Advanced Application of Using Progression-Free Survival to Make Optimal Go-No Go Decision in Oncology Drug Development

Linda Z. Sun and Cong Chen
Biostatistics and Research Decision Sciences, Merck Research Laboratories (MRL), UG1C-46, 351 Sumneytown Pike, Upper Gwynedd, PA 19454, USA

Abstract

With cancer being a leading cause of death worldwide, there is an urgent need to accelerate oncology drug development so that more new therapies can be available to cancer patients. Seamless Phase II/III designs hold the promise for acceleration because they remove the white space between Phase II and Phase III. Almost all published seamless Phase II/III design papers assume Phase II and Phase III have the same endpoint. However, in oncology drug development, the Phase III endpoint is usually a clinical endpoint, i.e., overall survival (OS), which takes longer time to observe. The Phase II endpoint is usually a shorter term surrogate endpoint, e.g., progression-free survival (PFS). Because Phase II and Phase III use different primary endpoints, it is challenging to pre-specify a Go-No Go (GNG) decision rule from Phase II to Phase III. This is one of the reasons why seamless designs are less used in practice than expected in oncology drug development. In this paper, we would like to address the following issues: 1) how to effectively incorporate surrogate biomarker (e.g. PFS) data into the decision criteria; 2) how to derive objective GNG criteria from a benefit-cost ratio perspective to streamline the decision making process; 3) how to fully realize the potential of a seamless design with proper risk mitigation. This work is based on a real example in the oncology therapeutic area. However, the general approach is equally applicable to various other areas.

Key Words: Seamless design; Surrogate; Decision analysis; Benefit-cost ratio; Go-No Go decision, Progression-free survival

1. Introduction

There are many challenges in oncology drug development. One challenge is how to accelerate the development program. With cancer being a leading cause of death worldwide, there is an urgent need to accelerate oncology drug development so that more new therapies can be available to cancer patients. In addition to the urgency of unmet medical need, fierce competition is another reason that pharmaceutical companies seek to accelerate oncology drug development. Past experiences have shown that the maximum tolerated dose identified in Phase I studies may not be the dose with the best benefit/risk profile. However, sometimes in order to catch up with the competition a development program has to be moved forward without knowing the best dose to be used. In such scenarios, seamless Phase II/III designs seem to be a natural choice, as they remove the

white space between Phase II and Phase III while allow testing multiple doses in the Phase II part.

Another challenge in oncology drug development is how to make the Go-No Go (GNG) decision from Phase II to Phase III. On one hand, the number of new drug candidates and opportunities exploded with more understanding of the signalling pathways. On the other hand, there is limited resource for a pharmaceutical company to invest in these candidates. To address this problem, a good strategy to make GNG decision is needed. That is, an objective mechanism should be used to stop development of those futile candidates and to continue the promising candidates. Good GNG decisions are needed not only for portfolio management but also for a single drug candidate development program. One phenomenon in oncology drug development is that, even though many drug candidates were developed based on good science and showed exciting early efficacy signals, the Phase III success is very low. The conventional endpoint for a Phase III confirmatory trial in oncology is overall survival (OS), defined as time from randomization to death due to any cause. In recent years, there is an increasing interest in progression-free survival (PFS) as an endpoint, i.e., time from randomization to disease progression or death due to any cause, whichever occurs first. Because PFS takes shorter time to observe, often it is used as the Phase II endpoint in oncology. One of the reasons for the low success rate in oncology Phase III trials is due to this endpoint change from Phase II to Phase III, which makes the Go-No Go criteria difficult to define in sequential Phase II and Phase III programs, let alone in seamless Phase II/III designs. So it is not a total surprise that seamless designs are less used in practice than expected in oncology drug development.

In this paper, we use a real example as a motivation to address the following issues: 1) how to effectively incorporate surrogate biomarker (e.g. PFS) data into the decision criteria; 2) how to derive objective GNG criteria from a benefit-cost ratio perspective to streamline the decision making process; 3) how to fully realize the potential of a seamless design with proper risk mitigation. To focus on the main issues, many details of the real example are generalized or skipped. Section 2 will present the motivating example and the methodology used to address the issues. Section 3 will be summary and discussion.

2. Motivating Example and Methodology

2.1 Utilizing seamless design for acceleration

The motivating example is about the development of a drug candidate to be tested in platinum resistant ovarian cancer patients. Because this test drug is a targeted therapy, it has better safety profile than chemotherapy, even though the efficacy may be comparable or superior to the chemotherapy. The primary hypothesis of the Phase III study is that:

- The test drug is non-inferior to the comparator (chemotherapy) in terms of overall survival (OS) at the 1.1 hazard ratio margin and superior to the comparator in terms of safety profile.
- Or the test drug is superior to the comparator in terms of OS

Hierarchy testing procedure will be used to control the type I error rate. That is, the non-inferiority will be tested first, and once passed, the superiority will be tested.

When this test drug's MTD was defined, several competing drug candidates in the same class had completed single arm Phase II studies or were almost finishing randomized

Phase II trials. In this case, a sequential Phase II/III program, which will take long time to develop, became commercially less viable and a seamless Phase II/III design was considered to accelerate the program.

There are two types of seamless designs, inferentially seamless and operationally seamless. The inferentially seamless designs (Stallard and Todd 2003, Posch et al 2005) combine Phase II data and Phase III data with some multiplicity adjustment to control type I error rate in the final analysis. Although the statistical methodology is valid and in existence for decades, such designs are deemed as less well understood adaptive designs in FDA's adaptive design guidance paper (FDA 2010). Operationally seamless designs only use Phase III data in the final analysis, but the enrollment is seamless between Phase II and Phase III. In addition to health authority's concern about inferentially seamless design, several other factors led the development team to choose operationally seamless Phase II/III instead of inferentially seamless one in this motivating example. One factor is the difficulty of using surrogate biomarker, in this case PFS, to make GNG decision while the Phase III endpoint is OS. Another factor is about which decision body to make the dose selection based on Phase II data. If inferentially seamless design is chosen, the dose selection has to be made by an external data monitoring committee (eDMC), because otherwise the Phase II data may be unblinded and can not be utilized in the final analysis. Dose selection is usually a complicated decision; totality of the within-trial data including efficacy and safety, internal and external information all has to be assessed. Even though the guidelines for dose selection can be pre-specified in the study protocol, not all scenarios can be foreseen or simulated. Therefore, the development team preferred to make the dose selection by a joint effort of internal and external experts. Because internal team will be unblinded to the Phase II data, to keep the integrity of the study, Phase II data cannot be combined to Phase III data in the primary final analysis. Operationally seamless design can be considered as a middle ground between sequential Phase II and Phase III strategy and the inferentially seamless strategy.

2.2 Study design of the motivating example

The final design of the motivating example is shown schematically in Figure 1. In the Phase II portion, patients will be randomized to three (3) treatment groups with equal allocation: test drug high dose, test drug low dose, and control drug. The primary endpoint for Phase II is progression-free survival (PFS). The sample size for Phase II is to enroll about 210 patients and accumulate 135 PFS events to have sufficient power for each dose of the test drug to demonstrate superiority to the control in terms of PFS. After Phase II is completed, one dose will be selected to move into Phase III. In the Phase III portion, patients will be randomized to two treatment groups: test drug and control drug. The primary endpoint of Phase III is overall survival (OS). The sample size for Phase III is to enroll about 720 patients and accumulate 508 death events to have sufficient power to demonstrate that the test drug is non-inferior to the control drug. This sample size also provides sufficient power to demonstrate that the test drug is superior to the control drug in terms of event rate for a particular adverse experience (AE).

In order to realize the operationally seamless design, an interim analysis will be conducted in Phase II. The enrollment of Phase II will close when it is predicted that approximately 4 months after this time point there will be 135 PFS events. The interim analysis will take place approximately one month before the accrual completion. The purpose of this interim analysis is to determine whether Phase III enrollment can be initiated before final data of Phase II is available. If a Go decision is made, one arm of MK 4827 along with the control arm will be carried to Phase III. If a Go decision can not

be made at the interim analysis, Phase III will be on hold and a final decision will be made at end of Phase II. The Go criterion at this interim analysis is to have at least 80% conditional power to make a Go decision at the final analysis of Phase II. Since it will take about one month to conduct the interim analysis and make a decision, the timing of this interim analysis is set to be one month before Phase II accrual completes. This way, Phase III accrual will start seamlessly when Phase II accrual completes and a decision has come out from the interim analysis in Phase II.

2.3 Incorporating surrogate biomarker data in Go-No Go (GNG) decision making

The GNG decision for a drug candidate to move from Phase II to Phase III is a major decision in drug development. Ideally the decision should be made based on the data from the same endpoint which will be the primary endpoint of Phase III. In our discussion, since the primary endpoint of Phase III is OS, the most relevant data is the OS data in Phase II. However, since OS data usually take long time to observe, there are limited OS data by the end of Phase II.

A common practice in oncology drug development is to make GNG decision only based on the surrogate biomarker, PFS data, and ignore the OS data observed in Phase II. In Chen and Sun (2011), it is proposed to combine the PFS data and OS data for decision making so that no information is wasted. Before we explain how to combine PFS and OS data, we first discuss how to use PFS data from Phase II to estimate OS treatment effect.

The relative effect size (γ) between OS and PFS (in log-hazard-ratio scale) holds the key in such estimation (Chen, Sun, and Li 2011, Sun and Chen 2009). In this motivating example, it is assumed that this ratio is 0.6. It implies that the treatment effect in OS is 60% of the treatment effect in PFS, which represents a reasonable estimate based on published data of a variety of solid tumor in recent years. For example, if a drug has a treatment effect of hazard ratio (HR) = 0.8 in OS, it has a treatment effect of HR = 0.69 in PFS. In other words, if the treatment effect in PFS is 31% hazard reduction in Phase II, it implies that the treatment effect in OS is 20% hazard reduction for the test drug versus the control drug. Most GNG decisions between Phase II and Phase III in oncology drug development were made this way, even though often times the relative effect size were implicitly used and the decision makers may not even realize it. Is the translation from effect size in PFS to effect size in OS always a one-to-one translation? The answer probably is no. Therefore to adequately account for the uncertainty in effect size translation, we assume that the relative effect size (γ) has a normal distribution with mean of 0.6 and standard deviation of 0.2. This assumption covers a wide range of treatment effect ratio seen in the literature. With this variability, a 0.69 hazard ratio in PFS may translate into a range of hazard ratio in OS, and 95% of the estimated HR in OS fall between (0.69, 0.93). The low success rate in oncology Phase III studies may be partially explained by not adequately accounting this uncertainty in translating PFS effect in Phase II to the treatment effect in OS when making GNG decisions.

We then used a weighted method to combine the OS effect predicted from the observed PFS effect (Δ_{PFS}) and the observed OS effect OS (Δ_{OS}), both in log-hazard-ratio scale, using the formula below (Chen and Sun (2011).

$$S = -(w\Delta_{OS} + (1-w)\gamma\Delta_{PFS})$$

Since the number of OS events in Phase II is relatively small compared to the number of PFS events, less weight is put on the observed OS effect and more weight is put on the

predicted OS effect based on PFS. A weight of 0.15 (i.e., $w = 0.15$) is given to the observed OS effect in Phase II, and a weight of 0.85 is given to the predicted OS effect. This weighting scheme approximates the inverse-variance scheme when the true treatment effect is in the parameter space of interest while the actual numbers of PFS and OS events are reasonably close to the target. (See Appendix I for technical details of the characteristics of the test statistics.)

In the next section, we will discuss what value of S (an approximate measure of hazard reduction) will constitute a Go-to-Phase-III decision at the end of Phase II. That is how to set the GNG criteria between Phase II and Phase III.

2.4 A benefit-cost effective GNG criteria

Setting the GNG criteria between Phase II and Phase III is not an easy task. Even though objective criteria are sought after, most decisions made in the conventional paradigm are somewhat subjective, because the criteria only consider a single factor, for example, evidence of efficacy (p-value). As a matter of fact other factors are also important. If the Go bar is high, the probability of success (POS) for Phase III should be relatively high when conducted, but the chance of conducting a Phase III may be low and a good drug candidate may be missed. If the Go bar is low, the chance of conducting a Phase III may be high, but the POS of Phase III may be low and the investment may be wasted on a futile drug candidate. Using the principle in decision analysis, a utility function can incorporate different factors and considerations into one index. A natural utility function in drug development is the return of investment. In Chen and Beckman (2009), the utility function for return of investment is defined as the probability of success (POS) adjusted revenue per unit of cost of the development program. A GNG criterion can be set to maximize this return of investment function. To be more specific, the POS adjusted revenue is POS times the revenue, and the cost of the development is defined as per patient cost times the expected sample size of Phase II and Phase III, considering the likelihood that there may be No Go to Phase III (see below for more details). Because the return of investment is defined as a ratio, the actual monetary value of the revenue and cost don't matter.

We denote the Go criterion from Phase II to Phase III to be $S > C$, where S is the estimated OS effect from Phase II data defined in the previous section and C is a critical value to be solved so that the return of investment can be maximized. Then $P(S > C)$ is the probability of Go from Phase II to Phase III.

The POS of the program is the probability of Go times the power of the Phase III study. Under different assumptions of the treatment effect (i.e., hazard ratio HR of test drug over control drug), the probability of Go and the Phase III power will be different. In the Bayesian framework, we assume that the treatment effect has a discrete prior distribution, with π_1 probability of being better than the control (e.g. HR = 0.8), π_2 probability being equivalent to the control (i.e. HR = 1), and $(1- \pi_1 - \pi_2)$ probability of being worse than the control (e.g. HR = 1.1). With this prior, we can compute the predictive POS adjusted value of the test drug. Based on the industry benchmark data in oncology drug development, we used $\pi_1 = \pi_2 = 1/3$ in our example. That is, before conducting the Phase II and Phase III studies, we think that this test drug probably has equal chance of being superior, equivalent, and inferior to the control drug.

In this example, the Phase III is successful in two scenarios: (1) Superiority in efficacy is demonstrated; (2) Superiority in efficacy is not demonstrated; only non-inferiority is

demonstrated, but superiority in safety is demonstrated. The regulatory approvability and commercial value are different in these two scenarios. We also incorporated this consideration into our revenue calculation. In our example, stake holders and experts believe the relative approvability from health authority is 2:1 for scenario 1 vs. scenario 2, and the corresponding relative commercial value is 5:1. Let V be the relative value of the two scenarios, then $V = 2 \times 5 = 10$.

With the above set up, let B be the predictive POS adjusted value of the program in the motivating example,

$$B = M_B \sum_{i=1}^{2} \pi_i p_i (V q_{S,i} + q_{NI,i} q_{AE})$$

Where

- M_B is a constant. It is the monetary value of the test drug when only non-inferiority in efficacy is demonstrated and superiority of safety is demonstrated.
- π_i is the probability mass of the discrete prior distribution for the treatment effect (HR), $i = 1, 2, 3$. $\pi_1 + \pi_2 + \pi_3 = 1$. For example, the prior distribution of the treatment effect HR is P(HR = 0.8) = π_1 = 1/3, P(HR = 1) = π_2 = 1/3, and P(HR = 1.1) = π_3 = 1/3. Because there is no value of the test drug when it is inferior to the control, we don't include $i = 3$ in the value calculation.
- p_i is the probability of Go from Phase II to Phase III under the ith value of HR in the discrete prior distribution. For example, p_1 is $P(S > C)$ under HR = 0.8.
- V is the relative value of demonstrating superiority in efficacy vs. demonstrating non-inferiority in efficacy and superiority in safety. For example, $V = 10$.
- $q_{S,i}$ is the probability of demonstrating superiority in Phase III under the ith value of HR in the discrete prior distribution.
- $q_{NI,i}$ is the probability of only demonstrating non-inferiority and not superiority in Phase III under the ith value of HR in the discrete prior distribution.
- q_{AE} is the probability of demonstrating safety advantage of the test drug over control drug in Phase III.

Let D be the cost of the development program for Phase II and Phase III portion.

$$D = M_C (R + \sum_{i=1}^{3} \pi_i p_i)$$

Where

- M_C is a constant. It is the monetary cost of the Phase III study.
- R is the relative cost of Phase II portion to Phase III portion. In the motivating example, the operation team's estimate of R is 0.4 for this seamless design considering that the sample size is 210 patients in Phase II portion, 720 patients in Phase III portion, some Phase III sites needed to be set up at risk before final Phase II results are available (upfront cost), and various other factors.

Define the return of investment function as B/D. Notice that p_i = P($S > C$), the optimal GNG bar C is obtained by maximizing $U(C)$ below

$$U(C) = \frac{\sum_{i=1}^{2} \pi_i p_i (V q_{S,i} + q_{NI,i} q_{AE})}{(R + \sum_{i=1}^{3} \pi_i p_i)}.$$

The input variables that we need to give before solving for C are: the discrete prior distribution of treatment effect, the relative value of the superiority vs. non-inferiority Phase III results, and the relative cost of the Phase II portion vs. the Phase III portion. For the values of the input variables that we used in the motivating example, the optimal bar is $C = 0.09$. (Figure 2 illustrates how the utility function (U) changes with C.) Roughly speaking, this corresponds to a 9% hazard reduction based on the joint estimate of the OS effect by using both PFS data and OS data from Phase II as well as the estimate of relative effect size based on historical data.. The solid line in Figure 3 shows this optimal bar graphically in terms of PFS effect and OS effect at the end of Phase II. Table 1 shows the operating characteristics of this GNG bar under various assumptions of the true treatment effect.

When using the surrogate biomarker PFS data in decision making, we made an assumption about the relative effect size of PFS and OS. At the end of Phase II, to mitigate the risk of using a wrong assumption, we should check the relative effect size observed in Phase II. If the observed OS effect is smaller than the lower bound of the 95% confidence interval (CI) for the predicted OS effect from PFS effect ($\gamma\Delta_{PFS}$), we would be concerned, because it indicates that the observed OS effect is much smaller than the predicted effect from PFS data using the historical relationship of relative effect size. Therefore, our proposed GNG criteria at the end of Phase II are (1) the estimated OS effect (S) is greater than the optimal bar; (2) the observed OS effect is bigger than the lower bound of the 95% CI for the predicted OS effect. The dotted line in Figure 3 shows the boundary for criterion (2). Overall, it is a Go decision if the observed PFS effect and OS effect from Phase II falls below both solid and dotted lines, and a No Go decision otherwise.

Now we have the optimal GNG bar for the end of Phase II data, we can calculate the bar for the interim analysis (IA) in Phase II which gives 80% conditional probability that the Go bar will be passed at the end of Phase II. If the following criteria are met for the interim analysis data, seamless Phase III enrollment will be triggered.

(a) The observed OS effect and PFS effect can provide at least 80% conditional probability that criterion (1) will be met at the end of Phase II.
(b) The observed OS effect at IA is bigger than the lower bound of the 95% CI for the predicted OS effect based on observed PFS effect at IA.

Figure 4 shows the boundaries for criterion (a) and (b). If the observed OS effect and PFS effect at IA fall below both solid and dotted lines, Phase III enrollment will be triggered while waiting for the Phase II data to become mature. Appendix II shows the technical details of the conditional power calculation.

3. Summary and Discussion

In this paper, we used a motivating example in oncology to discuss and address a few challenging aspects of designing the strategy and making decision in drug development:
(1) How to use seamless design to accelerate development timeline? It was estimated that by using the operationally seamless design, the development time could be saved by 9 months, compared to sequential Phase II and Phase III in the motivating example. The strategy to realize the operationally seamless Phase II/III design by using an interim analysis in Phase II was also discussed in this paper.
(2) How to explicitly incorporate surrogate biomarker data in decision making and also incorporate the uncertainty of using surrogate biomarker data to predict treatment effect in a clinical endpoint? Underestimating the uncertainty of the decision is part of the

reason we see high attrition rate in Phase III.
(3) How to make objective GNG decision from Phase II to Phase III by maximizing a return-of-investment utility function, which is usually the implicit goal of decision making in drug development?

As explained in Section 2.4, a few input variables need to be specified in the utility function. We investigated how the values of the input variables impact the utility function and the optimal GNG criterion (C). Table 2 shows the optimal bar under different values of the input variables. For the prior belief of the test drug activity, we considered three discrete prior distributions. The probability mass (π_1, π_2, π_3) for HR = 0.8, HR = 1, HR = 1.1 of the OS endpoint are (11%, 22%, 67%), (1/3, 1/3, 1/3), (50%, 33%, 17%) representing weak prior belief, moderate prior belief, and strong prior belief of drug activity, respectively. For the cost structure we considered two scenarios, one is for sequential Phase II and Phase III design, and the other is for operationally seamless Phase II/III. The relative cost (R) for Phase II portion vs. the Phase III portion will be higher for the seamless design, because site ready activities for Phase III will need to start at risk before a Go decision is made, so that the seamless enrollment can be achieved. The two relative cost we investigated are 25% and 40%. The general observation from Table 2 is that the optimal GNG bar is not sensitive to the prior distribution of the treatment effect. This observation is assuring because we usually don't have much historical data to pin point the prior distribution of treatment effect before we embark the Phase II and III program. The other observation is that the optimal GNG bar is lower if the upfront cost of the Phase II portion is higher. This observation is somewhat intuitive. If the front loading cost (sunk cost) in Phase II is high, we may want to proceed to Phase III since stopping may not save much cost. This input variable, cost structure, usually can be estimated objectively before Phase II/III program. In summary, the optimal GNG doesn't depend on the values of the input variables much, rather it is more driven by the operating characteristics of the decision, e.g., the probability of Go under the null hypothesis (α for Phase II) and the probability of No Go under the alternative hypothesis (β for Phase II). As discussed in Chen and Beckman (2009), using a utility function to determine the optimal α and β can maximize the return of investment and balance different considerations, and it is more objective than heuristic or gut-feeling argument which may only focus on one aspect of the problem (e.g., either false positive risk or false negative risk).

The real motivating example was even more complex than what is presented in this paper. There were considerations of responder subgroup (Chen and Beckman 2009, Song and Chen 2011) and filing for accelerated/conditional approval based on PFS data in Phase III (Chen and Sun 2011). Let alone, the dose selection rules were not discussed in this paper at all. Such a complicated situation is not unusual in drug development nowadays. When facing such scenario in which many factors intertwines and need to be considered, statisticians as quantitative scientists can contribute more than just providing sample size and power calculation. Statisticians can first help team to formulate and reach consensus of the overall objective of the program, whether to maximize the return of investment or to have first-to-market at any cost. Then statisticians can incorporate the various factors into a utility function which is a direct measure of the overall objective. The optimal GNG criteria will be set to maximize the utility function. This paradigm of setting GNG criteria can be much more efficient to reach team consensus than the old paradigm that only focusing on one or two factors separately when setting the GNG criteria. Inevitably, some input information, which needs to be fed into the utility function, may not be 100% objective (e.g. prior belief of the drug activity). We

recommend carefully exploring the sensitivity of the optimal GNG bar to such assumptions.

References

Beckman R. A., Clark J., and Chen C. (2011). Integrating Predictive Biomarkers and Classifiers into Oncology Clinical Development Programs: An Adaptive, Evidence-Based Approach. *Nature Review Drug Discovery* **10**, 735-749.

Chen C, and Beckman, R. A. (2009). Optimal cost-effective Go-No Go decisions in late stage oncology drug development. *Statistics in Biopharmaceutical Research,* **1,** 159-169.

Chen C, and Beckman R. A. (2009). Optimal cost-effective Phase II proof of concept and associated Go-No Go decisions. *J. Biopharmaceutical Statistics,* **1,** 431-440.

Chen C, and Beckman R. A. (2009). Hypothesis testing in a confirmatory Phase III trial with a possible subset effect. *Statistics in Biopharmaceutical Research,* **1,** 431-440.

Chen C, and Sun, L. (2011). On quantification of PFS effect for accelerated approval of oncology drugs. *Statistics in Biopharmaceutical Research*, **3**, 434-444.

Chen C, Sun L, and Li C. (2012). Evaluation of Early Efficacy Endpoints for Proof-of-concept Trials, *Journal of Biopharmaceutical Statistics*, accepted.

FDA (2010). Draft Guidance for Industry: Adaptive Design Clinical Trials for Drug and Biologics.

Sun L, and Chen C. (2009). Evaluation of Early Endpoints for Go-No Go Decisions in Late-Stage Drug Development. *ASA Proceedings of the Joint Statistical Meetings*, 2273-2283

Song Y, and Chen C. (2012). Optimal Strategies for Developing a Late-stage Clinical Program with a Possible Subset Effect, *Statistics in Biopharmaceutical Research*, accepted.

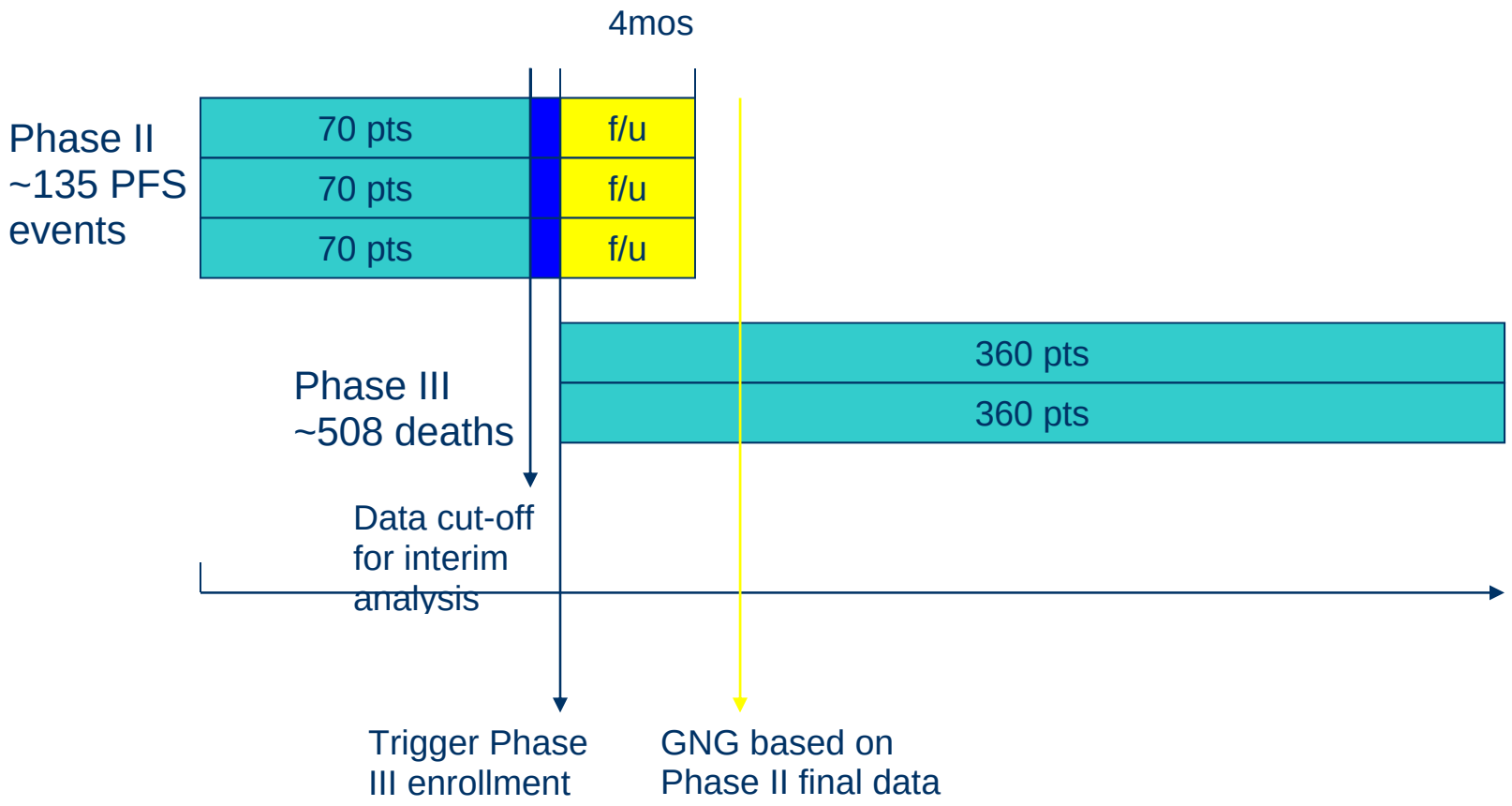
Song Y, and Chen C. (2009). Optimal strategies for developing a late-stage clinical program with a possible subset effect. *ASA Proceedings of the Joint Statistical Meetings*, 1408-1422.

Stallard N. and Todd S. (2003). Sequential Designs for Phase III Clinical Trials Incorporating Treatment Selection. *Statistics in Medicine* **22**:689–703.

Posch M. et al. (2005). Testing and Estimation in Flexible Group Sequential Designs with Adaptive Treatment Selection. Statistics in Medicine **24**: 3697-3714.

Figure 1

Study Design of the Motivating Example



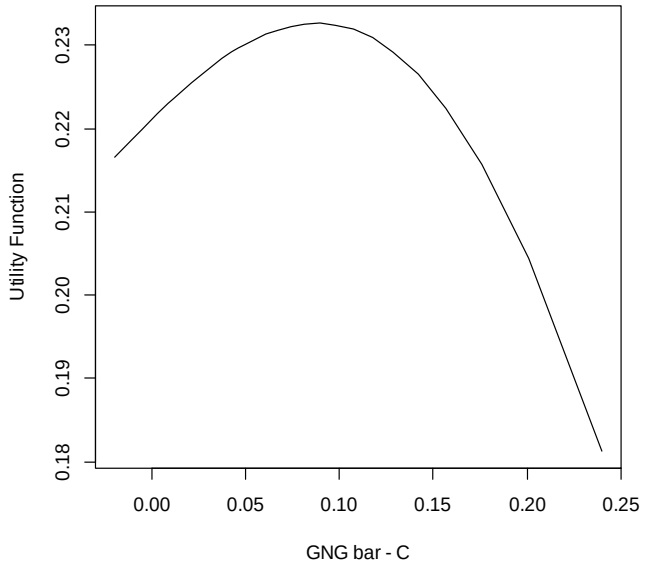


Figure 2 Illustration of How the Utility Function (U) Changes with the GNG Criterion (C)

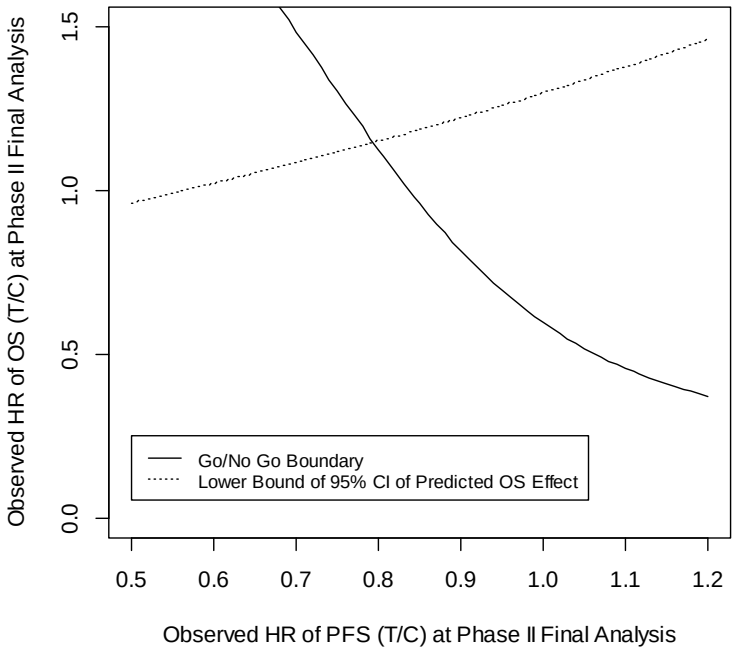


Figure 3 Optimal GNG Criteria at the end of Phase II. The lower bound of the 95% CI of predicted OS effect is the upper bound of the 95% CI in hazard ratio scale (test vs. control). The higher the HR the smaller the treatment effect.

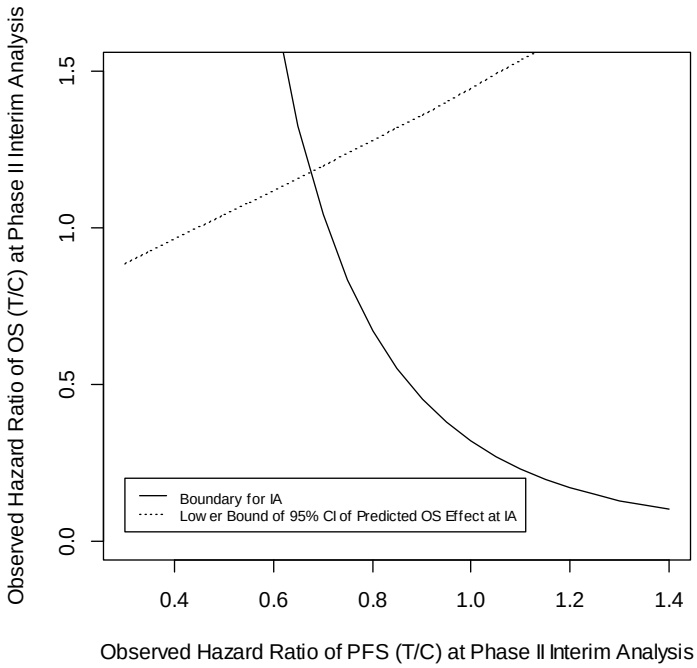


Figure 4 Criteria at Interim Analysis of Phase II to Trigger Phase III Enrollment. The lower bound of the 95% CI of predicted OS effect is the upper bound of the 95% CI in hazard ratio scale (test vs. control). The higher the HR the smaller the treatment effect.

Table 1
Operating Characteristics of GNG bar under Various Assumptions

HR_{PFS} (Test/Conrol)	HR_{OS} (Test/Control)	Probability of Go
1.17	1.1	10%
1	1	27%
0.69	0.8	81%

Table 2
Optimal GNG Bar (*C*) at the End of Phase II for Different Input Variable Values

Prior distribution of treatment effect			Relative cost of Phase II portion vs. Phase III portion	Optimal GNG bar in -log(HR) scale for the estimated OS effect (*S*)	Approximate optimal GNG bar in hazard reduction scale for the estimated OS effect
P(HR = 0.8)	P(HR = 1)	P(HR = 1.1)			
11%	22%	67%	25%	0.12	12%
			40%	0.10	10%
1/3	1/3	1/3	25%	0.12	12%
			40%	0.09	9%
50%	33%	17%	25%	0.11	10%
			40%	0.07	7%

Appendix I
Characteristics of the Test Statistics (S)

$$S = w\Delta_{OS} + (1-w)\hat{\gamma}\Delta_{PFS}$$

$$E(S) = w\log(HR_{OS}) + (1-w)\gamma\log(HR_{PFS})$$

$$Var(S)$$
$$= w^2\frac{4}{D_{OS}} + (1-w)^2(\gamma^2\frac{4}{D_{PFS}} + [\log(HR_{PFS})]^2\sigma_\gamma^2 + \sigma_\gamma^2\frac{4}{D_{PFS}})$$
$$+ 2w(1-w)\gamma\rho_{11}\sqrt{\frac{4}{D_{OS}}\frac{4}{D_{PFS}}}$$

If the weight is the inverse-variance weight,
$$w = (\gamma^2\frac{4}{D_{PFS}} + [\log(HR_{PFS})]^2\sigma_\gamma^2 + \sigma_\gamma^2\frac{4}{D_{PFS}})/(\gamma^2\frac{4}{D_{PFS}} + [\log(HR_{PFS})]^2\sigma_\gamma^2 + \sigma_\gamma^2\frac{4}{D_{PFS}} + \frac{4}{D_{OS}})$$

Appendix II
Conditional Power at Phase II Interim Analysis

Denote t_{OS} and t_{PFS} to be the information fraction of OS and PFS, respectively, at IA in Phase II.

Denote d_{OS} and d_{PFS} to be the observed log(HR) for OS and PFS, respectively, at IA in Phase II.

The second half of the data (from interim analysis to final analysis of Phase II) is
$$T = w(1-t_{OS})\Delta_{OS} + (1-w)\gamma(1-t_{PFS})\Delta_{PFS}$$

In condition power calculation,
$$E(\Delta_{OS}) = d_{OS}, \text{var}(\Delta_{OS}) = (4/D_{OS})/(1-t_{OS}))$$
$$E(\Delta_{PFS}) = d_{PFS}, \text{var}(\Delta_{PFS}) = (4/D_{PFS})/(1-t_{PFS}))$$

$$E(T) = w(1-t_{OS})d_{OS} + (1-w)\gamma(1-t_{PFS})d_{PFS}$$

$$Var(T)$$
$$= w^2(1-t_{OS})^2\ \text{var}(\Delta_{OS}) + (1-w)^2(1-t_{OS})^2\ \text{var}(\gamma\Delta_{PFS}) + 2w(1-w)(1-t_{OS})(1-t_{PFS})\gamma\rho\sqrt{\text{var}(\Delta_{OS})\,\text{var}(\Delta_{PFS})}$$

$$ConditionPower = \Pr(T > C - wt_{OS}d_{OS} - (1-w)\gamma t_{PFS}d_{PFS} \mid E(T), \text{var}(T))$$