Published in final edited form as:

Stat Med. 2017 November 20; 36(26): 4083–4093. doi:10.1002/sim.7422.

Inference for Multi Marker Adaptive Enrichment Trials

Richard Simon¹, Noah Simon²

¹Biometric Research Program, National Cancer Institute

²Department of Biostatistics, University of Washington

Abstract

Identification of treatment selection biomarkers has become very important in cancer drug development. Adaptive enrichment designs have been developed for situations where a unique treatment selection biomarker is not apparent based on the mechanism of action of the drug. With such designs, the eligibility rules may be adaptively modified at interim analysis times in order to exclude patients who are unlikely to benefit from the test treatment. The adaptive enrichment approach of Simon and Simon [1] is particularly flexible, permitting development of model based multi-feature predictive classifiers as well as optimized cut-points for continuous biomarkers. A single significance test, including all randomized patients is performed at the end of the trial of the strong null hypothesis that the expected outcome on the test treatment is no better than control for any of the subset populations of patients accrued in the K stages of the clinical trial.

In this paper we address two issues involving inference following a Simon and Simon [1] type of adaptive enrichment design. The first is specification of the intended use population and estimation of treatment effect for that population following rejection of the strong null hypothesis. The second issue is defining conditions in which rejection of the strong null hypothesis implies rejection of the null hypothesis for the intended use population.

Keywords

adaptive clinical trials; biomarker; enrichment; resampling

1. Introduction

Over the past 15 years it has become established that human cancers of the same primary site are generally heterogeneous with regard to the genomic alterations which cause them and determine, in large part, their sensitivity to treatment. This has led to major changes in the strategies of cancer treatment development, with increasing focus on drugs which target specific molecular alterations and the development of companion diagnostics to assist in selection of patients for specific drugs. Biotechnology has provided an increasingly broad range of assays for characterizing the state of the tumor genome and for developing effective treatment selection biomarker classifiers.

In some cases the appropriate predictive biomarker can be easily determined based on the mechanism of action of the drug; e.g. the BRAF inhibitor vemurafenib is used for tumors with a mutation in the BRAF gene [2]. In many cases, however, the situation is more

complex and it has not been possible to develop an appropriate predictive biomarker based on biological principals and phase II data [3]. In some cases the appropriate biomarker measurement is known by that time but an appropriate cut-point for positivity has not been established. In other cases several relevant assays have been identified by the start of the phase III clinical trial but it has not been established which is best or how to use them in combination.

A variety of adaptive enrichment designs have been described for adaptively determining the appropriate predictive biomarker during the course of a phase III clinical trial. In most cases these designs determine whether the target population of patients should be the overall initially eligible population or a pre-specified subset [4, 5, 6, 7]. Often there is a single interim analysis at which time a decision is made whether to continue accrual for the full population or just for the pre-specified subset. Most other designs involve selection of one or more subsets from a pre-specified stratification of the patient population [8, 9, 10]. Simon and Simon [1] described a more general adaptive enrichment framework in which the eligibility restrictions are based upon an adaptively developed model of patient response from multiple candidate features which may involve continuous, ordinal, categorical or binary features and may involve the establishment of cut-points. The biomarker classifier can be modified at each interim analysis using a pre-specified algorithm.

Because their framework is general, the Simon and Simon [1] designs do not use the multiple comparison framework of most other methods. A single test of the strong null hypothesis of no benefit for the new treatment is conducted at the end of the trial in a manner which preserves the type I error in a model independent manner. In this paper we address two issues involved with inference following the adaptive enrichment designs of Simon and Simon [1]. The first is specification of the intended use population and estimation of treatment effect for that population. Conventional clinical trials define the intended use population based on the unchanging eligibility criteria, although this often results in small average treatment effects, large NNT (number of patients to treat for each patient who benefits) and substantial over-treatment of patients. Since eligibility rules may change during the adaptive enrichment trial, the definition of the intended use population requires modification. The second issue we deal with here is testing the null hypothesis of no treatment effect for the intended use population.

2. Estimating treatment effect for the intended use population

We consider a clinical trial of a test treatment and a control (or standard of care) regimen. Each patient accrued is randomized with equal probability to one of the two arms. Let z_i indicate the treatment assignment for patient i — $z_i = 1$ for the new treatment and $z_i = 0$ for control. For each patient i we also record a covariate vector x_i and an outcome y_i . Let P_x denote the joint distribution of the covariate vector in the initially eligible patient population.

As we accrue more patients we would like to restrict enrollment to exclude those kinds of patients who are unlikely to benefit from the test treatment. Assume that there are $K - 1$ interim analysis points at which we can change eligibility criteria. Let $f_k(x)$ denote the eligibility restriction function that will be used for the k th stage of accrual. That is; $f_k(x) = 1$

indicates that a patient with covariate vector x is eligible, and 0 otherwise. The initial eligibility function $f_1(x)$ is pre-specified. For all later periods k , the function $f_k(x)$ is obtained by applying a pre-specified algorithm to the treatment, covariate, and outcome data available at the start of the k 'th period of accrual. In stage k , we randomize n_k patients where the $\{n_k\}$ constants are pre-specified. In this work we assume that K is fixed in advance.

Because the distribution of measured and unmeasured prognostic factors may change during the trial and because changes in eligibility criteria are outcome dependent, standard methods of analysis are not guaranteed to control the type 1 error. Simon and Simon ([1]) showed how to test the strong null hypothesis

$$\bigcap_{k=1}^K \{\delta(S_k) \leq 0\} \quad (1)$$

where $S_k = \{x : f_k(x) = 1\}$ and $\delta(S_k)$ denotes the average treatment effect for the patients eligible for block k . That is; if $F_T(x)$ and $F_C(x)$ denote the expected responses for a patient with covariate vector x under treatment and control respectively, then $\delta(S_k) = \int_{x \in S_k} (F_T(x) - F_C(x)) dP(x)$.

The final analysis is based on a test statistic of the form $T = \sum_k w_k t_k$ where the summation is over the blocks of patients accrued between interim analysis times, w_k is a pre-specified weight for block k , and t_k is a treatment effect statistic for all patients accrued in block k . t_k must have a known null distribution which does not depend on any information from patients in earlier blocks. For example, if each t_k has a standard normal distribution under the strong null then the null distribution of T will be normal with mean zero and variance $\sum_k (w_k^2)$. In many cases the t_k will be only asymptotically normal. For example, the t_k may be standard two-sample t-statistics comparing average outcomes in the two treatment groups for patients accrued during the k 'th period. Simon and Simon [1] had shown examples where their approach is very effective in improving the power of the clinical trial.

As with most sequential decision procedures, a short term endpoint must be used for interim decision making. In adaptive phase II trials in lung cancer, 8 weeks free of tumor progression as been used as the endpoint. The interim eligibility decisions at the k 'th interim analysis are made using the patients accrued during periods $1, \dots, k$ whose endpoints are observed by the end of period k . The patients whose endpoints are not observed in time are, however, included in the final analysis with those of other period k accrued patients.

With the conventional clinical trial, the initial eligibility function never changes and defines the intended use population. Here, we shall consider two possible ways of defining the intended use population. The first takes the eligibility function used for the last stage of accrual to define the intended use population. We denote this intended use population by S_K where

$$S_K = \{x : f_K(x) = 1\}$$

Let $\hat{\delta}(S_K, K)$ denote the maximum likelihood estimate of the treatment effect for the eligibility rules used during the final block using data only from the final block. This quantity is not biased by selection because data for the final period is not used for defining the intended use population.

Because $\hat{\delta}(S_K, K)$ is based only on patients accrued in the last block, it may have large variance and we will consider an additional estimate using earlier block data also.

Let $\hat{\delta}(S_K, -K)$ denote the mle of treatment effect in subset S_K using data for patients accrued prior to the final block. Generally this will be an optimistically biased estimate of $\delta(S_K)$ because the data available at the final interim analysis determined the selection of S_K . If there were some way of correcting this bias, we could potentially utilize estimators of the form

$$w\hat{\delta}(S_K, K) + (1 - w)[\hat{\delta}(S_K, -K) - \hat{b}] \quad (2)$$

where w is a weight and $\hat{b}$ is an estimate of $b = E[\hat{\delta}(S_K, -K) - \delta(S_K)]$.

We investigated estimating b using a parametric bootstrap approach. We first computed the maximum likelihood estimates of the response distributions $\hat{F}_T(x)$ and $\hat{F}_C(x)$ from the full dataset and the prevalence $P(x)$ from the first block of data. Using these distributions we simulate a replication of the full clinical trial applying the adaptive eligibility restriction algorithm at each interim analysis. That is, at each interim analysis, we determine the eligibility restriction for the next period of accrual and these eligibility restrictions may be different than those employed in the actual clinical trial. For the simulated clinical trial we determine the intended use population as the eligibility criteria used for the final period of accrual. We then computed for that simulated replication

$$\hat{b}^{boot} = \hat{\delta}(S_K^{boot}, -K) - \delta^{boot}(S_K^{boot})$$

The first term on the right hand side is the empirical treatment effect for the intended use population of that replication, using data for patients accrued before the final period in that simulation. The final term is the treatment effect for the intended use population computed using the parameters $\hat{F}_T(x)$ and $\hat{F}_C(x)$ that generated bootstrap samples. These bias estimates are then averaged over bootstrap replications.

3. Evaluating the bootstrap bias adjusted treatment effect

To evaluate this approach to estimation of treatment effect for the intended use population we performed monte-carlo simulations in which outcomes were binary and each patient was characterized by membership in one of J disjoint strata. Although quantitative or categorical covariate values can be transformed into strata, we generated the stratum identifiers directly. At each interim analysis, the strata accruing patients at that time were evaluated for whether they should continue accrual or not. If the log likelihood ratio for the hypothesis $p_T(j)$ $p_C(j) + \delta^*$ relative to the hypothesis $p_T(j) = p_C(j)$ was less than a threshold γ then further

accrual to that stratum ceased. $p_T(j)$ and $p_C(j)$ denote the true response probabilities for the treatment and control regimens for stratum j . δ^* and γ are pre-specified constants. $p_C(j)$ was not assumed known, log likelihoods were maximized over $p_C(j)$. We performed simulations of clinical trials under each of a variety of conditions. For each simulation of the clinical trial, we applied the adaptive enrichment algorithm described above and at the end tested the strong null hypothesis using test statistic (1) of Simon and Simon [1]. In cases where the null hypothesis was rejected, we computed our estimates of the treatment effect for the intended use population using 100 bootstrap replications.

Table 1 shows results of simulations with binary patient response and up to 6 equally prevalent non-overlapping strata. In each case, half the strata have no treatment effect (response probability 0.2 for control and treatment group) and half the strata have a large treatment effect (response probability 0.2 for control and 0.5 for treatment group). There were two periods with 100 patients total accrued in each period. In each case we simulated 1000 clinical trials and performed 100 bootstrap replications for each trial. The adaptive eligibility restriction was performed at the single interim analysis time using parameters $\delta^* = \gamma = 0.25$ as described above. Using a γ of 0.25 for the log likelihood is equivalent to using a threshold of 1.28 for the likelihood ratio. These values provide for relatively weak strata selection. For example in a stratum with 10 patients per treatment at the first interim analysis, if there are two responses on the control regimen the treatment regimen will be continued if it has four or more responses. These parameters δ^* and γ are not claimed to be optimal in any way; they are merely used to illustrate the de-biasing approach. δ^* and γ are parameters of the adaptive enrichment algorithm, not of the de-biasing method. Whatever algorithm is employed in the trial for adaptive enrichment should be used in the bootstrap to compute the de-biasing factor. For guidance on designing the adaptation procedure for an adaptive enrichment trial see [11]

For each of the 1000 clinical trial replications, the patient strata and binary response data was generated for the first period. At the interim analysis each stratum was examined for whether it qualified for inclusion in accrual during the second period. The pre-planned total number of patients to be accrued during the second period was not reduced based on eliminating strata; rather, the target number of patients was allocated to the eligible strata. Patient assignment to strata and treatment was generated by Bernoulli allocation so there was not always exact balance by treatment or strata. We did assume that the strata were equally prevalent.

For computing the table entries, 100 bootstrap samples were performed for each of the 1000 clinical trial replications. The second period eligibility of a trial determined the intended use population for that replication. The full dataset for that trial was used to estimate the response probabilities for each stratum and treatment and those estimates were used to generate the 100 bootstrap samples for that replication. Using those bootstrap samples the de-biasing factor was computed. Consequently, for each replication we computed two estimates of treatment effect for the intended use population; one based on the second period only, one based on a weighted average of the de-biased first period data and the second period. The tables show the effectiveness of the procedure for de-biasing and the resultant mean-squared errors.

The last column of Table 1 shows the estimated power for rejecting the strong null hypothesis of no treatment effect. The second column labeled $\hat{\delta}(S_K, K)$ is the average bias of the empirical treatment effect for the patients treated in the final (second) period. This empirical treatment effect for the second period was just the difference in observed response proportions for the set of patients included in the second period. The strata included in the final period are determined by the adaptive eligibility function computed at the interim analysis based on the first period data. That population may vary among simulated trial replications. The tabulated value shown is the average bias over the subset of the 1000 simulated trials in which the null hypothesis is rejected. The bias in a given clinical trial is the difference between the estimated treatment effect for the final period and the true treatment effect for mixture of strata included in eligibility for the final period. As can be seen in the table, there appears to be a small positive bias in these estimates. We would expect these estimates to be unbiased, but since we have averaged only over simulation replications in which the null hypothesis is rejected, a small positive bias results. Even in non-adaptive clinical trials, the estimated treatment effect for “positive” clinical trials are optimistically biased, the degree of bias depending on the size of the clinical trial and the nature of the interim analysis that may lead to early termination.

The third column in Table 1 labeled $\hat{\delta}(S_K, -K)$ is the average bias of the empirical treatment effect in the first period for the subset of strata selected for inclusion in the second period (the intended use population). As can be seen, this is highly optimistically biased and the bias increases with the number of strata. The fourth column shows that our bootstrap bias adjustment was very effective in correcting for this bias regardless of the number of strata. The bootstrap adjustment made the estimate of treatment effect for the intended use population computed from the first period almost unbiased. Hence, data from periods preceding the last can be used in a combined estimate of treatment effect for the intended use population if effectively bias adjusted. We examined simple averaging of the two estimates and found that the root mean squared error of the treatment effect for the intended use population was reduced by about 20 percent in this way for the cases we examined with two periods and equal weighting of the two estimates (Table 3). Clinical trials with more periods can be expected to provide greater reductions because a smaller proportion of the cases will be included in the final period and hence the estimate based on final period data alone would be expected to be less satisfactory.

Table 2 shows simulations under the same conditions but with 2 interim analyses and therefore 3 blocks of equal accrual. We included 67 patients per block to keep the total number of patients the same as for Table 1. In Table 2 the bias of estimating the treatment effect using data before the final periods unadjusted is less than in Table 1, but is still substantial when the number of strata is 4 or 6. The bootstrap adjustment is again effective here in removing the bias of $\hat{\delta}(S_K, -K)$. Simple averaging of the two estimates, $\hat{\delta}(S_K, K)$ and $\hat{\delta}(S_K, -K)$ reduced the root mean squared error of the treatment effect for the intended use population by about 40 percent for the cases considered with 2, 4, or 6 strata when there were 3 periods of accrual (Table 3).

Equal weighting of the component estimates is not necessarily optimal. Figure 1 shows for a two-period design with 6 strata the RMS error of the weighted average of treatment effect

for the intended use population as a function of the weight for the second period. The first term of expression (2) contains an estimate of treatment effect for the intended use population based on direct observation in the last (second) period in which all n_2 patients are members of the intended use population. The second term is more complicated. It involves an estimate of treatment effect for the intended use population based on a retrospective subsetting of the patients accrued in the previous periods. For a two period design, only a portion of first period patients are members of the subsequently determined intended use population. Hence if $n_1 = n_2$, the sample size for estimating the treatment effect of the intended use population in the first period will be less than for the second period. In addition, the estimate based on the first period must be de-biased, which adds an additional source of variability. Usually for estimation based on weighted averages of independent quantities, the optimal weights are based on the inverse of the variances. If one ignores lack of independence, one might expect that the optimal weight might be approximately $n_2 / (\eta n_1 + n_2)$ where η denotes the proportion of the first period population who would be eligible for the second period. For the example shown in Figure 1, $\eta = 0.5$ and $n_1 = n_2$ so the approximate optimal weight is computed as 0.67 in pretty good agreement with the figure. A similar guideline can be developed for three period designs.

4. Determining the intended use population using the full trial dataset

In the preceding sections we specified the intended use population based on the adaptive eligibility function computed at the final interim analysis. This was motivated by the desire to have a direct unbiased estimate of treatment effect for that population; the sample treatment effect observed in the final period of accrual. We could, however, consider defining the intended use population by applying the adaptive eligibility algorithm to the full dataset available at the end of the clinical trial. We will denote the eligibility function computed from the full dataset by $f_{K+1}(\cdot)$; that is:

$$S_{K+1} = \{x: f_{K+1}(x) = 1\}$$

where S_{K+1} denotes the intended use population determined by applying the algorithm at the conclusion of the trial to the full dataset. There is no $(K+1)$ st period in which treatment effect can be unbiasedly estimated for this population but we evaluated a bootstrap bias adjusted estimator for estimating treatment effect for S_{K+1} .

The empirical treatment estimate $\hat{\delta}(S_{K+1}, -(K+1))$ for patients in S_{K+1} estimated using the full dataset will be optimistically biased but we can potentially correct for that bias using the bootstrap in a manner similar to that described above. In particular we can use $\hat{\delta}(S_{K+1}, -(K+1)) - \hat{b}$.

We extended our simulations to evaluate this approach for determining the intended use population and the results are shown in Table 4. The third column gives the bias for the uncorrected estimate of treatment effect using the full dataset. The bias is fairly small for two strata but increases with more strata. The fourth column shows the bias of the bootstrap adjusted estimate. The bootstrap bias adjustment is effective with some evidence of a slight over-adjustment. Basing the future intended use population on a classifier developed using

the full dataset can potentially be more accurate than one based on excluding the last period data, at least for studies with a small number of periods. The last two columns of the table show the proportions of the strata that were correctly classified as having a null treatment effect or a positive treatment effect in the simulations. A stratum was considered correctly classified if it had a true positive treatment effect and was included in the intended use population by the classifier or was null and not included. The table compares these proportions for the classifier based on the full dataset to the one based on the data up until the final period of accrual. The full dataset classifier is seen to be more accurate.

5. Hypothesis testing for the intended use population

Rejecting the strong null hypothesis (1) establishes that there is a treatment effect for some subset of patients but it does not establish that there is a treatment effect for the subset taken as the intended use population in the adaptive trial. We will show here that under certain conditions, the final hypothesis test can be interpreted as a test of the null hypothesis of no treatment effect for the population determined by the eligibility criteria used for the final block of accrual; or even an indicated population based on the eligibility restriction function constructed using the full dataset.

We will consider the case in which there are p candidate binary markers $x_1, \dots, x_p$. This assumption is more general than it seems because a quantitative or ordinal biomarker with n possible cut-points can be represented by n binary markers; e.g. coded as (0,0,0) for values in the lowest quartile, (1,0,0) for values in the second quartile, (1,1,0) for values in the third quartile and (1,1,1) for the largest values. The regression coefficients associated with the binary variables determine the steepness of the response for the quantitative marker on the quantile scale. Categorical markers with no ordering of the levels cannot be represented in this formulation however because of the assumptions described below.

Define $\geq$ to be an element-wise comparison: $x \geq x'$ is defined to mean $x_j \geq x'_j$ for every j . Also let $P(x)$ denote the population proportion of patients with biomarker levels x . We make four assumptions:

- A1** The eligibility functions are boolean conjunctions; i.e. $f_k(x) = 1$ iff $x \geq c^k$ component-wise where each component of c^k is zero or one. For example, suppose that the x vector has 3 components; $x_1 = 1$ for EGFR amplified and 0 otherwise; $x_2 = 1$ for KRAS wild type and 0 otherwise; $x_3 = 1$ for NRAS wild type and 0 otherwise. Initially all patients are eligible so $c^1 = (0, 0, 0)$. Eligibility functions which can be represented as conjunctions include, but are not limited to: (i) EGFR amplified; (ii) EGFR amplified and KRAS wild type; (iii) EGFR amplified and KRAS wild type and NRAS wild type.
- A2** Markers are coded so that positive marker values (i.e. $x_j = 1$) are the candidates for larger treatment effects. In particular $\delta(x) \geq \delta(x')$ if $x \geq x'$ where $\delta(x)$ denotes the treatment effect for a patient with covariate vector x .
- A3** Eligibility sets are nested. That is $c^k \geq c^{k'}$ if $k < k'$.

A4 Binary markers have non-negative dependence: That is for any $j \leq p$, $c \in \{0, 1\}^p$, $s \in \{0, 1\}^{p-1}$, and $s \in \{0, 1\}^{p-1}$ we have $P(x_{-j} \geq s/x_{-j} \geq c, x_j = 1) \geq P(x_{-j} \geq s/x_{-j} \geq c, x_j = 0)$. This holds for independent markers.

Although assumption A4 is not intuitive, it is required for the proof of the following result to hold. To give intuition for this assumption, think of our 0–1 random variables as absence/presence of given markers. One understanding of this assumption is that presence of any given marker (ie. $x_j = 1$ for a given j) cannot decrease the probability of presence of any other set of markers.

Theorem 1—Under the above assumptions (A1, A2, A3, and A4), if the strong null hypothesis (1) is false, then the null hypothesis for the intended use population S_K or S_{K+1} is also false.

To prove this we first give the following lemma.

Lemma 1—Consider a threshold $c \in \{0, 1\}^p$, with $c_1 = 0$; the rest of the elements may have arbitrary values. Define c' by $c'_1 = 1$, and for $j \geq 1$, $c'_j = c_j$. Define $S = \{x \geq c\}$ and $S' = \{x \geq c'\}$. Under assumptions A2, and A4 we have $\delta(S) \geq \delta(S')$.

This lemma states that the average effect-size is non-decreasing if we swap a single element of our threshold from 0 to 1. We leave the proof of this lemma to the appendix.

Proof of Theorem 1—If (1) does not hold, then there exists some period k^* for which

$$\delta(S_{k^*}) > 0 \quad (3)$$

Since $k^* \leq K$, and our eligibility sets are nested we have $c^{k^*} \leq c^K$ by assumption A3. If $c^{k^*} = c^K$ then our result follows trivially. Otherwise we can get from c^{k^*} to c^K via a finite sequence of $0 \rightarrow 1$ swaps. More formally, there exist a non-decreasing sequence of thresholds $c^{k^*} = c_1 \leq c_2 \leq \dots \leq c_m = c^K$, with corresponding eligibility sets $S_{k^*} = S_1, S_2, \dots, S_m = S_K$, such that for each $j < m$, the thresholds c_j and c_{j+1} differ in only a single position (where c_j is 0 and c_{j+1} is 1). From here we can apply Lemma 1 and get that $0 < \delta(S_{k^*}) = \delta(S_1) \leq \delta(S_2) \leq \dots \leq \delta(S_m) = \delta(S_K)$. This completes the proof. An identical argument proves this for S_{K+1} .

5.1. Evaluation of Assumption A4

Assumption A4 can be evaluated prior to the start of the clinical trial if multivariate external data on the markers is available. We examined the validity of assumption A4 for a hypothetical investigation of an anti-EGFR agent (eg. cetuximab). Two biomarkers to consider here are: EGFR expression level in the tumor, and KRAS mutation status. Treatment is believed to have increased efficacy in tumors with increased EGFR expression; and in tumors that are KRAS wildtype. Using the colorectal cancer genomic data from The Cancer Genome Atlas [12] we evaluated the dependence between these two markers. This dataset contains 207 cases with data available on both KRAS mutation status and EGFR mRNA expression levels. (The supplemental methods section of the above Nature paper

states that for the expression profiling, sample and reference were co-hybridized on a Custom Agilent 244K Gene Expression Microarray (AMID019760). The expression data was Lowess normalized and the ratio of the Cy5 channel (sample) and Cy3 channel (reference) were log2 transformed to create gene expression values).

For the purpose of this analysis, cases with any type of KRAS mutation were classified as '0' (wildtype were '1'). We also created a binary variable from our continuous EGFR levels (indicating expression above the median). A 2x2 table of sample proportions is given as table 5. We see that in this case, A4 appears to be satisfied: KRAS-wildtype tumors are more likely than KRAS-mutants to be high EGFR expressers; and high EGFR expressers are more likely than low expressers to be KRAS-wildtype.

We defined the indicator variable for KRAS to have value 1 for wildtype because tumors bearing a KRAS mutation are unlikely to be effected by blocking the EGF receptor. The same is true for mutations in NRAS, a different member of the RAS family of genes. Defining a binary variable for NRAS with 1 denoting wildtype however would not satisfy assumption A4 because tumors that bear KRAS mutations are less likely to bear NRAS mutations than tumors that are wildtype for KRAS. In the TCGA data for colorectal cancer, there were 19 patients with mutant NRAS and 5 (26%) of them also had mutant KRAS. Of the 188 patients with wildtype NRAS, 79 (42%) had mutant KRAS. This negative correlation between NRAS and KRAS mutations also holds within the two levels of EGFR expression. Consequently if our intended use population at the end of the trial is over-expressed EGFR and both wildtype KRAS and NRAS, then we will be unable to use the theorem presented here to claim that the treatment effect is statistically significant for the intended use population. The theorem gives necessary but not sufficient conditions however. In fact, if these are the only three variables, the average treatment effect for the set S' characterized by high EGFR expression and wildtype of both RAS genes is greater than the average treatment effect for the set S characterized by high EGFR and wildtype KRAS. As in the first step of the proof of Lemma 1, these average treatment effects can be expressed as

$$\begin{aligned}\delta(S) &= \delta(0, 1, 1)Pr[x = (0, 1, 1) \mid x \geq (0, 1, 1)] + \delta(1, 1, 1)Pr[x = (1, 1, 1) \mid x \geq (0, 1, 1)] \\ \delta(S') &= \delta(1, 1, 1)\end{aligned}\quad (4)$$

where the first component of the x vector is for NRAS (1=wildtype) and the other two components are for KRAS and EGFR. Therefore

$$\begin{aligned}\delta(S) - \delta(S') &= \delta(0, 1, 1)Pr[x = (0, 1, 1) \mid x \geq (0, 1, 1)] - \delta(1, 1, 1)(1 - Pr[x = (1, 1, 1) \mid x \geq (0, 1, 1)]) \\ &= Pr[x = (0, 1, 1) \mid x \geq (0, 1, 1)](\delta(0, 1, 1) - \delta(1, 1, 1)) \\ &\leq 0.\end{aligned}\quad (5)$$

Consequently, with exactly three binary variables, if two of the variables are selected first and satisfy the four assumptions, then the theorem will be true for the classifier containing all three variables even if assumption A4 is not satisfied for the final variable. The reason, intuitively stated, is that when the final variable is added to the classifier, there are no remaining variables whose distribution may be of concern. In the future, we will attempt to

broaden the conditions under which significance for the intended use population can be established for problems with multiple important biomarkers.

6. Discussion

Phase III enrichment designs have been used in the past decade as the basis for regulatory approval of several oncology drugs. In other cases, however, the appropriate predictive biomarker has not been identified by the time that the pivotal phase III trials are initiated. The phase III trial itself often provides a more appropriate patient population and control group for identifying the appropriate predictive biomarker than do phase II studies. The adaptive enrichment approach provides a framework for the design of such studies. In addition to offering some of the efficiency advantages of the enrichment design ([13]), it can provide a more accurate identification of the intended use population than do conventional broad eligibility clinical trials.

The designs discussed here are very general in that the eligibility restriction algorithms can utilize information for multiple biomarkers although they are probably most useful in adaptively determining how to utilize candidate biomarkers which have been identified in previous studies. For the single statistical test used by the Simon & Simon [1] adaptive designs, all patients are included but the test statistic is a weighted average of period-specific test statistics, each having a known null distribution. The analysis plan can include group sequential interim analyses as with non-adaptive trials.

We have used a parametric bootstrap method to de-bias the estimated treatment effect. Adaptively changing eligibility rules limits the applicability of the more common non-parametric bootstrap. That is, the sampling distribution of covariate vectors in periods $k > 1$ of a bootstrap replication may not be the same as for that period of the actual trial. Using the parametric bootstrap we estimate prevalence of covariate vectors using first period data and fit a model for the response given covariate vector using the full dataset. The parametric bootstrap approach has worked well to reduce bias in the cases we have examined. In the case of disjoint patient strata, the estimated stratum-specific outcome distributions for each treatment are used to generate bootstrap samples. For the case where multiple quantitative markers are measured for each patient, the estimated parameters of regression models can be used for driving the bootstrap. This will have limitations for high dimensional covariate vectors, but generally in phase 3 clinical trials a small number of candidate markers will be of interest. Even in this case, however, care should be taken in the modeling of response. The type of classifier used for adaptive enrichment should be specified in advance, but the type of model used for bootstrap de-biasing need not be determined until the trial is complete and the full dataset available. The bootstrap estimate obtained in this manner should be consistent [14] if the model class used for the bootstrap is sufficiently broad as to contain the true model. The modeling may influence both the effectiveness of adaptive enrichment of the population and the accuracy of the de-biasing although the effect on the former may be greater than the latter. For example we performed a simulation using a linear logistic model with two quantitative variables for enrichment and bootstrap sampling. The true model, however, was a logistic model with no main effect of treatment or either covariate but with an interaction term indicating a treatment effect only if variable one exceed it's median

value. For 1000 replications of a two-period design with 150 patients per period, the average observed treatment effect (difference in response proportions) for the first period was 0.1399. The average treatment effect after enrichment in the second stage was 0.1981. The average re-substitution estimate of first period treatment effect for the population selected based on the interim analysis was 0.2366. The bias adjusted estimate of treatment effect for first period patients in the selected population was 0.2015, less than the biased 0.2366 re-substitution estimate and close to the unbiased 0.1981 second period average.

The bias in estimation of treatment effect after selection has been discussed previously in terms of selection of treatment arms, rather than patient populations [15]. Magnusson et al. [16] proposed a bootstrap approach for de-biasing estimates of treatment effect for disjoint strata two-period designs but did not evaluate its performance. Rosenkranz also discussed bootstrap debiasing of the estimate of maximum outcome for a treatment in several populations [17].

Our focus here is improving the point estimate in treatment effect for the intended use population. The bootstrap method used served to enable data from earlier blocks of patients to be used without bias and thereby improved the precision of the point estimate. The bootstrap can also be used to provide a confidence interval for the treatment effect in the intended use population, although evaluation of the coverage of such intervals awaits future investigation. That confidence interval could, if the coverage is correct, also be used to provide a significance test of the null hypothesis of no treatment effect for the intended use population in cases where the conditions of the Theorem in Section 5 are not satisfied.

A wide variety of adaptive eligibility restriction algorithms can be used. In the case of multiple disjoint strata the decision functions are most obviously based on estimates of treatment effects within each stratum. Simon & Simon [1] considered the case of a single quantitative biomarker and restricting eligibility to those with biomarker values above an adaptively chosen threshold. More generally, the eligibility restriction function can be based on regression modeling of outcome as a function of biomarkers for each treatment group and identification of regions where the treatment effect is consistent with a clinically useful benefit. The eligibility restriction functions can also be based on Bayesian modeling of outcome in the test treatment and control groups. This provides for effective use of Bayesian methods within a frequentist phase III clinical trial. This works because the adaptive enrichment framework used here preserves the experiment-wise type I error regardless of the adaptive enrichment algorithm used. The bootstrap de-biasing will reflect the effect of the particular adaptive enrichment method used.

When survival time is the primary endpoint for analysis, the development of an eligibility restriction may have to depend on an intermediate endpoint such a progression free survival or disease control at 8 weeks as has been used for other adaptive trials [18]. There is no requirement that the intermediate endpoint be a valid surrogate for survival but it is useful for it to be a “conditional surrogate” in the sense that a treatment effect on the intermediate endpoint is necessary but not sufficient for having a treatment effect on survival. With a survival endpoint the test statistic for a three period design would be of the form $T = w_1 z_1 + w_2 z_2 + w_3 z_3$ where z_i may be the normal quantile for a log-rank test of treatment effect for

patients accrued in period i and w_i is a pre-specified weight. The log-rank statistics are computed at the analysis time which will be after sufficient events have occurred but the eligibility adaptive function would be based on a short term endpoint.

The adaptive enrichment approach may be used in a variety of clinical circumstances. For example, the first stage may represent a phase II screening stage in which patient groups or biomarkers are selected for further accrual based on response to the test treatment. This becomes particularly relevant in cases where eligibility is based on a genomic alteration in the pathway of the molecular target of the test drug and there is uncertainty about which genes are most suitable and which mutations are functional. In other circumstances, patient groups may be de-selected for subsequent eligibility when there is greater evidence that they do not benefit from the test treatment. Bayesian predictive probabilities can play a role in this assessment.

There are many interesting open questions regarding how to optimally utilize the adaptive enrichment framework in clinical development. For example there are questions of how aggressive to restrict eligibility. There are also questions of adaptive sample size modification and establishment of group sequential monitoring boundaries. The results of this paper, however, have addressed two important issues that move this class of designs closer to applicability for major phase III clinical trials.

Acknowledgments

N.S. was supported by NIH Grant DP5OD019820.

References

1. Simon N, Simon R. Adaptive enrichment designs for clinical trials. *Biostatistics*. 2013; 14(4):613–625. [PubMed: 23525452]
2. Chapman PB, Hauschild A, Robert C. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *The New England Journal of Medicine*. 2011; 364:2507–2516. [PubMed: 21639808]
3. Majewski JJ, Bernards R. Taming the dragon: genomic biomarkers to individualize the treatment of cancer. *Nature Medicine*. 2011; 17:304–312.
4. Wang S, Hung HJ, O'Neill R. Approaches to evaluating treatment effect in randomized clinical trials with genomic subset. *Biopharmaceutical Statistics*. 2007; 6:227–244.
5. Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine*. 2009; 28(10):1445–1463. [PubMed: 19266565]
6. Rosenblum M, van der Laan MJ. Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika*. 2011; 98(4):845–860. [PubMed: 23049131]
7. Song JX. A two-stage patient enrichment adaptive design in phase II oncology trials. *Contemporary Clinical Trials*. 2014; 37:148–154. [PubMed: 24342820]
8. Jennison C, Turnbull BW. Adaptive seamless designs: selection and prospective testing of hypotheses. *Journal of Biopharmaceutical Statistics*. 2007; 17(6):1135–1161. [PubMed: 18027222]
9. Mehta CR, Gao P. Population enrichment designs: Case study of a large multinational trial. *Journal of Biopharmaceutical Statistics*. 2011; 21:831–845. [PubMed: 21516572]
10. Magnusson BP, Turnbull BW. Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine*. 2013; 32:2695–2714. [PubMed: 23315698]

11. Simon N, Simon R. Using Bayesian models in frequentist adaptive enrichment designs. *Biostatistics*. 2017
12. Network CGA, et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012; 487(7407):330–337. [PubMed: 22810696]
13. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*. 2004; 10(20):6759–6763. [PubMed: 15501951]
14. van der Laan M, Bryan J. Gene expression analysis with the parametric bootstrap. *Biostatistics*. 2001; 2(4):445–461. [PubMed: 12933635]
15. Bauer P, Koenig F, Brannath W, Posch M. Selection and bias: two hostile brothers. *Statistics in Medicine*. 2010; 29(1):1–13. [PubMed: 19844944]
16. Magnusson BP, Turnbull BW. Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine*. 2013; 32(16):2695–2714. [PubMed: 23315698]
17. Rosenkranz GK. Bootstrap corrections of treatment effect estimates following selection. *Computational Statistics & Data Analysis*. 2014; 69:220–227.
18. Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein GR, Tsao A, et al. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discovery*. 2011; 1(1):44–53. [PubMed: 22586319]

8. Appendix

Here we have the proof of Lemma 1:

Proof of Lemma 1

Let $c_{-1} = [c_2, \dots, c_p]$ and $x_{-1} = [x_2, \dots, x_p]$ (for arbitrary x). We first define $u = y^* z - y^* (1 - z)$. Now note that

$$\begin{aligned}
 \delta(S) &= E[u \mid x \geq c] \\
 &= \sum_{x_{-1} \geq c_{-1}} (E[u \mid x_{-1}, x_1 = 0] P(x_{-1} \mid x_{-1} \geq c_{-1}, x_1 = 0) P(x_1 = 0 \mid x_{-1} \geq c_{-1}) \\
 &\quad + E[u \mid x_{-1}, x_1 = 1] P(x_{-1} \mid x_{-1} \geq c_{-1}, x_1 = 1) P(x_1 = 1 \mid x_{-1} \geq c_{-1})) \\
 \delta(S') &= E[u \mid x \geq c'] \\
 &= \sum_{x_{-1} \geq c_{-1}} E[u \mid x_{-1}, x_1 = 1] P(x_{-1} \mid x_{-1} \geq c_{-1}, x_1 = 1)
 \end{aligned}$$

Taking the difference, and simplifying, we see that

$$\begin{aligned}
 \delta(S') - \delta(S) &= P(x_1 = 0 \mid x_{-1} \geq c_{-1}) \sum_{x_{-1} \geq c_{-1}} (E[u \mid x_{-1}, x_1 = 0] P(x_{-1} \mid x_{-1} \geq c_{-1}, x_1 = 0) - E[u \mid x_{-1}, \\
 &\quad x_1 = 1] P(x_{-1} \mid x_{-1} \geq c_{-1}, x_1 = 1))
 \end{aligned}$$

We aim to show that $\delta(S') - \delta(S) \geq 0$. This follows from stochastic dominance if we can show that, for every t we have

$$P(E[u \mid x_{-1}, x_1 = 0] \geq t \mid x_{-1} \geq c_{-1}, x_1 = 0) \leq P(E[u \mid x_{-1}, x_1 = 1] \geq t \mid x_{-1} \geq c_{-1}, x_1 = 1) \quad (6)$$

The remainder of the proof argues that (3) holds. Let t be fixed. Consider the set

$$S_t = \{x_{-1} \mid E[u \mid x_{-1}, x_1 = 0] \geq t\}$$

By assumption A2, we can see that this set must be the union of shifted positive cones. Thus there exist $s_{1(t)}, \dots, s_{m(t)}$ such that we can write S_t as

$$S_t = \{s \mid s \geq s_{1(t)} \text{ OR } s \geq s_{2(t)} \text{ OR } \dots \text{ OR } s \geq s_{m(t)}\}$$

Plugging in we see that

$$\begin{aligned} P(E[u \mid x_{-1}, x_1 = 0] \geq t \mid x_{-1} \geq c_{-1}, x_1 = 0) &= P(x_{-1} \geq s_{1(t)} \text{ OR } x_{-1} \geq s_{2(t)} \text{ OR } \dots \text{ OR } x_{-1} \geq s_{m(t)} \mid x_{-1} \geq c_{-1}, x_1 = 0) \\ &\leq P(x_{-1} \geq s_{1(t)} \text{ OR } x_{-1} \geq s_{2(t)} \text{ OR } \dots \text{ OR } x_{-1} \geq s_{m(t)} \mid x_{-1} \geq c_{-1}, x_1 = 1) \\ &= P(E[u \mid x_{-1}, x_1 = 0] \geq t \mid x_{-1} \geq c_{-1}, x_1 = 1) \end{aligned}$$

The inequality in the second line above follows from assumption A4.

We now note that, by A2, for all $s \in S_t$ we have

$$t \leq E[u \mid x_{-1}, x_1 = 0] \leq E[u \mid x_{-1}, x_1 = 1]$$

Thus,

$$P(E[u \mid x_{-1}, x_1 = 0] \geq t \mid x_{-1} \geq c_{-1}, x_1 = 1) \leq P(E[u \mid x_{-1}, x_1 = 1] \geq t \mid x_{-1} \geq c_{-1}, x_1 = 1) \quad (7)$$

Stringing all of this together we verify our stochastic domination claim:

$$P(E[u \mid x_{-1}, x_1 = 0] \geq t \mid x_{-1} \geq c_{-1}, x_1 = 0) \leq P(E[u \mid x_{-1}, x_1 = 1] \geq t \mid x_{-1} \geq c_{-1}, x_1 = 1)$$

Which gives us that $\mathcal{R}(S') - \mathcal{R}(S) \geq 0$.

Figure 1. RMS error of the weighted average of treatment effect for the intended use population as a function of the weight for the second period from a two-period design with 6 strata.

Table 1

Two Blocks of Accrual

Number of strata	Mean bias of $\hat{\alpha}(S_K)$	Mean bias of $\hat{\alpha}(S_{-K})$	Mean bias of $\hat{\alpha}(S_{-K}) - \hat{\alpha}$	Power
2	.00712	.0422	-.00435	.813
4	.0154	.0997	-.0098	.835
6	.0179	.1386	.00060	.844

Table 2

Three Blocks of Accrual

Number of strata	Mean bias of $\hat{\alpha}(S_K)$	Mean bias of $\hat{\alpha}(S_{-K})$	Mean bias of $\hat{\alpha}(S_{-K}) - \hat{\alpha}$	Power
2	.00569	.0237	-.0109	.795
4	.0183	.0547	-.0091	.858
6	.0128	.0845	.00113	.847

Table 3

RMS Error

Blocks of accrual	Strata	RMS Error $\hat{\delta}_{S_K}$	RMS Error weighted average
2	2	.0861	.0644
2	4	.0835	.0706
2	6	.0853	.0700
3	2	.1083	.0643
3	4	.1087	.0659
3	6	.1067	.0650

Table 4

Blocks of accrual	Strata	Mean Bias of		Proportion of strata correctly classified	
		$\delta(S_{K+1})$	$\delta(S_{K+1}) - \delta$	using full dataset	using K-1 blocks
2	2	.0182	-.00299	.989	.968
2	4	.0451	-.00706	.892	.837
2	6	.0659	.00352	.822	.772
3	2	.0180	-.00930	.994	.984
3	4	.0391	-.01200	.863	.846
3	6	.0510	-.01181	.793	.777

Table 5

	EGFR– (0)	EGFR+ (1)
KRAS-mutant (0)	51	33
KRAS-wildtype (1)	52	71