognize the recently discovered heterogeneity of many diseases previously considered to be fairly homogeneous. These treatments attack specific genetic pathways which are only dysregulated in some smaller subset of patients with the disease. Often this subset is only rudimentarily understood until well into large scale clinical trials. As such, standard practice has been to enroll a broad range of patients and run post-hoc subset analysis to determine those who may particularly benefit. This unnecessarily exposes many patients to hazardous side effects, and may vastly decrease the efficiency of the trial (especially if only a small subset benefit). In this manuscript we propose a class of adaptive enrichment designs which allow the eligibility criteria of a trial to be adaptively updated during the trial, restricting entry to patients likely to benefit from the new treatment. We show that our designs both preserve type 1 error, and in a variety of cases provide a substantial increase in power.

Key words: adaptive clinical trials; biomarker; enrichment; cut-point

1. Introduction

The literature on adaptive clinical trial design has focused on sample-size re-estimation, changing the plan for interim analyses, or modifying randomization weights (Chow and Chang (2007), Muller and Schafer (2001), Rosenberger and Lachin (1993), Karrison, Huo, and Chappell (2003), and Kim et al. (2011)). In oncology therapeutics development, attention has turned towards discovery of baseline predictive biomarkers to identify patients likely to benefit from the new treatment (Papadopoulos, Kinzler, and Vogelstein (2006), Schilsky (2007), and Sawyers (2008)). Tumors of most body sites have been found to be biologically heterogeneous with regard to their causal mutations and molecularly targeted drugs are unlikely to benefit most patients in the broad diagnostic categories traditionally included in clinical trials. When the pathophysiology of the disease and the mechanism of action of the drug are well understood, a binary predictive biomarker can be identified prior to or early in clinical development and used to restrict entry of patients to the pivotal phase 3 clinical trials comparing the new drug to a suitable control. Such “enrichment” designs can serve to magnify the treatment effect and thereby improve the efficiency of the clinical trial (Simon and Maitournam (2005), Maitournam and Simon (2005), and Mandrekar and Sargent (2009)).

Because of the complexity of cancer biology, it is frequently impossible to identify a single candidate predictive biomarker and a known threshold by the time the phase 3 trials are initiated (Sikorski and Yao (2009) and Sher et al. (2011)) Often, several candidate biomarkers are available and phase 2 information is not adequate to reliably select among them. Rather than making arbitrary decisions based on inadequate phase 2 data, we will describe a phase 3 design which begins without restricting entry based on any of the candidate biomarkers, and sequentially restricts entry in an adaptive manner. This gives much of the efficiency of the “enrichment” approach without the need to choose a subset beforehand.

There has been relatively little previous methodological work on adaptively changing the eligibility criteria during a clinical trail. The phase II Bayesian adaptive methods of Kim et al. (2011) involved a randomized comparison of several treatments within each of several biomarker strata. Although patient eligibility for the trial was not modified, in some cases a treatment arm would be discontinued from use within a stratum. Wang, O'Neill, and Hung (2007) considered a design which compared treatment to control with a single binary biomarker, allowing termination of the biomarker negative cohort at an interim analysis. Liu et al. (2009) and Follman (1997) describe designs for a single binary marker and a single interim analysis. Rosenblum and Van Der Laan (2011) permit several disjoint strata with a single interim analysis but assume that there are no data dependent period effects. We will consider the problem in greater generality.

In practice, changes to eligibility criteria are not unusual. Eligibility is sometimes narrowed as a result of toxicity experience or broadened to increase accrual rate. The eligibility criteria for a phase 3 clinical trial is often thought of as defining the target population for future use of the new treatment. This viewpoint is, however, problematic. The eligibility criteria, even without changes, may not adequately reflect the group of patients who actually participated in the trial. Also, many clinical trials establish a small average treatment effect for the eligible patients as a whole. Even an improvement in 5-year disease free survival from 70 percent to 80 percent for surgery with chemotherapy compared to surgery alone means that 70 percent of the patients did not need the new treatment and of the 30 percent of patients who did need some additional treatment, two-thirds did not benefit from the chemotherapy. Given the considerable expense and potentially serious adverse effects of many new treatments, using the eligibility criteria as a basis for indicating who should receive therapeutics is increasingly unsatisfactory.

In the next section we will present a general framework for adaptive enrichment. We will introduce two methods of analysis for binary response clinical trials which are guaranteed to preserve the type I error. In the section following that, we describe a simulation study we performed to evaluate adaptive enrichment of the threshold of positivity for a single biomarker/classifier and compare it to a standard design without adaptive enrichment. We then present methods of analysis that are available when adaption takes place in a group sequential manner. We discuss application of the methods to other endpoints and discuss generalization of the results to future patients.

2. Preserving type I error with adaptive enrichment for binary outcome

We first consider binary outcome. Assume we have a single new treatment we are comparing to control (or standard of care). We randomize each patient that we accrue with equal probability to one of the two arms. Let y_i be the treatment assignment for patient i: $y_i = 1$ for the new treament and $y_i = 0$ for control. Let x_i denote a vector of covariates measured on patient i. Finally, let z_i be the outcome for patient i where $z_i = 1$ for response and $z_i = 0$ for non-response.

As we accrue more patients we would like to restrict enrollment to those patients who will benefit from the treatment. Let $f(x)$ be the map from our covariate space to $\{0, 1\}$ which indicates whether a patient with covariate vector x will perform better on treatment or control:

$$f(x) = I\{p_T(x) > p_C(x)\}$$

where $p_T(x)$ and $p_c(x)$ are the probabilities of response for a patient with covariate vector x under treatment and control. For each m, let $\hat{f}_m(x)$ be our estimate of $f(x)$, computed after accrual of m patients. The data available for developing $\hat{f}_m(x)$ are $x_1, \ldots, x_{m-1}$, $y_1, \ldots, y_{m-1}$ and $z_1, \ldots, z_{m-1}$.

Consider the following procedure:

1. Randomize the first m_0 patients without exclusions to treatment or control to get a baseline estimate of $\hat{f}_{m_0}$. Now for each $m > m_0$:

2. Find $\hat{f}_m$ based on previous patients (covariates, treatment status, outcome).

3. Restrict entry into the clinical trial to only patients with $\hat{f}_m(x) = 1$.

4. Repeat until a total of n patients have been enrolled.

The enrichment classifier can be re-computed after each new outcome is obtained or in a group sequential manner. It can be based on modeling the unknown $p_T(x)$ and $p_C(x)$ functions or using other classification strategies. Our focus will not be on determining how best to estimate $f(x)$ but on demonstrating how to preserve the type I error when using adaptive enrichment. The null hypothesis for this setup is that no sub-populations benefits more from treatment than control; i.e.

$$p_T(x) = p_C(x) \quad \text{for all } x$$

Because the prognosis of patients included in the clinical trial may change sequentially due to our changing enrollment criteria, and because our change in enrollment criteria is outcome dependent, standard methods of analysis are not guaranteed to control the type 1 error. For example, in Section 6 we show in a simulation how a standard permutation test can give type 1 error in excess of 15%.

Here we propose using the test statistic

$$S = \sum_{i=1}^{n} [y_i z_i + (1 - y_i)(1 - z_i)]. \tag{2.1}$$

S is just the number of successes on the new treatment plus the number of failures on the control. It is straightforward to see that under the null, regardless of the values of $p_T(x) = p_C(x)$ and of how enrollment criteria change, we have

$$y_i z_i + (1 - y_i)(1 - z_i) \sim \text{Ber}(0.5).$$

Thus, under the null

$$S \sim \text{binomial}\,(n, 0.5)$$

Comparing S to the tails of this binomial is a valid test that protects the type 1 error regardless of the method used for adaptively modifying enrollment criteria.

If patients are accepted and randomized in pairs, one to each treatment arm, and enrollment criteria $\hat{f}$ updated no more frequently than after each pair, then the test statistic we proposed above has a familiar form. If we let $z_{i,C}$ and $z_{i,T}$ be the outcome for the control observation and treatment observation respectively from pair i, our statistic (2.1) after n pairs is equivalent to

$$\tilde{S} = \sum_{i=1}^{n} \left\{ I\left\{ z_{i,T} > z_{i,C} \right\} - I\left\{ z_{i,T} < z_{i,C} \right\} \right\} \tag{2.2}$$

This is the number of untied pairs favoring treatment minus the number of untied pairs favoring control. Under the null hypothesis, each untied pair is equally likely to favor treatment or control. If we continue to enroll patients until we have a prespecified number, u, of untied pairs, then under the null

$$\frac{\tilde{S} + u}{2} \sim \text{binomial}\left(u, 0.5\right).$$

The hypothesis test based on this statistic is exactly McNemar's test.

Several extensions to the above formulations are possible, some of which will be pursued later in this paper. For example, the paired approach is easily generalizable to non-binary endpoints using the same test statistic $\tilde{S}$. Our assumption that a single statistical significance test will be performed after a pre-specified n patients are randomized is also inessential. One can pre-specify K interim analysis points after $n_k, k = 1, \ldots, K$ patients or untied pairs have been treated on each arm and an interim analysis plan that allocates the type I error among the interim analyses (Pocock (1982), Lan and DeMets (1983), and Jennison and Turnbull (1999)).

3. Application - Adaptive Threshold Enrichment Design

One important application of adaptive enrichment is to the frequently occurring setting where a single candidate predictive biomarker is available but no cut-point has been determined (Jiang, Freidlin, and Simon, 2007). Drug developers would often like to use an "enrichment design" in which test negative patients are excluded, but early phase clinical trials are frequently too limited in size to reliably determine an acceptable cut-point. Regulators would also often prefer that the clinical trial not restrict entry initially based on the biomarker so that the value of the test can be more adequately evaluated. In such settings there are often a discrete set of candidate cut-points which we will denote by $\xi_1, \ldots, \xi_K$. These may represent possible values of semi-numerical assays or quantiles of numerical assays (e.g. 0^{th}, 25^{th}, 50^{th}, 75^{th} quantiles).

There are several reasonable ways of modeling the $f(x)$ function for this single biomarker setting. We will describe one approach here for a simple alternative — that the treatment effect $p_T(x) - p_C(x)$ for a patient with biomarker value x is either 0 or δ and that the treatment effect is monotone non-decreasing in x with a jump only at one of the candidate cut-points. At an interim analysis during the study, let $l(\xi_k)$ denote the log likelihood of the data maximized with regard to the unknown constants $p_0 \leqslant p_1$ subject to the constraints $p_C(x) = p_0$ for all x, $p_T(x) = p_0$ for $x \leqslant \xi_k$ and $p_T(x) = p_1$ for $x > \xi_k$. We take the candidate cut-point ξ_k at which the log likelihood is maximized as an estimate of the true cutpoint, x^* and restrict subsequent accrual to patients whose biomarker is greater than that value.

To illustrate our approach, We ran a simulation of the adaptive enrichment design under the single biomarker model above. The biomarker was uniformly distributed on (0,1) with K equally spaced potential cutpoints (at $1/(K+1), \cdots, K/(K+1)$). We used a single interim analysis at which change of enrollment criteria was considered. Before the interim analysis, n_1 simulated patients were randomly allocated to treatment T or control C with equal probability. Outcome was binary, 0 (non-response) or 1 (response) with response probability of p_0 for both the control group and patients in the treatment group with biomarker value below the true cut-point x^* and p_1 for patients on treatment with biomarker value above x^*.

At the interim analysis we found the candidate cut-point $\hat{x}^*$ which maximized the log-likelihood with the restriction that $p_0 \leqslant p_1$. If this log-likelihood did not exceed the null log likelihood (i.e. cut-point 1.0) by at least 0.25, accrual was teminated. Otherwise, accrual was restricted to patients with biomarker values greater than $\hat{x}^*$ for the remainder of the trial. The number of total patients N was determined in advance and $N - n_1$ patients were accrued after the interim analysis. The trial was analyzed using the test statistic (2.1) and a one-tailed 5% rejection region.

Column 5 of Table 1 shows the statistical power for the adaptive enrichment design as assessed by computer simulation for 10,000 replications of clinical trials with a total of 200 patients and 100 at the time of interim analysis. We vary p_0, p_1, x^*, and K (the number of candidate cut-points). Note that the marker values were $U(0, 1)$, so $x^* = 0.25$ indicates that 75 percent of patients are more likely to benefit from treatment. We used our adaptive procedure with statistic (2.1) and a single interim analysis. As shown in rows 1 and 2, however, the actual size of the test is somewhat less than the nominal 5 percent. Column 6 shows the simulated power of a contingency chi-square test with continuity correction for trials based on 200 patients but without adaptive modification of the eligibility criteria.

The adaptive enrichment procedure has much greater power than the standard clinical trial for most conditions addressed in Table 1. For example, in the simulations shown in the fourth row of the table the power for the adaptive enrichment approach with one interim analysis was 89.3% as compared to a power of 72.2% for a standard clinical trial without adaptive enrichment.

Table 2 shows the results for simulated clinical trials when the response probabilities for the control group and the new treatment group are different for patients entered prior to and following the interim analysis Under the null hypothesis that there is no treatment effect before or after the interim analysis, the type I error is preserved using the test statistic (2.1) . The large advantage of the adaptive enrichment procedure over the standard clinical trial is about the same in Table 2 as it was in Table 1.

While our simulations show a significant increase in power with adaptive enrichment, the more you restrict the eligibility criteria, the longer patient accrual will take. Adaptive enrichment is most powerful relative to the standard non-adaptive approach when only a small subset of patients benefit, however this is exactly when the accrual rate is most decreased. Column 7 of Table 1 shows the mean accrual time for the simulated adaptive trials assuming a total accrual rate of 100 unselected patients per year. The final column shows the accrual time for a standard non-adaptive clinical trial that has the same power as the adaptive design (column 5 of the corresponding row). The added sample size required for the non-adaptive design to achieve equivalent power in many cases negates the potential advantage of the standard design with regard to duration of accrual, but the standard design retains an advantage for some cases. The total sample size required for the non-adaptive design can be computed by multiplying the final column by 100.

4. Group Sequential Analysis

In large multi-center clinical trials continual re-analysis of the data is generally not practical even if it were desirable and the group sequential approach to interim analysis has been very popular (Pocock (1982), Lan and DeMets (1983), and Jennison and Turnbull (1999)). The group sequential approach was utilized in the previous section where there was a single interim analysis time at which the eligibility criteria could be modified as a function of the interim data. We showed in an earlier section that for any adaptive enrichment strategy, using the number of total responses on the new treatment plus the number of non-responses on the control as test statistic preserved the type I error. When the adaptiveness is performed in a group sequential manner, there are other analysis strategies that preserve type I error.

4.1 *General Statistics*

We will begin with a short discussion of a general class of statistics (and tests) which preserve the type 1 error. We start with some notation. For each block, k, let t_k be some statistic based on the data in that block. We will combine all of these statistics with some function $G\left(t_1, \ldots, t_K\right)$.

Let $\mathcal{L}_k$ denote all the data, outcomes, covariate vectors and treatment assignments, for blocks $1, \ldots, k$.

If we are careful to select our statistics, t_k, so that under the null the distribution of each t_k is known and independent of $\mathcal{L}_{k-1}$, then we may choose any G and construct a valid test which preserves the type 1 error. This test is straightforward to construct. Because t_k uses only observations from block k and its null distribution is independent of $\mathcal{L}_{k-1}$, it is independent of all previous t_i. Thus, under the null, we have $t_1, \ldots, t_k$ independent with known distributions. This in turn will induce a known null distribution for G.

One must define the t_k carefully to achieve independence of $\mathcal{L}_{k-1}$. For example, suppose outcomes are binary with equal numbers $n_k/2$ of subjects on each treatment in block k. Let r_{Tk} denote the number of responses on treatment T in block k and let r_{Ck} denote the number of responses on control. One might naively want to use

$$t_k = r_{Tk} - r_{Ck} \tag{4.3}$$

However, while under the null this will have mean 0, the variance will depend on the overall prognosis of the patients in the k^{th} block (which may depend on $\mathcal{L}_{k-1}$).

There are, however, some t_k which do satisfy this requirement. For continuous response data, the Mann-Whitney-Wilcoxon u statistic (within-block) is independent of $\mathcal{L}_{k-1}$. Also, any valid p-value based on continuous outcomes in the k^{th} block is distributed uniformly on 0 to 1 independently of $\mathcal{L}_{k-1}$. In the next sections we discuss specific choices which we recommend for t_k and G in several scenarios.

4.2 *Continuous Data*

For continuous data with only a single block, it is standard practice to use either a t-test or a Mann-Whitney-Wilcoxon test for comparing treatments. There are simple analogs to these for the adaptive enrichment design.

In a standard non-adaptive design one may use the t-statistic

$$t = \frac{\bar{y}_T - \bar{y}_C}{\sqrt{\hat{\sigma}_T^2/n_T + \hat{\sigma}_C^2/n_C}}. \tag{4.4}$$

where $\bar{y}_T$ and $\bar{y}_C$ denote the overall average outcomes on the new treatment and control and the denominator is the standard error of the difference between these two averages. For our adaptive design we instead propose

$$\frac{1}{\sqrt{n}} \sum_{k \leqslant K} \sqrt{n_k} \left(\frac{\bar{y}_{(T,k)} - \bar{y}_{(C,k)}}{\sqrt{\hat{\sigma}^2_{(T,k)}/n_{T,k} + \hat{\sigma}^2_{(C,k)}/n_{C,k}}} \right) \tag{4.5}$$

where $\bar{y}_{(T,k)}$, $\bar{y}_{(C,k)}$, $\hat{\sigma}^2_{(T,k)}$, $\hat{\sigma}^2_{(C,k)}$, $n_{T,k}$ and $n_{C,k}$ denote the treatment and control sample means, variances and sample sizes in the k^{th} block. n_k denotes the total sample size in the k^{th} block and the n is the total overall sample size.

The difference between the standard t statistic (4.4) and our statistic (4.5) is that we standardize by the variance in each block, rather than by the "pooled" variance. One might note that even if we assume a common variance for all of the blocks (which we definitely do not), (4.4) is still a poor choice. The estimate $\hat{\sigma}_T^2$ in (4.4) is

$$\begin{aligned} \hat{\sigma}_T^2 &= \frac{1}{n-1} \sum_{i \leqslant n} \left(y_{(T,i)} - \bar{y}_T \right)^2 \\ &= \frac{1}{n-1} \sum_{k \leqslant K} \sum_{i \leqslant n_k} \left(y_{(T,k(i))} - \bar{y}_{(T,k)} \right)^2 + \frac{n}{n-1} \sum_{k \leqslant K} \left(\bar{y}_{(T,k)} - \bar{y}_T \right)^2. \end{aligned}$$

Even in the common variance case this is an overestimate by roughly $\sum_{k \leqslant K} \left(\bar{y}_{(T,k)} - \bar{y}_T \right)^2$. If the overall prognosis varies among blocks, this may be a very large quantity.

One may also think of our statistic as the weighted sum of t statistics. For a given block, k, with n_k sufficiently large, we have

$$\sqrt{n_k} \left(\frac{\bar{y}_{(T,k)} - \bar{y}_{(C,k)}}{\sqrt{\hat{\sigma}^2_{(T,k)}/n_{T,k} + \hat{\sigma}^2_{(C,k)}/n_{C,k}}} \right) \dot{\sim} t_{n_k} \sqrt{n_k}$$

under the null (under very general regularity conditions) regardless of $\sigma^2_{(T,k)}$, $\sigma^2_{(C,k)}$, and $\mu_{(T,k)} = \mu_{(C,k)}$. This is key as the value of these parameters (even under the null) may depend on observed outcomes and assignments of patients in previous blocks. The null distribution of our test statistic is thus a linear combination of t statistics with weights fixed by the number of patients per block. For each n_k sufficiently large, under the null, our statistic is distributed

$$\left(\frac{1}{\sqrt{n}}\sum_{k\leqslant K} t_{n_k}\sqrt{n_k}\right) \dot{\sim} N(0,1)$$

We would reject our null hypothesis for particularly large or small values of our statistic.

One should also note that under a full-population alternative (ie. all sub-populations have the same distribution under the null, and identical change under the alternative) under suitable weak conditions for a fixed K (with $n_k \to \infty$ for all k) this statistic is asymptotically as efficient as the standard t-statistic against a mean shift. For the case of balanced treatment assignments where $n_{T,k} = n_{C,k} = n_k/2$, we can write eq (4.5) as

$$\frac{1}{\sqrt{n}}\sum_{k\leqslant K}\left(n_k/\sqrt{2}\right)\left(\frac{\bar{y}_{(T,k)} - \bar{y}_{(C,k)}}{\sqrt{\hat{\sigma}^2_{(T,k)} + \hat{\sigma}^2_{(C,k)}}}\right)$$

We also know that $\hat{\sigma}^2_{(T,k)} \to \sigma^2_T$, the common treatment variance, in probability and similarly $\hat{\sigma}^2_{(C,k)} \to \sigma^2_C$ (even under local alternatives). Thus by applying Slutsky's theorem (with local alternatives) we see that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{k\leqslant K}\left(n_k/\sqrt{2}\right)\left(\frac{\bar{y}_{(T,k)} - \bar{y}_{(C,k)}}{\sqrt{\hat{\sigma}^2_{(T,k)} + \hat{\sigma}^2_{(C,k)}}}\right)\\
&\quad \sim \frac{1}{\sqrt{n}}\sum_{k\leqslant K}\left(n_k/\sqrt{2}\right)\left(\frac{\bar{y}_{(T,k)} - \bar{y}_{(C,k)}}{\sqrt{\sigma^2_T + \sigma^2_C}}\right)\\
&\quad = \sqrt{n/2}\left(\frac{\bar{y}_T - \bar{y}_C}{\sqrt{\sigma^2_T + \sigma^2_C}}\right)
\end{aligned}$$

This last line is exactly what we get from applying Slutsky's Theorem to our usual t statistic — thus, the two statistics have the same limiting distribution for full-population [local] alternatives (and under the null). From this we see that our adaptive t-test is asymptotically efficient.

Although we have omitted some of the details for our application of Slutsky's theorem (to be fully precise one needs to discuss the limiting distribution of the numerator under local alternatives), these details are straightforward.

If we do not want to resort to asymptotic normality, as an alternative one could use a block Mann-Whitney-Wilcoxon test. If we assume the strong null hypothesis that treatment and control observations have the same distribution within each block and are absolutely continuous with respect to Lebesgue measure, then the ranks within a block are uniformly distributed (ties are a probability 0 event), independent of the exact distribution of the observations. Thus, one might use as a statistic

$$u = \sum_{k \leqslant K} w_k u_k$$

where u_k is the Mann-Whitney statistic for only block k, and w_k is a predefined weight. As we said, under the null, any ranking of the variables within a block is equally probable. This induces a null distribution for u, and can be used to construct a test which strictly controls type 1 error.

4.3 *Binary Data*

For binary data we would like to compare sample proportions between treatment and control. Again, we assume a balanced design with n total patients, and n_k in each block (though this can be generalized to unbalanced designs). In a non-adaptive design one often uses the statistic

$$z = \frac{\hat{p}_T - \hat{p}_C}{2\sqrt{\hat{p}_{pool}\left(1 - \hat{p}_{pool}\right)/n}} \tag{4.6}$$

where $\hat{p}_T$ and $\hat{p}_C$ are sample success proportions in treatment and control respectively, and $\hat{p}_{pool} = \left(\hat{p}_T + \hat{p}_C\right)/2$. For our adaptive design we propose

$$z = \frac{1}{\sqrt{n/2}} \sum_{k \leqslant K} \sqrt{n_k/2} \left(\frac{\hat{p}_{(T,k)} - \hat{p}_{(C,k)}}{2\sqrt{\hat{p}_{(pool,k)}\left(1 - \hat{p}_{(pool,k)}\right)/n_k}} \right) \tag{4.7}$$

This is the binary analog to (4.5) and is asymptotically $N(0,1)$ (though it can be better approximated in small samples as a linear combination of t-distributions). These asymptotics are independent of the actual value of $p_{(T,k)} = p_{(C,k)}$ (so long as they are non-degenerate).

As before, one can use Slutsky's theorem to show that this statistic asymptotically loses no efficiency versus the non-adaptive z-statistic for full-population alternatives.

4.4 *Survival Data*

Time-to-event data can also be handled fairly simply but one must take care to account for the fact that the probability of censoring in later blocks may be a function of earlier outcomes and assignments.

Let $\ell_k\left(\beta\right)$ denote the log-likelihood of the Cox model for the k^{th} strata (with β the coefficient for the treatment indicator), with first and second derivatives $\ell_k^{'}$ and $\ell_k^{''}$. Now we may use as a statistic

$$T = \sum_k w_k \frac{\ell_k^{'}\left(0\right)}{\sqrt{-\ell_k^{''}\left(0\right)}}$$

where the w_k are pre-specified non-negative weights. This is just the sum of the weighted, signed, normalized scores of each block. Since each of these scores is asymptotically $N(0,1)$, we have a valid N(0,W) test where W is the sum of squares of the weights.

Updating eligibility criteria is less effective for survival data because censoring reduces the information available at interim analysis points. The enrichment classifier can be based on an observed intermediate endpoint while the final analysis remains based on the survival endpoint.

While the statistics proposed up to this point seem very straightforward, as discussed earlier we have been careful to choose only statistics whose distributions are invariant under the null. In the next section we will discuss permutation methods and illustrate why our previous approach was key for protecting type 1 error.

5. Failure of the Permutation Test

In general, permutation tests provide a flexible, robust way to test hypotheses with few parametric assumptions. One might consider permuting class labels within each block to find a conditional null distribution of any statistic of interest. While this seems like a reasonable approach, unfortunately, in this case it does not strictly protect type 1 error.

The permutation test is derived by conditioning on the outcomes and considering the induced distribution of treatment assignments under the null. In examples where observations are independent, the induced distribution on the assignments under the null is the permutation distribution — every set of assignments with n_C patients randomized to control and n_T patients randomized to treatment is equally likely. This in turn induces a null distribution on our test statistic, and we can compare the original value of the statistic to the tails of this "permutation null".

In our case, however, even under the null, the outcomes in the later blocks are dependent on the treatment assignments and outcomes of the earlier patients — eg. some combinations in block 1 may make us choose a better prognosis sub-population for block 2, while others may not. So simple rerandomization tests (even within block) do not preserve type 1 error control. This is particularly pronounced when interim differences in outcome between the treatment groups lead to major changes in the prognosis of subsequent patients.

To illustrate we ran several simulations. We assumed binary outcomes with an initial probability of response p_0 for both treatment groups. Patients were accrued in a group sequential manner with a balanced n patients per block for each treatment. At the end of each block of accrual the difference in cumulative number of responses on each treatment divided by the standard error of the difference was computed. In computing the standard error, the underlying true response rate for the block was used. If the absolute value of this standardized difference was greater than a pre-specified critical value $z*$, then the common response probability for patients accrued in the next block changed to p_{00}; otherwise it remained as p_0. The statistic used for testing the null hypothesis was the total number of successes on the new treatment. The null hypothesis was rejected if the test statistic was greater than the 97.5^{th} percentile or less than the 2.5^{th} percentile of the permutation distribution. Treatment labels were permuted within block. For each clinical trial simulated, 1000 permutations were performed. For each set of parameters considered, 5000 clinical trials were simulated.

We simulated clinical trials with 5 blocks and 20 patients per block for each treatment, with $p_0 = 0.5$, $p_{00} = 0.01$ and $z^* = 1.5$. The two-sided type I error of the permutation test under these conditions, based on 5000 simulations was estimated as 17.32% instead of the nominal 5%. If the patients were accrued in 10 blocks of 10 patients per treatment instead of 5 blocks of 20 patients, the estimated type I error of the permutation test increased to 21.46%. With less extreme changes in the prognostic makeup of the patients, the degree of anti-conservatism of the permutation test was reduced. The simulation demonstrated, however, that with interim outcome dependent changes in eligibility, the permutation test is not guaranteed to preserve type I error.

6. Identifying the Target Population

The adaptive enrichment approach can provide substantial improvements in power for detecting whether a new treatment is effective for some subset of the patients initially eligible for the clinical trial. However at the end of the trial there is a question of which subset actually benefits? This is a difficult question, and providing recommendations for future use of the new treatment may depend on additional analyses.

It is important to note, however, that this is just as big a problem in a standard clinical trials where the analysis is based on the initially eligible population supplemented by post-hoc subset analysis. The problem is more explicit with adaptive enrichment and is more tractable because the algorithm for calculating the enrichment classifier is pre-specified. In standard trials "global efficacy" is frequently driven by a small subset of patients who benefit, and yet the medication becomes broadly approved with many or most of the patients achieving no benefit.

In order to minimize uncertainty in the intended population, we recommend use of the group sequential approach with a small number (1-2) of interim eligibility changes. The function $\hat{f}_m(x)$ used for the final stage of accrual might be taken as the indication for future use. Table 3 shows that for the adaptive threshold enrichment designs described earlier this approach is quite effective when the number of candidate cut-points is limited. We believe that the rejection of the global null hypothesis and the development of an "indication classifier" for providing guidance for future use of the new treatment should be seen as two different aspects of the analysis of phase III clinical trials. The multiple testing framework is not necessarily the most appropriate one for developing a classifier to guide future use of a new regimen to maximize net benefit for a population of patients (Zhang et al., 2012).

As for methods for estimating $\hat{f}_m(x)$, we have given some suggestions in the threshold case. For stratified populations without covariates, classification can be based on the estimates of treatment effect for each stratum. Development of enrichment classifiers with low or high dimensional covariates is an important topic for further research. The classifiers should be evaluated with regard to their effect on the operating characteristics of the clinical trial, their accuracy of classification and their net effect on outcomes for future patients.

7. Discussion

We have introduced an adaptive enrichment strategy for randomized clinical trials that enables eligibility criteria to adapt to exclude patients who appear unlikely to benefit from the new treatment. Such designs can both increase the efficiency of the clinical trial and protect patients from exposure to treatments with serious toxicities to which they may have little likelihood of benefit. It is well known that the statistical power of a clinical trial is critically dependent on the size of the treatment effect in the eligible population. The sample size or number of events required often varies as the reciprocal of the square of the treatment effect. That relationship is responsible for the potential efficiency of the enrichment design. The fixed eligibility enrichment design has limited applicability — it is difficult to have available at the start of a phase III trial a single candidate predictive classifier and a well documented appropriate cut-point. Often, phase I and II trials may provide one or more candidate predictive biomarkers but without adequate data to confidently define cut-points of positivity. The framework we have developed here enables the refinement of this information during the course of the phase III trial. When benefit of a drug is restricted to a small, but initially undetermined, sub-population we have shown that our adaptive enrichment design can preserve studywise type I error, provide substantial improvements in statistical power, and suffer little statistical power loss against global alternatives.

We have described a broad class of significance tests that will preserve type I error for group sequential adaptive enrichment designs with binary, continuous, and time-to-event outcomes, and given examples of common, intuitive tests which are not level preserving. There are many significance tests that do preserve type I error under adaptive enrichment and future research should evaluate them from the perspective of statistical power. The generality of the formulation under which we have demonstrated preservation of studywise type I error also suggests important future research on the types of enrichment classifiers to use for interim and final analyses.

References

Chow, S. and Chang, M. (2007). Adaptive design methods in clinical trials. *Chapman & Hall.*

Follman, D.(1997). Adaptively Changing Subgroup Proportions in Clinical Trials. *Statistic Sinica*, **7**,1085–1102.

Jennison, C., and Turnbull, B.(1999). Group Sequential Methods With Applications to Clinical Trials. *Chapman and Hall.*

Jiang, W., Freidlin, B. and Simon, R. (2007). Biomarker adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute*, **99**,1036–1043.

Karrison, T., Huo, D., and Chappell, R. (2003). A group sequential, response-adaptive design for randomized clinical trials. *Controlled Clinical Trials*, **24**,506–22.

Kim, E., Herbst, R., Wistuba, I., Lee, J., Blumenschein, G., Tsao, A., Stewart, D. and Hicks, M. (2011). The battle trial: Personalizing therapy for lung cancer. *Cancer Discovery*, **1**, 44–53, 2011.

Lan, K., and DeMets, D. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, **70**, 659–63, 1983.

Liu, A., Li, Q., Yu, K. and Yuan, V. (2009). A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations. *Statistics in Medicine*, (In Press).

Maitournam, A., and Simon, R.(2005). On the efficiency of targeted clinical trials. *Statistics in Medicine*, **24**, 329–339, 2005.

Mandrekar, S., and Sargent, D.(2009). Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. *Journal of Clinical Oncology*, **27(24)**, 4027–34.

Muller, H., and Schafer, H.(2001). Adaptive group sequential designs for clinical trials. combining the advantages of adaptive and classical group sequential approaches. *Biometrics*, **57**, 886–891.

Papadopoulos, N., Kinzler, K. and Vogelstein, B.(2006). The role of companion diagnostics in the development and use of mutation-targeted cancer therapies. *Nature Biotechnology*, **24(8)**, 985–995.

Pocock, S.(1982). Interim analyses for randomized clinical trials. *Biometrics*, **39**, 153.

Rosenberger, W. and Lachin, J.(1993). The use of response-adaptive designs in clinical trials. *Controlled Clinical Trials*, **14(6)**, 471–84.

Rosenblum, M. and Van Der Laan, MJ.(2011). Optimizing Randomized Trial Designs to Distinguish Which Subpopulations Benefit from Treatment. *Biometrika*, **98**,845–860.

Sawyers, C.(2008). The cancer biomarker problem. *Nature*, **452**, 548–552.

Schilsky, R.(2007). Target practice: oncology drug development in the era of genomic medicine. *Clinical Trials*, **4**, 163–166.

Sher, H., Nasso, S., Rubin, E. and Simon, R.(2011). Adaptive clinical trials designs for simultaneous testing of matched diagnostics and therapeutics. *Clinical Cancer Research*, **17**, 6634–6640.

Sikorski, R. and Yao, R.(2009). Parallel paths to predictive biomarkers in oncology: Uncoupling of emergent biomarker development and phase iii trial execution. *Science Translational Medicine*, **1**, 10-11.

Simon, R. and Maitournam, A.(2005). Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*, **10**, 6759–6763.

Wang, S. J., O'Neill, R. and Hung, H.(2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics*, **6**, 227–244.

Zhang, B., Tsiatis, AA., Laber EB. and Davidian, M.(2012) A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics*, **68**, 1010–1018.

8. tables

p_0	p_1	K	x^*	Power Adapt	Power nonAdapt	Adapt Accrual	Equiv Accrual
0.2	0.2	5	0.5	0.034	0.033	2.42	
0.5	0.5	5	0.5	0.035	0.038	2.49	
0.2	0.5	1	0.5	0.898	0.717	2.48	3.25
0.2	0.5	3	0.5	0.893	0.722	3.07	3.25
0.2	0.5	5	0.5	0.897	0.726	3.19	3.25
0.2	0.5	9	0.5	0.892	0.724	3.25	3.25
0.2	0.5	3	0.25	0.971	0.952	2.47	2.25
0.2	0.5	5	0.25	0.968	0.955	2.55	2.25
0.2	0.5	5	0.75	0.768	0.281	3.97	4.75
0.2	0.45	5	0.5	0.761	0.579	3.23	3.0
0.2	0.45	3	0.5	0.761	0.582	3.05	3.0
0.2	0.45	3	0	0.959	0.979	2.22	1.7
0.4	0.7	5	0.5	0.896	0.637	3.12	4.0
0.1	0.3	5	0.5	0.581	0.568	3.22	2.1
0.1	0.25	5	0.5	0.376	0.385	3.22	2.0

Table 1. Power and duration for adaptive vs nonadaptive methods in a variety of scenarios. "Power Adapt" and "Power nonadapt" are the simulated power estimates for the adaptive and non-adaptive procedures. Power was calculated by simulation (with $10,000$ replications). p_0 is the response probability for all patients on control, and for patients on treatment with biomarker value $x < x^*$ (where x is marginally $U(0,1)$). p_1 is the response probability for patients on treatment with $x \geqslant x^*$. K is the number of candidate cutpoints. Adapt Accrual, is the average adaptive trial duration measured in years based on an accrual rate of 100 patients per year. Equiv Accrual is the accrual time for a non-adaptive design based on increasing the sample size to match power for the adaptive design.

$p_{(C,\text{before})}$	$p_{(T,\text{before})}$	$p_{(C,\text{after})}$	$p_{(T,\text{after})}$	Power Adapt	Power nonadapt
0.2	0.2	0.5	0.5	0.035	0.037
0.5	0.5	0.2	0.2	0.035	0.033
0.2	0.5	0.5	0.8	0.897	0.646
0.2	0.45	0.5	0.75	0.757	0.502
0.1	0.3	0.5	0.7	0.590	0.347

Table 2. Power and type one error for adaptive and nonadaptive tests when the population changes after interim analysis. $p_{(C,\text{before})}$ and $p_{(C,\text{after})}$ are the simulated response probabilities before and after interim analyses for patients on control. $p_{(T,\text{before})}$ and $p_{(T,\text{after})}$ are the response probabilities before and after interim analyses for patients on treatment.

# of Candidate Cutpoints	x^*	Distribution of Selected Cutpoints			
		0	**0.33**	**0.5**	**0.67**
1	0	0.93		0.07	
1	0.5	0.08		0.92	
2	0	0.87	0.10		0.03
2	0.33	0.12	0.79		0.09
2	0.67	0.05	0.09		0.86

Table 3. Simulated estimate of how often each cutpoint is chosen with a single interim check. Preselection block has 100 patients. Biomarker is uniform $(0, 1)$. Control patients and treated patients with biomarker below x^* have response probability of 0.2. Treated patients with biomarker above x^* have response probability of 0.5.