Two-stage futility analysis in phase III adaptive trials with time-to-event endpoint

Xiaoyun (Nicole) Li[1] and Cong Chen[1]
[1]Merck Research Laboratory, 351 N. Sumneytown Pike, North Wales, PA, 19454

Abstract

Interim futility analysis is a critical component in oncology phase III studies due to high failure rate in spite of strong efficacy signals observed in phase II trials. For phase III trials with a time-to-event endpoint, a typical interim futility analysis is performed. We propose a two-stage futility analysis in phase III adaptive trials with a time-to-event endpoint to decide whether to continue enrollment or stop the trial completely in the early portion of the study without any requirement on minimum follow-up time. This approach is appropriate if the proportional hazard assumption is valid. However, it is rarely true in practice that proportional hazard assumption holds. We propose a two-stage interim futility design that allows a third option: pause enrollment and wait for mature follow up. To prevent operational challenge and save time, a stage-one futility analysis is performed without requirement on minimal follow-up. In case the futility bar has not crossed, we pause the study and wait for data to mature to make a final decision at stage-two. This approach mitigates the risk of stopping early for futility when it takes time for the study drug to differentiate from the control. We perform a simulation study to illustrate and compare this proposed design with conventional futility designs.

Key Words: Benefit-cost ratio, survival analysis, adaptive design, oncology phase III trials

1. Introduction

The goal of a phase III randomized clinical trial (RCT) is to provide sufficient evidence that the benefit-risk profile of an experimental drug is better than the standard of care. A randomized phase III trial is usually based on hundreds or thousands of patients and takes several years to recruit patients and to complete. At the same time, there is uncertainty on treatment effect prior to conducting a phase III confirmatory trial. For instance, in Oncology, many phase II trials which provide the efficacy proof-of-concept (PoC) for phase III confirmatory trials are single arm studies without comparison and the sample sizes are small. At the same time, early endpoints such as response rate are used in the phase II studies while the ultimate objective in phase III trial is usually to prolong overall survival with the relationship between early endpoints and ultimate endpoint unclear. The use of intermittent endpoints and the inevitable disadvantages of phase II study design have brought in challenges of evaluating the benefit of the experimental drug prior to conducting a phase III trial. Therefore, it is worthwhile considering close monitoring of the study and stop early after examining interim data, to avoid recruitment and additional patient exposure to ineffective treatment. In case the treatment effect of the experimental group is not as good as has expected, the trial should stop early to prevent patients from exposing to ineffective experimental treatment. Stopping early for futility also has other advantages including savings in financial resources and staff, which are important in today's pharmaceutical environment.

The most common approach of futility analysis implementation is group sequential methods (Whitehead and Matsushita, 2003; Lanchin JM, 2009, Pocock, 2006). It uses early data from patients recruited up to a specific time point and make futility decision under the assumption that future data follow the same pattern. However, the assumption has its limitations and may even be scientifically flawed. For example, with time-to-event endpoints, the assumption of proportional hazards is often made but could be violated during the trial. In cancer immune-therapy where response rate is relatively low while the response duration is long, there tends to be no separation in survival between the experimental arm and the control arm at the beginning of the treatment. In this case, the assumption of proportional hazards would no longer be valid. In the pivotal study of ipilimumab in patients with metastatic melanoma (Hodi et al. 2010), the survival curves of ipilimumab and the control arm (gp100) did not separate until after 4 months. There is a much bigger magnitude of survival effect in the later time than that in the first 4 months. Should the futility analysis be conducted at an earlier time with majority of events fall into first four months in the ipilimumab study using traditional futility approach, the study may have stopped for futility though the final result is positive. Jital et al. (2012) performed a retrospective analysis and showed that studies with positive treatment effect may have stopped early for futility using conventional approaches. Another disadvantage of conventional futility analysis approach is that it does not take into account the cost (including patient exposure cost, financial cost and time cost) and the potential benefit (including treatment effect size and market revenue) explicitly in the futility decision. The decision of stopping recruitment for a trial early is a complicated decision (unless there is safety concerns) and should incorporate not only treatment effect size observed up to a specific time point, but also consider the adjusted probability of success (POS) of the experimental drug at the time of the interim analysis, the cost has already spent and additional cost needed for the rest of the study (Jital et al. 2012; Chen and Bechman 2009).

In this paper, we propose a novel two-stage futility analysis design with potential pausing of recruitment to provide opportunity to attain mature data. This would ease some of the issues of falsely stopping the trial early for futility while the treatment effect is indeed significant but delayed. We also incorporate benefit and cost in terms of sample size and explicitly take these into futility consideration. Since overall survival is the most commonly used endpoint in Phase III oncology trials, we will use it throughout this paper. Without loss of generality, the same approach can be extended to other time-to-event endpoints in different therapeutic areas.

2. Proposed Design

We start this section by introducing some concepts and notations. We are interested in a time-to-event endpoint (say, overall survival). Let Δ (>0) denote the target effect size in the phase III confirmatory trial (log hazard ratio scale), α denote the overall type I error rate of the trial and β denote the type II error rate of the trial without incorporating any interim analysis. Therefore, the number of events needed to detect treatment effect Δ with type I error and type II error doublet (α, β) is $D = 4 * \frac{(Z_\alpha + Z_\beta)^2}{\Delta^2}$. We can further assume a constant enrollment rate, total enrollment period and study period and an exponential failure rate in the control arm. Therefore, sample size needed to reach D events can be calculated (Lachin J. M and Foulkes M. A. 1986).

For a phase III oncology trial, type I error rate is usually set at 2.5% level (one-sided). Consider a phase III study in the first line non-small cell lung cancer (NSCLC) setting, where a 25% hazard reduction in survival is clinically meaningful. We design a study to have 90% power (type I error rate 10%) to detect a 25% hazard reduction at 2.5% type I error level, without adjusting for interim analysis. To allow flexibility in decision making, the futility rules are set up as non-binding, i.e., whether to follow the futility rules or not would not inflate the overall type I error. This would provide flexible decision makings.

2.1 Optimal interim futility bound

We consider only one interim analysis for the study. Let *t (0<t<1)* be the portion of information at interim analysis. Let $(\alpha_{iA}, \beta_{IA})$ be the type I error rate and type II error rate at the interim analysis. After incorporating the interim analysis, the overall power of the study is:

$$1-\beta^* = \Pr(X_{IA} > Z_{1-\beta_{IA}}, X > Z_{1-\beta})$$

Let C_3 denote the total cost of the phase III trial, and C_{IA} denote the cost that is spent up to the interim analysis. We assume that the experimental arm has a prior probability p of being active with treatment effect Δ and probability $1-p$ being inactive. In many cases, p is also noted as probability of success (POS). This assumption is straightforward and we usually use the industry benchmark for estimating p. To be more sophisticated, we can incorporate a prior distribution of the drug being active based on information from previous studies, instead of a single value p, which we will not elaborate in detail in this paper. Chen and Beckman (2009) proposed an optimal futility boundary that maximizes the expected benefit per expected resource unit expended, i.e., benefit-cost ratio. The optimization is performed by maximizing the utility function of benefit-cost ratio, where the numerator is the expected power of getting a positive phase III result adjusted for probability-of-success (POS), multiplied by the benefit of the drug; and the denominator is the expected phase III trial cost, adjusted for possibly stopping for futility. Let *B* denote benefit (e.g., size of patient population or market revenue as appropriate). The benefit-cost ratio *R* can be expressed as the following:

$$R = \frac{Bp(1-\beta^*)}{C_{IA} + (C_3 - C_{IA})[p(1-\beta_{IA}) + (1-p)\alpha_{IA}]}.$$

Since *B* depends on factors other than the phase III trial itself, such as the landscape and competitions from other similar drugs in development, it is hard to estimate upfront. However, it doesn't play any role in deriving the futility boundary and we can treat it as a nuisance parameter.

Chen and Beckman (2009) have discussed in length in their paper how *p* and *t* could impact the futility bound and type II error spent at the interim analysis and the robustness of the futility bound. Since the futility bound is set up based on the assumption that future data follow same pattern as observed, we assume the futility rules using the optimal benefit-cost ratio are based on mature survival follow up data.

2.2 Two-stage futility design

We set up the two-stage futility analysis design as the following: A futility bar is set up based on mature follow-up data at d_2 events, using Chen and Beckman's (2009) approach. Based on this futility bar, a stage-one futility analysis is conducted when a portion of the d_2 events have occurred. At this stage-one futility analysis, we do not require sufficient follow up time, i.e., the data is preliminary comparing to data with mature follow up. There are two possible outcomes of this stage-one futility analysis: If the data show high confidence that the study would pass the futility bar based on d_2 events, the study will continue enrollment without stage-two futility analysis; If the data is not able to show high confidence of passing the futility bar at d_2 events, the enrollment will be paused and the study will be on hold, and a stage-two futility analysis will be conducted after data become mature with d_2 events. At the stage-two futility analysis, the decision is to either continue enrollment or to stopping the study completely.

Conditional power is used for decision making at stage-one futility analysis. Let θ denote the treatment effect of interest and Z_1 the test statistic at stage-one futility analysis. Conditional power can be formulated as

$$CP(Z_1, \theta) = \Pr\{\text{Pass stage-two futility analysis} \mid Z_1\},$$

where θ is the target treatment effect. Depending on how stringent and how much confidence is required at the stage-one futility analysis, different conditional power thresholds to pause enrolment can be set up for the stage-one futility bound. Simulations are recommended and operating characteristics should be provided in order to select an appropriate bound. As an illustration, we use at least 70% conditional power of passing the stage-two futility analysis under observed treatment effect as our the criteria of continuing enrollment at stage-one futility analysis.

3. An Example: A Phase III Oncology Study

In this section, we consider an example of a phase III oncology pivotal study. A total of 660 patients will be enrolled with a 1:1 ratio into the experimental arm and the control arm. The study will complete when 510 events have occurred. With this sample size, the study has 90% power (without accounting for interim analyses) to detect a hazard ratio of 0.75 (25% hazard reduction) when controlling the type I error at 2.5% (one-sided). Assume the median survival in the control arm is 6 months, the study would require enrollment duration of 22 months and a minimum follow up of 6 months. We assume 20% initiation cost, i.e., 20% resource is spent prior to the study start. The rest of 80% resource is spent uniformly on each of the patient enrolled. Suppose a futility interim analysis is conducted at 30% information. It is projected that approximately 60% of the total cost has been spent at the 30% information time. There are two approaches to attain 30% deaths: the first approach is to enroll approximately 50% patients and to pause enrollment and follow up them until 30% deaths have occurred. The second approach is to continue enrollment until the target number of deaths have occurred. The latter approach is commonly used since it is more operational feasible and would not jeopardize study timeline. However, the trade off is that the data at interim analysis may have different pattern from the future data. In the two-stage futility analysis, we allow the use of preliminary data without mature follow up to make an earlier decision. Therefore, it provides a middle ground between the two approaches. We assume 60% and 70% of total cost has been spent prior to the interim analysis under the first and second approach,

respectively. Table 1 shows the optimal futility boundaries at two-stage interim analysis after 60% cost has been spent.

Table 1 Optimal futility boundaries at two-stage interim analysis after spending 60% of total cost

POS (p)	Information at stage-one interim (t_1)	Information at stage-two interim (t_2)	Futility bound at stage-two interim (hazard ratio)	Futility bound at stage-one interim (hazard ratio)	Empirical bound at final analysis (hazard ratio) for a positive trial
0.3	0.2	0.3	0.93	0.88	0.84
0.5	0.2	0.3	0.95	0.91	0.84
0.7	0.2	0.3	0.98	0.93	0.84
0.3	0.3	0.4	0.90	0.86	0.84
0.5	0.3	0.4	0.92	0.89	0.84
0.7	0.3	0.4	0.94	0.90	0.84

4. Simulations

In this section, we conduct simulations to evaluate the characteristics of the two-stage futility interim analysis design.

We consider four different futility designs: 1) Two-stage futility analysis. A preliminary futility analysis is conducted without minimum follow up when 20% information has achieved. If there is more than 70% confidence (in terms of conditional probability) that our data will pass the futility analysis with mature follow up when there is more information, the study will continue without pausing enrollment. If the confidence level is less than 70%, the study would pause, patients will continue survival follow up and the stage-two interim analysis will be conducted when 30% events have occurred. 2) Conventional futility analysis. The futility is conducted when 30% events occurred. No minimum follow-up time required. 3) Conventional futility analysis. The futility is conducted when 20% events occurred. No minimum follow-up time required. 4) No futility interim analysis and the study would continue until the end.

We use the same data set up as that in the example. A total of 660 patients are randomized into either the experimental arm or the control arm with 1:1 ratio. An enrollment period of 22 months and a minimum follow up of 6 months after enrollment completion is needed. The target hazard ratio is 0.75. The study is completed when 510 deaths have occurred.

In the first scenario, we simulate data under proportional hazard assumption. True hazard ratio (HR) in the data simulations is set as 0.7, 0.75, 0.8, 0.9 and 1, respectively. We evaluated each of the four designs on the simulated data and decide whether the study will be stopped early for futility or continue till the end, and whether the study result will be positive. We performed 1000 simulations and summarized the probability of achieving positive result at the end of the study in order to compare the operating characteristics of four designs. We also evaluate the relative cost in each design. Table 2 and Table 3

provide the probability of positive result and the relative cost under different scenarios, respectively. We can see that when the proportional hazard assumption is valid, all three interim analysis designs perform similarly well.

Table 2 Probability of positive trial under different hazard ratios (based on 1000 simulations)

True HR	Two-stage futility analysis	One-stage futility analysis at 30% info	One-stage futility analysis at 20% info	Without futility analysis
0.7	0.92	0.93	0.91	0.98
0.75	0.85	0.85	0.83	0.90
0.8	0.63	0.64	0.61	0.71
0.9	0.19	0.19	0.19	0.22
1	0.023	0.024	0.023	0.026

Table 3 Expected cost relative to total trial cost under different hazard ratios (based on 1000 simulations)

True HR	Two-stage futility analysis	One-stage futility analysis at 30% info	One-stage futility analysis at 20% info	Without futility analysis
0.7	0.96	0.97	0.96	1
0.75	0.9	0.91	0.91	1
0.8	0.73	0.75	0.72	1
0.9	0.40	0.44	0.40	1
1	0.38	0.43	0.36	1

Secondly, we consider the case that survival curves between the experimental arm and the control arm is unlikely to separate at the beginning but have large separation at the later portion of the study. Let λ be the hazard ratio between the experimental arm and the control arm. We let $\lambda=1$ for the first four months, i.e., represents no treatment difference in the first 4 months. We let $\lambda=1$, 0.8, 0.7, 0.6 respectively, from month 5 to the end. When $\lambda=1$, it represents the scenario that there is no treatment effect. When $\lambda=0.8$, 0.7 or 0.6, it represents the scenarios where the study is positive. In Table 3, we can see that the type I error is comparable across four designs and when the true hazard ratio is 0.7 or 0.6 after four months, our proposed futility design would gain more power. Table 4 shows the expected cost in each design. The relative costs among three interim designs are similar, while without interim analysis, we always have to spend the total cost.

Table 4 Probability of positive trial under different hazard ratios (based on 1000 simulations)

True HR	Two-stage futility analysis	One-stage futility analysis at 30% info	One-stage futility analysis at 20% info	Without futility analysis
0.6	0.7	0.65	0.63	0.83
0.7	0.46	0.43	0.4	0.56
0.8	0.2	0.2	0.2	0.26
1	0.02	0.021	0.02	0.024

Table 5 Expected cost relative to total trial cost under different hazard ratios (based on 1000 simulations)

True HR	Two-stage futility analysis	One-stage futility analysis at 30% info	One-stage futility analysis at 20% info	Without futility analysis
0.6	0.81	0.84	0.80	1
0.7	0.65	0.70	0.65	1
0.8	0.49	0.53	0.5	1
1	0.39	0.50	0.47	1

5. Conclusion

Current futility analysis with time-to-event endpoint performs futility evaluations using study data at a snapshot of the data under the assumption that the future data follows same pattern as the observed data. There are limitations especially when proportional hazards assumption is in doubt. In this paper, we proposed a two-stage futility analysis which gives us an opportunity to look at the data at a later time when there is mature follow up information. Simulations showed that when proportional hazard assumption is violated, the two-stage futility analysis approach can save cost and have higher power comparing to other approaches. At the same time, we build in the optimal boundary in the two-stage futility analysis so that the return on investment is optimized after explicit incorporation of the drug development assumptions, such as benefit, cost and probability of success.

References

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

Chen, C, Beckman RA. Optimal cost-effective designs of proof of concept trials and associated Go-No Go decisions. Proceedings of the American Statistical Association, Biometrics Section, (2007).

Hodi, F. S. et al. Improved Survival with Ipilimumab in Patients with Metastatic Melanoma, *New England Journal Medicine*, 8: 711-723, 2010.

Jitial, M, Khan, I, Lee, SM. and Hackshaw, A., Stopping clinical trials early for futility: retrospective analysis of several randomized clinical studies. British Journal of Cancer 107: 910-917, 2012.

Lachin J. M. Futility interim monitoring with control of type I and II error probabilities using the interim Z-value or confidence limit. Clin Trials 6: 563-573, 2009.

Lachin, J. M. and Foulkes, M. A.. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up noncompliance, and stratication. Biometrics, 42:507-519, 1986.

Pocock, S. J. Current controversies in data monitoring for clinical trials. *Clin Trials* 3:513-521, 2006.

Whitehead. J and Matsushita, T. Stopping clinical trials because of treatment ineffectiveness: a comparison of a futility design with a method of stochastic curtailment. *Statistics in Medicine* 22 (5): 677-687, 2003.