Paper # 2010-14

Impact of Model Mis-Specification on Conditional Power in Interim Survival Analysis

Chao Chen, PharmaNet Development Group, Inc., Princeton, NJ

ABSTRACT

For survival analysis used in clinical trials and other types of experiment, constant hazard ratio model is commonly used to compare the survival rate of active treatments with that of placebo group. If active treatments do not become effective immediately after treatment initiation, constant hazard model is not exactly applicable. Nevertheless, with the expectation that as the trial duration becomes longer, the effect of non-constant hazard ratio at the initial stage of the trials will diminish and becomes negligible. Even this argument seems intuitively acceptable for a long-term trial, in an interim analysis with data collected and analyzed during the course of the trial, impact of non-constant ratio may not be ignored, especially when the interim statistical results is to be used for the calculation of conditional power, a projected probability of final study outcome based on existing dataset. SAS procedure PHREG is a versatile tool to model non-constant hazard ratio where there are time-dependent covariates in the model. We used this tool to evaluate the impact of modeling a non-constant hazard ration as constant. Simulated data from non-constant hazard ratio models with different censoring time will be presented to demonstrate the impact of model mis-specification in the calculation of conditional power.

INTRODUCTION

The motive of this study originates from the statement “all models are wrong, but some are useful” from Box [1]. In occasions that interim analysis can be performed at various stages of the trial or experiment, the same model can be applied repetitively with various level of validity. This viewpoint is exemplified in the current study using pseudo random variables with matched or mismatched models in an interim survival analysis setup. Following a review of related theory, various datasets analyzed by compatible or incompatible SAS codes will be presented. The results indicated that even the mismatched model can have an acceptable power for final analysis; the power can be drastically distorted in the early-stage interim analysis.

THEORETICAL BACKGROUND

Theory from survival analysis, random walk model, and conditional power are summarized in this section for future reference. For a general review of conditional power, please refer to Lachin [2, 3].

MODEL FOR HAZARD RATIO

In the analysis of survival time data, hazard (λ) at time *t* is defined as the conditional probability that death or other endpoint event will occur at the next time unit when the study object is still alive at time *t*. For continuous time, it can also be heuristically explained as the instantaneous risk at time *t*. When there are two or more study groups involved in the study, the ratio of the hazard from various groups can be used to evaluate the difference between groups. In additional to treatment groups, other covariates with potential influence on the hazard ratio can be included in a Cox proportional model for the evaluation of covariates effect. Hazard ratio can be treated as constant or non-constant (time-dependent) across the study period. The focus of current study is on the situation when hazard ratio varies over time, and the only covariate considered is treatment group. Statistical test results from a case when the true hazard ratio is non-constant but was treated as constant will be examined and compared with cases when model is correctly specified.

RANDOM WALK MODEL

As in the hazard model, there is a time component involved in the theory of random walk model. Unlike the survival analysis, time in the random walk model can be rescaled as ‘information time. Loosely speaking, it is the amount of information collected during a trial or experiment. The information time ranges from 0 to 1, with 0 means the start and 1 the end of the trial.

In a discrete-time model, a particle at position 0 moves up (assigned as 1) or down (assigned as -1) with certain pre-specified probabilities at each discrete time point *t = 0, 1, 2, 3, … n*. Suppose the movement in the next step is independent of previous movement history and with equal probability (one half) of moving up or down. The variance of the position at time *n* can be viewed as the variance of the sum of *n* independent binary random variables with equally possible outcome of 1 or -1, which is just the sum of individual variance at each step. In other words, the variance of the final position from origin is proportional to the time the particle travelled. The longer time the particle travels, the larger the variance will be. This is also true for continuous time model. If the particle moves up or down with unequal probability, this is random walk with a drift, that means the particle has a tendency to move to one direction. In a random walk model without drift, the *expected* position of a particle at any time is always 0, however, the conditional expectation of the particle location at the next move is just the current position. In a random walk model with drift, since there is a tendency to move to one side, the expectation at the end is non-zero. The expectation value of the end position is called drift parameter (or non-centrality parameter in statistical terms). A plot of particle location vs. time is called a sample path.

Distribution theory regarding the position of the particle can be used to calculate probabilities related to interim statistical analysis.

TEST STATISTICS AS A RANDOM WALK

In a typical clinical trial, patients are enrolled sequentially, receive treatment according to a pre-determined randomization code list, patients are monitored for clinical outcome measurement or study endpoint. Theoretically, statistical test procedures can be repetitively performed with data from each patient added. Under regular statistical assumptions (including i. i. d. samples in the parametric model) and with certain rescaling, the series of test statistics from hypothesis testing with each additional patient added can be viewed as a sample path follows the property of random walk model.

To be more specific, we have the following:
For a test procedure (test of equal hazard in our study), under the null hypothesis (H_0) that the difference is zero between two treatment groups, the test statistics *S* follows a normal distribution with mean zero and variance σ_0^2 / n, where σ_0^2 is the known variance and *n* is the sample size. The rescaled test statistics $Z = S \cdot \sqrt{n} / \sigma_0$ follows a standard normal distribution. Reject the null hypothesis H_0 when $Z > Z_{1-\alpha}$ will provide a test procedure with type I error at α which is usually set as 0.05. When this general statistical argument applies to interim analysis performed at information time $t \in (0,1)$, we have the following:

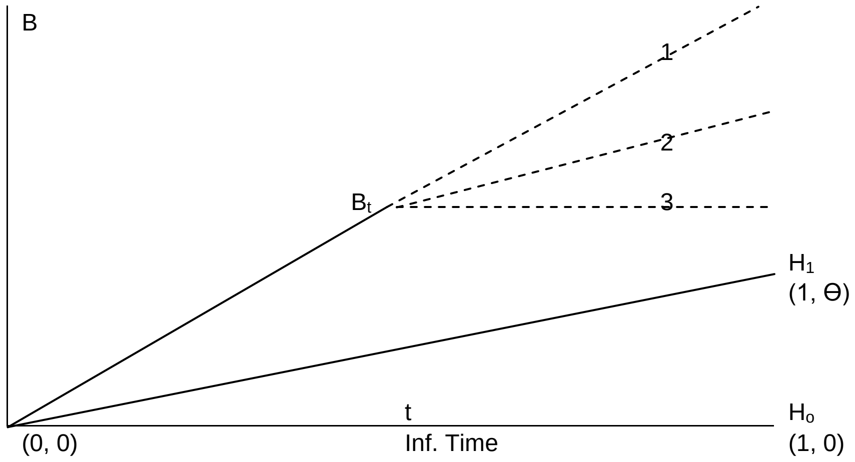
$$B_t = Z_t \cdot \sqrt{t} \sim N(0,t)$$

under H_0 . Notice that B_t is the standardized test statistics Z_t rescaled by a factor equal to square root of information time, which is the proportion of sample size used in interim analysis as compared with the final sample size for fixed sample size trial with no censored observations, or the proportion of number of endpoint events observed at the time of interim analysis vs. total number of endpoint events planned for fixed endpoint trial with censored observations.

The reason that random walk theory is applicable to statistical testing is that B_t , derived for the test statistics, shares the property of random walk in the sense that the variance is proportional to information time. At the end of the study when all the data are collected for analysis, i. e., when *t = 1*, B_t is just the standardized test statistics which follows standard normal distribution with mean zero when the null hypothesis is true. When the null hypothesis is not true, in other words, there is a difference between two groups, B_t is still normal with variance *t*, but the mean becomes $t \cdot \theta$ where θ is the drift (non-centrality) parameter. This is comparable to the case of random walk with a drift parameter equivalent to the expected location at the end of random walk.

In reality, interim analysis is performed only at certain information time points *t*, not all the information time point *1/n, 2/n, 3/n, … n/n = 1*. When B_t are calculated at all or certain selected information time *t*, a sample path can be created by plotting B_t vs. information time *t*. From a theoretical viewpoint, the expected sample path should be right on the straight lime between (0, 0) and (1, 0) under the null hypothesis H_o , under the alternative hypothesis H_1 , the expected final point is (1, θ) as shown in the solid straight lines in Figure 1. The straight horizontal line or the linear increase trend results derived from random walk theory can be affected by various reasons including model mis-specification investigated in the current study.

Figure 1: Expectation of sample path under three scenarios 1, 2, and 3



CONDITIONIAL POWER DERIVED FROM RANDOM WALK MODEL

Conditional power is the probability of detecting the true difference given the rescaled interim test statistics B_t . As the final outcome of hypothesis testing depends on $B_1 = Z_1$ and at the information time *t*, B_t is the *realized* part of B_1, the final outcome actually depends totally on the *unrealized* part $U_t = B_1 - B_t$. Under the regular assumption of independent trial and the fact that the sum of two independent normal random variables (B_t and U_t in the current case) is again normal with mean and variance equal to the sum of the corresponding components, it can be verified that

$$U_t = B_1 - B_t \sim N(\theta \cdot (1-t), 1-t)$$

The calculation of the conditional power depends on the conditional mean of the drift parameter θ at information time *t*, which is equal to the sum of realized part B_t and the conditional mean of the un-realized part U_t . The predicted trajectory of the future sample path is usually based on one of the three scenarios and presented as dotted line in Figure 1.

1. The current trend of the data collected by interim analysis will continue in the future analysis. This will be the scenario used in of the calculation of conditional power in the following sections.
2. The future outcome will be generated from alternative hypothesis H_1, therefore, the slope of the projected sample path will be the same as the slope under H_1.
3. Disregard the data collected from interim analysis, the null hypothesis will still govern the future outcome, the projected future sample path will be horizontal.

A graphical representation of the three scenarios was shown in Figure 1. In terms of the utilization of existing data, scenario 1 and 3 are at the extreme side of each other. Using both approaches provides the scope of possible future study outcomes. Scenario 2 is in the middle of these two extremes, empirical Bayesian approach can also be used to find a balance between these two extreme scenarios.

NUMERICAL SETUP AND RESULTS

In the current study, we consider two sets of data, one from constant hazard ratio model and the other from non-constant model. Similarly, SAS codes can be developed with or without the consideration of non-constant hazard ratio. In all, there are four possible combinations of data and SAS codes:

1. constant ratio data - constant ratio SAS code,
2. constant ratio data – non-constant ratio SAS code,
3. non-constant ratio data – constant ratio SAS code,
4. non-constant ratio data – non-constant ratio SAS code.

Combinations 2 and 3 represent mismatch between data and SAS code. The goal is to verify computational results with the theoretical development based on theory summarized in previous sections when data and code match, and to quantitatively evaluate the effect of mismatch (model mis-specification) when interim data were used to project final study outcome. Combination 1 is a simpler case that pretty much follows the theory so the focus is on non-constant model.

CONSTANT AND NONCONSTANT HAZARD RATIO

Two cases, one constant hazard, the other one non-constant hazard were considered in our simulation study.

Case 1:

$$\lambda_a = 0.166,$$
$$\lambda_b = 0.042$$

Case 2:

$$\begin{aligned}\lambda_a &= 0.083, \\ \lambda_b &= 0.166\, if \;\; t < 2 \\ &= 0.042\, if \;\; t \geq 2\end{aligned}$$

where *a* and *b* are the two groups to be compared. For each of the two cases, 300 pseudo random variables were generated for each treatment group. Kaplan-Meier estimates were calculated and plotted, the results were shown in Figures 2 and 3.

As expected, group b has a higher survival rate throughout the time span in Case 1. As for Case 2, since the hazard rate is higher for group b when *t < 2*, the higher survival rate for group b was not observed until *t* is greater than 7. Case 2 is a case that the risk of group *b* is actually twice of group *a* when *t < 2*, however, the risk of group *b* is only half that of group *a* when *t* is equal to or greater than 2. The cut-off point at *t = 2* was artificially set to investigate the effect of model mis-specification.

There is no guarantee that the hazard ratio will remain constant across the complete time span. In clinical trials, the full treatment effect may not occur at the initial stage [4]. In a comparison study between a surgical group and a medically treated group, there can be an initial stage of higher risk before a lower risk can be achieved in the surgical group. In both situations, Case 1, a typical constant hazard model, though can be easily coded, does not reflect the actual underlying mechanism the data came from. As long as the underlying data-generating mechanism does not deviate too much form the model, the assumption of constant hazard rate is sometimes ignored. It is the situation when drastic changes in hazard ratio, or even a reverse trend as shown in Case 2 that deserves more attention. In the following sections, SAS codes taking non constant hazard ratio into consideration was developed, test statistics for the null hypothesis of equal hazard ($\lambda_a = \lambda_b$) was calculated using fixed censored time set at t = 1, 2, 4, 8, 16 and 32.

Figure 2: Survival Curve for simulated data from Case 1.

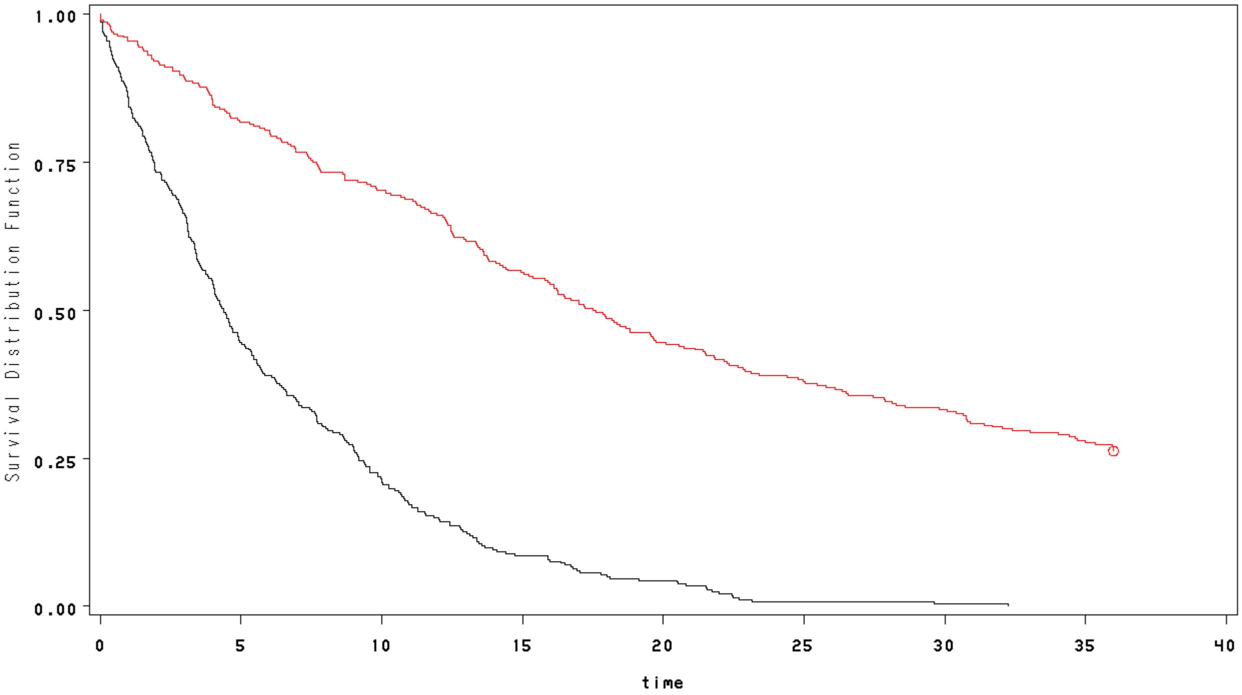
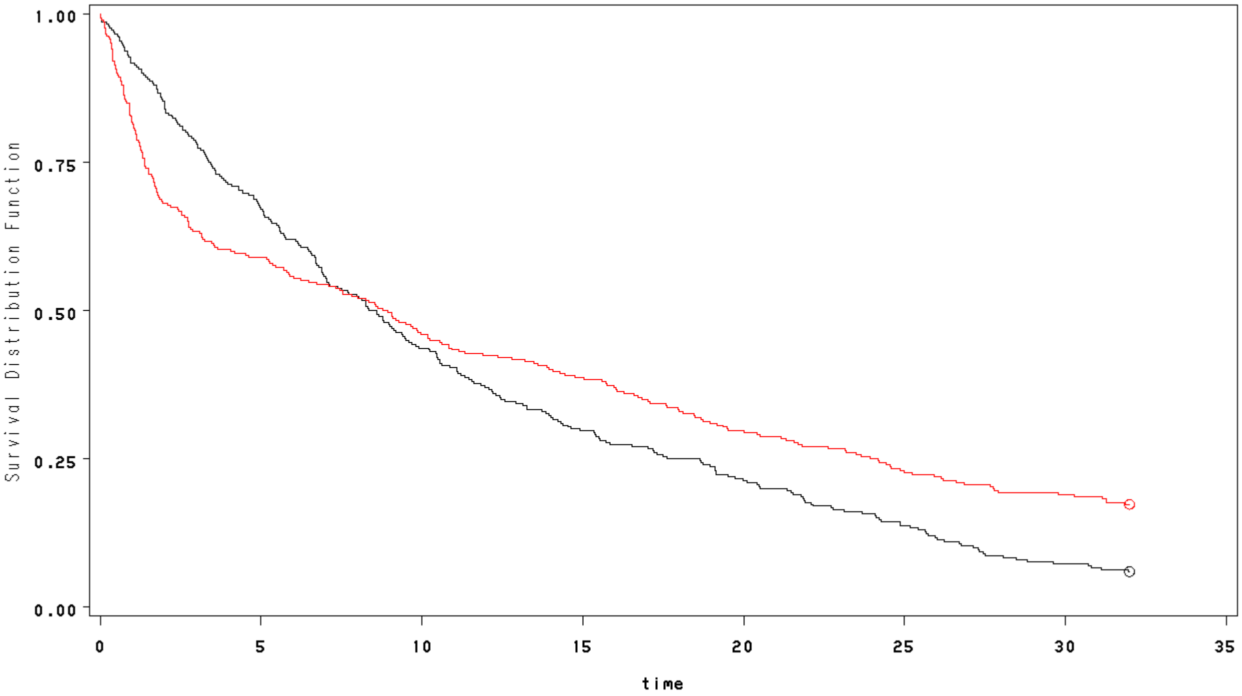


Figure 3: Survival Curve for simulated data from Case 2



SAS PHREG CODE FOR HAZARD RATIO

The very basic SAS PHREG code is as follow:

SAS Code 1:

```
   proc phreg data = … ;
   model time*flag(0)=treatment;
   run;
```

where treatment is either *a* or *b*. This simple approach has an implicit assumption that hazard ratio are constant. With some minor modifications, SAS PROC PHREG can handle non-constant hazard ratio models. If the cut-off time point $t = 2$ can be treated as *known* as in our study, the SAS code are a follow:

SAS Code 2

```
   proc phreg data = … ;
   model time*flag(0)=effect;

   if treatment = a then effect = 0;

   if treatment = b and time lt 2 then effect = 1;
   if treatment = b and time ge 2 then effect = -1;
   run;
```

The code specifies risk ratios before and after $t = 2$ are just reciprocal of each other though the magnitude was treated as unknown. In terms of regression coefficient in the Cox model, they are the negative value of each other. As shown in the code, the effect, either 1 or -1, is treated as a time-dependent covariate. The corresponding regression coefficient estimate can be converted into hazard. All the typical SAS statements in SAS DATA step can be used after the model statement to accommodate more general and complicated situations. Our hypothetical example is artificially chosen to provide a clear trend of model mis-specification impact on hypothesis testing. The resulting parameter estimates, including the estimates of regression coefficient β in Cox model, hazard ratio and p-values of test statistics can be generated from standard SAS output and are summarized in Table1.

Table 1

Censor time		2	4	8	16	32
Estimate of ß	Code 1 Code 2	-0.82 0.82	-0.46 0.68	-0.13 0.70	0.09 0.67	0.20 0.65
Hazard ratio	Code 1 Code 2	1/2.28 2.28	1/1.59 1.97	1/1.13 2.02	1/0.91 1.96	1/0.82 1.96
P for equal hazard ratio	Code 1 Code 2	<0.001 <0.001	0.001 <0.001	0.288 <0.001	0.36 <0.001	0.02 <0.001

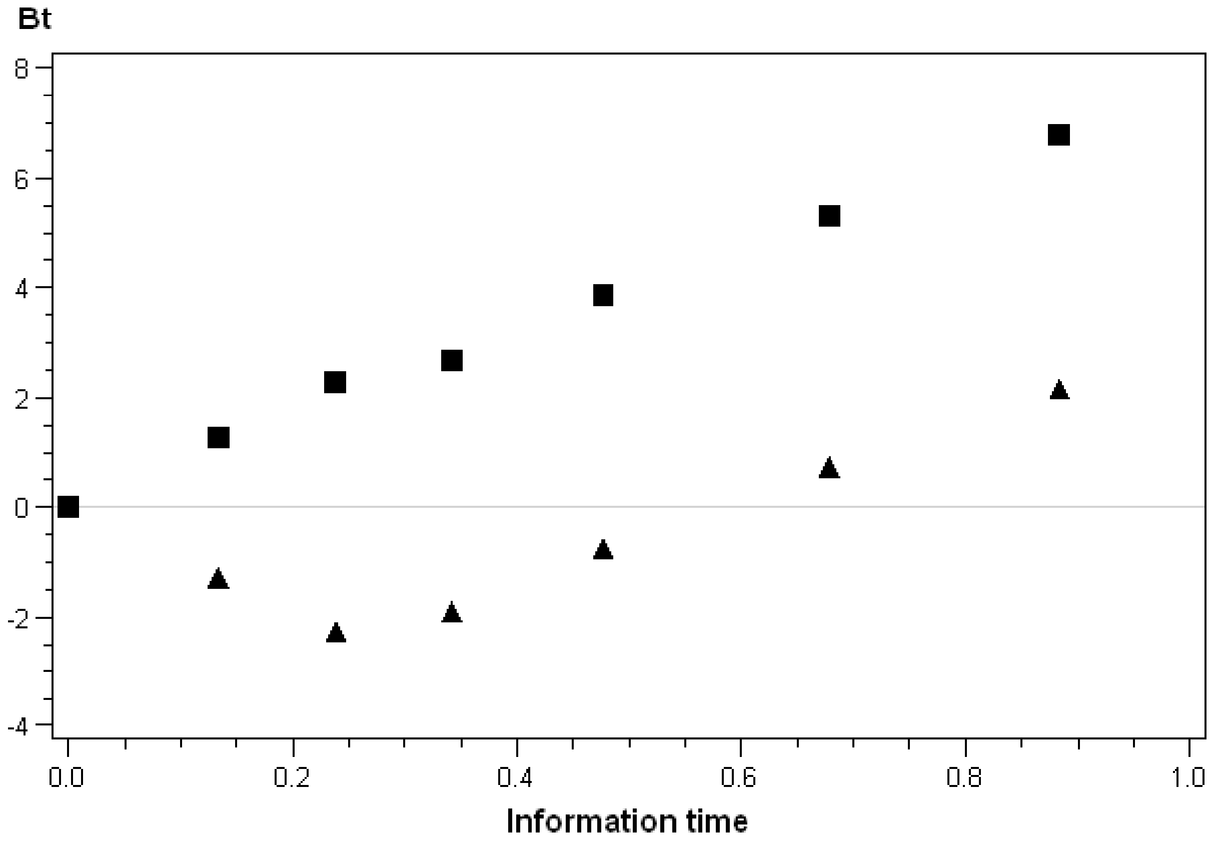
The results from Code 2 must be explained with care. Though the SAS code treat group a as baseline, as there is an effect statement in Code 2 to reverse the treatment effect, hazard ratio presented as 2.28 in the table is equivalent to using group b as baseline so a pretty stable hazard ratio around was achieved in Code 2. Results for information time *t*, Z_t and B_t derive from the random walk model are summarized in Table 2 below.

Table 2

Censor time		2	4	8	16	32
Inf. Time t (# event/N)		143/600	205/600	286/600	407/600	530/600
Z_t : Est of ß / SE	Code 1 Code 2	-0.82/0.18 0.82/0.18	-0.46/0.14 0.68/0.15	-0.13/0.12 0.70/0.13	0.09/0.10 0.67/0.10	0.20/0.09 0.65/0.09
B_t : Z_t * sqrt(t)	Code 1 Code 2	-2.28 2.28	-1.91 2.68	-0.73 3.86	0.74 5.29	2.15 6.79

Plot of B_t vs. information time from Table 2 was shown in Figure 4.

Figure 4: Plot of B_t vs. information time



When the SAS Code 2 is applied to data generated from the compatible Case 2, numerical results above are pretty consistent with the random walk theory, that is, the sample path indicated by the solid square (■) closely follows a straight line, this reflects the fact that expectation of B_t is $t \cdot \theta$. On the other hand, the sample path under the mis-specified constant hazard model, as presented by the solid triangle ▲ has an initial dip followed by a late-stage surge. Though not included in this report, the linear trend is also observed when SAS Code 1 was applied to Case

1, a further indication that the linear trend is dependent on whether the model matches the data or not, not the model alone.

DISCUSSION AND CONCLUSION

Interim analysis based on data collected up to information time less than 0.5 must be examined with great care, especially when the interim results is to be used to project final study outcome when the information time reaches 1. Conditional power based on the assumption that current trend will continue until the end of the study can be used for the calculation of conditional power. However, as illustrated in Figure 4. Current trend can be distorted by model mis-specification which in terms leads to the wrong conclusion.

To demonstrate this viewpoint, the analysis using the mismatched SAS code and data from Case 2 in the previous section was repeated 100 times. An overall of 60000 pseudo random variables from the two treatment groups in Case 2 were generated and divided equally into 100 sets of 600 observations. Survival analysis using the incorrectly specified constant hazard model was repeated for these 100 data sets, all the associated estimates, test statistics, the corresponding values from random walk models were calculated. With the classical geometrical argument applied to Figure 1, estimates of the drift parameter $\widehat{\theta}$ under scenario 1 (continuation of existing trend to the end of the study) can be calculated as

$$\widehat{\theta} = B_t + (\frac{1-t}{t}) \cdot B_t = B_t \,/\, t$$

As $\widehat{\theta}$ has a conditional variance of (*1-t)*, the rescaled quantity $\widehat{\theta} \,/\, \sqrt{(1-t)}$, follows standard normal distribution, this result can be used for hypothesis testing. The null hypothesis of equal hazard ratio is rejected when the absolute value of this rescaled value is greater than 1.96, and significant difference in hazard ratio in the two groups is claimed. Significant difference was observed in only 33 out the 100 times at the fixed censoring time = 8 (information time approximately 0.5). Similar calculation shows when the fixed censoring time is set as 16 (information time approximately 0.66), 94 out of the 100 simulated cases are projected to be significant. When the fixed censoring time reaches 32, which is about at the information time of 0.9, 99 out of the 100 cases are significant by the actual test statistics without using conditional power.

The example above indicates that even with the incorrectly specified model, a reasonable empirical power (99%) can be achieved when the information time approaches 1. On the other hand, there is a sharp drop of projected empirical power to only 33% based on test result at information time of 0.5. Termination of a trial based on lack of conditional power during interim analysis can be a misguidance when the underlying model is not correctly specified. Various assumptions regarding the future sample path as mentioned in earlier sections, together with model checking tools, which is not the topic of this current study, should be used to guard against such mistakes.

We do want to point out that fixed censoring time is purposely chosen in the study with the expectation that it will accentuate the deviation from linear trend under model mis-specification. A plausible conjecture is that when patients were enrolled over a longer time span, the untoward effect of deviation for linear trend will be alleviated. Fixed censoring time seems to have a more drastic effect, but further study is needed to confirm this conjecture. Certainly there are other factors that will push the expected sample path away from a straight line. These factors, including homogeneity of the patients enrollment, delayed treatment effect as discussed by Lakatos [4], study population due to amendment of the trial procedures, seasonal and geographical effects, to name just a few, may intertwined and the net effect can be difficult to isolate. Some of these factors are controllable (such as randomization procedure to assure the homogeneous patient enrollment), some are not (such as delayed treatment effect). Whenever possible, clinical trials should be conducted with the mindset to minimize the digression from statistical assumptions. If this is not feasible, statistical assumptions and the resulting statistical method, programming code can be revised to reflect the true data pattern. This practice is especially important when interim data analysis result is to be used for the projection of final study outcome.

REFERENCES

1. **Box, GEP; Norman, RD.** *Empirical Model-Building and Response Surfaces*. Wiley, 1987
2. **Lachin, JM,** Introduction to Sample Size Determination and Power Analysis for Clinical Trials. *Controlled Clinical Trials* 2: 93-113, 1981.

3. **Lachin, JM,** A review of methods for futility stopping based on conditional power. *Statistics in Medicine* 24: 2747-2764, 2005
4. **Lakatos, E,** Designing complex group sequential survival trials. *Statistics in Medicine* 21: 1969-1989, 2005.

ACKNOWLEDGMENTS

The author would like to acknowledge Mary Johnson and Tricia Yeh from PharmaNet Development Group, Inc. for bringing the issue of conditional power to the author’s attention and their support of this study.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chao Chen
Director of Biostatistics
PharmaNet Development Group, Inc.
504 Carnegie Center
Princeton, NJ 08540
Phone: 609-951-6783
Email: chchen@pharmanet.com
Website: www.pharmanet.com