A Design Adapting Between Clinical Trial Phases

Keaven M. Anderson [*] Xiaoyun (Nicole) Li [†]

Abstract

At the time a study is designed there is imperfect information and it may not be completely clear that a drug is worth the investment required for a large, pivotal trial required for regulatory approval. An adaptive phase II/III study with an early futility analysis is sometimes chosen in order to shorten the development timeline and limit unneccessary patient exposure and investment. However, in a trial with a time-to-event endpoint, there may be many patients with relatively little follow-up if a trial is stopped early. Here we consider transformation of a trial from Phase III to a smaller Phase II at an interim analysis as an additional option. The intent is to allow collection of follow-up on patients already enrolled in order to get the best data possible for future decisions surrounding development of the drug. The objectives and required observed treatment effect and Type I error for a positive Phase II adaptation may be different from the Phase III design. Compared to stopping a Phase III trial for futility, transforming to a Phase II trial with an intermediate treatment effect may be more cost-effective. The methods used here are related to group sequential design theory. The critical criteria for and timing of phase selection at interim analysis are discussed.

Key Words: Group sequential design, adaptive design, time-to-event endpoint, gsDesign, interim analysis

1. Introduction

We begin with an example which we try to keep relatively simple. Consider a drug that is hypothesized to extend the lives of patients with lung cancer. Early data is available that suggests a high response rate to therapy compared to historical results with standard therapy. While response to therapy is considered a "proof-of-concept" that the drug is active for lung cancer patients, there is doubt as to how long patients' lives might be extended. We might start with an assumption of a proportional hazards model where the true hazard ratio of death among patients on the new treatment compared to standard is 0.7; in this case, this is equivalent to an increase in median survival from 6 to 8.6 months. For 85% power and 2.5% (one-sided) Type I error, a two-arm trial following patients until 283 deaths have been observed would be required [9]. We assume further that there is a constant enrollment rate, exponential failure rate with median 6 in the control group and exponential dropout rate of 2 percent per month. Assuming 28 months of enrollment and follow-up of 12 months for the last patient, such a trial would require 368 patients [6].

Because of the doubts about the translation of early trial results into a mortality benefit, a futility analysis might be planned to stop the trial early if results are not sufficiently promising. Group sequential methods can be applied for this purpose [11, 5] using, for example, the gsDesign R package [1]. A typical group sequential design provides 3 options at an interim analysis: 1) stop for a successful (highly positive) result, 2) stop for an unsuccessful result or 3) continue. Because many

[*] Merck & Co.

[†] Merck & Co.

patients would have little follow-up at the time of such an analysis, we consider an additional interim decision given "intermediate" results. We divide the "continue" decision into: 3a) continue to complete a definitive Phase III trial or 3b) limit(or stop enrollment) and continue to complete a more exploratory Phase II trial. Option 3b is less expensive than continuing to Phase III and allows a less accelerated path for development of a drug with moderately promising results. With this strategy, the rule to stop the trial immediately can be less aggressive. It allows a full exploration of the Phase II study data prior to better enable the decision to carry out a Phase III trial. The purpose of this paper is to extend group sequential methods to accomplish this last adaptation.

The paper is organized as follows. In the next section we provide a design description followed by sections developing testing and statistical bounds for decision making, Type I error and power, decision strategies and criteria for trial phase selection, and discussion.

2. Design description

2.1 Test statistics and distributional assumptions

The intent is to set up two group sequential designs that are, for some $d \geq 1$, identical prior to analysis d and to choose between the two designs if the trial continues to analysis d. We will refer to analysis d as the adaptation point or analysis. Let k_1, k_2 (both $> d$) denote the total number of planned analyses for these two designs. Corresponding to this, we consider two sequences of multivariate normal test statistics $Z_{m1}, Z_{m2}, \ldots, Z_{mk_m}$, $m = 1, 2$. We assume the canonical form for a group sequential design as laid out by Jennison and Turnbull [4] where for $m = 1, 2$, some $0 < n_{m1} < n_{m2} < \ldots n_{mk_m}$ and $1 \leq i \leq j \leq k_m$

$$\mathrm{E}\{Z_{mi}\} = \theta\sqrt{n}_{mi}$$

$$\mathrm{Cov}\{Z_{mi}, Z_{mj}\} = n_{mi}/n_{mj}.$$

In this notation, n_{mi} could represent, for example, a number of observations, a number of events (for a trial with a time-to-event endpoint) or statistical information.

2.2 Testing boundaries

For each of the two group sequential designs we define bounds in the usual fashion for the test statistics Z_{im} just defined, but with some restrictions. We assume a null hypothesis of $\theta = 0$ and for design m, $m = 1, 2$, an alternate hypothesis of $\theta = \theta_m$ for some fixed $\theta_m > 0$. This means both designs seek an alternative in the same direction. This is not necessary, nor is it necessary to divide into only 2 designs. However, for simplicity and the application of interest we make these restrictions here. For design m, $m = 1, 2$, we assume a set of lower and upper bounds a_{mi}, b_{mi} where $a_{mi} < b_{mi}$ for $i < k_m$ and $a_{mk_m} = b_{mk_m}$. This final restriction is not necessary, but it is useful for the situations where we have 2 possible decisions at the end of the trial for each design (reject $\theta = 0$ or reject $\theta = \theta_m$). An intermediate region could be allowed at the end of the trial with $a_{mk_m} < b_{mk_m}$ if an additional decision region were desired.

Since the test statistics are identical for the two designs for $i \leq d$, we define $Z_i \equiv Z_{1i} = Z_{2i}$. For $i < d$ we would assume a common set of bounds for the two designs; *i.e.*, $a_{1i} = a_{2i} = a_i < b_i = b_{1i} = b_{2i}$. We reject $\theta = 0$ in favor of $\theta = \theta_m$

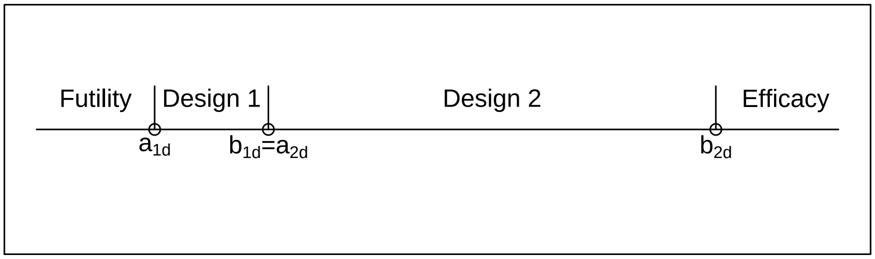

Figure 1: Test statistic at adaptation point.

if $Z_{mi} \geq b_{mi}$ and reject $\theta = \theta_m$ in favor of $\theta = 0$ if $Z_{mi} < a_{mi}$. Figure 1 diagrams decision regions at the adaptation point. We assume $a_{1d} < b_{1d} = a_{2d} < b_{2d}$. We reject both $\theta = \theta_1$ and $\theta = \theta_2$ if $Z_d < a_{1d}$. We reject $\theta = 0$ in favor of $\theta \geq \theta_2$ if $Z_d \geq b_{2d}$. We are left with two intermediate regions for Z_d: $[a_{1d}, b_{1d} = a_{2d})$ and $[b_{1d} = a_{2d}, b_{2d})$. For Z_d in the first of these regions, the trial continues with design 1 after analysis d, while for the second, the trial continues with design 2.

For interim analyses i after the adaptation analysis d, where $i > d$, the decisions are based on bounds for each design m, $m = 1, 2$. The stopping rules for all scenarios are summarized in Table 1.

Table 1: Planned stopping rules for the adaptive design.

Analysis	Sample size	Stop for futility	Continue trial	Stop for efficacy
$i < d$	n_i	$Z_i < a_i$	$a_i \leq Z_i < b_i$	$Z_i \geq b_i$
$i = d$	n_d	$Z_d < a_{1d}$	$a_{1d} \leq Z_i < b_{2d}$	$Z_d \geq b_{2d}$
Continue with design 1 if $a_{1d} \leq Z_d < b_{1d} = a_{2d}$				
$i = d+1 \ldots k_{1d}$	n_{1i}	$Z_{1i} < a_{1i}$	$a_{1i} \leq Z_{1i} < b_{1i}$	$Z_{1i} \geq b_{1i}$
Continue with design 2 if $b_{1d} = a_{2d} \leq Z_d < b_{2d} = b_{1d}$				
$i = d+1 \ldots k_{2d}$	n_{2i}	$Z_{2i} < a_{2i}$	$a_{2i} \leq Z_{2i} < b_{2i}$	$Z_{2i} \geq b_{2i}$

3. Type I error and power

In this section, we calculate the type I error and power for the whole study as well as for each phase. First, we define the events leading to decisions to follow designs 1 and 2, respectively, as

$$D_1 = \{a_{1d} \leq Z_d < a_{2d}\} \cap_{i=1}^{d-1} \{a_i \leq Z_i < b_i\},$$

$$D_2 = \{a_{2d} \le Z_d < b_{1d}\} \cap_{i=1}^{d-1} \{a_i \le Z_i < b_i\}.$$

Following are mutually exclusive results which will result in a positive trial (including design 1 and design 2) assuming decision rules are followed:

- For $i \le d$,
$$B_i = \{Z_i \ge b_{2i}\} \cap_{j=1}^{i-1} \{a_j \le Z_j < b_j\}$$
- For $d < i \le k_1$
$$B_{1i} = D_1 \cap_{j=d+1}^{i-1} \{a_{1j} \le Z_{1j} < b_{1j}\} \cap \{Z_{1i} \ge b_{1i}\}$$
- For $d < i \le k_2$
$$B_{2i} = D_2 \cap_{j=d+1}^{i-1} \{a_{2j} \le Z_{2j} < b_{2j}\} \cap \{Z_{2i} \ge b_{2i}\}.$$

Since these events are mutually exclusive, the total probability of a positive finding as a function of θ is obtained by summing as follows:

$$\alpha(\theta) = \sum_{i=1}^{d} \mathrm{P}\{B_i|\theta\} + \sum_{m=1}^{2} \sum_{i=d+1}^{k_m} \mathrm{P}\{B_{mi}|\theta\} \tag{1}$$

Total Type I error for the design is $\alpha(0)$ while for $\theta > 0$ the total power for the trial is $\alpha(\theta)$. This type of adaptation has been used previously by [3], [10].

We define the probability of crossing an upper bound for design 2 as the Phase III Type I error or power:

$$\alpha_2(\theta) = \sum_{i=1}^{k_2} \mathrm{P}\{\{Z_{2i} \ge b_{2i}\} \cap_{j=1}^{i-1} \{a_{2j} \le Z_{2j} < b_{2j}\}|\theta\}. \tag{2}$$

Thus, the Type I error for the Phase III adaptation assuming decision rules are obeyed is $\alpha_2(0)$. To allow flexibility in decision-making, we may assume the interim futility bounds may be ignored when computing the Phase III Type I error. Phase 3 Type I error assuming non-binding interim rules will be denoted by $\alpha_2^+(0)$ where

$$\alpha_2^+(\theta) = \sum_{i=1}^{k_2} \mathrm{P}\{Z_{2i} \ge b_{2i} \cap_{j=1}^{i-1} Z_{2j} < b_{2j}|\theta\}. \tag{3}$$

We set up Type I error computations to allow maximum flexibility in decision making and conclusions at the end of the trial. While using this Type I error computation with the Phase II/III selection and futility criteria will use less than the full Type I error for each Phase, the additional flexibility is important to maintain the Type I even when futility and decision rules are not followed.

If the trial continues only until 160 events, we can still allow for a positive Phase III finding with a very positive result. We achieve this flexibility by setting the analyses for design 2 at every number of events where an analysis is planned for design 1. Separate bounds are set for a Phase III 2.5% Type I error and Phase II 10% Type I error.

4. Choice of timing for analyses and bounds

The timing and bounds for decisions may be chosen in consideration of a number factors representing tradeoffs. One one hand, a larger sample size makes it possible to lower Type I and Type II error rates or to have higher confidence that decisions are based on a meaningful estimate of treatment effect. On the other hand, waiting for a large sample size may be very expensive and take a lot of time. Note that enrollment rates relative to the rate of endpoint accumulation are critical to completing an analysis to alter sample size. Getting sites open quickly and enrolling at a relatively constant rate should help maximize potential savings; this is not a minor issue, but will not be discussed at length here. Tradeoffs will be different depending on the situation at hand. Flexible spending functions such as those provided by [2] allow choosing bounds to fit a desired level of significance at each analysis. By doing this and experimenting with timing of the analysis also allows selection of approximate observed treatment effects corresponding to the bounds selected.

4.1 An Example

Continuing our example, we choose between a 2- and 3-analysis design at the first interim analysis. Then $k_1 = 2$, $k_2 = 3$, $d = 1$, $n_{11} = n_{21} = n_1$, $Z_{11} = Z_{21}$. We also let $n_{12} = n_{22}$. For, say, a trial with a binomial or normal outcome, n_{12} and n_{23} would represent different total sample sizes. Here we consider a trial with a time-to-event outcome. In this case, n_{11}, n_{12}, n_{21}, n_{22} and n_{23} denote the number of events at each analyis. If the smaller Phase II design 1 is selected, enrollment may be discontinued while continuing treatment and follow-up for patients already enrolled. If the larger Phase III design 2 is selected, further patients are enrolled to power the result for a definitive finding. Thus, although we assume the number of events $n_{12} = n_{22}$ are the same whether we choose a Phase II or Phase III trial at the first analysis after the adaptation, the patient population from which these endpoints are derived would be different. Under these assumptions we have $Z_{11} \equiv Z_{21}$, but $Z_{12} \neq Z_{22}$. However, under the usual assumptions of proportional hazards for a clinical trial [11], the pairs Z_{11}, Z_{12} and Z_{21}, Z_{22} are asymptotically identically distributed with the canonical form of Section 2.1.

For our example, we chose to do the first interim analysis after 120 deaths and enrollment of approximately 220-250 patients. This allowed a reasonable tradeoff between Type I error, Type II error and the approximate treatment effect at decision boundaries. Two-parameter spending functions that allow specifying cumulative spending for the analyses after 122 and 176 events. Even though final analyses are planned for Phase II after 176, we chose a spending function that planned for the Phase III number of events. This will allow specification of p-values and confidence intervals using the methods of [7], which will be documented elsewhere. A Cauchy spending function was chosen for design 1 and a logistic spending function for design 2. As seen in Figure 2, this allows the spending function for design 1 (Phase II) to be greater than that for design 2 at all points.

The properties of the decision regions at the first interim analysis are summarized in Table 2.The study would continue as a phase III after the first interim analysis if the interim Z-statistic is greater than $b_{11} = a_{21}$ =1.23 which corresponds to a p-value of 0.109 and approximately to an empirical hazard ratio of 0.9. If the Z-statistic is between 0.58 and 1.23, the enrollment will stop and the study will convert to a Phase II study. A Z-statistic of 0.58 corresponds to a p-value of 0.282 and approximately to an emprical hazard ratio of 0.9.

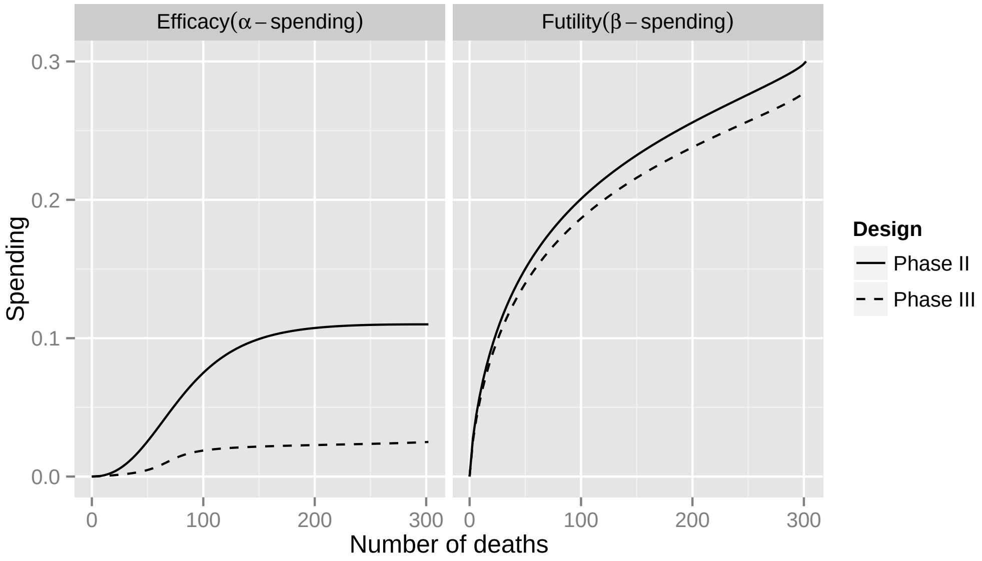

Figure 2: Spending functions for Phase II and Phase III.

In this case, consider 2 prior distributions for predictive power: one informative and one non-informative. For the informative prior, we assume the log of the hazard ratio has a mean of the logarithm of 2/3 and a variance of 4/38; this is equivalent to the uncertainty for a frequentist estimate of the log hazard ratio with 38 events. This distibution suggests the hazard ratio has prior probability of 2/3 of being in the interval (0.58, 0.77). For the non-informative prior we assume the log hazard ratio is centered about 0 (no treatment effect) and having a standard deviation of 1. This suggests a prior probability of 2/3 the the hazard ratio is between 0.65 and 1.54. We consider conditional power both for the estimated treatment effect and at the original alternate hypothesis, the latter as recommended by [8]. Note that while the choice of a prior is somewhat arbitrary, using conditional power places probability 1 on a single value of the treatment effect. In any case, for our example the predictive probability of a positive trial at the phase III decision bound is 0.56 under the informative prior and 0.48 under the non-informative prior.

Setting this adaptation bound so high makes sense within the described context. We set the futility bound to stop the trial as a_{11} =0.58 which produces Type II error for the development program at the first analysis of 0.083 and corresponds approximately to a hazard ratio of 0.9.

5. Decision strategy

With a moderately positive test statistic ($b_{1d} = a_{2d} \leq Z_d < b_{2d}$) results in adapting to Design 2. While this is not necessary, for the purposes of this work we assume Design 2 is a ”large” Phase III design.

For a less positive test statistic ($a_{1d} \leq Z_d < b_{11} = a_{2d}$) results in adapting to Design 1. In our case this adaptation will be to a smaller, Phase II design; another alternative would be to employ this adaptation to increase sample size and have a larger Phase III trial [3].

Table 2: Decision probabilities for first interim analysis.

		Stop for Futility	Select Phase II	Select Phase III	Stop for Efficacy
Z-value		≤0.58	(0.58, 1.23)	(1.23, 3.8)	>3.8
$\widehat{\text{HR}}$ (approx.)		≥0.9	(0.9, 0.9)	(0.9,0.5)	>0.5
			Decision Probabilities		
True HR	.5	0	0.01	0.49	0.5
	.6	0.01	0.05	0.78	0.16
	.7	0.08	0.15	0.74	0.03
	.8	0.26	0.24	0.49	0.01
	.9	0.5	0.24	0.26	0
	1	0.72	0.17	0.11	0
	1.2	0.94	0.05	0.01	0
	Informative prior				
	Non-informative prior				

Table 3: Conditional and predictive probabilities for Phase II and Phase III options at interim analysis phase selection bound.

		Phase Selected	
		Phase II	Phase III
Predictive	Informative		
	Non-informative		
Conditional	$\widehat{\text{HR}}$		
	HR=.7		

6. Discussion

While the actual bounds for this design were selected using spending functions, this was not essential to the presentation here. Selecting bounds in a way such that p-values and confidence intervals can be computed is worthy of future investigation. The methods of Liu and Anderson [7] should be applicable.

We hope the methods presented here may be of use to those designing adaptive clinical trials. The fact that the method is a straightforward extension of group sequential design should make it reasonably understandable to reviewers. The flexibility added by creating an additional interim decision possibility will allow meaningful adaptation to accelerate or moderate drug development in accord with the strength of an interim treatment effect evaluation. Continuing the Phase II design after a decision not to immediately pursue Phase III allows a deliberate review of the data before making additional development decisions.

References

[1] Keaven M. Anderson. gsdesign r package. http://cran.r-project.org/web/packages/gsDesign/.

[2] Keaven M. Anderson and Jason B. Clark. Fitting spending functions. *Statistics in Medicine*, 29:321–327, 2010.

[3] Sarah C. Emerson, Kyle D. Rudser, and Scott S. Emerson. Exploring the benefits of adaptive sequential designs in time-to-event endpoint settings. *Statistics in Medicine*, 30:1199–1217, 2010.

[4] Christopher Jennison and Bruce W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, Boca Raton, FL, 2000.

[5] Kyungmann Kim and Anastasios A. Tsiatis. Study duration for clinical trials with survival response and early stopping rule. *Biometrics*, 46:81–92, 1990.

[6] John M. Lachin and Mary A. Foulkes. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*, 42:507–519, 1986.

[7] Qing Liu and Keaven M. Anderson. On adaptive extensions of group sequential trials for clinical investigations. *Journal of the American Statistical Association*, 103:1621–1630, 2008.

[8] Qing Liu and George Y. Chi. On sample size and inference for two-stage adaptive designs. *Biometrics*, 57:172–177, 2001.

[9] David Schoenfeld. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68:316–319, 1981.

[10] P. Teal, D. Selchen, and et al. Sherman, D. Rreact study: rapid response with an astrocyte modulator for the treatment of acute cortical stroke. Presented at the American Stroke Association 29th International Stroke Conference; San Diego, California, USA., February 2007.

[11] Anastasios A. Tsiatis. Repeated significance testing for a general class of statistics use in censored survival analysis. *Journal of the American Statistical Association*, 77:855–861, 1982.