To Appear: Statistics in Medicine, 2014

Biomarker Driven Population Enrichment for Adaptive Oncology Trials with Time to Event Endpoints

Cyrus Mehta[(†)], Helmut Schäfer[(*)], Hanna Daniel[(*)], Sebastian Irle[(⊕)]

[(†)]Cytel Corporation, Cambridge MA, USA
[(‡)]Harvard School of Public Health, Boston MA, USA
[(*)] University of Marburg, Marburg, Germany
[(⊕)]German Central Bank, Frankfurt, Germany

June 29, 2014

Abstract

The development of molecularly targeted therapies for certain types of cancers has led to the consideration of population enrichment designs that explicitly factor-in the possibility that the experimental compound might differentially benefit different biomarker subgroups. In such designs, enrollment would initially be open to a broad patient population with the option to restrict future enrollment, following an interim analysis, to only those biomarker subgroups that appeared to be benefiting from the experimental therapy. While this strategy could greatly improve the chances of success for the trial, it poses several statistical and logistical design challenges. Since late-stage oncology trials are typically event driven, one faces a complex trade-off between power, sample size, number of events and study duration. This trade-off is further compounded by the importance of maintaining statistical independence of the data before and after the interim analysis and of optimizing the timing of the interim analysis. This paper presents statistical methodology that ensures strong control of type-1 error for such population enrichment designs, based on generalizations of the conditional error rate approaches of Müller and Schäfer [8] and Irle and Schäfer [12]. The special difficulties encountered with time-to-event endpoints are addressed by our methods. The crucial role of simulation for guiding the choice of design parameters is emphasized. Although motivated by oncology, the methods are applicable as well to population enrichment designs in other therapeutic areas.

Keywords: Subgroup selection, targeted therapies, adaptive design, clinical trial, conditional error function, multiple comparisons, survival endpoints, precision medicine, predictive biomarker

1 Introduction

This paper presents a method for designing two-stage adaptive trials that permit biomarker driven population enrichment at the interim analysis. The approach is applicable to situations where a single binary biomarker partitions the population into two subgroups. The focus is on oncology though the methods can also be applied in other therapeutic areas. Nearly 60% of oncology trials fail in phase 3 testing [1]. A major cause of this attrition is inability to identify the appropriate population with the greatest potential to benefit from the test drug. Treatment effects can differ greatly between subsets of patients with different genomic characteristics for tumors in the same disease class. For example, in several recent trials of metastatic colorectal cancer, the benefit of anti-EGFR antibodies was shown to be limited to patients with KRAS wild-type tumors [2]. Table 1 displays several targeted therapeutic agents that were approved in the United States for specific subgroups of patients. In many of these

Table 1: Oncology Products Approved in the U.S. for Selected Populations

Compound	Target	Indication
Crizotinib (Xalkori®)	ALK	ALK-rearranged non-small cell lung cancer
Vemurafenib (Zelboraf®)	BRAF	BRAF mutant advanced melanoma
Trametinib (Mekinist®)	MEK	BRAF mutant advanced melanoma
Trastuzumab (Herceptin®)	Her 2	Her 2 expressing breast cancer
Lapatinib (Tykerb®)	Her 2	Her 2 expressing metastatic gastric cancer
Rituximab (Rituxan®)	CD20	CD20(+) B-cell lymphomas
Cetuximab (Erbitux®)	EGFR	$KRAS^{wt}$, EGFR(+) metastatic colorectal cancer
Panitumumab (Vectibix®)	EGFR	$KRAS^{wt}$, EGFR(+) metastatic colorectal cancer

cases the investigation of tumor sensitivity to the new therapeutic agent was obtained retrospectively from prospective trials in which patients were randomized to the treatment or control arms without regard to biomarker status. For regulatory approval, however, it is necessary to demonstrate the efficacy of the new agent in the targeted population through a confirmatory phase 3 trial.

The dilemma for the investigator planning a phase 3 confirmatory trial for a targeted therapy is whether to open enrollment to all patients regardless of biomarker status or to restrict enrollment to a targeted subgroup based on a biological understanding of the mechanism of action. Restricting enrollment to the targeted subgroup without sufficient empirical evidence of lack of efficacy in the non-targeted subgroup may deny a large segment of the population access to a potentially beneficial treatment. On the other hand by running a large trial in a heterogeneous population the treatment effect may be diluted, resulting in an underpowered study. One way to resolve this dilemma is to start out by enrolling all patients, regardless of biomarker status. At a suitable time point an interim analysis is performed and a decision is taken to either continue enrollment to both subgroups, continue enrollment to the targeted subgroup only (population enrichment), or terminate the trial for futility. In the phase 2 setting, the data obtained at the end of such a trial can inform the investigator concerning a follow-on phase 3 trial. If the evidence suggests that the biomarker is predictive of treatment effect, that would justify investing in the development of a validated companion diagnostic test, a regulatory requirement, and launching a phase 3 trial in the targeted subgroup only.

In this paper we will develop statistical methodology for designing two-stage adaptive population

enrichment trials of the type described above. We will focus on oncology trials with time-to-event endpoints. The methods are, however, also applicable to other therapeutic areas. The population F is partitioned into two non-intersecting subgroups S and $\overline{S}$ on the basis of a single binary biomarker. The primary endpoint is a time-to-event endpoint such as progression free survival (PFS) or overall survival (OS). This ensures that the conclusions obtained from the phase 2 setting are relevant for the follow-on phase 3 trial where the primary efficacy endpoint is also time-to-event. Data dependent adaptations of two-stage time-to-event trials can be problematic because decisions made at the interim analysis could affect the total number of events contributed to the final analysis by the stage 1 recruits (some of whom are still censored at the end of stage 1). This problem was first observed by Bauer and Posch [3] in the context of adaptive time-to-event trials with sample size re-estimation. Our method will, however, permit full utilization of all available interim data (complete as well censored observations) for the decision to either continue with the full population or the targeted subgroup. It is based on a generalization of the conditional error rate approaches of Müller and Schäfer [8] and Irle and Schäfer [12], and guarantees strong control of type-1 error.

Alternative approaches that are applicable to subgroup selection in time-to-event trials have been proposed by Brannath et. al. [4], Jenkins, Stone and Jennison [5] and Friede, Parsons and Stallard [6]. Jenkins et. al. [5] were the first to develop a suitable design for time-to-event endpoints where it is permissible to utilize all available information from the first stage, including early outcome information such as early tumor response or PFS in patients who are still censored for their OS outcome. They combined the data from the two stages with pre-specified weights, in the manner of Bauer and Köhne [7] to control the error due to subgroup selection at the interim, and utilized Simes test to control the error due to multiple testing among subgroups. Friede et. al. [6] improved on this approach by applying the conditional rejection probability (CRP) principle of Müller and Schäfer [8] for the subgroup selection problem and utilized a more powerful intersection hypothesis test due to Spiessens and Debois [9] for the multiple testing problem. Their final test statistic, however, is also constructed by combining the data from the two stages with pre-specified weights.

Our paper differs from the preceding ones in terms of the hypotheses being tested. The previous methods tested null hypotheses of no treatment effect in population F and subgroup S. In contrast we will be testing null hypotheses of no treatment effect in subgroups S and $\overline{S}$. We believe that this is the appropriate family of hypotheses for which control of type-1 error is required. In oncology trials of targeted therapies there is an a priori assumption that the biomarker is predictive of treatment efficacy. That is, there is considerable treatment efficacy in subgroup S but little or none in subgroup $\overline{S}$. This assumption typically has a strong biological basis and may be supported by pre-clinical studies, phase 1 testing, or retrospective analysis of completed trials. As pointed out by Buyse et. al. [10], although the predictive potential of a putative biomarker can be suggested by these methods, the ultimate proof that a biomarker is truly predictive comes from a randomized clinical trial. The direct way to verify predictivity in such a trial is to perform tests of hypothesis in S and $\overline{S}$ rather than in F and S. This is discussed further and supported by simulations in Section 5.

The remainder of this paper consists of four sections. In Section 2 we describe the basic statistical principle underlying our method. In Section 3 we illustrate this principle through a numerical example. Section 4 consists of an extensive simulation study. Section 5 contains some final conclusions.

2 The Statistical Methodology

Our goal is to construct a two-stage design in which an experimental arm (E) is compared to a control arm (C) with respect to a time-to-event endpoint, say survival. Patients arriving in staggered fashion from some population F are screened and stratified on the basis of a binary biomarker into subgroup S or subgroup $\overline{S}$, and then randomized to one of the two treatment arms. Let θ_S and $\theta_{\overline{S}}$ denote the negative log hazard ratio of E relative to C in subgroups S and $\overline{S}$, respectively. We shall be interested in testing the null hypotheses H^S: $\theta_S \leq 0$ and $H^{\overline{S}}$: $\theta_{\overline{S}} \leq 0$ against one-sided alternatives, with strong control of the family wise error rate (FWER). At an interim analysis the data are unblinded and a decision is taken to either continue with F for the remainder of the trial, drop $\overline{S}$ and continue with S for the remainder of the trial, or terminate the trial for futility. We are not considering the option of dropping S and continuing with $\overline{S}$ for the remainder of the trial. The underlying belief, supported by biological, pre-clinical and retrospective clinical evidence, is that treatment E is targeted at subgroup S. Therefore, if the interim data support this assumption, we wish to maximize the power to reject H^S by enriching the remainder of the trial with subgroup S patients only. We shall compare this adaptive strategy with the a non-adaptive strategy in which there is no interim analysis but a multiplicity adjusted test of H^S is performed at the end of the study. We next show how to achieve strong control of the FWER in the presence of multiple testing and possible subgroup selection.

2.1 Closed Testing with the Conditional Error Rate Approach

We first consider the problem of testing H^S and $H^{\overline{S}}$ simultaneously at FWER equal to α, without any interim analysis. To adjust for multiplicity we apply the closed testing principle of Marcus et. al. [11]. This implies that the three hypotheses $H^S, H^{\overline{S}}$ and $H^{\{S,\overline{S}\}} = H^S \cap H^{\overline{S}}$ must each be controlled at level α. It is convenient to formulate this requirement in terms of the decision functions $\varphi^S, \varphi^{\overline{S}}$ and $\varphi^{\{S,\overline{S}\}}$ such that: $\varphi^S = 1$ if H^S is rejected and 0 otherwise; $\varphi^{\overline{S}} = 1$ if $H^{\overline{S}}$ is rejected and 0 otherwise; $\varphi^{\{S,\overline{S}\}} = 1$ if $H^{\{S,\overline{S}\}}$ is rejected and 0 otherwise. We require

$$E_0(\varphi^S) = E_0(\varphi^{\overline{S}}) = E_0(\varphi^{\{S,\overline{S}\}}) = \alpha \tag{1}$$

where $E_0(.)$ denotes expectation under the appropriate null hypothesis. To this end let $T^S_{k^S}$ ($T^{\overline{S}}_{k^{\overline{S}}}$) be the logrank score for testing the null hypothesis H^S ($H^{\overline{S}}$) after observing k^S ($k^{\overline{S}}$) deaths in subgroup S ($\overline{S}$). Then the decision functions φ^S and $\varphi^{\overline{S}}$ are indicator variables

$$\varphi^S = I(T^S_{k^S} > c^S) \text{ and } \varphi^{\overline{S}} = I(T^{\overline{S}}_{k^{\overline{S}}} > c^{\overline{S}}) \tag{2}$$

for suitable critical boundaries c^S and $c^{\overline{S}}$, respectively, that satisfy the level requirement (1). The decision function for the intersection hypothesis is the indicator variable

$$\varphi^{\{S,\overline{S}\}} = I\left((T^S_{k^S}, T^{\overline{S}}_{k^{\overline{S}}}) \in R)\right) \tag{3}$$

where R is a rejection region of the form

$$R = \left\{(t^S, t^{\overline{S}})|(t^S > d^S) \vee (t^{\overline{S}} > d^{\overline{S}})\right\} \tag{4}$$

for critical boundaries d^S and $d^{\overline{S}}$ that satisfy the level requirement (1). Suppose $\alpha = 0.05$. Since, asymptotically, $T^S_{k^S} \sim N(0, k^S)$ under H^S and $T^{\overline{S}}_{k^{\overline{S}}} \sim N(0, k^{\overline{S}})$ under $H^{\overline{S}}$, we have $c^S = 1.6448\sqrt{k^S}$

and $c^{\overline{S}} = 1.6448\sqrt{k^{\overline{S}}}$. Again, setting $d^S/\sqrt{k^S} = d^{\overline{S}}/\sqrt{k^{\overline{S}}}$ we have, by the independence of $T^S_{k^S}$ and $T^{\overline{S}}_{k^{\overline{S}}}$, that $d^S = 1.9545\sqrt{k^S}$ and $d^{\overline{S}} = 1.9545\sqrt{k^{\overline{S}}}$. We have chosen this intersection hypothesis test for ease of exposition but other, more powerful, tests could just as well be adopted.

We next show how these tests may be extended to permit possible subgroup selection at an interim analysis. The type-1 error of the modified design will be protected if the conditional error rates of the tests of H^S and $H^{\{S,\overline{S}\}}$ in the modified design are bounded by the corresponding conditional error rates of the original design. This is the conditional rejection probability (CRP) principle of Müller and Schäfer [8]. To be specific, if it is decided to drop subgroup $\overline{S}$ at the interim analysis, and possibly increase the number of events for subgroup S from k^S to $\tilde{k}^S$, we must define a new final decision function ψ^S for testing H^S and $H^{\{S,\overline{S}\}}$ that preserves the conditional rejection probabilities

$$E_0(\psi^S|X) \leq E_0(\varphi^S|X) \tag{5}$$

and

$$E_0(\psi^S|X) \leq E_0(\varphi^{\{S,\overline{S}\}}|X) \tag{6}$$

where X is the set of all interim information on patients in S and $\overline{S}$ used for the decision on the design modification. It may be impossible to explicitly specify the vector X, which includes observed times-to-event as well as preliminary information correlated with time-to-event from patients who have not yet reached the endpoint. According to Irle and Schäfer [12] it is sufficient to condition on a random vector Y for which you can compute the conditional expectations $E_0(\varphi^S|Y), E_0(\varphi^{\{S,\overline{S}\}}|Y)$, and $E_0(\psi^S|Y)$, and which has the property that X is stochastically independent of the decisions functions $\varphi^S, \varphi^{\{S,\overline{S}\}}$ and ψ^S given Y. The CRP principle will then require the new decision function to satisfy

$$E_0(\psi^S|Y) \leq E_0(\varphi^S|Y) \tag{7}$$

and

$$E_0(\psi^S|Y) \leq E_0(\varphi^{\{S,\overline{S}\}}|Y) \tag{8}$$

The new decision function will be an indicator variable of the form

$$\psi^S = I(T^S_{\tilde{k}^S} \geq \tilde{c}^S)$$

where $\tilde{c}^S$ is defined implicitly by (7) and (8). By closed testing and the CRP principle, this decision function will maintain the test of H^S at level α.

If subgroup $\overline{S}$ is not dropped at the interim analysis, then of course ψ^S will not be computed and H^S will be rejected by a closed test in accordance with the decision functions $\varphi^S, \varphi^{\overline{S}}$ and $\varphi^{\{S,\overline{S}\}}$ given by (2) and (3). In the next section we specify what Y should be when the logrank test is applied to H^S and $H^{\{S,\overline{S}\}}$.

2.2 Application to Logrank Tests

At the calendar time of the interim analysis a subset $S' \subseteq S$ of patients has been randomized, of whom a subset of patients S'_{dead} has already died, while its complement S'_{risk} consists of patients in S' still at risk. We define subsets $\overline{S}', \overline{S}'_{\text{risk}}$ and $\overline{S}'_{\text{dead}}$ similarly. Our method permits the use of all available information in S' and $\overline{S}'$, including even the information about early outcomes like PFS or

tumor regression in S'_{risk} and $\overline{S}'_{\text{risk}}$, for the interim decision making. Without this flexibility one would be hard pressed to make an informed decision to drop $\overline{S}$ in settings where survival events are slow to arrive. It is necessary, however, to specify the following quantities prior to unblinding the interim data.

1. We must specify k^S and $k^{\overline{S}}$, the total number of events to be obtained from subgroups S and $\overline{S}$, respectively, at the time of the final analysis under the original design. In the phase 2 setting realistic choices for k^S and $k^{\overline{S}}$ are typically dictated by the sample size a sponsor is prepared to commit to the trial. Their impact on power and study duration is best evaluated by simulation under a range of scenarios for the alternative hypotheses, as shown in Section 4.

2. We must specify $k^{\overline{S}'}$, the contribution from the subset $\overline{S}'$ to $k^{\overline{S}}$. We shall see shortly that this specification is needed to ensure that the conditioning event Y in (7) and (8) will be properly defined even if recruitment to $\overline{S}$ is stopped after the interim analysis. Ideally $k^{\overline{S}'}$ should be so chosen that the arrival of the last of the $k^{\overline{S}'}$ events in $\overline{S}'$ is closely aligned in calendar time with the arrival of the last of the $k^{\overline{S}}$ events in $\overline{S}$. While this requirement is not essential for preservation of type-1 error, adherence to it will minimize unused events from $\overline{S}$ for the final analysis if this subgroup is retained after the interim analysis. We can use the blinded data available prior to the interim analysis to achieve this alignment as nearly as possible. The Appendix shows how.

Having fixed the values of $k^S, k^{\overline{S}}$ and $k^{\overline{S}'}$, we can compute the conditioning event Y needed to evaluate the conditional expectations (7) and (8) of the new decision function ψ^S. This conditioning event consists of two logrank statistics, one computed from subset S' and the other computed from subset $\overline{S}'$.

The Conditioning Event from S': The conditioning event is $T^{S'}_{k^S}$, the logrank statistic calculated from patients belonging to subset S' at the time of the arrival of the k^Sth event from subgroup S. This implies, of course, that the conditioning event is not observed at the time of the interim analysis but rather at the time of the pre-planned final analysis for H^S under the original design. Let $S'' = S \setminus S'$ denote the subset of patients in S that are enrolled after the interim analysis. Let $k^{S'}$ be the contribution from patients in subset S' to the k^S events required from subgroup S. Then $k^{S''} = k^S - k^{S'}$ is the contribution from patients in subset S'' to the k^S events required from subgroup S. We shall require this variable for the evaluation of the conditional rejection probabilities defined by (7) and (8).

The Conditioning Event from $\overline{S}'$: Let $\overline{S}'' = \overline{S} \setminus \overline{S}'$ denote the subset of patients in $\overline{S}$ that are enrolled after the interim analysis under the original design. We have pre-specified that the total number of events required from $\overline{S}$ is $k^{\overline{S}}$ with the first $k^{\overline{S}'}$ of these events to be contributed from subset $\overline{S}'$. Therefore the number of events to be contributed from subset $\overline{S}''$ must be

$$k^{\overline{S}''} = k^{\overline{S}} - k^{\overline{S}'} \ . \tag{9}$$

Note that, since $k^{\overline{S}}$ and $k^{\overline{S}'}$ are pre-specified, $k^{\overline{S}''}$ is well defined even if recruitment to subgroup $\overline{S}$ is stopped after the interim analysis. The conditioning event is $T^{\overline{S}'}_{(k^{\overline{S}'}, k^{\overline{S}''})}$, a logrank statistic computed from patients belonging to subset $\overline{S}' \subseteq \overline{S}$ as follows:

- If $\overline{S}$ is dropped at the interim analysis, $T^{\overline{S}'}_{(k^{\overline{S}'},k^{\overline{S}''})}$, is computed at the calendar time that $k^{\overline{S}'}$ events have arrived from $\overline{S}' \subseteq \overline{S}$
- If $\overline{S}$ is not dropped at the interim analysis, $T^{\overline{S}'}_{(k^{\overline{S}'},k^{\overline{S}''})}$, is computed at the **later** of the two calendar times when either $k^{\overline{S}'}$ events have arrived from $\overline{S}'$ or $k^{\overline{S}}$ events have arrived from $\overline{S}$, *with only the first* $k^{\overline{S}'}$ *events from* $\overline{S}'$ *contributing to the calculation of the statistic*

The statistic $T^{\overline{S}'}_{(k^{\overline{S}'},k^{\overline{S}''})}$ is well defined and can be computed whether or not recruitment to subgroup $\overline{S}$ is halted after the interim analysis. In either case it assumes the value $T^{\overline{S}'}_{k^{\overline{S}'}}$. The notation $T^{\overline{S}'}_{(k^{\overline{S}'},k^{\overline{S}''})}$ has been used because $k^{\overline{S}''}$ represents the additional events that would arrive from subset $\overline{S}''$ if recruitment to $\overline{S}$ continuew after the interim. We shall see that this variable, whose value can be evaluated by equation (9), is required for the evaluation of the conditional rejection probabilities defined by (7) and (8). It is also important to point out that if recruitment to $\overline{S}$ continues after the interim analysis, the test of $H^{\overline{S}}$ at the time of the final analysis may not be able to avail of all available events; for example, by the time that $k^{\overline{S}}$ events have arrived from subgroup $\overline{S}$, the pre-specified quota of $k^{\overline{S}'}$ events from subset $\overline{S}' \subseteq \overline{S}$ may be exceeded. The additional events in $\overline{S}'$ cannot be used. Hence the importance of trying to estimate in advance the value of $k^{\overline{S}'}$ such that, on average, the calendar time by which all $k^{\overline{S}'}$ events have arrived from subset $\overline{S}'$ will coincide with the calendar time by which all $k^{\overline{S}}$ events have arrived from subgroup $\overline{S}$.

The conditioning event Y is thus the pair of logrank statistics $(T^{S'}_{k^S}, T^{\overline{S}'}_{(k^{\overline{S}'},k^{\overline{S}''})})$. Suppose the trial is modified at the interim analysis, by discontinuing enrollment to subgroup $\overline{S}$, and possibly increasing the number of events in S from k^S to $\tilde{k}^S$ for the final analysis. In order to preserve the type-1 error the new critical value $\tilde{c}^S$ for the test of H^S must satisfy the CRP conditions (7) and (8). In terms of logrank statistics these conditions reduce to

$$P_0(T^S_{\tilde{k}^S} > \tilde{c}^S | T^{S'}_{\tilde{k}^S}) \leq \min\left\{ P_0(T^S_{k^S} > c^S | T^{S'}_{k^S}), P_0\left((T^S_{k^S}, T^{\overline{S}}_{k^{\overline{S}}}) \in R | T^{S'}_{k^S}, T^{\overline{S}'}_{(k^{\overline{S}'},k^{\overline{S}''})} \right) \right\} \tag{10}$$

where the subscript $P_0(.)$ denotes probability under the appropriate null hypothesis and $T^S_{\tilde{k}^S}$ is the logrank statistic computed at the time of the final analysis, when $\tilde{k}^S$ events have arrived from patients in subgroup S.

The evaluation of the conditional probabilities in (10) utilizes Theorem 1 of Irle and Schäfer [12] which, in this setting, implies that under the null hypothesis H^S

$$\begin{pmatrix} T^{S'}_{k^S} \\ T^S_{k^S} - T^{S'}_{k^S} \end{pmatrix}$$

is asymptotically

$$\mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k^{S'}/4 & 0 \\ 0 & k^{S''}/4 \end{pmatrix} \right] \tag{11}$$

and under the null hypothesis $H^{\overline{S}}$

$$\begin{pmatrix} T^{\overline{S}'}_{k^{\overline{S}}} \\ T^{\overline{S}}_{k^{\overline{S}}} - T^{\overline{S}'}_{(k^{\overline{S}'},k^{\overline{S}''})} \end{pmatrix}$$

is asymptotically

$$\mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k^{\overline{S}'}/4 & 0 \\ 0 & k^{\overline{S}''}/4 \end{pmatrix}\right] . \tag{12}$$

Equation (11) ensures stochastic independence of the $T^{S'}_{k^S}$ and $T^{S}_{k^S} - T^{S'}_{k^S}$ at information time k^S. Similarly equation (12) ensures stochastic independence of $T^{\overline{S}'}_{k^{\overline{S}}}$ and $T^{\overline{S}}_{k^{\overline{S}}} - T^{\overline{S}'}_{k^{\overline{S}}}$ at information time $k^{\overline{S}}$. It is this stochastic independence that permits interim decisions to be based on all available stage 1 data without influencing the final number of events to be realized from the stage 1 recruits. An alternative way to preserve stochastic independence is to actually pre-specify the final number of events to be realized from the stage 1 recruits prior to unblinding the interim data. This was the approach advocated by Jenkins et. al. [5].

Our procedure requires any modification of the pattern of patient recruitment to be independent of the unblinded interim data. Otherwise it would be possible to inflate the type-1 error. To see why, consider the following extreme example constructed by a referee for the simpler case of a single hypothesis H^S. Suppose the interim data can predict perfectly the value of $T^{S'}_{k^S}$. Then we can forecast at the time of the interim analysis itself whether the event $T^{S'}_{k^S} > c^S$ will occur or the event $T^{S'}_{k^S} \leq c^S$ will occur. If $T^{S'}_{k^S} > c^S$ is forecast, we can simply stop all further recruitment, set $\tilde{k}^S = k^S$, wait for additional events to arrive from the S' cohort until $k^{S'} = k^S = \tilde{k}^S$, and perform the final analysis with the test statistic $T^{S}_{\tilde{k}^S} = T^{S}_{k^S} = T^{S'}_{k^S}$. In this situation $k^{S''} = 0$. The critical limit $\tilde{c}^S$ will be obtained from the CRP condition

$$P_0(T^{S}_{\tilde{k}^S} > \tilde{c}^S | T^{S'}_{\tilde{k}^S}) \leq P_0(T^{S}_{k^S} > c^S | T^{S'}_{k^S}) \tag{13}$$

resulting in $\tilde{c}^S = c^S$, because this obviously fulfills the above inequality. At the end of the study we will reject H^S because, as forecast in advance, $T^{S'}_{k^S} > c^S$. On the other hand if $T^{S'}_{k^S} \leq c^S$ is forecast, we will continue recruiting patients from the S'' cohort and possibly increase the information time of the final analysis from k^S to $\tilde{k}^S$. The overall type-1 error of our adaptive strategy is therefore

$$\begin{aligned} &P_0(T^{S'}_{k^S} > c^S \text{ is forecast }) \times P_0(\text{rej } H^S | T^{S'}_{k^S} > c^S \text{ is forecast }) \quad + \\ &P_0(T^{S'}_{k^S} \leq c^S \text{ is forecast }) \times P_0(\text{rej } H^S | T^{S'}_{k^S} \leq c^S \text{ is forecast }) \quad = \\ &\qquad \alpha \times 1 + (1-\alpha) \times P_0(\text{rej } H^S | T^{S'}_{k^S} \leq c^S) > \alpha \ . \end{aligned}$$

This type of adaptation, however, violates the requirement that the CRP on the right hand side of (13) must be computed **under the initial (unmodified) design**. By setting $T^{S}_{k^S} = T^{S'}_{k^S}$, the right hand side of (13) produces a conditional rejection probability under an already modified design. (Modified by altering the recruitment process such that we will no longer be recruiting any patients from the S'' cohort.) To calculate the CRP under the initial design we should continue recruiting patients from the S'' cohort under the original pattern of patient recruitment until the pre-specified k^S events have arrived. Modifications of the original recruitment pattern are permitted as long as they are independent of the observed interim data. Since the right hand side of (13) is computed by the formula $1 - \Phi[(c^S - T^{S'}_{k^S}(\text{obs}))/k^{S''}]$, the only way to systematically bias the CRP is to control the individual components $k^{S'}$ and $k^{S''}$ of $k^S = k^{S'} + k^{S''}$. If, however, changes in the pattern of recruitment are made independently of the unblinded interim data, such control cannot be imposed on the CRP. In practice this independence requirement is automatically satisfied since decisions that affect the rate of recruitment are made by the trial sponsor and not by the indpendent data monitoring committee responsible for implementing the adaptive changes. The sponsor makes

decisions that affect the recruitment process such as opening or closing of clinical sites, limiting the percentage of patients enrolled from specific geographic regions, or modifying the inclusion/exclusion criteria **without access to unblinded interim data**. These decisions are influenced by operational considerations and are not affected by the unblinded interim results.

Finally, Irle and Schäfer [12] (Remark 1, page 344) implies that $T^S_{\tilde{k}^S} - T^{S'}_{\tilde{k}^S}$ and $T^{S''}_{\tilde{k}^S}$ are asymptotically equivalent. The implication is that events from cohort S' that arrive between information times k^S and $\tilde{k}^S$ do not contribute to the calculation of the revised critical cut-off $\tilde{c}$. This loss of information is one of the trade-offs from having a procedure that permits the use of all available data for interim decision making. An recent paper by Magirr, Jaki, Koenig and Posch [14] attempts to recover this lost information under a conservative assumption that guarantees type 1 error conrol.

3 Example: Non-Small Cell Lung Cancer Trial

In this section we present a numerical example of a Phase 2 trial to illustrate the methods discussed in Section 2. While the trial itself is hypothetical the design inputs are realistic, being based on a collaboration with oncologists at a major pharmaceutical company [15]. The numerical results in this example were obtained by simulating the trial once with the design inputs given below.

We consider a clinical trial comparing an experimental drug (Treatment E) to standard of care (Treatment C) for second line therapy in patients with metastatic non-small cell lung cancer (NSLC). The primary endpoint is progression free survival (PFS). The median PFS for the control arm is estimated from historical data to be 5 months. The experimental drug is targeted at an epidermal growth factor receptor (EGFR) that partitions the patient population into subgroup S (EGFR positive) and subgroup $\overline{S}$ (EGFR negative). There is increasing evidence that NSLC patients with EGFR mutations have higher response rates and longer survival [13]. Thus the prior belief is that the hazard ratio for treatment E versus treatment C is between 0.5 and 0.6 in subgroup S, whereas in subgroup $\overline{S}$ it is not expected to be any lower than 0.8.

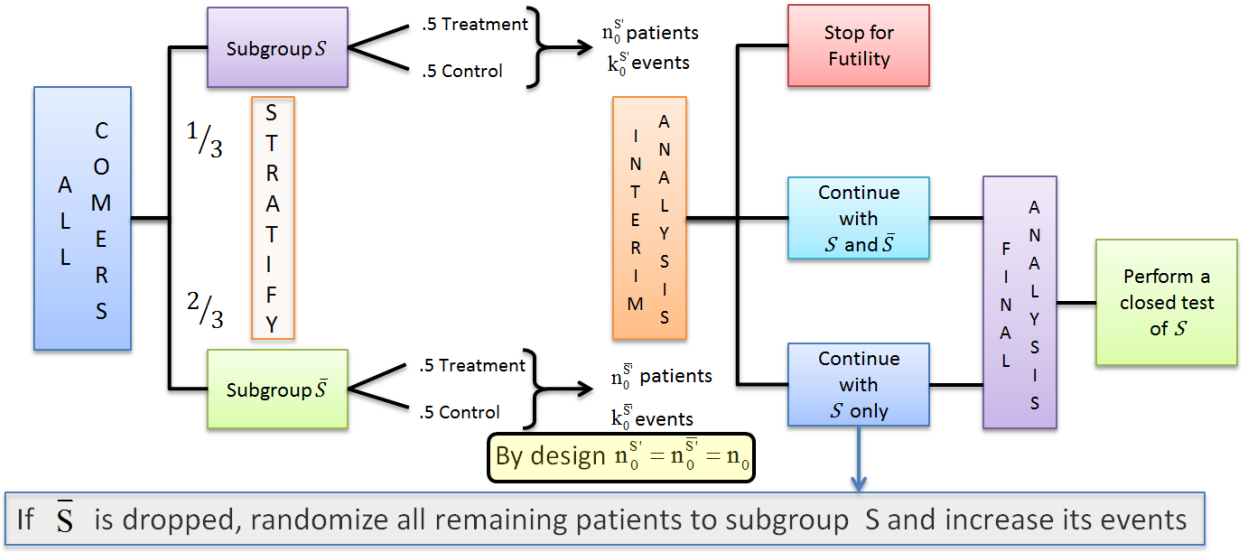
The sample size allocated to this trial is 160 patients, a standard commitment for a phase 2 trial in this disease class. Patients are expected to arrive at the rate of 15 per month with approximately two patients from subgroup $\overline{S}$ arriving for each patient from subgroup S. Randomization to treatment E or C will be stratified by EGFR status. The interim analysis will be performed after 80 patients have been recruited into the trial. However, despite the 2:1 ratio of patients in $\overline{S}$ relative to S, we will require an equal number, 40 patients, to be recruited from each subgroup for the interim analysis. While this is not essential for the statistical methodology, it is a way to ensure that there will be an adequate number of events available from subgroup S for the interim decision making. Operationally this means that after 40 patients have been recruited from subgroup $\overline{S}$, further enrollment will be restricted, temporarily, to subgroup S. Additional arrivals from subgroup $\overline{S}$ will be classified as screen failures. This will continue until the requisite number of 40 patients have arrived from subgroup S, at which time the interim analysis will be performed. After a thorough examination of all the data, unblinded by treatment, one of three decisions will be taken:

- Recruit the remaining 80 patients in equal numbers from each subgroup so that 40 patients are enrolled from subgroup S and 40 are enrolled from subgroup $\overline{S}$
- Drop subgroup $\overline{S}$ and recruit the remaining 80 patients from subgroup S only

- Terminate the trial for futility

Figure 1 is a schematic representation of the design. In this figure $n_0^{S'}$ and $n_0^{\overline{S}'}$ denote the number of

Figure 1: Schematic Representation of Adaptive Population Enrichment Design



patients recruited from subsets S' and $\overline{S}'$, respectively, by the time of the interim analysis. By design $n_0^{S'} = n_0^{\overline{S}'} = n_0$. In our example, $n_0 = 40$. The corresponding number of events arriving from subsets S' and $\overline{S}'$ by the time of the interim analysis are denoted by $k_0^{S'}$ and $k_0^{\overline{S}'}$, respectively.

Prior to unblinding the interim data it is necessary to specify the number of events $k^S, k^{\overline{S}}$ and $k^{\overline{S}'}$ to be obtained from subgroup S, subgroup $\overline{S}$ and subset $\overline{S}' \subseteq \overline{S}$, respectively. These choices will impact the power and study duration. If we choose large values for k^S and $k^{\overline{S}}$, the power will increase but the study duration will be prolonged. For this example we set $k^S = k^{\overline{S}} = 70$ as the initial design specification. If it is decided at the interim analysis to stop additional recruitment of patients from subgroup $\overline{S}$, we will recruit the remaining 80 patients from subgroup S, thereby ending up with a total of 120 patients belonging to subgroup S for the final analysis. In that case we will also increase the total number of events required for the final analysis, from $k^S = 70$ to $\tilde{k}^S = 110$. We shall obtain the operating characteristics of this design, including estimates of average study duration, by simulation in Section 4.

It remains only to specify a value for $k^{\overline{S}'}$, the commitment of events from patients belonging to subset $\overline{S}' \subseteq \overline{S}$, at the time of the final analysis under the original design. This quantity must be estimated prior to unblinding the interim data. Assuming exponential survival, the hazard rate for the control arm is $\ln(2)/5 = 0.139$ in subgroup $\overline{S}$. If we assume that the hazard ratio is $\text{HR}_{\overline{S}} = 0.8$ for subgroup $\overline{S}$, then the hazard rate for the experimental arm is $0.8 * 0.139 = 0.111$. With these estimates, and the further assumption that patients are recruited from subgroup $\overline{S}$ at the rate of 10/month, we can obtain a reasonable estimate for $k^{\overline{S}'}$. Calculations based on uniform enrollment and exponential survival show that if we set $k^{\overline{S}'} = 37$ then, on average, $k^{\overline{S}'}$ and $k^{\overline{S}}$ will be aligned in calendar time. That is, the arrival of the 37th event from subset $\overline{S}'$ will coincide on average with the arrival of the 70th event from subgroup $\overline{S}$. The details of this calculation are given in the Appendix. Although the

calculation depends on an assumption about the hazard ratio in subgroup $\overline{S}$, we can show that the estimate of $k^{\overline{S}'}$ is very robust to misspecification of the hazard ratio. For example, $k^{\overline{S}'}$ would not change even if we assumed that $\text{HR}_{\overline{S}} = 0.5$, and under $\text{HR}_{\overline{S}} = 1.0$ we would get $k^{\overline{S}'} = 38$.

At the time of the interim analysis, when 40 patients have been enrolled from each of the two subgroups, let k_0^S denote the number of events obtained from subgroup S and $k_0^{\overline{S}}$ denote the number of events obtained from subgroup $\overline{S}$. Let $T^S_{k_0^S}$ and $T^{\overline{S}}_{k_0^{\overline{S}}}$ be the corresponding logrank statistics generated by the interim data from subgroups S and $\overline{S}$, respectively. At this point all the data are inspected, including data on tumor regression available from patients for whom the PFS event has not yet arrived. Suppose that, on the basis of this exhaustive data inspection, it is decided to stop further recruitment of patients belonging to subgroup $\overline{S}$, enroll all the remaining 80 patients from subgroup S, and increase the number of events required for the final analysis from $k^S = 70$ to $\tilde{k}^S = 110$. In order to prevent inflation of type-1 error resulting from this adaptive population enrichment, it is necessary to change the critical cut-off for the final analysis from c^S to $\tilde{c}^S$. The new cut off can be computed by application of the Müller and Schäfer CRP principle. If the adaptive decisions had been made solely on the basis of the observed values of the logrank statistics $T^S_{k_0^S}$ and $T^{\overline{S}}_{k_0^{\overline{S}}}$, without inspecting any other features of the interim data, then the CRP principle would imply that

$$P_0(T^S_{\tilde{k}^S} > \tilde{c}^S | T^S_{k_0^S}) \leq \min\left\{ P_0(T^S_{k^S} > c^S | T^S_{k_0^S}), P_0\left((T^S_{k^S}, T^{\overline{S}}_{k^{\overline{S}}}) \in R | T^S_{k_0^S}, T^{\overline{S}}_{k_0^{\overline{S}}} \right) \right\} \tag{14}$$

and the conditional probabilities on the right hand side of (14) could be evaluated immediately, using the independent increments structure of the sequentially computed logrank statistics. Since, however, we wish to have the additional flexibility to examine all the available data, and not merely base the adaptive decision on $T^S_{k_0^S}$ and $T^{\overline{S}}_{k_0^{\overline{S}}}$, the value of $\tilde{c}^S$ must be obtained as the solution to (10). But the conditioning events on the right hand side of (10) are $T^{S'}_{k^S}$ and $T^{\overline{S}'}_{(k^{\overline{S}'}, k^{\overline{S}''})}$, statistics whose values are not yet observable. The evaluation of these statistics must therefore be postponed until the time of the planned final analyses when they can be observed, and the corresponding conditional probabilities can be evaluated.

After the interim analysis the trial only recruits patients from subgroup S. However, the arrival of events from patients in subgroup S, subset S' and subset $\overline{S}'$ continue to be monitored. We extract the following statistics from the emerging patient data:

1. When $k^S = 70$ events have arrived from subgroup S we record the value of the conditioning event, $T^{k^{S'}}_{k^S}$, and note the value of $k^{S'}$, the contribution to these 70 events from the patients in subset S'. For the current example we have observed $T^{S'}_{k^S} = 3.9654$ and $k^{S'} = 33$. We can now compute $P_0(T^S_{k^S} > c^S | T^{S'}_{k^S})$ since, by equation (11), $T^S_{k^S} - T^{S'}_{k^S}$ is independent of $T^{S'}_{k^S}$ and has variance $k^{S''} = 70 - 33 = 37$. Recall that $c^S = 1.6448\sqrt{70}$. Thus

$$P_0(T^S_{k^S} > c^S | T^{S'}_{k^S}) = 1 - \Phi(\frac{1.6448\sqrt{70} - 3.9654}{\sqrt{37}}) = 0.05365.$$

   Next, for computing the intersection hypothesis, we require $P_0(T^S_{k^S} > d^S | T^{S'}_{k^S})$. Recall that $d^S = 1.9545\sqrt{70}$. Thus

$$P_0(T^S_{k^S} > d^S | T^{S'}_{k^S}) = 1 - \Phi(\frac{1.9545\sqrt{70} - 3.9654}{\sqrt{37}}) = 0.02085.$$

2. When $k^{\overline{S}'} = 37$ events have arrived from subset $\overline{S}'$ we record the value of the conditioning event $T^{\overline{S}'}_{(k^{\overline{S}'}, k^{\overline{S}''})}$. We have observed $T^{\overline{S}'}_{(k^{\overline{S}'}, k^{\overline{S}''})} = 5.1934$. We can now compute $P_0(T^{\overline{S}}_{k^{\overline{S}}} > d^{\overline{S}} | T^{\overline{S}'}_{(k^{\overline{S}'}, k^{\overline{S}''})})$ since, by equation (12), $T^{\overline{S}}_{k^{\overline{S}}} - T^{\overline{S}'}_{(k^{\overline{S}'}, k^{\overline{S}''})}$ is independent of $T^{\overline{S}'}_{(k^{\overline{S}'}, k^{\overline{S}''})}$ and, by equation (9), has variance $k^{\overline{S}''} = 80 - 37 = 33$. Recall that $d^{\overline{S}} = 1.9545\sqrt{70}$. Thus

$$P_0(T^{\overline{S}}_{k^{\overline{S}}} > d^{\overline{S}} | T^{\overline{S}'}_{(k^{\overline{S}'}, k^{\overline{S}''})}) = 1 - \Phi(\frac{1.9545\sqrt{70} - 5.1934}{\sqrt{33}}) = 0.02604. \tag{15}$$

It follows that the conditional rejection probability of the intersection hypothesis is

$$\begin{aligned} &P_0\left((T^S_{k^S}, T^{\overline{S}}_{k^{\overline{S}}}) \in R | T^{S'}_{k^S}, T^{\overline{S}'}_{(k^{\overline{S}'}, k^{\overline{S}''})}\right) \\ &= P_0(T^S_{k^S} > d^S | T^{S'}_{k^S}) + P_0(T^{\overline{S}}_{k^{\overline{S}}} > d^{\overline{S}} | T^{\overline{S}'}_{(k^{\overline{S}'}, k^{\overline{S}''})}) - P_0(T^S_{k^S} > d^S | T^{S'}_{k^S}) P_0(T^{\overline{S}}_{k^{\overline{S}}} > d^{\overline{S}} | T^{\overline{S}'}_{(k^{\overline{S}'}, k^{\overline{S}''})}) \\ &= 0.02085 + 0.02604 - 0.02085 \times 0.02604 \\ &= 0.04635 \end{aligned}$$

3. When $\tilde{k}^S = 110$ events have arrived from subset S we perform the final analysis. We have now observed $\tilde{k}^{S'} = 39$, $T^{S'}_{\tilde{k}^S} = 5.8742$ and $T^S_{\tilde{k}^S} = 13.4888$. By equation (10) the critical cut-off for the final analysis, $\tilde{c}^S$, must satisfy

$$P_0(T^S_{\tilde{k}^S} > \tilde{c}^S | T^{S'}_{\tilde{k}^S} = 5.8742) = \min(0.05365, 0.04635) = 0.04635.$$

By equation (12), $T^S_{\tilde{k}^S} - T^{S'}_{\tilde{k}^S}$ is independent of $T^{S'}_{\tilde{k}^S}$ and has variance $\tilde{k}^{S''} = k^{\tilde{S}} - k^{S'} = 110 - 39 = 71$. It follows that $\tilde{c}^S$ must satisfy

$$1 - \Phi(\frac{\tilde{c}^S - 5.8742}{\sqrt{71}}) = 0.04635$$

whereupon $\tilde{c} = 20.0415$. Thus H^S will be rejected if $T^S_{\tilde{k}^S} > 20.0415$. Expressing this condition on the standardized Wald statistic scale, H^S is rejected if $Z^S_{\tilde{k}^S} > 20.0415/\sqrt{110} = 1.9109$, or equivalently, if the final p-value is less than 0.028. In fact we obtained $T^S_{\tilde{k}^S} = 13.4888$ which corresponds to 1.286 on the Wald statistic scale, or to a p-value of 0.0992. Hence H^S cannot be rejected.

4 Simulation Guided Design

Suppose we are still at the design phase of the NSCLC clinical trial described in Section 3. We assume that patients arrive at the rate of 15/month and the ratio of arrivals from subgroup $\overline{S}$ relative to arrivals from subgroup S is 2:1. The primary endpoint is PFS and the median PFS on the control arm is 5 months. The total sample size is 160 subjects, and we have decided to take an interim look after 40 subjects have been randomized to each of the two subgroups. We have pre-specified that $k^S = k^{\overline{S}} = 70$ and $k^{\overline{S}'} = 37$. We wish to evaluate the operating characteristics of this study by simulation. Let $\text{HR}_S = \exp(-\theta_S)$ and $\text{HR}_{\overline{S}} = \exp(-\theta_{\overline{S}})$ be the hazard ratios of the experimental arm

relative to the control arm in subgroups S and $\overline{S}$, respectively We shall be interested in investigating scenarios in which HR_S is small (0.5 to 0.6) and $\text{HR}_{\overline{S}}$ is large (0.8 to 1). This is the setting in which the biomarker would be regarded as predictive. We shall also be interested in investigating scenarios in which $\text{HR}_S = \text{HR}_{\overline{S}}$ for hazard ratios lower than 0.6. This is the setting in which the new therapy is effective independent of biomarker status. The performance of the design in these two settings will depend on decision rules for continuing with both subgroups, continuing with subgroup S only, or terminating the trial for futility after the interim analysis.

The interim decision rules can be quite complex. As already shown, the statistical methodology supports the utilization of all available data including the data in the censored observations. For illustrative purposes, however, we shall use simple decision rules, based on conditional power, that are easy to simulate. (More complex decision rules based on stochastic models of tumor regression and PFS are under development but outside the scope of this paper.) Let CP_S and $\text{CP}_{\overline{S}}$ denote the conditional power, under the original design, to reject H^S and $H^{\overline{S}}$. Then

1. If $\widehat{\text{HR}_S}$, the estimate of the hazard ratio for treatment versus control in subgroup S, is less than A, terminate the trial for futility
2. If $\text{CP}_S > B$ and $\text{CP}_{\overline{S}} < C$, stop further enrollment to subgroup $\overline{S}$ and enroll all remaining patients to subgroup S
3. Otherwise continue to the end of the trial with both subgroups

We have simulated the operating characteristics of this design for a range of decision parameters A, B and C. For each choice of these decision parameters we investigated scenarios with $\text{HR}_S = (0.5, 0.6)$ and $\text{HR}_{\overline{S}} = (0.5, 0.55, \ldots 1.0)$. Each scenarios is simulated 100,000 times. The procedure described in Section 2 is used to control the family wise error rate at one-sided level $\alpha = 0.05$. Figure 2 displays simulation results for power versus $\text{HR}_{\overline{S}}$, given $\text{HR}_S = 0.5$, under three decision rules for $(\text{CP}_S, \text{CP}_{\overline{S}})$ and no futility stopping. In Figure 2(a) subgroup $\overline{S}$ is dropped after the interim look if $\text{CP}_{\overline{S}} < 0.5$. In Figure 2(b), subgroup $\overline{S}$ is dropped after the interim look if $\text{CP}_{\overline{S}} < 0.5$ and $\text{CP}_S > 0.5$. In Figure 2(c), subgroup $\overline{S}$ is never dropped; one performs closed tests for both H^S and $H^{\overline{S}}$ at the end of the trial. Thus Figures 2(a) and 2(b) depict adaptive population enrichment designs with different decision rules for dropping subgroup S, whereas Figure 2(c) depicts a non-adaptive design in which both subgroups are retained to the end. Figure 2(d) compares all three designs by superimposing them onto a single plot.

The plots in Figure 2 provide valuable insights for the design of the follow-on phase 3 design. To be specific, we assume that the phase 2 outcome will be used to guide the design of the follow-on phase 3 trial in the following manner:

Win S, Win $\overline{S}$. This phase 2 outcome is depicted by the red lines in Figure 2 and represents the situation in which both H^S and $H^{\overline{S}}$ are rejected. In this case we would conclude that the new therapy is effective indpendent of biomarker status and would open the enrollment to both subgroups in the follow-on phase 3 trial.

Win S, Lose $\overline{S}$. This phase 2 outcome is depicted by the green lines in Figure 2 and represents the situation in which H^S is rejected but $H^{\overline{S}}$ is accepted. In this case we would conclude that the biomarker is predictive and would restrict the enrollment to subgroup S in the follow-on phase 3 trial.

Other For all other phase 2 outcomes the follow-on phase 3 trial will be postponed pending further investigation. The outcome where $H^{\overline{S}}$ is rejected but H^S is accepted implies that the biological basis for biomarker predictivity is in question. The outcome where both H^S and $H^{\overline{S}}$ are accepted implies that the drug may be ineffective. In either case further investigation is necessary.

Suppose that, in truth, $\text{HR}_S = 0.5$. Then the plots in Figures 2(a) – 2(d) provide the following insights.

- For values of $\text{HR}_{\overline{S}}$ close to 0.5, the probability of rejecting both H^S and $H^{\overline{S}}$, thereby concluding that new therapy is effective indpendent of biomarker status (red lines), exceeds the probability of rejecting H^S only, thereby concluding that the biomarker is predictive (green lines). This is desirable since, in this setting, one would want to include both subgroups in the follow-on phase 3 trial.

- However, even at $\text{HR}_S = \text{HR}_{\overline{S}} = 0.5$, the probability of concluding that the new therapy is effective indpendent of biomarker status is only about 58% under the decision rule of Figure 2(a), while the corresponding probability of concluding that the biomarker is predictive is almost 33%. This means that there is about a one in three chance of concluding, falsely, that the biomarker is predictive when in fact the new therapy is effective in both subgroups, thereby depriving future patients belonging to subgroup $\overline{S}$ of a beneficial treatment. This risk diminishes under the decision rules implemented in Figure 2(b) and Figure 2(c).

- For larger values of $\text{HR}_{\overline{S}}$, the reverse is true. Now the probability of concluding that the biomarker is predictive increases rapidly with $\text{HR}_{\overline{S}}$ and reaches 90% in 2(a), 85% in 2(b) and 80% in 2(c), while the probability of concluding that the new therapy is effective in both subgroups declines to below 5%. This underscores the desirability of having a strong a priori biological basis for assuming that the biomarker is predictive. If indeed that is the case, there is a high probability that the phase 2 trial will provide the necessary empirical evidence.

- Figures 2(a), 2(b) and 2(c) depict the performance of three different decision rules for dropping $\overline{S}$. In order to compare these three decision rules, Figure 2(d) superimposes all three figures onto a single plot.

  - It is seen that as we move from Figure 2(a) to Figure 2(c), the red and green lines shift in such a way that the probability of concluding that the biomarker is predictive declines while the probability of concluding that the new therapy is effective indpendent of biomarker status increases. The point where the red and green lines intersect shifts from $\text{HR}_{\overline{S}} = 0.565$ to $\text{HR}_{\overline{S}} = 0.675$.
  - The black lines in Figures 2(d) show that the probability of rejecting H^S is always at least 80% for all values of $\text{HR}_{\overline{S}}$ and all decision rules for dropping $\overline{S}$. This shows that the typical phase 2 sample size limit of 160 patients will suffice to show efficacy in subgroup S, if $\text{HR}_S = 0.5$, regardless of the value of $\text{HR}_{\overline{S}}$.
  - Furthermore at $\overline{S} = 1$ the probability to reject H^S declines by about 13% between the adaptive design depicted by the dashed black line (corresponding to the decision rule of Figure 2(a)) and the non-adaptive design depicted by the dotted black line (corresponding to the decision rule of Figure 2(c)).

Thus Figure 2 has shown that the choice of decision rule has a major impact on the operating characteristics of the design and will clearly depend on the investigator's prior assumptions about the predictivity of the biomarker.

The operating characteristics in Figure 2 assume that there is no early stopping for futility. The impact on power of imposing an early stopping decision rule is shown in Figure 3 where we compare the decision rules to drop $\overline{S}$ if $\mathrm{CP}_{\overline{S}} < 0.5$ and $\mathrm{CP}_S > 0.5$, with and without early futility stopping, simulating each scenario 100,000 times. In these simulations the criterion for early futility stopping is $\widehat{\mathrm{HR}}_S > 1.2$. Power losses between 3% and 6% are observed. It may be necessary, however, to impose futility stopping despite these power losses in order to avoid unnecessary prolongation of the trial if the treatments are ineffective. Table 2 displays the impact of futility stopping on study duration and FWER. In this table the trial was simulated 1,000,000 times under the global null hypothesis $\mathrm{HR}_S = \mathrm{HR}_{\overline{S}} = 1$. We utilized the decision rule $(\mathrm{CP}_{\overline{S}} < 0.5, \mathrm{CP}_S > 0.5)$ for dropping $\overline{S}$ at the interim analysis. The trial was stopped for futility if $\widehat{\mathrm{HR}}_S > 1.2$ at the interim analysis. A very slight

Table 2: Type-1 Error and Study Duration with and without Futility Stopping

Futility Criterion	FWER	Study Duration	Starting Seed
No Futility Stopping	0.051580	25.7 months	47513
Stop if $\widehat{\mathrm{HR}}_S > 1.2$	0.040358	19.34 months	48106
Based on 1,000,000 simulated trials			

inflation of FWER ($\alpha = 0.05158$) was obtained in the absence of a futility boundary. This was entirely due to the small trial size whereby the conditions for asymptotic convergence of the logrank statistic to the standard normal distribution were not fulfilled. The conservatism induced by the futility stopping rule resulted in a lowering of the FWER to $\alpha = 0.040358$. The presence of futility stopping reduced the average study duration from 25.7 months to 19.3 months.

5 Discussion

We have proposed a design that could be useful for testing new targeted agents in subgroups classified by biomarker status. We identified a simple but realistic setting in which a single biomarker partitions the patient population into two subgroups and there is a biological basis, possibly supported by limited empirical evidence, for supposing that the new agent targets just one of the two subgroups. This setting is specific to the oncology therapeutic area where there already exists a strong biological basis for assuming that the biomarker is predictive of treatment response. If this assumption can be validated at phase 2, the phase 3 trial can focus on a smaller, more homogeneous population, with a higher probability of success. In addition, success at phase 2 justifies the allocation of resources for the development of a validated companion diagnostic test to classify patients by biomarker status for inclusion in the phase 3 trial. The existence of a companion diagnostic is a regulatory requirement for phase 3 testing in oncology. However, since the proposed adaptive methodology ensures strong control of type-1 error it could, in principle, also be utilized in the confirmatory phase 3 setting. Even if the phase 2 trial has demonstrated that the biomarker is predictive for subgroup S, it might nevertheless be prudent to open phase 3 enrollment initially to both subgroups, thereby confirming its predictivity in a large well-controlled trial before eliminating the $\overline{S}$ subgroup from further consideration.

While the proposed approach will permit full utilization of all available interim data (complete as well as censored) for the enrichment decision, it is not true that all available data can be used at the time of the final analysis. Suppose subgroup $\overline{S}$ is retained after the interim analysis. In this case the final analysis for the test of $H^{\overline{S}}$ cannot be performed until all $k^{\overline{S}}$ events have arrived from subgroup $\overline{S}$ and all $k^{\overline{S}'}$ events have arrived from subset $\overline{S}' \subseteq \overline{S}$. Although, on average, the arrival of these two sets of events is aligned in calendar time (as per the Appendix), in fact one set of events will arrive before the other, resulting in some inefficiency due to unused events. For example, in Section 3 we have set $k^{\overline{S}'} = 37$ and $k^{\overline{S}} = 70$. Therefore if the 37th event from $\overline{S}'$ arrives ahead of the 70th event from $\overline{S}$ any additional events from $\overline{S}'$ that arrive while waiting for the full quota of 70 events from $\overline{S}$ cannot be used for the hypothesis test. Another important limitation is that while the value of k^S may be increased after the interim analysis, it cannot be decreased even if the interim results are highly promising for subgroup S. The only time that k^S may be decreased is if the trial is terminated for futility after the interim analysis.

The duration of the phase 2 trial depends crucially on the distribution of patients between subgroup S and subgroup $\overline{S}$. The smaller the proportion of patients in subgroup S, the longer the expected trial duration. This is because a sufficient number of events must arrive from subgroup S in order to make an informed interim decision, and also have sufficient power to reject H^S at the time of the final analysis. For the simulations in Section 4 we assumed that 33% of the patients belonged to subgroup S. With this assumption Table 2 showed that if patients enroll at the rate of 15/month, the average study duration is about 26 months without futility stopping and 19 months with futility stopping. If, instead, only 20% of patients belong to subgroup S, the average study duration will be prolonged to 33 months without futility stopping and 27 months with futility stopping. There would not, however, be any loss of power since the number events is not being reduced.

In Section 1 we argued that it is more appropriate to test $(H^S, H^{\overline{S}})$, rather than (H^F, H^S), when there is a strong a priori belief that the biomarker is predictive of treatment effect. It would be desirable to investigate this conjecture empirically. Accordingly we simulated the performance of our procedure (MDSI procedure) with that of Jenkins, Stone and Jennison [5] (JSJ procedure) for rejecting H^S and H^F. All hypothesis tests were designed for strong control of one-sided type-1 error at $\alpha = 0.05$. For the MDSI procedure we tested H^F only if both H^S and $H^{\overline{S}}$ were rejected, thus ensuring strong control at $\alpha = 0.05$ for all three null hypotheses. We observed, however, under a variety of scenarios that H^F was always rejected whenenever H^S and $H^{\overline{S}}$ were both rejected. Thus the test of H^F entailed no further loss of power once H^S and $H^{\overline{S}}$ had both been rejected. The comparison between the JSJ and MDSI adaptive procedures cannot, however, be made under identical starting conditions because, independent of the different adaptive approaches (conditional error rate versus combination p-values), they utilize different closed tests for the multiplicity adjustments. Nevertheless we attempted to make the comparison as fair as possible by:

- assuming that patients from S and $\overline{S}$ would enroll in equal proportions
- utilizing the same interim decision rules for dropping subgroup $\overline{S}$

Patients arrived at the rate of 5 per month. As in Sections 3 and 4, we pre-specified sample sizes of 80 patients per arm with $k^S = k^{\overline{S}} = 70$ and $k^{\overline{S}'} = 37$. An interim analysis was performed after enrolling 40 patients per arm, and subgroup $\overline{S}$ was dropped if $\text{CP}_{\overline{S}} < 0.5$ and $\text{CP}_S > 0.5$. (Under this decision rule the population is enriched about 50% of the time.) If $\overline{S}$ was dropped at the interim analysis, the sample size for subgroup S was increased to 120 patients and $k^S = 70$ was increased to $\tilde{k}^S = 120$. The

results for 100,000 simulated trials under $\text{HR}_S = 0.5$ and $\text{HR}_{\overline{S}}$ ranging between 0.5 and 1 are displayed in Figure 4. As anticipated, the MDSI method performs better in the predictive setting (values of $\text{HR}_{\overline{S}}$ closer to 1) while the JSJ method performs better when the new therapy is effective indpendent of biomarker status (values of $\text{HR}_{\overline{S}}$ closer to 0.5). For example, if $(\text{HR}_S = 0.5, \text{HR}_{\overline{S}} = 1)$, it is clearly undesirable to reject H^F since the new agent does not work in the $\overline{S}$ subgroup, which constitutes half the population. Yet the JSJ method rejects H^F about 30% of the time, compared to only 3% of the time for the MDSI method. On the other hand it is clearly desirable to reject H^S alone, which MDSI does with 83% probability compared to a corresponding probability of 56% for JSJ. In contrast, when both HR_S and $\text{HR}_{\overline{S}}$ equal 0.5 (the new therapy is effective indpendent of biomarker status), the JSJ method is preferred, since it has a higher probability for rejecting H^F (75% for JSJ; 61% for MDSI) and a lower probability for rejecting H^S alone (23% for JSJ; 28% for MDSI). Similar qualitative results were obtained over a wide range of interim decision rules thereby confirming that, in situations where the biomarker is believed to be predictive, it is dersirable to formulate the hypothesis testing framework in terms of $(H^S, H^{\overline{S}})$ rather than (H^F, H^S).

We focused the numerical example and the simulations on metastatic non-small cell lung cancer because design inputs like sample size, baseline hazard rate, clinically meaningful hazard ratios and rates of enrollment are fairly well established for this disease. Thus the results presented in Sections 3 and 4 are realistic even though the trial itself is hypothetical. Additionally, since metastatic NSCLC is characterized by short PFS and OFS, it is possible to observe a sufficient number of primary endpoints in order to make an informed decision on dropping or retaining the non-targeted subgroup at the time of the interim analysis. The methodology is applicable more generally, however, to all types of advanced cancers where sufficient data relevant to the decision to drop subgroup $\overline{S}$ are available at the interim analysis.

We have not exploited fully the freedom to use all relevant data for interim decision making, relying instead on relatively simple decision rules based on conditional power. However, the methodology of this paper has paved the way for utlizing more sophisticated interim decision rules through the use of mixture models that capture the relationship between tumor response and PFS or OS [16]. These models can then be combined with more sophisticated Bayesian decision rules such as those of Thall et. al. [16] [17] or Glimm and Di Scala [18]. This is currently a work in progress.

An extension of the statistical methodology to more than one interim analysis, possibly permitting early efficacy stopping based on group sequential boundaries, can be achieved by repeated application of the distribution theory of Irle and Schäfer [12] as implemented in equations (11) and (12). This may not be so useful for phase 2 trials, given their small sample sizes. In the phase 3 setting, however, the approach could play an important role. Another natural extension is to the problem of more than two subgroups. Multiple subgroups could arise either because more than one biomarker is being targeted or because the biomarker is not binary. For example the biomarker may be measuring the expression of a specific tumor antigen on a continuous scale. The classification of the biomarker levels into just two categories based on a single cut-point may not be realistic. It may also be necessary to find optimal values for the cut-points. All these problems require further investigation.

Acknowledgement:
The authors thank Dr Ashwin Gollerkeri for providing an oncologist's perspective on the right statistical problem to tackle. They thank Mr. Pranab Ghosh for expert computing help. They thank an anonymous referee for important insights that have improved this paper.

Appendix

This appendix shows how one may estimate the value of $k^{\overline{S}'}$ such that, on average, the arrival of the $k^{\overline{S}'}$th event from subset $\overline{S}'$ coincides in calendar time with the arrival of the $k^{\overline{S}}$th event from subgroup $\overline{S}$. We assume that patient enrollment occurs at a uniform rate of a patients per unit of time for subgroup S and $\overline{a}$ patients per unit of time for subgroup $\overline{S}$. Since S is the targeted subgroup we expect that $a < \overline{a}$. Thus, as explained in Section 3, there will be a pause in further enrollment from $\overline{S}'$ after n_0 patients have been enrolled from this subset, while we wait for n_0 patients to be enrolled from subset S'. Thus enrollment of patients from subset $\overline{S}'$ is halted at calendar time $t_1 = n_{0/\overline{a}}$ and the enrolled patients are followed up to calendar time $t_2 = n_0/a$ at which point the interim analysis is performed. We assume that the survival distributions are exponential with hazard rates $\overline{\lambda}_1$ and $\overline{\lambda}_2$ for the control and treatment arms, respectively. Then, the expected number of events from subset $\overline{S}'$ and treatment i at the time of the interim analysis is

$$E_{0i}^{\overline{S}'} = \int_{x=0}^{t_1} \overline{a} dx \int_{s=0}^{t_2-x} \overline{\lambda}_i e^{-\overline{\lambda}_i s} ds \ ,$$

and the expected number of patients still at risk of failure from subset $\overline{S}'$ and treatment i is $0.5n_0 - E_{0i}^{\overline{S}'}$. At any calendar time $l > t_2$, the expected number of events from subgroup $\overline{S}$ and treatment i will be

$$E_{li} = (0.5n_0 - E_{0i}^{\overline{S}'})(1 - e^{-\lambda_i(l-t_2)}) + \int_{x=0}^{Sa} a dx \int_{s=0}^{l-x} \lambda_i e^{-\lambda_i s} ds \ .$$

The first term is the contribution of new events between time t_2 and l from the $(0.5n_0 - E_{0i}^{\overline{S}'})$ patients of subset $\overline{S}'$ expected to still be at risk at time t_2. The second term is the contribution of events from newly enrolled patients after the interim analysis; that is, the patients enrolled from subset $\overline{S}''$. Thus $E_0^{\overline{S}'} = E_{01}^{\overline{S}'} + E_{02}^{\overline{S}'}$ is the expected total number of events that have arrived from subset $\overline{S}'$ by calendar time t_2, when the interim analysis occurs, and $E_l = E_{l1} + E_{l2}$ is the expected total number of additional events that arrive from subsets $\overline{S}'$ and $\overline{S}''$ between calendar time t_2 and calendar time l. Thus $E_0^{\overline{S}'} + E_l$ is the total number of events from subgroup $\overline{S}$ by calendar time l. Let l^* be the solution to

$$E_0^{\overline{S}'} + E_{l^*} = k^{\overline{S}}$$

Thus l^* is the calendar time at which the pre-specified $k^{\overline{S}}$ events from subgroup $\overline{S}$ are expected to arrive. If we set

$$k^{\overline{S}'} = E_0^{\overline{S}'} + (n_0 - E_0^{\overline{S}'})(1 - e^{\overline{\lambda} l^*})$$

then, on average, the $k^{\overline{S}'}$th event from subset $\overline{S}'$ and the $k^{\overline{S}}$th event from subgroup $\overline{S}$ will both be aligned at calendar time l^*.

References

[1] Kola I, Landis J (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews (Drug Discovery)* **3**, 711-715.

[2] Toll J, Punt CJ (2010). Monoclonal antibodies in the treatment of metastatic colorectal cancer: a review. *Clinical Therapeutics*. 32:3, 437-453.

[3] Bauer P, Posch M (2004). Letter to the editor. *Statistics in Medicine* **23**, 1333-1335.

[4] Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, Racine-Poon A (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* **28**, 1445-1463.

[5] Jenkins M, Stone A, Jennison C (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* **10**, 347-356.

[6] Friede T, Parsons N, Stallard N (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine*, on-line version.

[7] Bauer P, Köhne K (1994). Evaluation of experiments with adaptive interim analysis. *Biometrics* **50**, 1029-1041.

[8] Müller HH, Schäfer H (2004). A general statistical principle for changing a design any time during the course of a trial *Statistics in Medicine* **23**, 2497-2508.

[9] Spiessens B, Debois M (2011). Adjusted significance levels for subgroup analysis in clinical trials. *Contemporary Clinical Trials* **31**, 647-656.

[10] Buyse M, Michiels S, Sargent DJ, Grothey A, Matheson A, De Gramont A (2011). Integrating biomarkers in clinical trials. *Expert Rev. Mol. Diag.* (11), 2.

[11] Marcus R, Peritz E, Gabriel K (1974). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655-660.

[12] Irle S, Schäfer H (2012). Interim design modifications in time-to-event studies. *Journal of the American Statistical Association* **107**, 341-348.

[13] Bonomi, PD, Buckingham L, Coon J (2007). Selecting patients for treatment with epidermal growth factor tyrosine kinase inhibitors. *Clin Cancer Res* 2007; 13(15 Suppl).

[14] Magirr D, Jaki T, Koenig, F, Posch M (2014). Adaptive survival trials. *Stat. AP*. In press.

[15] Gollerkeri A (2013). Personal communication from Ashwin Gollerkeri, M.D., Executive Director, Cambridge, Massachusetts.

[16] Inoue LT, Thall PF, Berry DA (2002). Seamlessly expanding a randomized phase II trial to phase III. *Biometrics*. 58, 823-831

[17] Thall PF, Nguyen HQ, Wang X, Wolff JE (2012). A hybrid geometric phase II-III clinical trial design based on treatment failure time and toxicity. *Journal of Statistical Planning and Inference*. 142, 944-955.

[18] Di Scala L, Glimm E (2011). Time-to-event analysis with treatment arm selection at interim. *Statistics in Medicine* **30**, 3067-3081.

Figure 2: Power versus $HR_{\overline{S}}$ for Various Decision Rules, given $HR_S = 0.5$

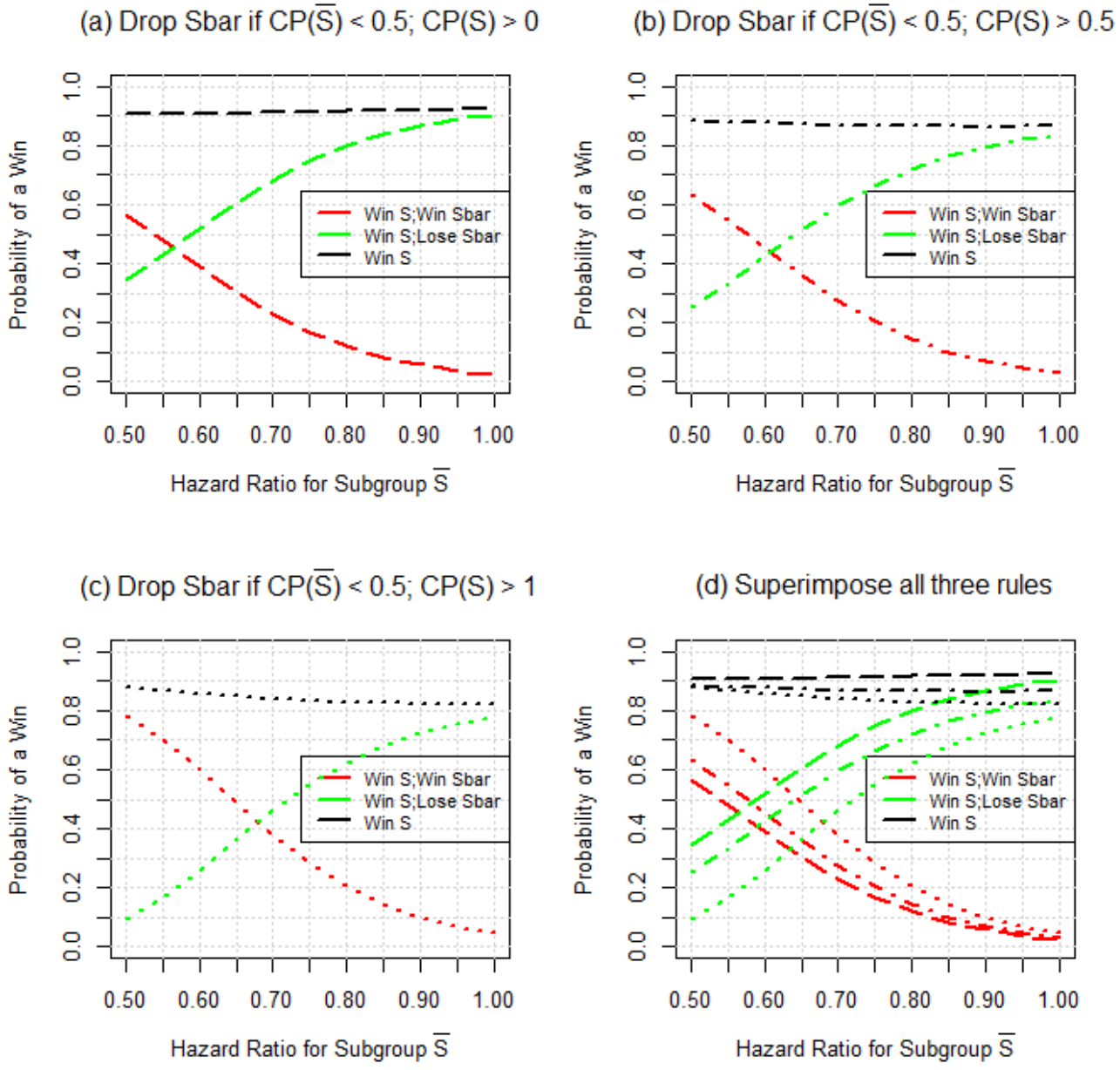


Figure 3: Power versus $\mathrm{HR}_{\overline{S}}$ under decision rule to drop $\overline{S}$ if $\mathrm{CP}_{\overline{S}} < 0.5$ and $\mathrm{CP}_{S}.0.5$ with and without futility stopping. (Stopping for futility occurs if $\widehat{\mathrm{HR}}_{S} > 1.2$)

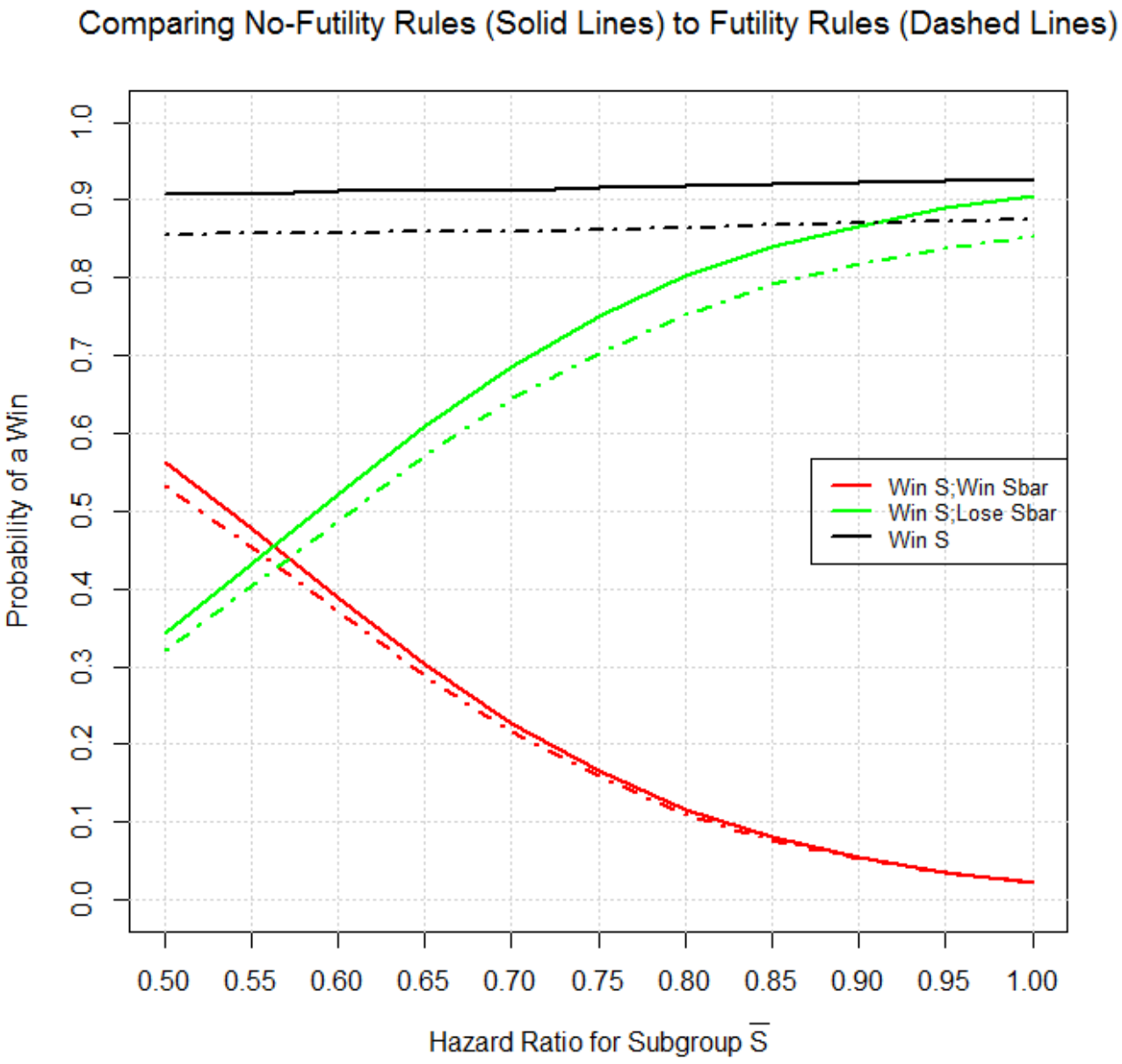


Figure 4: Comparing JSJ and MDSI for rejecting H^F and H^S when $\text{HR}_S = 0.5$ and $0.5 \leq \text{HR}_{\overline{S}} \leq 1$