Abstract

KUNG, MEIFEN. Information-based Group Squential Tests With Lagged Or Censored Data. (Under the direction of Professor Anastasios A. Tsiatis.)

Conventionally, values of nuisance parameters given in a statistical design are often erroneous, thus may result in overpowering or underpowering a test using traditional sample size calculations. In this thesis, we propose to use Fisher Information data monitoring in group sequential studies to not only allow an early stopping in a clinical trial but also maintain the desired power of the test for all values of nuisance parameters. Simulation studies for the simple case of comparing two response rates are used to demonstrate that a test of a single parameter of interest with a specified alternative achieves the desired power in information-based monitoring regardless of the value of the nuisance parameters, provided that this parameter of interest can be estimated efficiently. The emphasis in this part is to show how information-based monitoring can be implemented in practice and to demonstrate the accuracy of the corresponding operating characteristics in some simulation studies.

When there is lag time in reporting, standard statistical techniques often lead to biased inferences on interim data. A maximum lag estimator ensures complete information by using data before a lag time period. The estimator is unbiased but less powerful. We propose an inverse probability weighted estimator which accounts for censoring and is consistent and asymptotically normal in estimating mean of dichotomous variables. The joint distribution of test statistics at different times have the covariance structure of a sequential process with independent increments. This allows the use of information-based monitoring. Simulation study shows that our estimator preserves the type I and type II errors, and reduces the number of participants required in a trial. Future approach in finding an efficient estimator is also suggested in chapter 3.

INFORMATION-BASED GROUP SEQUENTIAL TESTS WITH LAGGED OR CENSORED DATA

by

MEIFEN KUNG

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

STATISTICS

Raleigh

2000

APPROVED BY:

Chair of Advisory Committee	

To my parents and grandparents

Biography

Meifen Kung was born in Taipei, Taiwan, to parents Ken-Tang Kung and Gui-Huan Tsui on June 10, 1970. After receiving a B.S. in statistics from Tamkang University in June 1992, her curiosity propelled her to seek higher education abroad. She went to Michigan State University for graduate school and had great pleasure in seeing snow in winters. She received a M.S. in May 1996. To serve her interest in applied fields, Meifen then joined the statistics program at North Carolina State University and loved the design of the program. Upon completion of her Ph.D. in May 2000, she will return home to Taipei, Taiwan, where she will work at the Center for Drug Evaluation.

Acknowledgements

I always consider myself a very fortunate person. People offered great advices and companionships at each stage of my life. I can not acknowledge them all but I am forever grateful at heart.

To my advisor Dr. Anastasios A. Tsiatis, I can not thank you enough for taking me under your supervision, leading me through my interest in applied fields, providing guidance when I needed yet giving me plenty of freedom in trying my hands. It's a great pleasure working with you.

To my first year academic advisor Dr. Sastry Pantula, thank you for all the support and help. Without you, I might not be able to hold on at the first year.

To Dave & Kath, thank you for all the wine tastings and delicious dinners. To Volkan & Rukiye, thank you for sharing the mystery of pregnancy and the cutest baby boy Alper. His smile always warm my heart. To Ann, Noi and my long term friend Christine, your friendships are precious to me. To Chu & and Fang, thank you for sharing your home with me.

Last but not least, to my parents Ken-Tang & Gui-Huan whom are also my best friends, thank you for always supporting me 100%. I could not ask for any better. And to my grandmother whom is 103 years old, thank you for helping raise me up and keeping your promise that you will wait to see me get my Ph.D.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

In this thesis, we focus on promoting the use of information-based group sequential monitoring in phase III time-to-event clinical trials. The research was motivated by resolving obstacles encountered in real data analyses. We propose the use of information to first free the effect the nuisance parameters may cause to over– or under– power a test. For a common scenario of data coming in lag, we suggest a suitable estimator for the parameter of interest, then verify and incorporate it into the information-based monitoring procedure. In this chapter we will give a general background of sequential methodologies. They were originated from the allowance of early stopping in clinical trials then developed through years to generalized group sequential methods.

In a fixed sample statistical study, data are collected to a predetermined sample size to obtain the expected level of significance and power. At the end of sampling, a test statistic is computed and compared to a boundary value that can be observed from a standard normal table such that the probability of type I error is as desired for the study. In clinical trials, a phase III study, human participants are subjects to an intervention treatment group and a control group to assess the effectiveness of the intervention and its role in clinical practice. For ethical as well as practical reasons it is necessary to have a committee monitoring data during the trial period and a

trial may be terminated early upon the decision of the committee. For example, in a study on drug toxicity or efficacy, it is the responsibility of the monitoring board to avoid harmful effects in participants if they are detected. On the other hand, it is also unethical to keep it away when response data show a clear benefit of the drug. A trial can also be stopped early if partway through the trial the information obtained prove to be so unimpressive in drug effect that it is infeasible to continue. The main goal is to minimize the number of participants in a trial when there is sufficient significance for the purpose for which the experiment is designed, yet preserve the credibility of the test in statistical analyses.

Different from a fixed sample study, the trial now produces multiple tests during accrual time. The boundary, as the criterion in decision-making at each interim analysis, depends on the nominal level of significance. For the naive use of the boundary from a fixed sample study, the more analyses done during the monitoring process the higher the probability of rejecting the null hypothesis incorrectly. For example, given the overall level of significance 0.05, if the monitoring board decides to have a look at the data accumulated every half year during the 2 year trial period, the type I error after 4 interim analyses will increase to 0.126, more than twice the expected. See Table 1.1. The type I error could increase to the degree of unacceptable as the number of looks increases. The solution to maintaining a desired significance level is to adjust the boundary values such that all nominal levels add to the overall. For the past three decades, researchers have proposed many methodologies to deal with the problems arised by early stopping of a trial. Naturally, the first idea is to monitor data continuously so a trial can be terminated at its earliest time.

1.2 Classical Sequential Methods

Classical sequential methods were first applied to clinical trials by developers such as Armitage. It involves continuous testing in an experiment. For treatment comparison studies, data are analyzed in pairs, one observation from each treatment group. After each pair, a test statistic is computed and compared with a derived stopping boundary. If the test statistic crosses the boundary, then the trial is terminated;

otherwise, the trial continues until the next analysis. Since the decision to terminate a trial depends solely on whether the outcome of one treatment is significantly better, or worse, than the other, there is an uncertainty of ever reaching a point where the study would be stopped. The requirement in pairing is also not appealing in a sense that the two participants may be very different and the pair may not be well matched in the important prognostic variables. Most of the sequential methods also assume that the response variable outcome is known in a relatively short time. To studies of chronic diseases, classical sequential methods will not be very useful. Nevertheless, with the possibility of early stopping, sequential methods do give the benefit of early stopping and minize the number of participants that must be entered into the study, giving rise to a modification of sequential designs.

1.3 Group Sequential Methods

One of the reasons that classical sequential designs are not favorable in clinical trials is the continual assessments of data after each pair of outcomes or events. Pocock (1977) developed a group sequential method that divided data into equal-sized groups with sample size even for both treatments. The maximum number of groups K is pre-set along with the power $1-\gamma$ and level of significance α, so that at the Kth analysis the trial is stopped whether there is sufficient significance detected or not.

For data accumulated up to i^{th} group, the test statistic $T_i = \frac{1}{\sqrt{i}} S_i$, where S_i is the partial sum of independent identically distributed standard normal random variables, and has the covariance structure of independent increments. The independent increments property is a very important feature to group sequential methods in finding the joint distribution of sequential test statistics. The stopping boundary Z for each analysis is computed numerically to maintain the overall significance level α when early stopping has not occurred under the null hypothesis. That is,

$$P(|T_1| < Z, |T_2| < Z, \ldots, |T_{K-1}| < Z, |T_K| \geq Z | H_0) = \alpha. \tag{1.1}$$

The decision of stopping a trial is a sequential process that at the first interim analysis

the study is stopped if $|T_1| \geq Z$; the study is continued after the first and stopped at the second analysis if $|T_1| < Z$, $|T_2| \geq Z$; or continued after the first and the second but stopped at the third analysis if $|T_1| < Z$, $|T_2| < Z$, and $|T_3| \geq Z$. The study is terminated at the i^{th} analysis if the test statistic at the i^{th} time exceeds the stopping boundary but at all previous looks are less than boundary Z. To compute the type I error spent at each analysis, the nominal level of significance, $\alpha(i) = \text{P}(|T_1| < Z, |T_2| < Z, \ldots, |T_{i-1}| < Z, |T_i| \geq Z \mid H_0)$, where i=1, 2, ... , K; $\alpha(K) = \alpha$.

O'Brien & Fleming (1979) proposed a similar method using Pearson Chi-square tests where the test statistics are proportioned by the number of groups. Wang & Tsiatis (1987) also proposed approximately optimal one-parameter boundaries that have the form of c(α, K, Δ)$\cdot i^{\Delta-0.5}$, i=1, ... , K, where c(α, K, Δ) is derived from the joint probability given in (1.1). Pocock (Δ=0.5) and O'Brien-Fleming (Δ=0) boundaries are included within this class of boundaries that minimize the number of subjects required for detecting significance difference at given level of significance α, and power $1 - \gamma$. In comparison, Pocock method has boundaries fixed through all group analyses while the value of O$'$Brien-Fleming boundaries decreases as the number of groups increases. O$'$Brien-Fleming method has less chance of early stopping than Pocock method since its acceptance regions are wider in the early analyses and narrower when the trial is close to completion. Thus O$'$Brien-Fleming method is more conservative in early stopping than Pocock method. Figure 1.1 shows the boundaries for both Pocock and O'Brien-Fleming methods when K=5, α =0.05.

Group sequential method reduces the average sample size considerably and avoids many of the limitations of classical sequential methods such as mentioned previously. Yet it does not go without obstacles of its own. One is the need to specify K, the maximum number of interim analyses, in advance. Another is the requirement of equal sample size for each treatment, this means the time to do interim analysis is also restricted. A more flexible group sequential method by Lan & Zucker (1993) suggested the use of alpha-spending function. It allows investigators free control of how much of the type I error they want to spend on one interim analysis, and still guarantee the desired overall level of significance α at the last analysis. With the use of alpha-spending function, data monitors will not have to fix either on the number

or the time of analyses. This approach generalizes the previous group sequential methods so that Pocock and O$'$Brien-Fleming monitoring procedure become special cases. For example, the alpha-spending function for Pocock boundaries is set as $\alpha(\pi_i) = \alpha \cdot log[1 + (e-1)\pi_i]$. For O$'$Brien-Fleming spending function in two-sided test, $\alpha(\pi_i) = 4 - 4\Phi(Z_{\frac{\alpha}{4}}/\sqrt{\pi_i})$, where $0 \leq \pi_i \leq 1$. The proportion π can be defined as the fraction of information observed up to some time. For example, $\pi_i = \frac{i}{K}$ at the i^{th} group with maximum analysis time K. It can also be approximated by the variance up to time divided by the total variance of the test.

Once a spending function is selected, the information fractions π_1, π_2, ..., π_M, determine the sequential boundary values b_1, b_2, ..., b_M, under the criteria

$$P(|T_1| < b_1, |T_2| < b_2, \ldots, |T_{j-1}| < b_{j-1}, |T_j| \geq b_j | H_0) = \alpha(\pi_j) - \alpha(\pi_{j-1}) \qquad (1.2)$$

where $1 \leq j \leq M$, $\pi_0 = 0$, $\pi_M = 1$, $\alpha(\pi_0) = 0$, $\alpha(\pi_M) = 1$. The sequential boundaries b_1, b_2, ..., b_M can be determined numerically using the recursive integration formula by Armitage, McPherson & Rowe (1969). Follow the sequential process, at each interim analysis, a test statistic is compared with a stopping boundary and the action of stopping or continuing a trial will be made.

1.4 Outline

A recent monitoring procedure using information fraction in clinical tests was proposed by Lan & DeMets (1983) and characterized by Sharfstein, Tsiatis & Robins (1997), where the problem can be cast into a test of a single scalar parameter of interest under a specific statistical model. In chapter 2, we propose using this information-based group sequential monitoring procedure to release unreliable guessings on nuisance parameters when the design model is correct. This methodology gives robust values on the power of tests. At this time we assume data are observed completely. In chapter 3, when there is lag time in reporting, we first derive an unbiased estimator of the parameter of interest that takes censoring into account, then prove that the joint distribution of these test statistics are asymptotically multivariate normal and its covariance structure has the property of independent increments. By simulations

we compare type I and II errors with other traditionally used methodologies–one ignores cesored data and the other uses maximum lag time to ensure completely observed data. The results demonstrate that our proposed method achieves the expected power and reduces the number of participants needed in a trial.

Figure 1.1: Group Sequential Boundaries for Two-sided Tests with K=5 and α =0.05: Δ =0 and Δ =0.5

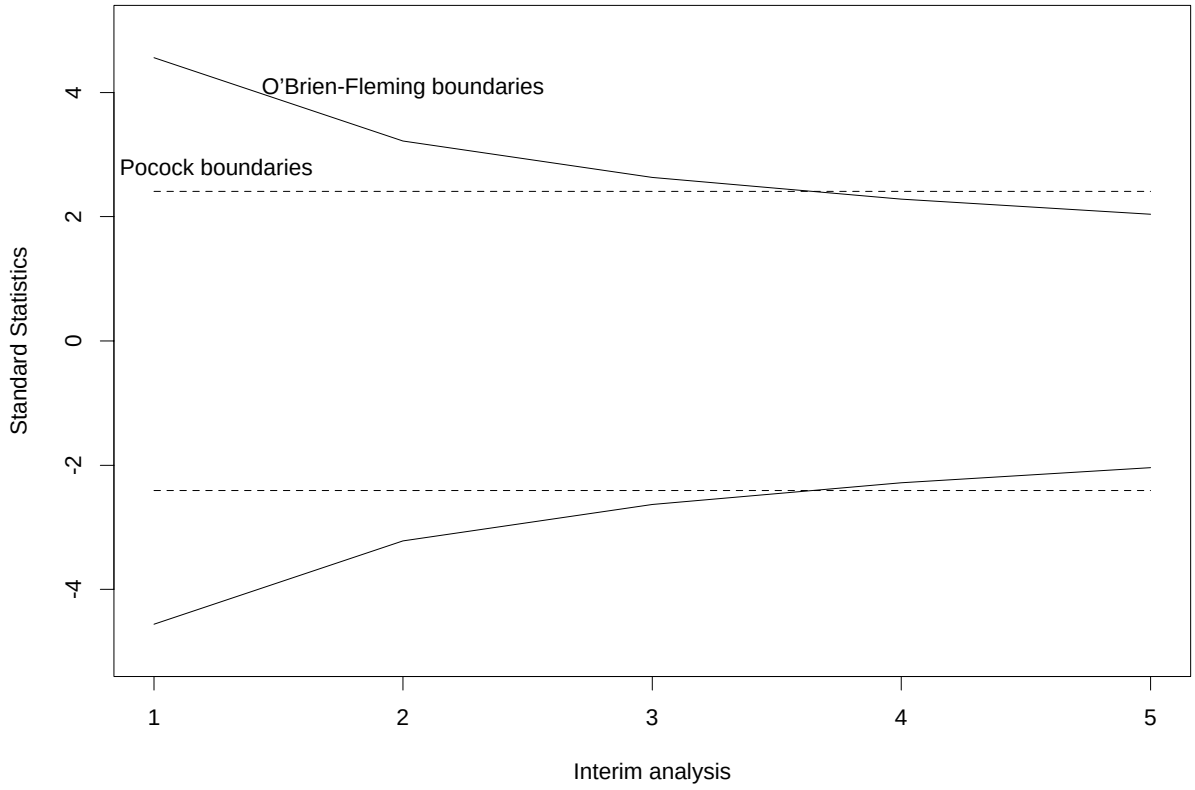

Table 1.1: Effect of Multiple Looks at the Nominal Level

K	False Positive Rate
1	0.050
2	0.083
3	0.107
4	0.126
5	0.142
10	0.193
20	0.246
50	0.320
100	0.374
1000	0.530
⋮	⋮
∞	1.000

Chapter 2

Nuisance Parameters Free IBM

2.1 Introduction

In the design phase of a clinical trial, the total number of participants is traditionally computed based on the values of nuisance parameters to achieve a certain level of significance and power. In practical applications, nuisance parameter values are often provided by unreliable guesses founded on little or no available past history. As a result, if the initial guesses for the nuisance parameters are far from the truth, then the study may be under– or over– powered to detect the desired treatment difference.

For example, in a clinical trial comparing dichotomous responses among two treatments, suppose the investigators want to design the study to detect a 10% difference in the response rate between two treatments with 90% power using a test at the .05 (two-sided) level of significance. The traditional approach would necessitate that an initial guess be made for the response rate among controls. Using the best available data, say, this is given to be 20%. Using standard tables for binomial responses, it is then a simple exercise to obtain the necessary sample size so that a binomial test, at the .05 level of significance, would have 90% power to detect a difference from 20% to 30%. In this case, a total of 783 participants randomized with equal probability to the two treatment would be necessary. In truth, however, suppose the initial guess was wrong and the response rate for the control group was 40%. Then with a sample size of 783, the power of the binomial test at the .05 level of significance to detect a

10% difference, or a difference from 40% to 50%, would have been diminished to 80%.

In the previous example, the treatment effect was given as the difference in response rates. In some cases, investigators may be interested in assessing relative risk, that is, the ratio of response rates between treatments. For example, a treatment may be deemed important if the relative risk is 1.5 or greater (i.e. a 50% relative increase in response). Hence, if the control group response rate were 10%, then a treatment that had a response rate of 15% or greater would be clinically important, whereas, if the control group response rate were 20%, then a 30% response rate or greater would be necessary for the treatment group. When evaluating sample sizes during the design stage to detect relative risk differences with some prespecified power, the intial guess for the control group response rate plays even a more critical role as deviations from the initial guess will have a greater effect on power.

For ethical as well as practical considerations, most clinical trials are designed with formal sequential stopping rules. That is, the data from a clinical trial are monitored periodically, usually by an external data monitoring board, and, if the treatment difference becomes sufficiently large during one of the interim analyses, then the study may be stopped. Formal sequential boundaries have been derived that dictate how large the treatment difference must be at different interim analyses before a study is stopped. These boundaries are constructed so that the overall test has the desired operating characteristics. That is, the resulting sequential test will have the desired level of significance and power to detect a clinically important treatment difference. Issues regarding the effect on power, of misspecifying the nuisance parameter during the design stage, are similar to those discussed earlier for the fixed sample procedure. However, when data are monitored periodically, the parameters in the model can be estimated with the available data and if these estimates are sufficiently far from the values used in the design stage, then the investigators have the ability to alter the design; i.e. adaptively increase or decease the sample size. It is the feasibility of this strategy that we will investigate in this paper.

We propose using what is referred to as nuisance–parameter–free information-based monitoring for designing and monitoring clinical trials with dichotomous endpoints. As is well known, the power of a test to detect a clinically important treatment

difference (usually parametrized through a single parameter) is directly related to the Fisher Information available in the data for that parameter. Consequently, a test will have the desired power to detect a clinically important difference, if the necessary information were available at the end of the study, independent of the nuisance parameters. Since the Fisher Information can be estimated by the observed information, we suggest monitoring a study using the observed information rather than the sample size. Also, since for most estimators of treatment difference, the observed information and the standard error of the parameter estimate of treatment difference are related to each other, this means that we would continue monitoring a study until the standard error becomes sufficiently small. Such a strategy may necessitate changing the sample size from that initially guessed during the design stage. This would be the case if the parameters used in designing the study differed from the truth.

Recent results given by Sharfstein & Tsiatis (1997) show how Fisher Information can be used in conjunction with sequential monitoring. The well known theory that an efficient parameter estimator is asymptotically normally distributed with variance equal to the inverse of a Fisher Information has been generalized to the joint sequential distribution of test statistics computed at interim times. It is shown that the joint distribution of the sequentially computed test statistics is multivariate normal with a covariance structure which again only depends on the Fisher Information at each of the interim analyses. This result allows the construction of group-sequential level α tests. The Fisher Information can be easily estimated by the inverse of the sample variance that involves only the data estimates of the parameters. So with information-based monitoring, the power can be achieved regardless of the initial guesses on nuisance parameters, provided that the parameter estimator for treatment difference is efficient. We will illustrate this concept in the next section and briefly describe the steps necessary to use the information-based monitoring procedure to a group sequential study, so to allow the possibility of an early stopping in a clinical trial. These methods will be applied to the problem of comparing dichotomous response rates between two treatments. To further demonstrate our proposed idea, we will consider three different scenarios for defining treatment difference; i.e. absolute difference, relative risk, and relative odds. Extensive simulation studies are run under

these three scenarios and tables of results on power and type I error are given in 2.3. Finally, we give an account of summary in 2.4.

2.2 Information-based Monitoring in Group Sequential Studies

The general theory for information-based monitoring has been described in detail by Sharfstein & Tsiatis (1997). In this section, we recap the steps for implementing this in a group sequential study which is generalized by the use of a type I error spending function (Lan & DeMets 1983) fit for calculating O'Brien-Fleming boundaries.

As always, we posit a statistical model for our data in terms of parameters (β, θ), where β denotes the parameter of interest. For our problem, the parameter β will denote treatment difference, which we take to be single valued, and θ will denote a vector of nuisance parameters necessary to adequately describe the probability distribution of the data. We will focus on testing the null hypothesis of no treatment difference, usually parametrized as $H_0 : \beta = 0$, against a two-sided alternative $\beta \neq 0$. The clinically important treatment difference will be denoted by β_A. That is, if, in truth, $\beta \geq \beta_A$, then we would want to detect such a difference with power at least equal to $1 - \gamma$ using a test at the α level of significance. We wish to emphasize that the manner in which we choose to parameterize the problem has important implications. For example, when we were comparing dichotomous response rates between two treatments, we realized that for some studies the interest might focus on detecting absolute differences, whereas, for other studies relative risk might be important. The choice for the parameter of treatment difference β, e.g. absolute difference or relative risk, implies that the clinically important difference, $\beta = \beta_A$, is independent of the other nuisance parameters θ. Thus, we believe it is important at the design stage to ascertain, from the investigators, the magnitude of difference in response rates that is deemed clinically important to detect over a wide range of nuisance parameters. By so doing, one may more realistically determine how

to characterize treatment difference, whether it be absolute difference, relative risk, relative odds, or some other more appropriate measure of treatment difference.

We will consider tests that are efficient such as Wald tests, score tests, or likelihood ratio tests. For concreteness, let us consider the level α (two-sided) Wald test that rejects the null hypothesis when the statistic $W = \hat{\beta}/\text{se}(\hat{\beta})$ exceeds, in absolute value, $Z_{\alpha/2}$, where $\hat{\beta}$ denotes the maximum likelihood estimator for β, $\text{se}(\hat{\beta})$ denotes the standard error of the estimator, and Z_x denotes the $(1-x)$th quantile of a standard normal distribution.

The Fisher Information matrix is defined as the expected value of minus the Hessian (i.e. matrix of second partial derivatives with respect to all the parameters) of the log likelihood. Under the usual regularity conditions, the asymptotic variance of the maximum likelihood estimators $(\hat{\beta}, \hat{\theta})$ is equal to the inverse of the information matrix. The asymptotic variance for $\hat{\beta}$ is the corresponding element on the diagonal of the inverse of the information matrix. The Fisher Information for the parameter β is defined as the inverse of the asymptotic variance for $\hat{\beta}$. If a single test of the null hypothesis were conducted using the level α Wald test, then by standard asymptotic theory, one could show that the information for β, denoted by I, necessary to detect the clinically important treatment difference β_A, with power $1-\gamma$, must equal

$$I = \left(\frac{Z_{\frac{\alpha}{2}} + Z_{\gamma}}{\beta_A} \right)^2 .$$

This value of information is independent of the nuisance parameters θ, although the nuisance parameters come into play in the definition of the information. The key point is that Fisher Information can be estimated using the data. In fact, the standard error of $\hat{\beta}$ is generally obtained from the observed information matrix, which estimates the Fisher Information matrix. Consequently, $\text{se}^{-2}(\hat{\beta})$ can be used as an estimate for the information I. This suggests that the desired power of a level α Wald test, to detect a clinically important difference, would be achieved when

$$\text{se}^{-2}(\hat{\beta}) = \left(\frac{Z_{\frac{\alpha}{2}} + Z_{\gamma}}{\beta_A} \right)^2 .$$

If the data are to be monitored periodically, say, at times $t_1, \ldots, t_K$, then Wald tests could be computed at each of these times using the available data. We denote

these test statistics as $W(t_j), j = 1, \ldots, K$. A group sequential test would reject the null hypothesis and stop the study at the first time t_j for which $|W(t_j)| \geq c_j, j = 1, \ldots, K$, where the boundary values c_j are chosen so that the group sequential tests have the appropriate operating characteristics. Namely, the values c_j must satisfy the following equalities:

$$1 - P_{\beta=0}\{\bigcap_{j=1}^{K} |W(t_j)| < c_j\} = \alpha$$

and

$$P_{\beta=\beta_A}\{\bigcap_{j=1}^{K} |W(t_j)| < c_j\} = \gamma.$$

In order to derive such group-sequential boundaries that satisfy the above equalities we must be able to characerize the joint distribution of the sequentially computed Wald test statistics $\{W(t_1), \ldots, W(t_K)\}$. The key result that allows us to compute such boundaries in terms of Fisher Information, comes from the general theorem given by Sharfstein et al. (1997), which may be stated as follows:

Theorem 1. *If $\hat{\beta}$ is an efficient estimate of β then, regardless of the model generating the data, the asymptotic joint distribution of the sequentially computed K-dimensional statistic $\{W(t_1), \ldots, W(t_K)\}$ is multivariate normal with mean vector*

$$\beta\{I(t_1)^{1/2}, \ldots, I(t_K)^{1/2}\}$$

and $K \times K$ covariance matrix, whose $(j,k)_{th}$ element, $j \leq k$, is given by

$$\left[\frac{I(t_j)}{I(t_k)}\right]^{1/2} \quad j, k = 1, \ldots, K,$$

where $I(t_j)$ denotes the Fisher Information at time t_j.

We notice from the above theorem that the joint distribution is completely characterized by the parameter β and the Fisher Information available at the interim times. Different stategies for deriving sequential boundaries have been suggested in the literature. A class of boundaries given by Wang & Tsiatis (1987) used power functions. Namely, the boundaries c_j were chosen to be proportional to $I(t_j)^{\Delta - .5}$. The proportionality constant and the maximum information MI, defined as the information at the final analysis, (i.e., $MI = I(t_K)$), were derived by satisfying the two

equalities above necessary for the level and power of the test to be respected. When the power $\Delta = 0$, the resulting boundaries are similar to those originally proposed by O'Brien & Fleming (1979) and we will refer to such boundaries as O'Brien-Fleming boundaries. Another popular class of boundaries is when $\Delta = .5$. These are similar to boundaries suggested by Pocock (1977). It has become standard practice in clinical trials to use O'Brien-Fleming type boundaries. Therefore, we will restrict our numerical studies to these boundaries only. It is often convenient to characterize the maximum information MI, necessary to have power $1 - \gamma$ to detect the clinically important difference β_A, as

$$MI = \left(\frac{Z_{\frac{\alpha}{2}} + Z_{\gamma}}{\beta_A} \right)^2 IF,$$

where IF, defined as the inflation factor, repesents the relative increase in the information necessary for a group sequential test to achieve the same power as a fixed sample test. This is a useful construct since it is often the case that IF remains relatively stable for certain classes of group-sequential boundaries. For example, it is shown in Wang & Tsiatis (1987) that an IF of 1.03 is associated with O'Brien-Fleming boundaries that are monitored at least five times.

In practice, the information at the times of analysis are not generally known in advance. Lan & DeMets (1983) suggested that one specify a spending function that dictates the amount of the significance level to be used as a funtion of the maximum information, with the constraint that all the significance level α be used at the end of the study, i.e. when $I(t_K) = MI$. By so doing, the boundary values can be computed adaptively at the time of an interim analysis, t_j, by estimating the Fisher Information using $\text{se}^{-2}(\hat{\beta}(t_j))$. Spending functions that result in O'Brien-Fleming and Pocock type boundaries are given by Lan & DeMets (1983).

At each interim analysis time in a group sequential study, the information $I(t)$ is estimated by the inverse sample variance of the parameter estimator $\text{se}^{-2}(\hat{\beta}(t))$, where $\hat{\beta}(t)$ is an efficient estimator of the parameter of interest β at an interim time t. The stopping boundary at the time is computed sequentially and numerically (Armitage et al. 1969) using the result shown by Sharfstein et al. (1997), namely that the joint distribution of the time-sequential statistics is multivariate normal with a covariance

structure that only depends on the information at each of the interim analyses. In a maximum information trial, a study is terminated when a test statistic exceeds the stopping boundary or the maximum information is attained, i.e. $\text{se}^{-2}(\hat{\beta}) = \text{MI}$.

Information-based monitoring (IBM) is free of nuisance parameters in the sense that the monitoring number, $\text{se}^{-2}(\hat{\beta})$, uses only data–indicated nuisance parameter values and does not involve any initially guessed values. Therefore, with IBM we can obtain robust values of power over any set of nuisance parameters given a specified value of interest. For the example mentioned in 2.1, the primary interest is 10% difference between the two response rates. The information number at each time is estimated by the inverse sample variance, which involves only observed sample response rates and sample sizes at the time. Therefore, the test should obtain a 90% power whether the potential set of response rates is (20%, 30%) or (50%, 40%).

To demonstrate our point numerically, in the next section we will show simulation results of power and significance level of tests also on other types of interest such as log ratio or logit of the two response rates. We adopt the standard assumption that subjects enroll in a study according to a Poisson process with a pre-specified rate. And the outcome of interest are all observed and relatively soon so that no censoring occurrs.

2.3 Simulation Studies

We first give a general definition on the example of comparing two treatment responses. Since individual responses are independent and dichotomous, the number of responses for treatment i are binomially distributed with a response probability π_i, where $i = 1, 2$. The test is cast as $H_0 : \beta = 0$ vs. $H_1 : \beta = \beta_A$, where β is the difference of the two response rates, π_1, π_2; and β_A is the interest that is considered clinically important. We chose three common relationships of response rates for the identification of β_A:

1. The original interest in response difference. $\beta_A = \pi_1 - \pi_2$.

2. Log ratio of response rates. $\beta_A = log\frac{\pi_1}{\pi_2}$.

3. Log odds ratio (logit) of response rates. $\beta_A = \log \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}$.

To emphasize that information-based group sequential tests all show 90% power under 0.05 significance level regardless of the value (π_1, π_2), let's assign π_2 to be 0.1, 0.2, 0.3, 0.4, or 0.5. In order to make comparisons, we accommodate the value of π_1 and β_A such that at least one pair of (π_1, π_2) is equal in each case of interest. To illustrate this, let the pair of matching equality be (0.4, 0.3). Then for case 1, $\beta_A = 0.4 - 0.3 = 0.1$ and for $\pi_2 = 0.1$, 0.2, 0.4, 0.5, π_1 is computed by $\pi_1 = \pi_2 + \beta_A$ which is 0.2, 0.3, 0.5, and 0.6, accordingly. For case 2, with the same pair, $\beta_A = 0.29$ and π_1 is calculated based on log ratio function with π_2 value known. Similarly for case 3. Repeat the steps again for another matching pair (0.35, 0.3) to find the values of β_A and π_1 for each fixed π_2. The complete value of π_1 and π_2 for both matching pairs and for all cases are listed on Table 2.2 and 2.3.

In data simulations, we assume that patient entry time is a Poisson process with total entry of 100 patients per year. The treatment indicator is a uniform (0,1) random variable, each patient having 50% chance of receiving either treatment. Each case of interest produces a different number of MI since the value of β_A varies. Below is a table of all MI numbers in all three cases for both matching pairs. For 8000

Table 2.1: MI values in all 3 cases of interests for both matching pairs.

	pair (0.4, 0.3)	pair (0.35, 0.3)
1. $\beta_A = \pi_1 - \pi_2$	1081	4325
2. $\beta_A = log\pi_1 - log\pi_2$	131	455
3. $\beta_A = \log \frac{\pi_1}{1-\pi_1} - log\frac{\pi_2}{1-\pi_2}$	56	204

simulations, we averaged the number of tests rejected under the null hypothesis or under the alternative hypothesis, and also computed the average sample size (ASN). The simulation results are displayed in Table 2.2 and 2.3 as well.

We will first look at the results of case 1, interest in response differences, from Table 2.2. For the nuisance parameter set (π_1, π_2)=(0.3, 0.2), simulated tests give 90% power and on average the trials will require only 571 subjects, which is 212 less than

the number 783 in a fixed sample study in §1. This is one of the benefits in employing group sequential methods. For the same interest in detecting 10% difference, given (π_1, π_2)=(0.5, 0.4), we can see clearly that the tests obtain the same power. The reason behind it is that rather than using total sample size as a criterion to end a study, in information-based monitoring a study ends if MI is exhausted. For the same value of interest MI stays the same. Thus a test would be expected to maintain the same power. We find the same conclusions for other interests in testing log ratio and logit of the response rates. Table 2.3 shows that detecting smaller values of interest requires larger sample sizes. For nuisance parameter values of (0.35, 0.3), tests for all three cases of interest are equivalent and thus require similar number of subjects.

2.4 Summary

Conventional statistical tests inquire a total sample size that depends on potentially unreliable values of nuisance parameters. This paper suggested monitoring a trial from a different angle using Fisher Information and MI number. Under a correct model assumption, an information-based monitoring study maintains expected power under a significance level regardless of the values of nuisance parameter sets, provided that the single hypothesis parameter of interest can be estimated efficiently and the alternative value of interest is preset. The idea is simulated via a simple scenario of comparing dichotomous responses among two treatments. Generally, information-based monitoring procedure is applicable to any type of model for any type of group sequential study given necessary conditions.

Information-based monitoring can also be used to check on the assumption of a correct design model. Without this assumption, design parameters are added to the line of nuisance parameters. For the simulation example, the set of nuisance parameters would be the two treatment response rates, accrual time, and the Poisson process parameter. The parameter of the accrual rate and the length of time can be easily adjusted by looking into the information at interim times. For complex studies, **?**) proposed using the Bootstrap method to update the initial values of design parameters.

Table 2.2: Simulation results on 3 cases of parameter interests. Case1: response difference, Case2: log ratio, Case3: logit. The matching pair is (0.4, 0.3). α, $1-\gamma$ are the computed type I error and power. ASN is the averaged sample size over 8000 simulations.

TEST $\beta_A = 0.1$ π_1, π_2	CASE 1. α or $1-\gamma$, ASN	TEST $\beta_A = 0.29$ π_1, π_2	CASE 2. α or $1-\gamma$, ASN	TEST $\beta_A = 0.44$ π_1, π_2	CASE 3. α or $1-\gamma$, ASN
0.10, 0.10	0.06, 379	0.10, 0.10	0.05, 4690	0.10, 0.10	0.05, 2443
0.20, 0.10	0.90, 389	0.13, 0.10	0.90, 2855	0.15, 0.10	0.89, 1500
0.20, 0.20	0.06, 685	0.20, 0.20	0.05, 2075	0.20, 0.20	0.06, 1371
0.30, 0.20	0.90, 571	0.27, 0.20	0.90, 1260	0.28, 0.20	0.90, 888
0.30, 0.30	0.05, 901	0.30, 0.30	0.05, 1212	0.30, 0.30	0.05, 1047
0.40, 0.30	0.90, 693	0.40, 0.30	0.90, 723	0.40, 0.30	0.90, 708
0.40, 0.40	0.05, 1030	0.40, 0.40	0.05, 780	0.40, 0.40	0.05, 915
0.50, 0.40	0.90, 756	0.53, 0.40	0.90, 454	0.51, 0.40	0.90, 645
0.50, 0.50	0.05, 1073	0.50, 0.50	0.05, 521	0.50, 0.50	0.05, 878
0.60, 0.50	0.90, 759	0.67, 0.50	0.90, 290	0.61, 0.50	0.90, 650

Table 2.3: Simulation results on 3 cases of parameter interests. Case1: response difference, Case2: log ratio, Case3: logit. The matching pair is (0.35, 0.3). α, $1-\gamma$ are the computed type I error and power. ASN is the averaged sample sizes over 8000 simulations.

TEST $\beta_A = 0.05$ π_1, π_2	CASE 1. α or $1-\gamma$, ASN	TEST $\beta_A = 0.15$ π_1, π_2	CASE 2. α or $1-\gamma$, ASN	TEST $\beta_A = 0.23$ π_1, π_2	CASE 3. α or $1-\gamma$, ASN
0.10, 0.10	0.05, 1541	0.10, 0.10	0.06, 16543	0.10, 0.10	0.05, 9255
0.15, 0.10	0.90, 1328	0.12, 0.10	0.90, 10805	0.12, 0.10	0.90, 6055
0.20, 0.20	0.05, 2741	0.20, 0.20	0.05, 7336	0.20, 0.20	0.05, 5172
0.25, 0.20	0.90, 2116	0.23, 0.20	0.90, 4725	0.24, 0.20	0.90, 3466
0.30, 0.30	0.05, 3604	0.30, 0.30	0.05, 4257	0.30, 0.30	0.05, 3928
0.35, 0.30	0.90, 2657	0.35, 0.30	0.90, 2696	0.35, 0.30	0.90, 2674
0.40, 0.40	0.05, 4118	0.40, 0.40	0.06, 2710	0.40, 0.40	0.05, 3432
0.45, 0.40	0.91, 2958	0.47, 0.40	0.89, 1715	0.46, 0.40	0.90, 2386
0.50, 0.50	0.05, 4287	0.50, 0.50	0.06, 1809	0.50, 0.50	0.05, 3291
0.55, 0.50	0.90, 3036	0.58, 0.50	0.90, 1119	0.56, 0.50	0.89, 2360

Chapter 3

IBM with Lagged or Censored Data

3.1 Introduction

The use of sequential methods in clinical trials is now fairly common, and there is methodology and software available for their implementation. Specifically, information-based methods (Lan & Zucker 1993, Sharfstein & Tsiatis 1997) coupled with the use of alpha-spending functions (Lan & DeMets 1983) allow a flexible and unified approach for sequential monitoring of clinical trials.

One of the major stumbling blocks that still remains is the problem of lag time in reporting. During the interim stages of large-scale clinical trials, data are not completely up-to-date. For example, if interest focuses on a patient's response status, i.e., the dichotomous variable that a patient either responds or not, this information may not be available for analysis in the statistical center until some time after the patient enters the trial. The duration of time between a patient's enrollment and the moment when his or her response status is known to the statistical center is referred to as the "lag time". Lag may be simply due to administrative delays in transferring information from the clinical site to the statistical center or may be a consequence of biological factors that determine the time it takes to observe whether or not a patient responds to treatment. It may even be related to the design features of the

clinical trial such as timing of clinic visits. The lag time may be due to a complicated combination of all of the above factors. Regardless of the exact cause, the lag time itself is of no scientific interest, but merely a factor that needs to be taken into account when inference is made from the data on response rate.

Lag may also be considered as censored at an interim time of analysis. The consequence is that the data that are actually available to the statistical center at any point in time may not be representative of the results that will eventually be available. Some lagged data may actually not be observed by the time a trial ends. In other words, the interim data may be biased in some way. For example, if information on responders came into the statistical center more quickly on the average than it did for nonresponders, then it would appear at an interim analysis as if the response rate was higher than it would be when all the data are complete. Also if the distribution of lag times was differential both by treatment and response status, then, even if the response rates were equal for two treatments, it would appear that there is a difference at an interim analysis.

There are several standard ways of dealing with this type of incomplete data. A naive way is to ignore the problem and conduct the statistical analysis with the available data. Alternatively, the analysis may be limited to those patients who entered the trial early enough to guarantee that their data were complete by the analysis time (this assumes that there is a maximum possible lag time). The results of this limited analysis should be unbiased. The problem with this method, however, is that it discards the information on some patients whose data are available, and thereby reduces the power of the analysis to detect a treatment difference. We propose a general methodology for analyzing all the available data, while still taking into account the effect of lag time. This is accomplished by considering weighted tests and estimators.

The bias problem is exacerbated during interim analyses when the data are most likely to be incomplete and the potential for bias is the greatest. The goal for the first part of this paper is to derive unbiased tests and estimates that account for the lag in reporting. We will use these unbiased tests as the basis for sequential monitoring. In order to do so, we must characterize the joint distribution of our

test statistics and estimators at the different monitoring times. As we illustrate in 3.3, the test statistics that we propose may be represented as stochastic integrals of counting process martingales. This representation, together with the powerful machinery developed for such processes, will enable us to derive the joint sequential distribution in 3.4. This will then allow us to compute stopping boundaries that preserve the desired operating characteristics by using information-based methods and alpha-spending functions. In 3.5, we applied the results derived from previous sections to an example of comparing two response rates. In 3.6, we performed simulations on data with dichotomous responses to demonstrate that our proposed inverse weighted estimator is consistent and requires fewer individuals entered into a study than the other two commonly used estimators. Finally the conclusions and future approach for the study of efficiency in 3.7.

3.2 Notation

Before going into specific details, we first introduce some notation. For individual i, the outcome variable of interest will be denoted by Y_i. There is a follow-up time, T_i, that is necessary before the variable of interest is determined. Follow-up time is measured from patient entry and is referred to as patient time. The data that are collected over patient time will be denoted by $\{Z_i(u), u \leq T_i\}$, where $Z_i(u)$ denoted the data available at time u. For example, for objective response of cancer patients on a given treatment, T_i represents the time that the ith patient undergoes treatment, $Z_i(u)$ represents tumor size if it is measured at time u, and an objective response, Y_i, may be defined as the indicator of whether the minimum value of $Z_i(u)$ at measured values of $(u \leq T_i : Z_i(u) \geq 0)$ is less than .5 times $Z_i(0)$. Ultimately, we are interested in making inference regarding the outcome variable Y_i.

With complete follow-up, we would be able to calculate the outcome variable Y_i, on all patients and be able to use standard statistical techniques for inference. However, in many instances, we will not be able to observe the entire data history $\{Z_i(u), u \leq T_i\}$ due to right censoring. Censoring may result from incomplete follow-up at the time of analysis due to staggered entry. This is referred to as administrative

censoring, as the reason for censoring does not involve patient choice. In contrast, censoring may occur when a patient withdraws from the study and data on that individual are no longer collected. To accommodate censoring, we define a time to censoring as a positive random variable C_i. For a sample of n patients, the data, which include the possibility of being censored, can be described for the ith individual by $V_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$, and $Z_i^H(V_i)$, $i = 1, \ldots, n$, where $I(\cdot)$ is the indicator function, $Z_i^H(x)$ denotes the history of the Z process for the ith individual up to time x; that is, $Z_i^H(x) = \{Z_i(u), u \leq x\}$. For practical purpose, we need to restrict the time we evaluate each individual up to time L. That is, entry time T_i is the minimum of the time from treatment initiation to outcome observed or L.

3.3 A Consistent Estimator And Its Martingale Structure

Our aim is to estimate the mean of outcome interest Y when the history of data is subject to right censoring. The kernel of the methodology is based on the use of the inverse selection probability weighted estimator. To motivate this estimator, we note that with complete information for each individual, an obvious estimator for the expected value of Y, which we denoted by μ_Y, is $n^{-1}\sum_{i=1}^n Y_i$. With censoring, there is a potential censoring time C_i, which is assumed to be independent of the data-collection process as well as of the time T_i. Let the distribution function of the censoring time be given by $1-K(u)$, where $K(u) = \mathrm{P}(C > u)$. The data-collection process would be censored if $C_i < T_i$. Therefore, at the time of ascertainment T_i, the ith individual has a probability $K(T_i)$ of not being censored. Therefore, each person who has complete information (i.e., uncensored) represents, on average, $1/K(T_i)$ individuals who might have been censored. Hence, a weighted estimator using only uncensored individuals, i.e., complete cases, is given by $n^{-1}\sum_{i=1}^n \frac{\Delta_i Y_i}{K(T_i)}$, where $\Delta_i = I(T_i \leq C_i)$. That this is an unbiased estimator of μ_Y is a consequence of the following equality:

$$E\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i Y_i}{K(T_i)}\right\} = E\left[E\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i Y_i}{K(T_i)}\Big|T_i, Z_i^H(T_i)\right\}\right]$$

$$
\begin{aligned}
&= E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{K(T_i)}E\{I(C_i \geq T_i)|T_i, Z_i^H(T_i)\}\right] \\
&= E\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right) = \mu_Y
\end{aligned}
$$

Because the underlying survival distribution of censoring $K(T_i)$ is unknown, we propose to estimate it using the Kaplan-Meier estimator (Kaplan & Meier 1958). This is done by simply reversing the roles of censoring time C_i and ascertainment time T_i; i.e. here the time T_i censors the censoring time C_i. That is, with the data $\{V_i = min(T_i, C_i), \Delta_i, i = 1, \ldots, n\}$, the Kaplan-Meier estimator for the distribution of the censoring time C_i is defined by

$$\hat{K}(t) = \prod_{u<t}\left\{1 - \frac{dN^c(u)}{Y(u)}\right\},$$

where $N^c(u) = \sum_{i=1}^{n} I(V_i \leq u, \Delta_i = 0)$, and $Y(u) = \sum_{i=1}^{n} I(V_i \geq u)$. The probability weighted complete-case estimator is defined as

$$\hat{\mu}_{WT} = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i Y_i}{\hat{K}(T_i)}.$$

This estimator is similar to that used by Zhao & Tsiatis (1997) to find a consistent estimator for the distribution for Quality-Adjusted Life. Consequently, consistency and asymptotic normality would follow exactly as outlined in that paper.

For our problem, we let the filtration $\mathcal{F}(u)$ be the set of σ-algebras generated by $\sigma\{I(C_i \leq t), t \leq u; T_i, Z_i(x), 0 \leq x < \infty, i = 1, \ldots, n\}$, which is the increasing information derived by the entire data process, including ascertainment times for all individuals together with the censoring times that are observed up to time u. Although this is not the usual filtration used for censored survival problems, we find that choosing an appropriate filtration is often key to being able to solve problems that involve censoring. Because our methodology depends on estimating the distribution of censoring rather than the distribution of failure times, we will use a counting process that counts the number of individuals censored over time instead of counting the number of uncensored events. Therefore, we consider the martingale process $M_i^c(u)$, which can be written as $M_i^c(u) = N_i^c(u) - \int_0^u \lambda^c(t)\mathcal{Y}_i(t)dt$, where $N_i^c(u) = I(V_i \leq u, \Delta_i = 0)$, $\mathcal{Y}_i(u) = I(V_i \geq u)$, and $\lambda^c(u)$ is the hazard function for the

censoring distribution. We also define $M^c(u) = \sum M_i^c(u)$, $N^c(u) = \sum N_i^c(u)$ and $\mathcal{Y}(u) = \sum \mathcal{Y}_i(u)$.

Using some well-known martingale integral equations, the probability weighted estimator can be expanded as

$$n^{1/2}(\hat{\mu}_{WT} - \mu_Y) = n^{-1/2}\sum_{i=1}^{n}(Y_i - \mu) - n^{-1/2}\sum_{i=1}^{n}\int_0^L \frac{dM_i^c(u)}{K(u)}\{Y_i - G(Y,u)\} + o_P(1) \quad (3.1)$$

where

$$\begin{aligned} G(Y,u) &= \frac{1}{S(u)}E\{Y_i I(T_i \geq u)\}, \\ S(u) &= P(T > u), \end{aligned}$$

and $o_P(1)$ is a term that converges in probability to zero as the sample size increase.

Since Y_i is $\mathcal{F}(0)$ measurable, the two terms in (3.1) are uncorrelated and we use the result of covariance structure of martingale process, the variance of $n^{1/2}(\hat{\mu}_{WT} - \mu_Y)$ is equal to

$$\mathrm{var}(Y_i - \mu) + \int_0^L \{G(Y_i^2, u) - G^2(Y_i, u)\}S(u)\frac{\lambda^c(u)}{K(u)}du \quad (3.2)$$

The detailed derivation of (3.1) and (3.2) is shown in Appendix A.

For large samples, the martingale version of central limit theorem (Fleming & Harrington 1991) can be used to show that $\hat{\mu}_{WT}$ is consistent and asymptotically normal, and that the asymptotic variance of $n^{1/2}(\hat{\mu}_{WT} - \mu_Y)$ can be estimated by

$$\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i(Y_i - \hat{\mu}_{WT})^2}{\hat{K}(T_i)} + \frac{1}{n}\int_0^L \frac{dN^c(u)}{\hat{K}(u)}\{\hat{G}(Y^2, u) - \hat{G}(Y, u)\}\right] \quad (3.3)$$

The second peice of the estimate is derived by taking the Nelson estimator $dN^c(u)/\mathcal{Y}(u)$ for $\lambda^c du$, the Kaplan-Meier estimators $\hat{K}(u)$ and $\hat{S}(u)$ for $K(u)$ and $S(u)$, and $\hat{G}(Y_i, u)$, $\hat{G}(Y_i^2, u)$ for $G(Y_i, u)$, $G(Y_i^2, u)$; where

$$\hat{G}(Y,u) = \frac{1}{n}\frac{1}{\hat{S}(u)}\sum_{i=1}^{n}\frac{\Delta_i Y_i I(T_i \geq u)}{\hat{K}(T_i)}$$

The key result, given by (3.1), is that the weighted estimator minus the estimand can be approximated by a sum of identically and independently distributed mean zero random variables. This representation allows us to establish the asymptotic properties using the standard central limit theorem. Most estimators are asymptotically

linear; that is, the estimator minus the estimand can be approximated by a sum of identically and independently distributed mean zero random variables, which are referred to as influence functions. The asymptotic properties for such asymptotically linear estimators are therefore directly related to their influence function. In particular, the asymptotic variance of the estimator is the variance of the influence function for a single observation.

3.4 Sequential Joint Distribution with Censored Data

The problems associated with incomplete follow-up are most pronounced during interim analyses when the trial participants have varying amounts of follow-up and completeness. In most clinical trials sequential stopping rules are used to stop a study early if treatment difference become sufficiently large during an interim analysis. In order to develop such sequential stopping rules with desired operating characteristics (i.e., type I and type II errors) we must be able to characterize the joint distribution of relevant test statistics and estimators over calendar time. In the previous section, we described a comprehensive approach for testing and estimating at a single analysis time with incomplete follow-up data. We will generalize these results in order to derive the joint distribution of the sequentially computed tests or estimates. Once we have derived the joint sequential distribution we will then be able to construct stopping boundaries for group sequential tests that have the desired type I error and power. The key to deriving the joint sequential distribution is to identify the influence function of the estimators at the interim monitoring times. The joint distribution of the sequentially computed estimators would then be asymptotically normal with the correlation structure of their corresponding influence functions.

To accommodate interim monitoring, we define a random variable E_i to denote calendar entry time into the study. By convention, the study will start at calendar time zero and patients will enter into the study at various calendar times measured from time zero. In its most general form, data from such a study can be visualized

as a realization of a sequence of independent random vectors $\{Y_i, T_i, Z_i^H(T_i), E_i\}$, $i = 1, \ldots, n$. It is important to distinguish between two different time scales, patient time versus calendar time. Entry into the study occurs at calendar time as does the analysis of the data at different interim times. However, the time T_i and the process $Z_i^H(x)$ are measured from the time patient i enters into the study. The variable E_i will be assumed to be independent of $\{Y_i, T_i, Z_i^H(T_i)\}$. In this representation, n is the total number of patients that will eventually enter the trial at random times E_i, $i = 1, \ldots, n$, during a fixed accrual period. Asymptotic theory, which is predicated on letting n go to infinity, is a consequence of an increasing number of individuals arriving during the accrual period. This type of asymptotic theory most closely represents the situation in a large scale clinical trial, and has been used to derive the sequential theory for clinical trials based on calendar time, (Tsiatis 1981,1982,1995). Of course, at any time t, the analysis is limited to only those $n(t) = \sum_{i=1}^n I(E_i \leq t)$ individuals who entered the study prior to that time. Also, if the analysis is conducted at calendar tiem t, then an individual would have incomplete follow-up (i.e, censored data) if $T_i \leq t - E_i$. We therefore define the censoring variable at analysis time t as $C_{ti} = t - E_i$.

In many instances, the research hypothesis under investigation can be cast as a hypothesis testing question involving a single scalar parameter β. For example, if tumor response is the primary outcome, β may denote the difference in the response rates between two treatments. The model is usually parametrized so that the primary question being tested is the null hypothesis $H_0 : \beta = 0$. In a group-sequential analysis, the null hypothesis is tested at different interim times with the possibility of early stopping if the test statistic becomes sufficiently large at any of the analyses. We will focus on the use of a Wald test for this purpose, although a parallel development could also be used for score tests. Towards that end, we denoted the estimator for a parameter of interest calculated using only the data available up to calendar time t, by $\hat{\beta}_t$. Most estimators are asymptotically linear; that is, they can be written as

$$n^{1/2}(\hat{\beta}_t - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \phi_t(Y_i, T_i, Z_i^H(\cdot), E_i) I(E_i \leq t) + o_p(1),$$

where $\phi_t(Y_i, T_i, Z_i^H(\cdot), E_i) I(E_i \leq t)$ is a mean zero random variable defiend as the ith

influence function of $\hat{\beta}_t$. By the central limit theorem, the asymptotic distribution of $n^{1/2}(\hat{\beta}_t - \beta_0)$ is normally distributed with mean zero and variance equal to the variance of $\phi_t(Y_i, T_i, Z_i^H(\cdot), E_i)I(E_i \leq t)$.

If the data are monitored at interim times $t_1, \ldots, t_K$, then the joint distribution of

$$n^{1/2}(\hat{\beta}_{t_1} - \beta_0), \ldots, n^{1/2}(\hat{\beta}_{t_K} - \beta_0)$$

equivalent to the joint distribution of their corresponding influence functions

$$n^{-1/2}\sum_{i=1}^{n}\phi_{t_1}(Y_i, T_i, Z_i^H(\cdot), E_i)I(E_i \leq t_1), \ldots, n^{-1/2}\sum_{i=1}^{n}\phi_{t_K}(Y_i, T_i, Z_i^H(\cdot), E_i)I(E_i \leq t_K) \quad (3.4)$$

By the multivariate central limit theorem, the K-dimensional random vector (3.4) converges to a multivariate normal random vector with mean zero and K×K covariance matrix Ω with elements

$$\Omega_{jk} = \mathrm{cov}[\phi_{t_j}(Y_i, T_i, Z_i^H(\cdot), E_i)I(E_i \leq t_j), \phi_{t_k}(Y_i, T_i, Z_i^H(\cdot), E_i)I(E_i \leq t_k)]; \quad (3.5)$$
$$j, k = 1, \ldots, K.$$

3.5 Application in Comparing Two Response Rates

As an example, we illustrate how this method would be applied if we wanted to evaluated the sequential distribution of an estimator for response rate with incomplete follow-up. Let θ denote the probability of response. This would correspond to the E(Y), where Y is a dichotomous indicator of response. Use of the inverse selection probability weighted estimator for θ at calendar time t would result in

$$\hat{\theta}_t = n^{-1}(t)\sum_{i=1}^{n}\frac{\Delta_{ti}}{\hat{K}_t(T_i)}Y_i I(E_i \leq t)$$

where $\Delta_{ti} = I(C_{ti} \geq T_{ti})$, $C_{ti} = t - E_i$, and $\hat{K}_t(u)$ is the Kaplan-Meier estimator for C_{ti} using the censored data $\min(T_i, C_{ti})$ for $\{\mathrm{i} : E_i \leq t\}$. Using the martingale results presented in previous section and the facts that $n(t)/N$ converges to $\pi(t) = \mathrm{P}(E_i \leq t), \hat{G}(Y, u)$ converges to $G(Y, u) = E[Y_i I(T_i \geq u)]/S(u)$, and $\hat{K}_t(u)$ converges

to $K_t(u) = \mathrm{P}(C_{ti} \geq u)/\pi(t)$ we may write

$$\begin{aligned} n^{1/2}(\hat{\theta}_t - \theta) &= n^{-1/2} \sum_{i=1}^{n} \pi(t)^{-1} \{Y_i - \theta\} I(E_i \leq t) \\ &\quad - n^{-1/2} \sum_{i=1}^{n} \pi(t)^{-1} \int \frac{dM^{C_{ti}}(u)}{K_t(u)} \{Y_i - G(Y,u)\} + o_p(1), \end{aligned}$$

where L is the maximum time to evaluation. This implies that the influence function for $\hat{\theta}_t$ is given by

$$\begin{aligned} \phi_t(Y_i, T_i, Z_i^H(.), E_i) I(E_i \leq t) &= \pi(t)^{-1} \{Y_i - \theta\} I(E_i \leq t) \\ &\quad - \pi(t)^{-1} \int_0^L \frac{dM^{C_{ti}}(u)}{K_t(u)} \{Y_i - G(Y,u)\} \quad (3.6) \end{aligned}$$

From (3.4), the asymptotic joint distribution of the estimator for θ evaluated at times $t_1, \ldots, t_K$ is a multivariate normal with mean zero and covariance matrix given by (3.5).

It remains now to evaluate the covariance structure. It suffices to consider the influence function at two arbitrary points $s < t$; that is, we wish to compute

$$\mathrm{cov}\Big(\phi_s(Y_i, T_i, Z_i^H(.), E_i) I(E_i \leq s), \phi_t(Y_i, T_i, Z_i^H(.), E_i) I(E_i \leq t)\Big)$$

The important useful feature is that the influence functionis is expressed as a stochastic integral of a martingale process. If we additionally note that the martingale process $M^{C_{si}}(u) = M^{C_{ti}}(u + (t - s))$, then the influence functions at times s and t may be written as stochastic integrals with respect to a common martingale process, $M^{C_{ti}}(u)$. Invoking some simplifying formula for the covariance of stochastic integrals with respect to the same counting process, then after some algebra we obtain

$$\begin{aligned} &\mathrm{cov}\Big(\phi_s(Y_i, T_i, Z_i^H(.), E_i) I(E_i \leq s), \phi_t(Y_i, T_i, Z_i^H(.), E_i) I(E_i \leq t)\Big) \\ &\qquad = \mathrm{var}\Big(\phi_t(Y_i, T_i, Z_i^H(.), E_i) I(E_i \leq t)\Big) \quad (3.7) \end{aligned}$$

A detailed proof is shown in Appendix B.

The covariance structure given by (3.7) is the structure of a sequential process with independent increments. This is important because most sequentially computed tests and estimates in the literature have this property. This allows us to use "information

based design and monitoring" for this problem. Statistical information is equated with the inverse of the variance of the estimator, which grows over time as we accumulate more data. The group-sequential stopping rules can then be constructed according to the methods proposed by Lan & DeMets (1983), whereby an alpha-spending function is specified that dictates how much of the significance level can be used as a function of the statistical information. This in turn may be translated into corresponding stopping boundaries that preserve the overall type I and type II error.

One immediate consequence of the above result involves testing the equality of response rates between two treatments. The test would be based on the difference between the treatment specific estimates of response rates computed above. A simple calculation shows that the difference of the two estimators of responses, calculated sequentially over calendar time, also has the independent increments structure. The monitoring process for this problem may then be carried out routinely.

3.6 Simulations

A numerical simulation is performed to test the equality of response rates between two treatments, A, B, with information-based monitoring procedure. Let π_i denote the probability of response for each treatment. The null hypothesis is $H_0 : \theta = \pi_A - \pi_B = 0$. A set of data are generated under the assumption that individual entry time E_i follows a Poisson process with total entry of 500 subjects per time unit. The outcome of response Y is a dichotomous variable. And individual response or non-response time to each treatment is exponentially distributed with parameter λ varied. That is, the lag time T_{Ai} for individual i to respond to treatment A has $\lambda_A = 1$ while for non-responding to treatment A, the parameter $\lambda_A^N = 2$. Similarly for treatment B, the lag time T_{Bi} follows exponential distributions with $\lambda_B = 2$ for responders and $\lambda_B^N = 1$ for nonresponders. The rates of information available at time t not only vary by response time but also by treatments. This design signifies the bias that is caused by incomplete follow-up in a trial. The data are divided into groups according to treatments. Let Y_{li} be the response outcome of an individual taking treatment l, $l =$ A, B.; $n_l(t) = \sum I(E_i \leq t) I(\text{Trt} = l)$, where $I(\text{Trt} = l)$ is the indicator for subjects

taking treatment l; $\Delta_{tli} = I(C_{ti} \geq T_{li})$, where $C_{ti} = t - E_i$, the censored time.

Before carrying out the information-based monitoring procedure, we will first show the estimate and and asymptotic variance for each method. Naively, we could ignore the problem of incomlete follow-up and use the data available at time t to estimate the probability of responses. The naive estimator is denoted by $\hat{\theta_{NV}}$.

$$\hat{\theta}_{NV} = \hat{\pi}_A^{NV} - \hat{\pi}_B^{NV} = \frac{1}{n_A(t)} \sum_{i=1}^{n_A(t)} Y_{Ai}\Delta_{tAi} - \frac{1}{n_B(t)} \sum_{i=1}^{n_B(t)} Y_{Bi}\Delta_{tBi},$$
$$\text{where } \hat{\pi}_A^{NV} = \frac{1}{n_A(t)} \sum_{i=1}^{n_A(t)} Y_{Ai}\Delta_{tAi},\ \hat{\pi}_B^{NV} = \frac{1}{n_B(t)} \sum_{i=1}^{n_B(t)} Y_{Bi}\Delta_{tBi}.$$

And the asymptotic variance for $\hat{\theta}_{NV}$ is computed impirically. That is,

$$\begin{aligned}\hat{\text{Var}}(\hat{\theta}_{NV}) &= \frac{1}{n_A(t)}\frac{1}{n_A(t)-1} \sum_{i=1}^{n_A(t)} \{I(E_i \leq t)I(\text{Trt} = A) - \hat{\pi}_A^{NV}\}^2 \\ &\quad + \frac{1}{n_B(t)}\frac{1}{n_B(t)-1} \sum_{i=1}^{n_B(t)} \{I(E_i \leq t)I(\text{Trt} = B) - \hat{\pi}_B^{NV}\}^2\end{aligned}$$

For a desired 5% type I error and 90% power, the naive estimate is expected to have a much higher risk of making a type I error in a trial.

To take into account of the lag time variable, we assume the maximum lag time to be 1 time unit. That is, at time t the information before time $(t-1)$ will be complete. Thus at each interim analysis, we draw inference using data only one unit before time. The maximum lag time estimator

$$\hat{\theta}_{ML} = \hat{\pi}_A^{ML} - \hat{\pi}_B^{ML} = \frac{1}{n_A(t-1)} \sum_{i=1}^{n_A(t-1)} Y_{Ai} - \frac{1}{n_B(t-1)} \sum_{i=1}^{n_B(t-1)} Y_{Bi},$$
$$\text{where } \hat{\pi}_A^{ML} = \frac{1}{n_A(t-1)} \sum_{i=1}^{n_A(t-1)} Y_{Ai},\ \hat{\pi}_B^{ML} = \frac{1}{n_B(t-1)} \sum_{i=1}^{n_B(t-1)} Y_{Bi}.$$

And the asymptotic variance

$$\hat{\text{Var}}(\hat{\theta}_{ML}) = \frac{\hat{\pi}_A^{ML}(1-\hat{\pi}_A^{ML})}{n_A(t-1)} + \frac{\hat{\pi}_B^{ML}(1-\hat{\pi}_B^{ML})}{n_B(t-1)}$$

$\hat{\theta}_{ML}$ is a consistent estimate. But since this method discards partial information available at hands, it will require more subjects entered into the study to compensate for the loss of information.

The inverse weighted estimator for the probability of response for each treatment group is

$$\hat{\pi}_l^{IW} = \frac{1}{n_l(t)} \sum_{i=1}^{n_l(t)} \frac{\Delta_{tli} Y_{li}}{\hat{K}_{tl}(T_{li})},\ l = A, B.$$

Thus the inverse weighted estimator for θ is

$$\hat{\theta}_{IW} = \frac{1}{n_A(t)} \sum_{i=1}^{n_A(t)} \frac{\Delta_{tAi} Y_{Ai}}{\hat{K}_{tA}(T_{Ai})} + \frac{1}{n_B(t)} \sum_{i=1}^{n_B(t)} \frac{\Delta_{tBi} Y_{Bi}}{\hat{K}_{tB}(T_{Bi})}$$

The asymptotic variance of $\hat{\theta}_{IW}$ is the sum of the asymptotic variance based on (3.3) under each treatment group . Simulations should demonstrate that this method maintains the desired type I and type II errors, and reduces the number of individuals required in a trial than the maximum lag time method.

With the estimated variances, we can compute the proportion of information, I/MI, at each interim time where I is the inverse variance and MI is the maximum number. A study ends when a clinical difference is identified or when all information is exhausted, i.e. I/MI=1. To futher enhance the phenomena that incomplete follow-up may lead to biased results in the early interim analysis, we choose the alpha-spending function that resembles the Pocock boundary which is more optimistic in early stopping. For the probability of response ranging from 0.1 to 0.5, we would like to detect a 0.05 difference under the alternative hypothesis. Given the expected 5% level of significance and 90% power, the standard error under 1000 simulations for each hypothesis is equal to $\sqrt{\frac{0.05(1-0.05)}{1000}} = 0.006892$. Therefore the simulation type I error should be within the interval of (0.043, 0.057), and power within (0.893, 0.907). We also give the average sample number (ASN) for 1000 simulations to compare on all three methods. See Table 1.

It is clear that the naive method has type I error as high as 26% when it is expected to have only 5% of chance. This is caused by the bias of data representation at interim times. After considering lag time effect, the maximum lag time method has type I error in the range of expectation. For the same or higher type I error, the proposed inverse weighted method has ASN less than the maximum lag time method, which will reserve the resource and save on the management cost. The powers of the last

two methods show lower than expected because Pocock boundaries are more liberate in stopping in the early stage of a trial, therefore increasing the chance of making type II error. For the use of O'Brien–Fleming boundaries, The powers obtained from the maximum lag and inverse weighted methods are indeed inside in expected range.

3.7 Future Approach

We have demonstrated that our proposed estimator is consistent and has the property of independent increments in sequential monitoring even when data are lagged or censored. The numerical study also shows that this estimator maintains desired type I and type II errors and reduces particapants entered into a trial. For the study of efficiency, Bang & Tsiatis (1999) have built an improved estimator at one time point using the theory of semiparameteric efficiency for missing data processes given by Robins & Rotnitzky (1992), and Robins, Rotnitzky & Zhao (1994). Our next approach is to implement this improved estimator in sequential monitoring, and to show its status in type I and type II errors numerically. It will be very complicated to theoretically prove that this improved estimator has the property of independent increments. It may even turn out to be impossible to obtain a performable form. But for the benefit of practical applicance, if the numerical simulation based on the property proved to be holding up, we will be able to assume that the improved estimator is at least very closed to have independent increments covariance structure.

Table 3.1: Simulation result on 3 parameter estimators with incomplete follow-up data. Pocock boundaries are used.

Test π_A, π_B	Naive $\alpha/(1-\gamma)$, ASN	Max. lag time $\alpha/(1-\gamma)$, ASN	Inverse weighted $\alpha/(1-\gamma)$, ASN
0.10, 0.10	0.18, 2012	0.05, 2332	0.05, 2151
0.15, 0.10		0.90, 1812	0.90, 1565
0.20, 0.20	0.24, 3117	0.07, 3663	0.05, 3568
0.25, 0.20		0.90, 2514	0.90, 2307
0.30, 0.30	0.21, 4181	0.06, 4634	0.05, 4488
0.35, 0.30		0.90, 2978	0.90, 2796
0.40, 0.40	0.22, 4723	0.05, 5224	0.05, 5103
0.45, 0.40		0.90, 3241	0.90, 3124
0.50, 0.50	0.19, 5150	0.05, 5224	0.06, 4436
0.55, 0.50		0.89, 3451	0.90, 3181

Table 3.2: Simulation result on 3 parameter estimators with incomplete follow-up data. O'Brien-Fleming boundaries are used.

Test π_A, π_B	Naive $\alpha/(1-\gamma)$, ASN	Max. lag time $\alpha/(1-\gamma)$, ASN	Inverse weighted $\alpha/(1-\gamma)$, ASN
0.10, 0.10	0.13, 2119	0.05, 2219	0.06, 1834
0.15, 0.10		0.90, 1980	0.88, 1455
0.20, 0.20	0.23, 3201	0.05, 3292	0.05, 2972
0.25, 0.20		0.90, 2736	0.89, 2138
0.30, 0.30	0.14, 3144	0.05, 4159	0.06, 3793
0.35, 0.30		0.89, 3262	0.90, 2573
0.40, 0.40	0.13, 3479	0.06, 4645	0.04, 4291
0.45, 0.40		0.91, 3508	0.90, 2808
0.50, 0.50	0.11, 4765	0.05, 4806	0.05, 4436
0.55, 0.50		0.90, 3544	0.89, 2868

Bibliography

Armitage, P., McPherson, C. K. & Rowe, B. C. (1969), 'Repeated significance tests on accumulating data', *Journal of the royal statistical society* **132**, A:235–244.

Bang, H. & Tsiatis, A. A. (1999), 'Estimating medical costs with censored data', *Biometrika*.

Fleming, T. R. & Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, Wiley, New York.

Friedman, L. M., Furberg, C. D. & DeMets, D. L. (1998), *Fundamentals of Clinical Trials*, 3rd edn, Springer, New York.

Gill, R. D. (1980), *Censoring and Stochastic Integrals*, Mathematisch Centrum, Amsterdam. Mathematical Centre Tracts 124.

Jewell, N., Dietz, K. & Farewell, V., eds (1992), *AIDS Epidemiology-Methodological Issues*.

Kaplan, E. L. & Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of The American Statistical Association* **53**, 457–481.

Lan, G. K. K. & DeMets, D. L. (1983), 'Discrete sequential boundaries for clinical trials', *Biometrika* **70**, 659–663.

Lan, G. K. K. & Zucker, D. M. (1993), 'Sequential monitoring of clinical trials: The role of information in brownian motion', *Statistics in Medicine* **12**, 753–765.

O'Brien, P. C. & Fleming, T. R. (1979), 'A multiple testing procedure for clinical trials', *Biometrics* **35**, 549–556.

Pocock, S. J. (1977), 'Group sequential methods in the design and analysis of clinical trials', *Biometrika* **64**, 191–199.

Robins, J. M. & Rotnitzky, A. (1992), Recovery of information and adjustment for dependent censoring using surrogate markers, *in* Jewell, Dietz & Farewell (1992), pp. 297–331.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of The American Statistical Association* **89**, 846–866.

Sharfstein, D. O. & Tsiatis, A. A. (1997), 'The use of simulation and bootstrap in information-based group sequential studies', *Statistics in medicine* **17**, 75–87.

Sharfstein, D. O., Tsiatis, A. A. & Robins, J. M. (1997), 'Semiparametric efficiency: implications for the design and analysis of group sequential studies', *Journal of the American Statistical Association* **92**(440), 1342–1350.

Wang, S. K. & Tsiatis, A. A. (1987), 'Approximately optimal one-parameter boundaries for group sequential trials', *Biometric* **43**, 193–199.

Zhao, H. & Tsiatis, A. A. (1997), 'A consistent estimator for the distribution of quality ajusted survival time', *Biometrika* **84**, 339–348.

APPENDICES

Appendix A.

Derivation of $n^{1/2}(\hat{\mu}_{WT} - \mu_Y)$ and its variance

Using an identity from Robins & Rotnitzky (1992, p.313)

$$\frac{\Delta_i}{K(V_i)} = 1 - \int_0^\infty \frac{dM_i^c(u)}{K(u)},$$

a well-known martingale integral representation Gill (1980, p.37) that

$$\frac{\hat{K}(t) - K(t)}{K(t)} = -\int_0^t \frac{\hat{K}(u^-)}{K(u)} \frac{dM^c(u)}{\mathcal{Y}(u)},$$

where $\hat{K}(u^-)$ is the left-continuous version of Kaplan-Meier estimator for censoring, and also

$$n^{-1}\mathcal{Y}(u) = \hat{K}(u^-)\hat{S}(u^-),$$

where $\hat{S}(u)$ is the Kaplan-Meier estimator for $S(u) = \mathrm{P}(T > u)$, the simple weighted estimator can be expanded as

$$\begin{aligned} n^{1/2}(\hat{\mu}_{WT} - \mu_Y) &= n^{-1/2}\sum_{i=1}^n \frac{\Delta_i Y_i}{K(T_i)} + n^{-1/2}\sum_{i=1}^n \frac{\Delta_i Y_i}{K(T_i)} \left\{ \frac{K(T_i) - \hat{K}(T_i)}{\hat{K}(T_i)} \right\} - n^{1/2}\mu_Y \\ &= n^{-1/2}\sum_{i=1}^n (Y_i - \mu) - n^{-1/2}\sum_{i=1}^n \int_0^L \frac{dM_i^c(u)}{K(u)} \{Y_i - \hat{G}(Y,u)\} \\ &= n^{-1/2}\sum_{i=1}^n (Y_i - \mu) - n^{-1/2}\sum_{i=1}^n \int_0^L \frac{dM_i^c(u)}{K(u)} \{Y_i - G(Y,u)\} + o_P(1) \end{aligned} \tag{A.1}$$

where

$$\begin{aligned} G(Y,u) &= \frac{1}{S(u)} E\{Y_i I(T_i \geq u)\}, \\ \hat{G}(Y,u) &= \frac{1}{n}\frac{1}{\hat{S}(u)}\sum_{i=1}^n \frac{\Delta_i Y_i I(T_i \geq u)}{\hat{K}(T_i)}, \end{aligned} \tag{A.2}$$

and $o_P(1)$ is a term that converges in probability to zero as the sample size increase.

Since the first two terms in $(A.1)$ are uncorrelated and we use the result of covariance structure of martingale process, the variance of $n^{1/2}(\hat{\mu}_{WT} - \mu_Y)$ is equal

to

$$\begin{aligned}
&\mathrm{var}(Y_i-\mu)+E\left[\int_0^L \frac{\{Y_i-G(Y,u)\}^2}{K^2(u)}\lambda^c(u)\mathcal{Y}_i(u)du\right]\\
&= \mathrm{var}(Y_i-\mu)+E\left[\int_0^L \{Y_i-G(Y,u)\}^2 I(T_i\geq u)\frac{\lambda^c(u)}{K^2(u)}du\right],
\end{aligned}$$

where $\mathcal{Y}_i(u)=I(V_i\geq u)=I(\min(T_i,C_i)\geq u)=I(T_i\geq u)I(C_i\geq u)$

and $\dfrac{\mathcal{Y}_i(u)}{K(u)}=\dfrac{I(T_i\geq u)I(C_i\geq u)}{\mathrm{P}(C_i\geq u)}=I(T_i\geq u)$

$$= \mathrm{var}(Y_i-\mu)+\int_0^L \{G(Y_i^2,u)-G^2(Y_i,u)\}S(u)\frac{\lambda^c(u)}{K(u)}du$$

Appendix B.

Proof of independent increment structure using influence function ϕ of two response rates.

Our goal is to show that

$$\begin{aligned}
&\mathrm{cov}\Big(\phi_s(Y_i,T_i,Z_i^H(.),E_i)I(E_i\leq s),\phi_t(Y_i,T_i,Z_i^H(.),E_i)I(E_i\leq t)\Big)\\
&\qquad = \mathrm{var}\Big(\phi_t(Y_i,T_i,Z_i^H(.),E_i)I(E_i\leq t)\Big)
\end{aligned}$$

To start with, we reform the influence functions (3.6) at the two time points and shorten the notation $\phi_t(Y_i,T_i,Z_i^H(\cdot),E_i)$ to ϕ_t, $\phi_s(Y_i,T_i,Z_i^H(\cdot),E_i)$ to ϕ_s. And note that

$$\begin{aligned}
dM^{C_{ti}}(u) &= dN_i^c(u)-\lambda_t^c(u)y_i(u)du\\
&= I(C_{ti}=u,T_i\geq u)-\lambda_t^c(u)I(C_{ti}\geq u,T_i\geq u)\\
&= \Big[I(C_{ti}=u)-\lambda_t^c(u)I(C_{ti}\geq u)\Big]I(T_i\geq u)\\
&= dM^{*C_{ti}}(u)I(T_i\geq u)\\
&\quad \text{, where } dM^{*C_{ti}}(u)=I(C_{ti}=u)-\lambda_t^c(u)I(C_{ti}\geq u)
\end{aligned}$$

$M^{*C_{ti}}(u)$ is also a martingale process at the time censoring occurred.
Thus at time t,

$$\phi_t I(E_i \le t) = \pi(t)^{-1}\{Y_i - \theta\} I(E_i \le t) - \pi(t)^{-1} \int_0^L \frac{dM^{*C_{ti}}(u)}{K_t(u)} \{Y_i - G(Y,u)\} I(T_i \ge u) \tag{B.1}$$

At time s, it can be shown following the definition that

$$dM^{*C_{si}}(u) = dM^{*C_{ti}}(u - (t-s))$$

$$K_s(u)\phi(s) = K_t(u)\phi(t)$$

Hence,

$$\begin{aligned}
\phi_s I(E_i \le s) &= \pi(s)^{-1}\{Y_i - \theta\} I(E_i \le s) - \pi(s)^{-1} \int_0^L \frac{dM^{*C_{si}}(u)}{K_s(u)} \{Y_i - G(Y,u)\} I(T_i \ge u) \\
&= \pi(s)^{-1}\{Y_i - \theta\} I(E_i \le s) \\
&\quad - \pi(t)^{-1} \int_0^L \frac{dM^{*C_{ti}}(u + (t-s))}{K_t(u + (t-s))} \{Y_i - G(Y,u)\} I(T_i \ge u) \\
&= \pi(s)^{-1}\{Y_i - \theta\} I(E_i \le s) \\
&\quad - \pi(t)^{-1} \int_{t-s}^{L+(t-s)} \frac{dM^{*C_{ti}}(u)}{K_t(u)} \{Y_i - G(Y, u - (t-s))\} I(T_i \ge u - (t-s))
\end{aligned} \tag{B.2}$$

Now to find the covariance of the influence functions at time s and t,

$$\operatorname{cov}\Big(\phi_s I(E_i \le s), \phi_t I(E_i \le t)\Big) = E\Big(\phi_s I(E_i \le s)\phi_t I(E_i \le t)\Big) - E\Big(\phi_s I(E_i \le s)\Big) E\Big(\phi_t I(E_i \le t)\Big)$$

There are four cross products from $E\Big(\phi_s I(E_i \le s)\phi_t I(E_i \le t)\Big)$.

- The cross product of 1st terms of $(B.1)$& $(B.2)$

$$E\Big[\pi(t)^{-1}\pi(s)^{-1}\{Y_i - \theta\}^2 I(E_i \le s) I(E_i \le t)\Big] = \pi(t)^{-1} E\{Y_i - \theta\}^2 \tag{B.3}$$

  Since s < t, $I(E_i \le s)I(E_i \le t) = I(E_i \le s)$, E_i is independent of Y_i, and $E\Big(I(E_i \le s)\Big) = \pi(s)$.

- The cross product of 1st term of $(B.1)$ & 2nd term of $(B.2)$

$$\begin{aligned}
&\pi(t)^{-2}E\left[\{Y_i-\theta\}I(E_i\le t)\int_{t-s}^{L+t-s}\frac{dM^{*C_{ti}}(u)}{K_t(u)}\{Y_i-G(Y,u-(t-s))\}I(T_i\ge u-(t-s))\right]\\
&= \pi(t)^{-2}E\left[\int\frac{dM^{*C_{ti}}(u)I(E_i\le t)}{K_t(u)}\{Y_i-\theta\}\{Y_i-G(Y,u-(t-s))\}I(T_i\ge u-(t-s))\right]\\
&= \pi(t)^{-2}E\left[\int\frac{dM^{*C_{ti}}(u)}{K_t(u))}\{Y_i-\theta\}\{Y_i-G(Y,u-(t-s))\}I(T_i\ge u-(t-s))\right] \qquad \text{(B.4)}
\end{aligned}$$

where $dM^{*C_{ti}}(u)I(E_i\le t)=dM^{*C_{ti}}(u)$.
The range $[t-s,\ \mathrm{L}+(t-s)]=[0,\ \mathrm{L}+(t-s)]\setminus[0,\ t-s)$. The integral from $[t-s,\ \mathrm{L}+(t-s)]$ in $(B.4)$ is then the difference of two stochastic integrals, hence has expectation zero. Therefore, $(B.4)=0$.

- The cross product of 2nd term of $(B.1)$ & 1st term of $(B.2)$

$$\begin{aligned}
&\pi(s)^{-1}\pi(t)^{-1}E\left[\{Y_i-\theta\}I(E_i\le s)\int_0^{L}\frac{dM^{*C_{ti}}(u)}{K_t(u)}\{Y_i-G(Y,u)\}I(T_i\ge u)\right]\\
&=\pi(s)^{-1}\pi(t)^{-1}E\left[\int_0^{L}\frac{dM^{*C_{ti}}(u)I(E_i\le s)}{K_t(u)}\{Y_i-\theta\}\{Y_i-G(Y,u)\}I(T_i\ge u)\right]
\end{aligned} \qquad \text{(B.5)}$$

$I(E_i\le s)=I(t-E_i\ge t-s)$

$$\begin{aligned}
dM^{*C_{ti}(u)}I(E_i\le s) &= \Big[I(C_{ti}=u)-\lambda_t^c(u)I(C_{ti}\ge u)\Big]I(E_i\le s)\\
&= I(t-E_i=u)I(t-E_i\ge t-s)-\lambda_t^c(u)I(t-E_i\ge u)I(t-E_i\ge t-s)\\
&= \begin{cases} I(t-E_i=u)-\lambda_t^c(u)I(t-E_i\ge u) & \text{, for } u\ge t-s\\ -\lambda_t^c(u)I(t-E_i\ge t-s) & \text{, for } u<t-s\end{cases}\\
&= \begin{cases} dM^{*C_{ti}}(u) & \text{, for } u\ge t-s\\ -\lambda_t^c(u)I(E_i\le s) & \text{, for } u<t-s\end{cases}
\end{aligned}$$

Therefore,

$$\begin{aligned}
(B.5)=\pi(s)^{-1}\pi(t)^{-1}E&\left[\int_0^{t-s}\frac{-\lambda_t^c(u)I(E_i\le s)}{K_t(u)}\{Y_i-\theta\}\{Y_i-G(Y,u)\}I(T_i\ge u)\right]\\
&+\pi(s)^{-1}\pi(t)^{-1}E\left[\int_{t-s}^{L}\frac{dM^{*C_{ti}}(u)}{K_t(u)}\{Y_i-\theta\}\{Y_i-G(Y,u)\}I(T_i\ge u)\right]
\end{aligned} \qquad \text{(B.6)}$$

The second item in $(B.6)$ is the expectation of a stochastic integral from [0, L] minus the same martingale process but from $[0,\ t-s)$, hence is equal to 0. For the first item we take the expectation inside the integral, since $I(E_i \le s)$ is independent of Y_i and $E\Big(I(E_i \le s)\Big) = \pi(s)$ which cancels out $\pi(s)^{-1}$.

$$\begin{aligned}(B.6) &= \pi(t)^{-1}\int_0^{t-s}\frac{-\lambda_t^c(u)}{K_t(u)}E\Big[\{Y_i-\theta\}\{Y_i-G(Y,u)\}I(T_i\ge u)\Big]\\ &= \pi(t)^{-1}E\left[\int_0^{t-s}\frac{-\lambda_t^c(u)}{K_t(u)}\{Y_i-G(Y,u)\}^2 I(T_i\ge u)\right] \qquad \text{(B.7)}\end{aligned}$$

where

$$\begin{aligned}&E\Big[\{Y_i-\theta\}\{Y_i-G(Y,u)\}I(T_i\ge u)\Big]\\ &= E\Big[\{Y_i-G(Y,u)+G(Y,u)-\theta\}\{Y_i-G(Y,u)\}I(T_i\ge u)\Big]\\ &= E\Big[\{Y_i-G(Y,u)\}^2 I(T_i\ge u)\Big]+G(Y,u)E\Big[Y_iI(T_i\ge u)\Big]-G^2(Y,u)S(u)\\ &\quad -\theta E\Big[Y_iI(T_i\ge u)\Big]+\theta G(Y,u)S(u)\\ &= E\Big[\{Y_i-G(Y,u)\}^2 I(T_i\ge u)\Big],\\ &\quad \text{note that } G(Y,u)=\frac{1}{S(u)}E\{Y_iI(T_i\ge u)\}\end{aligned}$$

- The cross product of integral terms in $(B.1)$, $(B.2)$

$$\begin{aligned}&\pi(t)^{-2}E\left[\int_0^L\frac{dM^{*C_{ti}}(u)}{K_t(u)}\{Y_i-G(Y,u)\}I(T_i\ge u)\right.\\ &\qquad\left.\times\int_{t-s}^{L+(t-s)}\frac{dM^{*C_{ti}}(u)}{K_t(u)}\{Y_i-G(Y,u-(t-s))\}I(T_i\ge u-(t-s))\right]\\ &= \pi(t)^{-2}E\left[\int_{t-s}^{L}\frac{\lambda_t^c(u)I(t-E_i\ge u)}{K_t^2(u)}\{Y_i-G(Y,u)\}\{Y_i-G(Y,u-(t-s))\}I(T_i\ge u)\right]\end{aligned} \qquad \text{(B.8)}$$

This is resulted by applying the covariance structure of stochastic integrals, also $I(T_i \ge u)I(T_i \ge u-(t-s)) = I(T_i \ge u)$.

We again take expectation inside the integral and since E_i is independent of Y_i, T_i, $E\Big(I(t-E_i\ge u)\Big) = K_t(u)\pi(t)$,

$$\begin{aligned}(B.8) &= \pi(t)^{-1}\int_{t-s}^{L}\frac{\lambda_t^c(u)}{K_t(u)}E\Big[\{Y_i-G(Y,u)\}\{Y_i-G(Y,u-(t-s))\}I(T_i\ge u)\Big]\\ &= \pi(t)^{-1}E\left[\int_{t-s}^{L}\frac{\lambda_t^c(u)}{K_t(u)}\{Y_i-G(Y,u)\}^2 I(T_i\ge u)\right] \qquad \text{(B.9)}\end{aligned}$$

where with adding and subtracting $G(Y,u)$ on $\{Y_i - G(Y,u-(t-s))\}$,

$$E\Big[\{Y_i - G(Y,u)\}\{Y_i - G(Y,u-(t-s))\}I(T_i \geq u)\Big] = E\Big[\{Y_i - G(Y,u)\}^2 I(T_i \geq u)\Big]$$

Now adding $(B.3)$, $(B.4)$, $(B.7)$, $(B.9)$, we obtain

$$\begin{aligned}
&E\Big(\phi_s I(E_i \leq s)\phi_t I(E_i \leq t)\Big) \\
&= \pi(t)^{-1}E\{Y_i - \theta\}^2 + 0 - \pi(t)^{-1}E\left[\int_0^{t-s} \frac{-\lambda_t^c(u)}{K_t(u)}\{Y_i - G(Y,u)\}^2 I(T_i \geq u)\right] \\
&\quad + \pi(t)^{-1}E\left[\int_{t-s}^{L} \frac{\lambda_t^c(u)}{K_t(u)}\{Y_i - G(Y,u)\}^2 I(T_i \geq u)\right] \\
&= \pi(t)^{-1}E\{Y_i - \theta\}^2 + \pi(t)^{-1}E\left[\int_0^{L} \frac{\lambda_t^c(u)}{K_t(u)}\{Y_i - G(Y,u)\}^2 I(T_i \geq u)\right] \\
&= \operatorname{cov}\Big(\phi_s I(E_i \leq s), \phi_t I(E_i \leq t)\Big)
\end{aligned}$$

since $E\Big(\phi_s I(E_i \leq s)\Big) = 0 = E\Big(\phi_t I(E_i \leq t)\Big)$.
We are now left to find the variance of $\phi_t I(E_i \leq t)$ and show it is equal to the covariance above.

$$\begin{aligned}
&\operatorname{var}\Big(\phi_t I(E_i \leq t)\Big) = E\Big(\phi_t I(E_i \leq t)\Big)^2 \\
&= \pi(t)^{-1}E\{Y_i - \theta\}^2 + \pi(t)^{-2}E\left[\int_0^{L} \frac{\lambda_t^c(u) I(t - E_i \geq u)}{K_t^2(u)}\{Y_i - G(Y,u)\}^2 I(T_i \geq u)\right] \\
&= \pi(t)^{-1}E\{Y_i - \theta\}^2 + \pi(t)^{-1}E\left[\int_0^{L} \frac{\lambda_t^c(u)}{K_t(u)}\{Y_i - G(Y,u)\}^2 I(T_i \geq u)\right] \\
&= \operatorname{cov}\Big(\phi_s I(E_i \leq s), \phi_t I(E_i \leq t)\Big)
\end{aligned}$$