Published in final edited form as:

J Biopharm Stat. 2009 ; 19(3): 494–508. doi:10.1080/10543400902802425.

Selecting Promising Treatments in Randomized Phase II Cancer Trials with an Active Control

Ying Kuen Cheung

Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, New York 10032, U.S.A.

Ying Kuen Cheung: yc632@columbia.edu

Abstract

The primary objective of phase II cancer trials is to evaluate the potential efficacy of a new regimen in terms of its antitumor activity in a given type of cancer. Due to advances in oncology therapeutics and heterogeneity in the patient population, such evaluation can be interpreted objectively only in the presence of a prospective control group of an active standard treatment. This paper deals with the design problem of phase II selection trials in which several experimental regimens are compared to an active control, with an objective to identify an experimental arm that is more effective than the control, or to declare futility if no such treatment exists. Conducting a multi-arm randomized selection trial is a useful strategy to prioritize experimental treatments for further testing when many candidates are available, but the sample size required in such a trial with an active control could raise feasibility concern. In this paper, we extend the sequential probability ratio test for normal observations to the multi-arm selection setting. The proposed methods, allowing frequent interim monitoring, offer high likelihood of early trial termination, and as such enhance enrollment feasibility. The termination and selection criteria have closed form solutions, and are easy to compute with respect to any given set of error constraints. The proposed methods are applied to design a selection trial in which combinations of sorafenib and erlotinib are compared to a control group in patients with non-small-cell lung cancer using a continuous endpoint of change in tumor size. The operating characteristics of the proposed methods are compared to that of a single-stage design via simulations: the sample size requirement is reduced substantially and is feasible at an early stage of drug development.

Keywords

Noninferiority test; Probability of correct selection; Sample size re-estimation; Sequential elimination; Sequential probability ratio test; Symmetric boundaries; Type I error

1. Introduction

Phase II cancer clinical trials are “proof of concept” studies in which a new regimen is evaluated for its potential efficacy in terms of antitumor activity in a given type of cancer. The objective has traditionally been achieved with a single-arm design using clinical response as the endpoint. The major drawback of this approach is that the interpretation of the results of a single-arm phase II trial relies on the specification of a reference response rate, which is often based on historical data from the literature or within an institution. As pointed out in several recent articles, assuming a known and constant reference is elusive mainly because of heterogeneity in the entire patient population; see for example Taylor et al. (2006) and Karrison et al. (2007). These authors reconsider and seek to revive the randomized phase II design in which the new regimen is compared prospectively against a standard treatment. Conducting a randomized trial with an active control no doubt requires an increased sample size from a

single-arm study, and the increase is at times deemed too great to be feasible at the early phase of development. However, Taylor et al. (2006) find that even with a moderate number of patients available, having a concurrent control group is advantageous in terms of the probability of correctly identifying an effective experimental agent. In addition, the sample size requirement can be alleviated in a number of ways due to advances in oncology therapeutics and statistical methodology. First, the Response Evaluation Criteria in Solid Tumors is under criticism for its lack of sensitivity in evaluating the new class of non-cytotoxic agents, and different outcome measurements are advocated in early-phase trials (Michaelis and Ratain, 2006). Consequently, Karrison et al. (2007) propose using the continuous endpoint of tumor size change at 8 weeks for a phase II trial of the combination of sorafenib and erlotinib in non-small-cell lung cancer (NSCLC) patients. This will retrieve the information loss due to dichotomization of the clinical response, and keep the required sample size to a feasible number. Second, the prospect of enrollment feasibility can be improved by adopting sequential designs with futility interim; see Jung (2007) for example. Finally, a relatively large type I error rate or significance level may be used in phase II. Specifically, most investigators are willing to accept one-sided test at level ranging from 10% to 30%.

Motivated by the above considerations, the paper addresses the problem of selecting a promising experimental agent in *multi-arm* randomized phase II trials with a prospective control. As pointed out in Karrison et al. (2007), oncology has had the advantage of having a large number of agents available for screening. While the large number of candidate agents indicates great potential in developing effective therapies, it poses an immediate challenge for investigators in prioritizing the agents for phase III testing. For this reason, in oncology and other disease areas, conducting a randomized selection trial among several experimental treatments has been endorsed as an efficient way to remove inferior agents from further consideration. For example, Simon et al. (1985) advocated randomized phase II treatment selection without control at a time when there was a clear benchmark for a good response rate for solid tumor (e.g. 25%). Cheung et al. (2006) also considered selection without an active control in a disease area where the patient population was homogeneous and historical control was relatively reliable. These assumptions may not stand in the current therapeutic approaches for cancer patients. This paper investigates the enrollment feasibility of treatment selection with a control group in the context of the NSCLC trial in Karrison et al. (2007) who consider two sorafenib/erlotinib combinations versus the control arm of erlotinib alone.

There are several proposals of two-stage designs that deal with the selection problem; see Thall et al. (1988), Schaid et al. (1990), and Bischoff and Miller (2005) for example. These designs typically choose an empirically superior experimental arm in a selection stage, and randomize additional patients to the chosen arm and the control group in a second stage for final comparison. In addition, these designs have provisions for early stopping after the selection stage if none of the experimental treatments seems promising. In this regard, a two-stage design lessens the sample size requirement when nothing works. On the other hand, as pointed out by Cheung (2008), the sample size advantage of a two-stage design disappears if there is an effective treatment among the experimental arms. In this paper, we study sequential selection boundaries that allow frequent interim looks and early stopping due to either selection or futility.

The paper is organized as follows. Section 2 formulates the hypotheses for the selection problem, presents a single-stage design, and describes two novel sequential selection designs. Calibration of the sequential designs will be further discussed in Section 3. In Section 4, we will apply the proposed methods to design the sorafenib/erlotinib combination trial in NSCLC patients, and illustrate the advantages offered by the sequential designs. Practical aspects of design implementation will be discussed in Section 5.

2. Methods

2.1. Problem formulation and single-stage design

Consider a set of treatments $\{0, 1, \dots, K\}$, where 0 represents the control group. Assume that the 8-week *decreases* in tumor size (on log scale) are normally distributed with respective means $\mu_0, \dots, \mu_K$ and a common variance σ^2 . For brevity, we focus on decision making under two hypotheses:

$$H_0: \mu_1 = \dots = \mu_K = \mu_0 \text{ and } H_1: \mu_1 = \dots = \mu_{K-1} = \mu_0, \mu_K = \mu_0 + \delta$$

where $\delta > 0$ is a prespecified clinically significant improvement. Note that the treatment labels are unknown. In other words, we do not know which treatment is associated with the arm K . Our design goal is to have high probability of correct selection (PCS) under the two hypotheses. Let P_0 and P_1 denote the respective PCS under H_0 and H_1 . When the null H_0 prevails, a type I error is committed when an experimental treatment is declared superior to the control; therefore, our goal is to keep P_0 , the selection probability for the control under H_0 , at or above $1 - \alpha$. Another design constraint is to maintain P_1 , the selection probability for treatment K under H_1 , at about $1 - \beta$.

To begin, consider a simple single-stage design that randomizes n subjects to each of the $K + 1$ arms. Formally, let $X_{ij} \sim N(\mu_i, \sigma^2)$ denote the outcome of the j th patient in arm i , and $\bar{X}_i = \bar{X}_i(n)$ be the observed mean response for arm i based on n subjects. A single-step test procedure (Hochberg and Tamhane, 1987) selects the control arm if

$$\frac{\bar{X}_i - \bar{X}_0}{\sigma \sqrt{\frac{2}{n}}} \leq c_\alpha \text{ for all } i=1, \dots, K,$$

and select arm k if

$$\bar{X}_k > \bar{X}_i \text{ for all } i \neq k \text{ and } \frac{\bar{X}_k - \bar{X}_0}{\sigma \sqrt{\frac{2}{n}}} > c_\alpha.$$

In data analysis, σ^2 in the above expressions is replaced with the pooled sample variance $\hat{\sigma}_n^2$. For design purposes, assuming σ^2 known and equal to σ_0^2 , we are to specify c_α and n so that $P_0 \geq 1 - \alpha$ and $P_1 \geq 1 - \beta$. Using Slepian's inequality (1962), we can show that $c_\alpha = z_{(1-\alpha)^{1/K}}$ satisfies the type I error constraint, and that

$$P_1 \geq \Phi^{K-1} \left(\frac{\sqrt{n}\delta}{\sqrt{2}\sigma_0} \right) \Phi \left(\frac{\sqrt{n}\delta}{\sqrt{2}\sigma_0} - c_\alpha \right),$$

where z_p is the p th percentile of standard normal, and Φ is the cdf. Consequently, the sample size (per arm) in a single-stage design may be determined as

$$\min \left\{ j \geq 1: \Phi^{K-1} \left(\frac{\sqrt{j}\delta}{\sqrt{2}\sigma_0} \right) \Phi \left(\frac{\sqrt{j}\delta}{\sqrt{2}\sigma_0} - z_{(1-\alpha)^{1/K}} \right) \geq 1 - \beta \right\}. \quad (2.1)$$

This design approach achieves the nominal type I error rate α (i.e. $P_0 \geq 1 - \alpha$) even if the assumed σ_0^2 is not the true variance. To protect P_1 against misspecification of the variance, one may re-estimate the sample size n' by applying (2.1) with σ_0^2 replaced by $\widehat{\sigma}_n^2$ after $(K+1)n$ subjects have been randomized and observed. Continue randomization with an additional $(K+1)(n'-n)$ subjects if $n' > n$; stop the study otherwise. Sample size re-estimation based on the sample variance will not inflate the type I error rate as long as it is done before the comparison of the sample means takes place.

The single-stage design, with or without sample re-estimation, will serve as the reference design for the proposed methods.

2.2. An extension of the sequential probability ratio test (SPRT)

Define a shifted outcome $Y_{ij} = X_{ij} + a_i$ for some specified a_i so that $Y_{ij} \sim N(\mu_i + a_i, \sigma^2)$, and let $\bar{Y}_i(n_i) = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}$. Consider the sequential probability ratio test (SPRT) statistics (Robbins and Siegmund, 1974) based on the shifted outcomes:

$$Z_{ki} = \frac{n_k n_i}{n_k + n_i} \left[\bar{Y}_k(n_k) - \bar{Y}_i(n_i) \right] \text{ for } i, k = 0, 1, \dots, K; i \neq k.$$

The multi-arm SPRT will terminate the trial and select arm k at an interim when $Z_{ki} \geq d$ for all $i \neq k$, where $d > 0$ is prespecified. For practical reasons, we consider truncated SPRT. That is, if the termination criterion is not reached when the total sample size reaches a prespecified maximum $N_{\max}$, the trial will be closed and the arm with the largest $\bar{Y}_i(n_i)$ will be selected. Subject allocation do not have to be exactly even among the $K+1$ arms at each interim, although the results of Robbins and Siegmund (1974) seem to suggest that the expected total number of observations will be smallest when $n_0 = n_1 = \dots = n_K$. We do not require the group sizes to be identical, because it is usually difficult to achieve in a real trial. However, we recommend using identical enrollment processes for all arms so that the group sizes are roughly equal. This can be achieved, for instance, by performing simple randomization to each new incoming subject.

THEOREM 1. Suppose the shifted outcome $Y_{ij} \sim N(\theta_i, \sigma^2)$ and $\theta_k \geq \theta_i$ for $i \neq k$. Then if the enrollment processes do not depend on θ_i 's, an open-ended SPRT (i.e., $N_{\max} = \infty$) will correctly select arm k with probability bounded below by

$$\left\{ \sum_{i=0}^K \exp \left[\frac{2d}{\sigma^2} (\theta_i - \theta_k) \right] \right\}^{-1}. \quad (2.2)$$

If we run the SPRT with $a_0 = \dots = a_K$ under H_0 , then $\theta_0 = \dots = \theta_K$ and the lower bound (2.2) is equal to $(1+K)^{-1}$. By symmetry, we can deduce that the probability of selecting arm k , or any arm, by the SPRT is equal to $(1+K)^{-1}$. Thus, the lower bound is exact in this case. Furthermore, we observe that the probability of selecting an experimental arm under H_0 (i.e., type I error rate) will equal $K/(K+1)$ which is apparently too large to be considered in practice.

Therefore, we need to choose the shifts $a_0, a_1, \dots, a_K$ differently so as to satisfy conventional error constraints.

COROLLARY 1. If the shifts are chosen such that $a_0 > a_i$ for $i = 1, \dots, K$, then the selection probability for the control under H_0 (i.e. P_0) is bounded below by

$$LB_0 = \left\{ \sum_{i=0}^K \exp \left[\frac{2d}{\sigma^2} (a_i - a_0) \right] \right\}^{-1}$$

which increases and converges to 1 as $d \rightarrow \infty$.

COROLLARY 2. If the shifts can be chosen such that $a_K + \delta > a_i$ for $i = 0, 1, \dots, K-1$, then the selection probability for treatment K under H_1 (i.e. P_1) is bounded below by

$$LB_1 = \left\{ \sum_{i=0}^{K-1} \exp \left[\frac{2d}{\sigma^2} (a_i - a_K - \delta) \right] + 1 \right\}^{-1}$$

which increases and converges to 1 as $d \rightarrow \infty$.

Corollary 1 and Corollary 2 guarantee we can always find a constant d for any given error constraints if the shifts are chosen to satisfy the conditions in the corollaries. Details and applications of these results are to be provided in Section 3. The two corollaries are easy consequences of Theorem 1. The proof of Theorem 1 can be undertaken in the same manner as in Levin and Robbins (1981) who extend the SPRT to the multi-arm selection problem with binomial data and equal sample sizes. The proof of Theorem 1 is available from the author upon request.

2.3. A sequential elimination procedure (ELIM₀)

In this subsection, we consider eliminating inferior arm(s) at an interim based on the SPRT stopping boundaries $\{Z_{ki} \geq d\}$ even though not all boundaries have been reached. Specifically, the elimination procedure (denoted as ELIM₀) proceeds as follows:

1. At an interim, close arm i if $Z_{ki} \geq d$ for some open arm k .
2. Terminate the trial under one of the following situations:
 - a. select the remaining arm when there is only one open arm;
 - b. select the arm with the largest $\bar{Y}_i(n_i)$ when the total sample size reaches $N_{\max}$;
 - c. select the arm with the largest $\bar{Y}_i(n_i)$ when the control arm is eliminated.

The motivation of sequential elimination is to remove arms that will unlikely emerge successful against the other arms given the information accrued thus far, and to allocate the remaining subjects among the more promising contenders. Therefore, a small probability may be expected for the event in which an eliminated arm would have eventually been the best arm according to the SPRT had it remained in the trial. We thus argue heuristically that the lower bound (2.2) and the results in Corollary 1 and Corollary 2 will serve as good approximation of P_0 and P_1 by ELIM₀.

Condition 2(c) in ELIM₀ is in place to avoid randomizing additional patients to two equally efficacious arms once the control is eliminated. This is a reasonable option for selection trials in which the objective is to identify a treatment that is comparatively more effective than the control. In situations where ranking the experimental regimens is of interest, the elimination procedure may be applied without condition 2(c). As explained in Cheung (2008), applying condition 2(c) does not inflate the type I error rate, but may decrease P_1 by selecting a suboptimal treatment under H_1 more often. However, the impact has been found practically negligible via extensive simulations.

The ideas of extending SPRT to multi-arm setting and sequential elimination are not new. Levin and Robbins (1981) propose and study the multi-arm SPRT and sequential elimination procedure for binomial data. In the sequel, Leu and Levin (1999a, 1999b) provide a rigorous proof for a lower bound formula analogous to (2.2) for sequential elimination. Cheung (2008) apply these previous results to treatment selection with an active control and propose an analogous version of ELIM₀ for binomial outcomes.

3. Design Calibration

3.1. Design parameters

In situations where the experimental regimens are exchangeable *a priori*, we may set $a_1 = \dots = a_K$. Then the lower bounds become $LB_0 = \{1 + K \exp[2d(a_1 - a_0)/\sigma^2]\}^{-1}$ and $LB_1 = \{\exp[2d(a_0 - a_1 - \delta)/\sigma^2] + (K - 1) \exp(-2d\delta/\sigma^2) + 1\}^{-1}$. We observe that LB_0 and LB_1 depend on a_0 and a_1 only through their difference, and therefore will set $a_1 = 0$ without loss of generality. As a result, we need $0 < a_0 < \delta$ in order to satisfy the conditions in Corollary 1 and Corollary 2, which then give

$$LB_0 = \left[1 + K \exp\left(-\frac{2da_0}{\sigma^2}\right)\right]^{-1} \text{ and } LB_1 = \left\{\exp\left[\frac{2d(a_0 - \delta)}{\sigma^2}\right] + (K - 1) \exp\left(-\frac{2d\delta}{\sigma^2}\right) + 1\right\}^{-1}. \quad (3.1)$$

This constraint on the choice of a_0 is intuitive: it needs to be positive so that the control will look favorable under H_0 with $\mu_0 + a_0 > \mu_i$, but smaller than δ so that a treatment with a clinically significant improvement remains superior in terms of the shifted mean under H_1 .

With the clinician-defined parameters δ , K and the error constraints α , β specified, the lower bounds LB_0 and LB_1 for the probability of correct selection depends on a_0 , d , and the true variance σ^2 . In addition, the lower bound formulae derived from (2.2) are based on the open-ended SPRT. While we expect the theoretical results will hold if the truncation $N_{\max}$ is sufficiently large, the choice of a sufficiently large $N_{\max}$ apparently depends also on the variability of the outcome, i.e., σ^2 . Assuming σ^2 known for design purposes, Section 3.2 will discuss the calibration of a_0 and d , and Section 3.3 determination of $N_{\max}$. We then propose using interim data to estimate σ^2 so that the design parameters (a_0 , d , $N_{\max}$) may be changed midcourse.

3.2. d -minimal design

For a given a_0 that is between 0 and δ , LB_0 and LB_1 are increasing functions of d . This is expected because a larger value of d invokes trial termination or treatment elimination when more information has been accrued, and hence the decision is less likely to be error-prone. For the same reason, SPRT and ELIM₀ with a smaller d are expected to conclude a trial with fewer patients than when a larger d is used. Therefore, we take the design approach in Cheung (2008) whereby the shift a_0 is chosen to be d -minimal: a shift a_0^* is d -minimal when it minimizes

the required termination constant d^* for given error constraints $LB_0 \geq 1 - \alpha$ and $LB_1 \geq 1 - \beta$ and a given set of clinician-defined parameters δ and K . In particular, we derive in the Appendix that

$$a_0^* = \delta \left[\frac{\log K - \text{logit}(\alpha)}{\log \left(\frac{K}{\alpha} - 1 \right) - \text{logit}(\beta)} \right]. \quad (3.2)$$

Correspondingly, the required termination constant is

$$d^* = \frac{\sigma^2}{2\delta} \left[\log \left(\frac{K}{\alpha} - 1 \right) - \text{logit}(\beta) \right]. \quad (3.3)$$

Since a_0^* does not depend on the true variance σ^2 , each observation in the control arm will be shifted by the same constant throughout the trial. Note that $0 < a_0^* < \delta$ when $\beta < 0.5$, or more precisely, $\beta < (K - \alpha) / \{K - \alpha + K(1 - \alpha)\}$. In other words, the d -minimal criterion can be applied if the research team sets its goal to identify the superior experimental arm under H_1 with a target probability greater than 0.5.

The termination constant d^* , on the other hand, does depend on σ^2 . When implementing the SPRT or $ELIM_0$, we could repeatedly estimate σ^2 with the unbiased pooled sample variance $\hat{\sigma}^2$ throughout the trial. Therefore, the termination criteria will be slightly different at each interim; empirically, we find that the estimate of d^* becomes quite stable when σ^2 is estimated with at least 30 degrees of freedom.

3.3. Sample size determination

Since the motivation of this work is to improve the enrollment feasibility on the single-stage design, we may initially set to truncate the sequential procedures at $N_{\max}$ so that $N_{\max}/(K + 1)$ is determined according to (2.1) with respect to an assumed σ_0^2 . This guarantees the sequential designs adopted will always enroll fewer patients than the single-stage design. As will be seen in the next section, the truncation $N_{\max}$ thus computed keeps the actual error rates at the target level, if the true variance is less than or equal to σ_0^2 . As in the single-stage design, we could re-estimate the truncation $N'_{\max}$ using (2.1) based on the pooled sample variance if the trial reaches $N_{\max}$ without reaching the termination criteria. That is, continue randomization with an additional $N'_{\max} - N_{\max}$ subjects if $N'_{\max} > N_{\max}$; stop the trial and select the arm with the largest observed mean if otherwise.

4. Application to the NSCLC Trial

4.1. Trial designs

We now apply the three designs to a selection trial with $K = 2$ combinations of sorafenib and erlotinib and a control arm of erlotinib alone. These regimens are considered for the second-line setting of NSCLC so that the subjects have been previously treated. As suggested in Karrison et al. (2007), we use the 8-week tumor shrinkage (on log scale) as the endpoint, and assume that it is normally distributed with mean -0.05 and standard deviation (σ_0) 0.346 in the subjects treated with the control. Based on two-thirds of the historical effect size of sorafenib in renal cell cancer patients, Karrison et al. (2007) posit a superior combination in the NSCLC population will have a mean tumor shrinkage of 0.13 on the log scale. Hence, $\delta = 0.13 - (-0.05)$

$= 0.18$. The error constraints are to keep the type I error rate at $\alpha = 0.1$ under H_0 and the selection probability for the superior combination $P_1 \geq 0.8$ under H_1 , i.e., $\beta = 0.2$.

Applying (2.1) gives $n = 46$ for the single-stage design with $c_{0.1} = 1.632$. Thus, the total sample size is 138. This calculation is very close to that in Karrison et al. (2007) who obtained 48 per group based on a Tukey adjustment for multiple comparison. This sample size appears to be larger than what is generally expected in phase II trials, but is arguably feasible. The goal of the sequential designs is to further reduce the enrollment figure.

According to (3.2) and (3.3), the d -minimal design parameters for SPRT and ELIM₀ are $a_0^* = 0.120$ and $d^* = 12.03\sigma^2$. Since σ^2 is generally unknown, we use $d^* = 12.03\hat{\sigma}^2$ as the termination cutoff, where $\hat{\sigma}^2$ is the pooled sample variance based on all data at an interim. To avoid unstable estimate of $\hat{\sigma}^2$, the termination criteria will be applied only when there are at least 10 patients in each arm. Thereafter, interim analysis will be conducted regularly after a small cohort of patients, until the termination criteria has been reached or when the total sample size has reached $N_{\max} = 138$.

4.2. Simulation study

The operating characteristics of all designs were examined via simulations under H_0 and H_1 . For the single-stage design, each of the three arms would receive 46 subjects in each simulated trial. For the SPRT and ELIM₀, each cohort of subjects would be allocated to the open arms by simple randomization between interims. We did not enforce exact balance among treatment arms within each cohort, because it would be difficult and impossible to accomplish in practice, and risk unblinding the treatment allocation when the cohort size was small. We considered cohort size ranging from 1 up to 10 in the current application. A small cohort size would represent a frequent monitoring schedule, but impose a great administrative burden.

In the simulated trials, patients' outcomes for arm i were generated as normal with mean μ_i and variance σ^2 , where $\mu_0 = \mu_1 = -0.05$ under both H_0 and H_1 , and $\mu_2 = -0.05$ under H_0 and $= 0.13$ under H_1 . To assess the effects of variance misspecification on the designs' performance, we considered various true σ in the simulation. For all three designs, we also considered the option of re-estimating the required sample size based on the pooled sample variance in the first 138 patients. Each method was run 20,000 times under each test scenario.

Table 1 gives the operating characteristics of SPRT and ELIM₀ with cohort size 6, and that of the single-stage design.

The SPRT and ELIM₀ truncated at $N_{\max} = 138$ achieve the target PCS under H_0 and H_1 when the true variance $\sigma^2 \leq \sigma_0^2$. This suggests that the single-stage sample size (2.1) provides an adequately large truncation for the sequential designs when σ_0 is a good estimate of the true variability. When $\sigma = 1.2\sigma_0 = 0.415$, the PCS of these truncated sequential designs fall below the target by about 5 percentage points. However, this problem disappears when we re-estimate sample size $N'_{\max}$ after randomizing 138 subjects. As indicated by the theoretical results, the single-stage design maintains the target PCS under H_0 (i.e., P_0) regardless of the true variance; however, when $\sigma = 0.415$, it suffers great loss of PCS under H_1 , with P_1 being 13 percentage points below the target. This problem, again, is resolved by re-estimating the sample size after randomizing 138 subjects.

The sequential designs have a clear advantage over the single-stage design in terms of the required sample size, with ELIM₀ being superior to SPRT in general. First, the truncated sequential designs (without sample size re-estimation) will never enroll more patients than the single-stage design. Second, the shift in the sample size distribution of the sequential designs

is remarkable. Take ELIM_0 under the scenario when $\sigma = \sigma_0 = 0.346$ as illustration. The median sample size reduces by 43% from 138 to 78; the upper quartile is smaller than 138 indicating a high likelihood that the trial will be terminated early; the lower quartile is equal to 54 (39% of the maximum 138) highlighting the occasionally substantial reduction due to a frequent interim schedule. Finally, the shift in the sample size distribution adjusts properly in accord with the true variability: when $\sigma = 0.9\sigma_0 = 0.311$, the median number of subjects required by ELIM_0 is 66, a 52% reduction from 138. In contrast, a single-stage design does not have this kind of adaptability, and will enroll more patients than necessary if σ is smaller than the assumed σ_0 .

Performing sample size re-estimation increases sample size. However, it affects only the extreme upper tail of the sample size distribution of the sequential designs when the original truncation $N_{\max}$ is adequately large, as we see that the interquartile ranges are unaffected for SPRT and ELIM_0 when $\sigma \leq \sigma_0$. When the true variance is larger than σ_0^2 and the original truncation $N_{\max}$ appears to be inadequate, sample size re-estimation causes an upward shift in the final sample size for the sequential designs. This is expected and indeed desirable: sample size re-estimation is intended to guarantee adequate enrollment for a certain variability in the outcomes. This being said, the sequential designs are quite efficient at utilizing information in the first batch of 138 subjects, as the median sample sizes remain unchanged for the designs with sample size re-estimation. In contrast, the median sample size required by the single-stage analysis consequent to sample size re-estimation increases to 198, a 43% increase from 138.

Figure 1 plots the operating characteristics of the sequential designs against the cohort size between interims for scenarios where the true $\sigma = 0.346$. The top panel gives the PCS of SPRT and ELIM_0 , with and without sample size re-estimation. All four designs achieve the nominal type I error rate by having P_0 staying about 0.90 for all cohort sizes, although the SPRT truncated at $N_{\max} = 138$ seems to yield P_0 slightly below the target. Likewise, the PCS under H_1 (i.e. P_1) is close to 0.8 in all four designs.

Sample size re-estimation improves the PCS of the SPRT and ELIM_0 . It is interesting to note the pattern that, with a fixed truncation, ELIM_0 has higher PCS than SPRT, and that the trend is reverse for the designs with sample size re-estimation. This illustrates the advantage of sequential elimination when resources (patients) are limited to an absolute maximum $N_{\max}$: a clearly inferior arm can be removed from consideration early on so that the remaining resources can be allocated to the promising arms for more precise comparison. On the other hand, when more resources can be made available via sample size re-estimation, the SPRT emerges superior to ELIM_0 by avoiding early, potentially erroneous, elimination decisions in the trial. The price for such caution is the larger sample size required by the SPRT than that by ELIM_0 (cf. Table 1).

The bottom panel of Figure 1 gives the sample size distribution of ELIM_0 with sample size re-estimation for various cohort sizes. We observe that the sample size distribution gradually shifts upwards as cohort size becomes large. Such an increase seems to be the price for administrative ease; however, the effects of cohort size on the designs' operating characteristics are quite small in this range. We have run simulations (not reported here) using random cohort sizes (uniformly from 1 to 10) within a trial, and obtained very similar operating characteristics. This is a desirable feature that enables investigators to choose a cohort size that is administratively convenient; investigators may even opt to set a monitoring schedule in terms of calendar time, instead of scheduling an interim after every fixed number of new enrollments. Above all, with median sample size below 80 for all cohort sizes, ELIM_0 imposes a feasible enrollment requirement at an early stage of drug development.

Figure 2 displays the results for the sequential designs for scenarios with true $\sigma = 0.415$. The figure shows similar patterns found in Figure 1, although the advantage of sample size re-estimation is clear in these scenarios. Re-estimating the sample size may lead to a number much larger $N'_{\max}$ than 138 in these scenarios, especially under H_1 where the variability in sample size seems to be large. However, this increase in the final sample size should be noted in light of the fact that a single-stage design with an assumed $\sigma_0 = 0.415$ requires $n = 67$ subjects per arm, thus a total of 201 subjects.

5. Discussion

This article proposes sequential designs for treatment selection with a prospective control group using a normal endpoint. We have illustrated through simulation the substantial gain in sample size by using sequential methods. In particular, we recommend using a truncated sequential elimination procedure, ELIM_0 , augmented with sample size re-estimation based on the variance estimate if the termination criteria cannot be reached within the prespecified $N_{\max}$ subjects.

The proposed methods are extremely easy to calibrate for practical use: We can specify the mean shift a_0^* and termination constant d^* using formulae (3.2) and (3.3) respectively, and a truncation $N_{\max}$ using (2.1).

Such simplicity arises from the lower bound (2.2) given in Theorem 1, which can be applied to construct PCS lower bounds under a variety of hypothesis configurations. For example, if we are comparing several new cytostatic agents with a standard cytotoxic agent on the basis of tumor shrinkage, it may be appropriate to set up noninferiority test with the following hypotheses:

$$H_0: \mu_1 = \dots = \mu_k = \mu_0 - \delta \text{ and } H_1: \mu_1 = \dots = \mu_{k-1} = \mu_0 - \delta, \mu_k = \mu_0$$

where $\delta > 0$ is the noninferiority margin. Analogous results to Corollary 1 and Corollary 2 can then be derived as a consequence of Theorem 1, and applied to calibrate a_0 and d for noninferiority tests.

The proposed methods are easy to implement. First, we have demonstrated that the prespecified level of PCS can be achieved with a flexible monitoring schedule in terms of cohort size between interims. Second, the termination criteria for the SPRT and ELIM_0 are characterized by symmetry decision rules with respect to the shifted outcomes Y_{ij} 's, and the interim pooled sample variance can be reported without providing information about the primary comparison. Therefore, it is possible to make interim decisions without unblinding the treatment labels and avoid selection bias. In contrast, in order for a two-stage design to continue to a second stage, it is known that the chosen experimental arm has to be empirically superior to the control.

There are practical considerations when implementing the proposed designs in a cancer trial, namely, patients may die prior to the 8-week tumor size evaluation. To get round the missing outcomes due to death, our recommendation is to separately monitor death and the primary endpoint (e.g. tumor shrinkage) in patients who are alive at the followup. This approach is advantageous in that the primary comparison will evaluate the regimens purely by their anticancer mechanism via the choice of the endpoint. This being said, it is imperative to compare survival experience among the treatment arms: if an experimental regimen is selected, we may informally compare the survival curve of the selected regimen to that of the control group, and examine whether there is a drastic difference between the two groups. This informal

approach seems to be adequate at the early stage of drug development. A second approach to the missing outcome problem is to assign the worst possible outcome for death and use a rank-based test for statistical inference, as suggested in Karrison et al. (2007). We are currently investigating analogous rank-based sequential elimination procedures. While this approach apparently solves the missing data problem, the interpretation of such a composite outcome is not always clear. Finally, in trials where a non-trivial number of subjects may die within a few months of diagnosis (e.g., patients with advanced pancreatic cancer), it may be appropriate to consider death or short-term progression-free survival as the primary endpoint and monitor the trial based on binary outcomes; see Cheung (2008). This approach is recommended only when there are reasons to believe the experimental treatments will provide a substantial improvement in survival; otherwise, the sample size requirement will prove to be prohibitive for a phase II trial.

Acknowledgments

This work was supported by NIH grant R01 NS055809-01.

Appendix

This Appendix derives a_0^* (Equation 3.2) and d^* (Equation 3.3) for the d -minimal design. Define $d_0(a_0) = \min\{d : LB_0 \geq 1 - \alpha\}$ and $d_1(a_0) = \min\{d : LB_1 \geq 1 - \beta\}$ as functions of a_0 , where LB_0 and LB_1 are given in Equation (3.1). Thus, $d^*(a_0) = \max\{d_0(a_0), d_1(a_0)\}$ is the smallest termination constant that satisfies the error constraints for a given a_0 .

By continuity and monotonicity of LB_0 and LB_1 in d , we can replace the inequality sign in the definitions of $d_0(a_0)$ and $d_1(a_0)$ with an equal sign. Then it is easy to show that $d_0(a_0)$ is decreasing in a_0 , whereas $d_1(a_0)$ is increasing in a_0 ; and that the functions $d_0(a_0)$ and $d_1(a_0)$ intersect at some a_0^* between 0 and δ if $\beta < 0.5$ (or more precisely, $\beta < (K - \alpha)/(K - \alpha + K(1 - \alpha))$). Consequently, we can deduce that $d^*(a_0)$ is minimized at the intersection a_0^* . In other words, $d^*(a_0^*) = d_0(a_0^*) = d_1(a_0^*)$.

Now solving

$$\left[1 + K \exp \left(-\frac{2d_0(a_0)a_0}{\sigma^2} \right) \right]^{-1} = 1 - \alpha$$

gives

$$d_0(a_0) = \frac{\sigma^2(\log K - \text{logit}(\alpha))}{2a_0}.$$

Then by definition of $d_1(a_0)$, we have

$$\left\{ \exp \left[\frac{2d_1(a_0)(a_0 - \delta)}{\sigma^2} \right] + (K - 1) \exp \left[-\frac{2d_1(a_0)\delta}{\sigma^2} \right] + 1 \right\}^{-1} = 1 - \beta$$

Since $d_0(a_0^*)=d_1(a_0^*)$, we can solve for a_0^* by solving (A.1) and (A.2) with a_0 replaced by a_0^* in the equations. The desired results follow after some algebraic manipulation.

References

- Bischoff W, Miller F. Adaptive two-stage test procedures to find the best treatment in clinical trials. *Biometrika* 2005;92:197–212.
- Cheung YK. Simple sequential boundaries for treatment selection in multi-armed randomized clinical trials with a control. *Biometrics*. 2008 in press.
- Cheung YK, Gordon PH, Levin B. Selecting promising ALS therapies in clinical trials. *Neurology* 2006;67:1748–1751. [PubMed: 17130405]
- Hochberg, Y.; Tamhane, AC. Multiple Comparison Procedures. Wiley; 1987.
- Jung SH. Randomized phase II trials with a prospective control. *Statistics in Medicine*. 2007 in press.
- Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non-small-cell lung cancer. *J Natl Cancer Inst* 2007;99:1455–1461. [PubMed: 17895472]
- Leu CS, Levin B. On the probability of correct selection in the Levin-Robbins sequential elimination procedure. *Statistica Sinica* 1999a;9:879–891.
- Leu CS, Levin B. Proof of a lower bound formula for the expected reward in the Levin-Robbins sequential elimination procedure. *Sequential Analysis* 1999b;18:81–105.
- Levin B, Robbins H. Selecting the highest probability in binomial or multinomial trials. *Proceedings of the National Academy of Sciences of the United States of America* 1981;78:4663–4666. [PubMed: 16593069]
- Michaelis LC, Ratain MJ. Measuring response in a post-RECIST world: from black and white to shades of grey. *Nature Reviews Cancer* 2006;6:409–414.
- Robbins H, Siegmund DO. Sequential tests involving two populations. *Journal of the American Statistical Association* 1974;69:132–139.
- Schaid DJ, Wieand S, Therneau TM. Optimal two-stage screening designs for survival comparisons. *Biometrika* 1990;77:507–513.
- Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. *Cancer Treatment Reports* 1985;69:1375–1381. [PubMed: 4075313]
- Slepian D. The one-sided barrier problem for normal noise. *Bell System Tech J* 1962;41:463–501.
- Taylor JMG, Braun TM, Li Z. Comparing an experimental agent to a standard agent: relative merits of a one-arm or randomized two-arm phase II design. *Clinical Trial* 2006;3:335–348.
- Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. *Biometrika* 1988;75:303–310.

Figure 1.

Effects of cohort size on the operating characteristics of the sequential designs when $\sigma = 0.346$.

The designs are specified with $\alpha_0^* = 0.120$ and $d^* = 12\sigma^2$. The top panel shows the PCS of SPRT (dashed line) and ELIM₀ (solid line) with a fixed $N_{\max} = 138$ (indicated by “F”) and with sample size re-estimation (indicated by “R”). The bottom panel shows the sample size distribution of ELIM₀ with sample size re-estimation. Each box spans the interquartile range, the darker line inside is the mean, and the lighter median; the dashed line in each figure indicates the median sample size of the single-stage design.

Figure 2.

Effects of cohort size on the operating characteristics of the sequential designs when $\sigma = 0.415$.

The designs are specified with $\alpha_0^* = 0.120$ and $d^* = 12\sigma^2$. The top panel shows the PCS of SPRT (dashed line) and ELIM₀ (solid line) with a fixed $N_{\max} = 138$ (indicated by “F”) and with sample size re-estimation (indicated by “R”). The bottom panel shows the sample size distribution of ELIM₀ with sample size re-estimation. Each box spans the interquartile range, the darker line inside is the mean, and the lighter median; the dashed line in each figure indicate the median sample size of the single-stage design.

Operating characteristics of the single-stage design (SS), the d -minimal SPRT and ELIM₀. The designs are calibrated with respect to $\delta = 0.18$ and $\sigma_0 = 0.346$ for a selection trial with $K = 2$, $\alpha = 0.1$, and $\beta = 0.2$. For SPRT and ELIM₀, $a_0^* = 0.120$, $d^* = 12\sigma^2$, and $N_{\max} = 138$.

Table 1

True σ	Design	Re-estimate		Under null [†]		Under alternative [‡]	
		$N_{\max}$	P_0	Median N (IQR)	P_1	Median N (IQR)	
0.311	SS	No	0.91	138 (138, 138)	0.88	138 (138, 138)	
		Yes	0.91	138 (138, 138)	0.88	138 (138, 138)	
	SPRT	No	0.92	72 (48, 108)	0.80	78 (54, 126)	
		Yes	0.92	72 (48, 108)	0.80	78 (54, 126)	
	ELIM ₀	No	0.92	66 (48, 90)	0.81	66 (48, 108)	
		Yes	0.92	66 (48, 90)	0.81	66 (48, 108)	
0.346	SS	No	0.91	138 (138, 138)	0.81	138 (138, 138)	
		Yes	0.91	138 (138, 150)	0.83	138 (138, 150)	
	SPRT	No	0.90	84 (54, 132)	0.79	90 (60, 138)	
		Yes	0.92	84 (54, 132)	0.81	90 (60, 138)	
	ELIM ₀	No	0.91	78 (54, 108)	0.80	78 (54, 120)	
		Yes	0.91	78 (54, 108)	0.80	78 (54, 120)	
0.415	SS	No	0.91	138 (138, 138)	0.67	138 (138, 138)	
		Yes	0.91	198 (183, 216)	0.81	198 (183, 216)	
	SPRT	No	0.85	120 (78, 138)	0.75	126 (78, 138)	
		Yes	0.94	120 (78, 180)	0.83	126 (78, 210)	
	ELIM ₀	No	0.86	102 (66, 138)	0.75	102 (66, 138)	
		Yes	0.90	102 (66, 144)	0.79	102 (66, 162)	

[†] Under the null, $\mu_0 = \mu_1 = \mu_2 = -0.05$.

[‡] Under the alternative, $\mu_0 = \mu_1 = -0.05$ and $\mu_2 = 0.13$.