INFERENCE, POWER AND SAMPLE SIZE FOR ADAPTIVE TWO-STAGE TREATMENT STRATEGIES

by

Wentao Feng

B. S. in Chemistry, Peking University, China, 2001

M. S. in Chemistry, Carnegie Mellon University, 2003

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Wentao Feng

It was defended on

April 3, 2008

and approved by

Dissertation Advisor:
Abdus S. Wahed, PhD
Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Jong-Hyeon Jeong, PhD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Howard E. Rockette, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Chung-Chou Ho Chang, PhD
Research Assistant Professor
Department of Medicine and Biostatistics
School of Medicine, Graduate School of Public Health
University of Pittsburgh

INFERENCE, POWER AND SAMPLE SIZE FOR ADAPTIVE TWO-STAGE TREATMENT STRATEGIES

Wentao Feng, PhD

University of Pittsburgh, 2008

An adaptive treatment strategy (ATS) is defined as a sequence of treatments and intermediate responses. ATS' arise when chronic diseases such as cancer and depression are treated over time with various treatment alternatives depending on intermediate responses to earlier treatments. For example, in two-stage adaptive treatment strategies, patients receive one of the induction treatments followed by a maintenance therapy given that the patients responded to the induction treatment they received. Clinical trials are often designed to compare adaptive treatment strategies based on appropriate designs such as sequential randomization designs. One of the main objectives of these trials is to compare two or more treatment strategies in terms of largest patient benefit, such as prolonged survival.

Statistical inference from such trials needs to account for the sequential randomization structure of the design. Recent literature suggests several methods of estimation. A comparative review of available inferential procedures for analyzing data from such trials is presented. A sample size formula is introduced for comparing the survival probabilities under two treatment strategies sharing the same initial treatment. The formula is based on the large sample properties of inverse-probability-weighted estimator. Monte Carlo simulation study shows strong evidence that the proposed sample size formula guarantees desired power, regardless of the true distributions of survival times.

To test for a difference in the effects of different induction and maintenance treatment combinations, a supremum weighted log-rank test is proposed. The test is applied to a dataset from a two-stage randomized trial and the results are compared to those obtained using a standard weighted log-rank test. A sample-size formula is derived based on the limiting distribution of the supremum weighted log-rank statistic. Simulation studies show that the proposed test provides sample sizes which are close to those obtained by standard weighted log-rank test under a proportional hazard alternative. However, the proposed test is more powerful than the standard weighted log-rank test under non-proportional hazard alternatives.

The public health significance of this work is to provide a practical guidance of sample size determination and a test procedure in clinical trials that adopt two stage randomization designs.

Keywords: Adaptive treatment strategy; Brownian motion; Counting process; Inverse-probability-weighting; Potential outcomes; Proportional hazards; Supremum log-rank statistic; Survival function; Two-stage randomization design.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

My first, and most earnest, acknowledgment must go to my advisor, Dr. Abdus S Wahed, whose kindness, enthusiasm, and support made all the difference in my academic career. His mentorship was paramount in providing me with the foundation for becoming a biostatistician and researcher. The insightful guidance he offered me has been as general as statistical methodologies and career goals, but also as detailed as scientific writing and presentation skills. He even read my terrible drafts and edited my grammar disasters without a complaint - I am sure he would have proofread this acknowledgment had I asked.

I gratefully acknowledge all the members of my committee, Dr. Rockette, Dr. Jeong and Dr. Chang, for their valuable advice, heartful encouragement and generous accessibility during my graduate career at University of Pittsburgh.

The faculty and staff at the Department of Biostatistics in University of Pittsburgh are the most dedicated people that I have ever met and I feel honored to have worked with them. Their guidance has served me well and I owe them my heartfelt appreciation.

I would like to extend my thanks to my colleagues and friends in Pittsburgh, who made me never feel alone and completed my life in so many ways. I consider the friendship with them as my most precious treasure.

A sincere and sweet thank-you goes to my loving husband, whose sacrifice and encouragement made all my achievement possible. My parents-in-law also deserve a special note of thanks, who have given me their unconditional support both emotionally and financially.

From the bottom of my heart, I am also indebted to my dearest elder sister, who has not only believed in me since we were both kids, but also has been keeping close company to our parents so that I can feel less guilty for being so far away from them, and be more focused on my graduate study.

No words can nearly express the full measure of my appreciation to my wonderful parents, who inspired me to love learning and who always made my education one of their top priorities. I owe them everything and I believe the only way to show them just how much I love and appreciate them is trying to be a better person.

Thanks to all the great people in my life, it has been a wonderful journey and I could not ask for more.

1.0 INTRODUCTION

1.1 ADAPTIVE TREATMENT STRATEGIES

An adaptive treatment strategy (also known as dynamic treatment regime) is an individually tailored series of decision rules specifying how treatment option should vary over time. The rule at each stage uses time-varying measurements of response, adherence, and other patient characteristics up to that point to determine the next treatment type and/or dosage. The decision rules comprising a treatment regime are made prior to the beginning of the course of treatment. Dynamic treatment regimes are widely used in the treatment of chronic or complex diseases such as cancer, AIDS, hepatitis and mental illness, where the presence of heterogeneity in response, potential for relapse, variability of patients characteristics and problems with adherence demands the adjustments of clinical decisions over time. The objective in developing such multistage decision-making strategies is to improve patient outcomes over time. The study of sequenced treatment alternatives to relieve depression (STAR*D) by Rush et al. [1] is one such example where patients were treated according to one of several available treatments (or different doses of same drug) for a fixed period of time and then based on the intermediate response were switched to a different treatment. The main objective of such trials is to compare different treatment strategies in search of the best one.

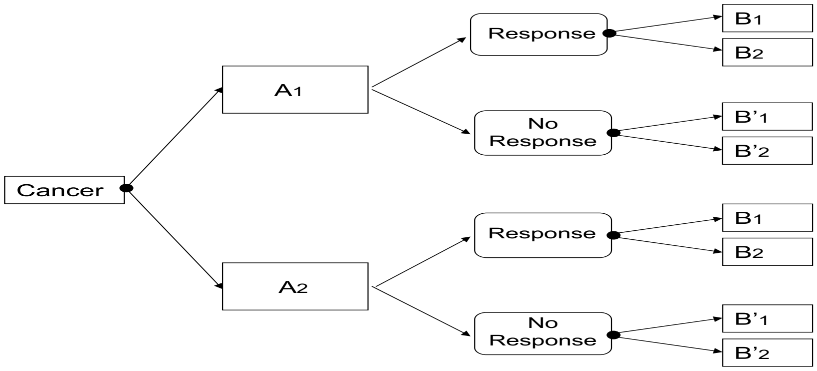


Figure 1.1: A typical two-stage randomization design: full circles, rectangles and arched rectangles represent respectively the time of randomization, available treatment arms and the intermediate outcome.

1.2 TWO-STAGE RANDOMIZATION DESIGNS

Randomized clinical trials comparing treatment strategies with randomization being done upfront to all possible strategies require large number of patients, even when the number of stages and the number of treatment choices at each stage are small. For instance, a clinical trial comparing treatment strategies with three stages and two possible treatment options at each stage requires randomization to $2^3 = 8$ possible regimes. By considering the natural course of treatment, one could randomize patients at the beginning of each stage once they become eligible. For example, to compare treatment strategies for a dynamic treatment regime with two stages and two treatment options at each stage, patients could be randomized to one of two possible therapies and depending on the intermediate response, could be randomized to further therapies at stage two. Such multistage randomization designs are referred to as sequential multiple assignment randomization trial or SMART [2]. A pictorial representation of a standard two-stage design is given in Figure 1.1. The treatment options B_j and B'_j , $j = 1, 2$ may be same or different depending on specific clinical trials. Unlike the situation described in Figure 1.1, where every patient receives some therapy at each stage, there may be cases where therapy may be stopped after the first stage if certain clinical conditions are not met. In the CALGB clinical trial described below, the non-responding patients did not receive further treatment in the second stage. For a two-stage design where therapy is stopped for patients not responding to the initial treatment, the branches involving B'_j , $j = 1, 2$ in Figure 1.1 will be missing. In such cases one could assume that the non-responding patients will receive a common treatment. Clinical trials employing two-stage randomization designs are commonly implemented in biomedical research. We describe two such clinical trials that motivated the methodologies in our research.

1.3 CALGB 8923 TRIAL

Cancer and Leukemia Group B (CALGB) conducted a two-stage clinical trial (Protocol 8923) to investigate the combination of different induction and maintenance therapies. As reported by Stone et al. [3], 388 AML (acute myelogenous leukemia) patients 60 years of age or older participated in this double-blind, placebo controlled trial. Following standard chemotherapy, in the first stage, 195 of these patients were randomly assigned to receive placebo and 193 receive granulocyte-macrophage colony-stimulating factor (GM-CSF). 79 in the GM-CSF group and 90 in the placebo group achieved complete remission and consented to further treatment. In the second stage, 37 GM-CSF and 45 placebo patients were randomly assigned to receive intensification therapy I, and the rest 42 GM-CSF patients and 45 placebo patients to intensification therapy II. The purpose of the trial was to examine the effects of infusions of GM-CSF after initial chemotherapy for elderly patients with AML.

1.4 E4494 CLINICAL TRIAL

The E4494 clinical trial conducted by the Eastern Cooperative Oncology Group (ECOG), CALGB and the Southwest Oncology Group (SWOG) and reported by Winter et al. [4] is another example of TSRD. This study was aimed to address the impact of the addition of rituximab to standard cyclophosphamide, doxorubicin, vincristine and prednisone (CHOP) therapy during induction with a second randomization to maintenance rituximab (MR) or observation on early and late treatment failures in diffuse large B-cell lymphoma (DLBCL) in elderly patients. Among the 632 previously untreated patients 60 years of age or older with DLBCL, 318 were randomized to the induction treatment with addition of rituximab(R) to CHOP, and 314 to standard CHOP. In the second stage, out of 415 responding patients, 207 were then randomized to MR and 208 to observation. After ineligibility exclusion, there were 267 R-CHOP and 279 CHOP patents in the induction stage, 174 MR and 178 observation patients in maintenance stage. The goal of the study was to compare the risk of treatment failure, time-to-treatment failure and overall survival among different treatment policies.

1.5 MOTIVATION AND PURPOSE

Traditional methods for analyzing data from two-stage trials separate the two stages, for example, first estimate and compare the survival distributions between two induction treatments for all patients in the study, ignoring the maintenance therapy, then for all responding patents, estimate and compare their survival distributions between two maintenance therapies conditioning on the response, regardless of the induction therapy they had received. The outcome of interest in the second stage is usually taken as the length of time from receiving the maintenance therapy to death or failure. Contrary to implementing intention-to-treat analysis which will be addressed in the following chapters, such methods discard information from the patients who could have potentially received the therapy and consequently reduces the effective sample size and makes the analysis inefficient. More importantly, such methods of analysis are limited to comparing different induction treatments or maintenance treatments, without being able to address the question of finding the best combination of induction and maintenance therapies.

For the cases where the outcome of the study is survival time, Lunceford et al. [5] proposed a class of consistent, asymptotically normal estimators for the survival distribution of treatment policies. Their framework allowed consistent estimation of survival distributions under intent-to-treat treatment policies. However, these estimators were not efficient and failed to use the auxiliary information collected in the form of covariates. Wahed and Tsiatis [6] obtained the most efficient semi-parametric regular asymptotically linear estimators for survival distribution and related quantities borrowing the idea of semi-parametric theory from Robins et al. [7]. The estimators proposed incorporated auxiliary time independent and time dependent covariates to gain efficiency. The cases of where the data may be right censored, were incorporated in Wahed and Tsiatis [8]. Considering the impractical nature of the most efficient estimator, they also proposed estimators that are easy to compute but are more efficient than Lunceford et al. [5] estimators. Lokhnygina and Helterbrand [9] employed Cox's proportional hazard model to derive a consistent estimator and score test for the log hazard ratio. Guo and Tsiatis [10] proposed a weighted risk set estimator (WRSE) for the survival distribution with right censoring using the concepts of counting process and risk sets described by Fleming and Harrington [11]. Recently, Guo [12] proposed a weighted log-rank test for testing the equality of two survival curves under two different strategies sharing the same maintenance therapy. However, as noted in Eng and Kosorok [13], this test has low power for detecting time-varying relative hazards.

There have been quite a few innovative procedures, some of which were mentioned above, to make inferences regarding adaptive treatment strategies based on the data collected from sequentially randomized designs. However, few techniques are available with respect to the design of such trials. For example, an important problem that has yet to be addressed is the power analysis and sample size determination to compare two or more strategies, or to detect a particular class of alternatives.

The notation and assumptions used throughout our research are introduced in chapter 2. The first part of this research provides an exhaustive and comparative review of analytical approaches available for the two-stage randomization designs with survival time as the primary outcome. Comparative conclusions are drawn based on the simulation studies. The results are presented in chapter 3. In the second part of the research, we present a sample size formula to compare the point-wise survival probabilities for different treatment strategies using Wald's test. The formula is based on the large sample properties of inverse-probability-weighted estimator. Simulation study provides strong evidence that the proposed sample size formula guarantees desired power, regardless of the true distributions of survival time. Results are presented in Chapter 4. In the final part of this thesis, for the purpose of testing the equality of survival distributions of two adaptive treatment strategies, a supremum weighed log-rank test is proposed, and a sample size formula is derived based on the limiting distribution of the supremum weighted log-rank test statistic, as elaborated in chapter 5. Simulation studies show that the proposed test provided sample sizes that are close to those obtained by standard weighted log-rank test under a proportional hazard alternative. Some remarks and potential future research are discussed in chapter 6.

2.0 MODEL FRAMEWORK AND NOTATION

Let us consider a two-stage clinical trial similar to the CALGB 8923 Study , where the induction treatment is A, with levels A_1 and A_2, and the maintenance treatment is B, with levels B_1 and B_2. The objective is to compare the survival distributions for different treatment policies $A_jB_k, j, k = 1, 2$, where A_jB_k stands for "treat with A_j followed by B_k if the patient is eligible and consents to subsequent maintenance therapy."

Let us assume that each patient i has an associated set of random variables, also referred to as potential outcomes, $\{R^*_{1i},\ R^*_{2i},\ (1 - R^*_{1i})T^*_{10i},\ (1 - R^*_{2i})T^*_{20i},\ R^*_{1i}T^{R*}_{1i},\ R^*_{2i}T^{R*}_{2i},\ R^*_{1i}T^*_{11i},\ R^*_{1i}T^*_{12i},\ R^*_{2i}T^*_{21i},\ R^*_{2i}T^*_{22i},\ V_i\}$, where R^*_{ji} is the eligible/consent status that patient i would achieve were s/he assigned to one of the two policies $A_jB_k,\ j, k = 1, 2,\ R^*_{ji} = 1$ if patient i was eligible and would consent to subsequent maintenance treatment, $R_{ji} = 0$ otherwise; T^*_{j0i} is the survival time of patient i if s/he received induction treatment A_j, and was not eligible or refused subsequent maintenance treatment, defined only when $R^*_{ji} = 0$; T^{R*}_{ji} is the time from initial randomization to A_j to the time s/he received maintenance therapy, defined only when $R^*_{ji} = 1$; T^*_{jki} is the survival time of patient i if s/he received induction treatment A_j, was eligible and consented to receive maintenance treatment and received B_k; V_i is a vector of auxiliary covariates including relevant baseline characteristics for patient i. From the definition above, we can see that T^{R*}_{ji} is defined only for those eligible and consent patients, and all of the ten variables $R^*_{1i}, R^*_{2i},\ T^*_{10i},\ T^*_{20i},\ T^{R*}_{1i},\ T^{R*}_{2i},\ T^*_{11i},\ T^*_{12i},\ T^*_{21i}$ and T^*_{22i} can not be observed for the same patient since a patient can receive only one of the two induction treatments, can not be both responder and non-responder, and can only receive one of the two maintenance treatments if s/he responds in the 1st stage. These variables, for such reason, are referred to as counterfactuals [14, 15] or potential random variables.

With the above notations, the survival time for patient i who received treatment policy A_jB_k would be $T_{jki} = (1 - R^*_{ji})T^*_{j0i} + R^*_{ji}T^*_{jki}$. Due to the fact that some patients eligible for maintenance therapy B_k may not consent to further treatment or may be randomized to the maintenance therapy $B_{3-k}, k = 1, 2$, the inference on features of these distributions addresses directly the "intent-to-treat" question of interest. With the above conceptualization, the primary goal is to estimate parameters and draw inference on the distribution of $T_{jk}, j, k = 1, 2$. Specifically, we consider the problem of estimating $S_{jk}(t) = \Pr(T_{jk} > t) = E\{I(T_{jk} > t)\}$, the survival probability beyond time t for treatment policy A_jB_k. In other cases, possible parameters of interest can be the mean or median restricted survival time.

If there is no censoring, the observed data can be represented as a set of i.i.d random vectors $\{Z_i, R_i, R_iT^R_i, (1 - R_i)T_{0i}, V_i, R_iX_i, T_i\}, i = 1, \cdots, n$, where Z_i denotes the A treatment randomization, i.e, $Z_i = 2 - j$ if the i^{th} patient is assigned to treatment A_j, $j = 1, 2$; $R_i = Z_iR^*_{1i} + (1 - Z_i)R^*_{2i}$ is the observed eligible/consent status for patient i; $T^R_i = Z_iT^{R*}_{1i} + (1 - Z_i)T^{R*}_{2i}$; $T_{0i} = Z_iT^*_{10i} + (1 - Z_i)T^R_{20i}$,; V_i is a vector of auxiliary covariates as defined before; X_i denotes the B treatment assignment indicator, defined only if $R_i = 1$, where $X_i = 2 - k$ if assigned to treatment B_k,$k = 1, 2$; and T_i is the observed survival time for patient i. Following stable unit treatment value assumption [16], we assume that the observed survival time for patient i is related to the potential outcomes through the relation

$$\begin{aligned} T_i &= Z_i\left\{(1 - R^*_{1i})T^*_{10i} + R^*_{1i}(X_iT^*_{11i} + (1 - X_i)T^*_{12i})\right\} \\ &\quad +(1 - Z_i)\left\{(1 - R^*_{2i})T^*_{20i} + R^*_{2i}(X_iT^*_{21i} + (1 - X_i)T^*_{22i})\right\}, \qquad (2.1) \end{aligned}$$

that is, for a patient who receives induction treatment A_j, if s/he is observed to be a non-responder, then his/her observed survival time T_i is equal to the corresponding potential survival time T_{j0i}; on the other hand if the patient is observed to be a responder and received treatment $B_1(B_2)$, his/her observed survival time T_i is equal to the corresponding counterfactual survival time $T^*_{j1i}(T^*_{j2i})$, $j = 1, 2$. In the presence of right censoring, the observed data can be summarized as the collection of i.i.d random vectors $\{U_i, \Delta_i, G^H_i(U_i)\}, i = 1, \cdots, n$, where $U_i = \min(T_i, C_i), \Delta_i = I(T_i \leq C_i), C_i$ is the censoring time and $G^H_i(U_i) = \{Z_i, R_iI(T^R_i \leq x), X_iR_iI(T^R_i \leq x), V_i(x), x \leq u\}$, where $V_i(x)$, similar to the V_i defined before, is a vector of auxiliary variables that may additionally be collected on patient i at time x. Thus $G_i^H(U_i)$ represents data-history collected on individual i prior to time u, which contains the information of the eligibility/consent status, the time of response if responded, the assignment of maintenance treatment and other auxiliary variables of interest of patient i.

Since in most clinical trials total follow-up time is limited, only restricted survival time up to time L can be considered, where L is some value less than the maximum follow-up time for all patients in the sample, in such cases, T_{jk} will actually represent $\min(T_{jk}, L)$.

The first goal of our research is to estimate and compare $S_{jk}(t)$ for the policy A_jB_k, $j, k = 1, 2$. Then, in the second part of the thesis, we develop a sample size formula for testing the hypothesis $H_0 : \ F_{11}(t) = F_{12}(t)$ vs. $H_1 : \ F_{11}(t) \neq F_{12}(t)$ where $F_{1k}(t) = \Pr(T_{1k} \leq t) = E\{I(T_{1k} \leq t)\}$, denotes the probability of failure before or at time t for treatment strategy $A_1B_k, k = 1, 2$. Furthermore, a supremum weighted log-rank test and corresponding sample size formula are derived in order to compare the distributions of T_{11} and T_{21}.

3.0 A REVIEW OF INFERENTIAL PROCEDURES

3.1 INTRODUCTION

In this chapter, we review and compare the currently available inferential procedures for the two-stage randomization designs with survival time as the primary outcome. Since the data from patients receiving induction treatment A_1 are independent of those from patients with induction treatment A_2, we focus only on the data from patients who received A_1, that is, patients with treatment policies A_1B_1 and A_1B_2. The methods for policies A_2B_1 and A_2B_2 follow analogously. Since we only consider the two treatment policies that are associated with the induction treatment A_1, we drop the subscript 1 in this chapter and chapter 4. For instance, in these two chapters, T_{ki} is short for T_{1ki}, $k = 1, 2$.

3.2 AVAILABLE INFERENTIAL PROCEDURES

3.2.1 NAÏVE ESTIMATOR

To estimate $S_k(t)$ for the policy A_1B_k, a naïve approach would be to construct an estimator only using the data from those patients who are treated consistently with that policy. If there was no censoring, this would mean that one could average the indicator function $I(T_i > t)$ over all the patients in the set: $\{i : 1 - R_i + R_iX_{ki} = 1\}$, where $X_{1i} = X_i$ and $X_{2i} = 1 - X_i$, to get

$$\hat{S}_k^{\text{NAÏVE}}(t) = \left\{\sum_{i=1}^{n}(1 - R_i + R_iX_{ki})\right\}^{-1} \times \sum_{i=1}^{n}(1 - R_i + R_iX_{ki})I(T_i > t). \tag{3.1}$$

This naïve estimator takes into account the patients who did not respond and those who were assigned to maintenance treatment B_k. However, it neglects those patients who responded and were randomized to treatment B_{3-k}, $k = 1, 2$, as a result, the naïve estimator is expected to underestimate $S_k(t)$ by overestimating the contribution of the non-responders to the survival distribution. Besides, the group of patients that has been used is no longer a random sample from those who could potentially follow the policy A_1B_k. In the cases where their data is censored, Kaplan-Meier estimator for the survival distribution from the group $\{i : 1 - R_i + R_iX_{ki} = 1\}$ could be used to calculate the naïve estimator.

3.2.2 CONSISTENT AND ASYMPTOTICALLY NORMAL ESTIMATORS

In order to make more efficient use of the information from patients who are inconsistent with the policy A_1B_k, Lunceford et al. [5] proposed three forms of consistent and asymptotically normal estimators. Assume that the assignment of B treatment is conditionally independent of the potential survival time given the induction treatment and the data collected prior to observing the response. Let the probability of randomization to the B_1-treatment be denoted by $\pi_1 = P(X_{1i} = 1|R_i = 1)$. Then this assumption could be interpreted as $X_{1i} \perp T^*_{1i}, T^*_{2i}|G^H_i(T^R_i), R_i = 1$. This, in turn, implies that $\Pr(X_{1i} = 1|T^*_{1i}, T^*_{2i}, G^H_i(T^R_i), R_i = 1) = \Pr(X_{1i} = 1|R_i = 1) = \pi_1$. This assumption is the "sequential randomization assumption" or the assumption of "no unmeasured confounders" as discussed in Robins (1997). The probability π_k can be allowed to depend on the data-history prior to the randomization including the induction treatment, but for simplicity, we avoid discussing it here. To be consistent with the examples in chapter 1, we take π_1 to be known by design. Let us define $\pi_2 = 1 - \pi_1 = \Pr(X_{1i} = 0|R_i = 1) = \Pr(X_{2i} = 1|R_i = 1)$, where $X_{2i} = 1 - X_{1i}$. Let $K(u) = \Pr(C_i > u)$ denote the survival distribution for the censoring time C_i. Assume also that the censoring time is independent of the observed data and counterfactuals.

The first estimator in the sequel of three is defined as the weighted average of the patients who are consistent with the treatment policy. Since by definition, non-responders are consistent to the policy A_1B_k, they were given unit weight in the construction. Responders who were assigned to B_k with randomization probability π_k are also consistent with the policy.

But, due to the fact that some of the responders were randomized to the other B treatment, each patient receiving B_k represents $\frac{1}{\pi_k} - 1$ other similar patients who could have potentially be assigned to B_1 treatment, and thus received the weight $\frac{1}{\pi_k}$. Combining both, the weight function takes the form $Q_{ki} = 1 - R_i + \frac{R_i X_{ki}}{\pi_k}, \quad k = 1, 2$. Additionally, since patients may be censored at any time, a second form of weighting was applied to account for the censored patients. Each uncensored patient with survival time U_i represents $\frac{1}{K(U_i)} - 1$ prognostically similar patients who survived beyond time U_i and thus receives a weight of $\frac{1}{K(U_i)}$. Thus the combined weight for a patient with complete survival time U_i becomes $\frac{\Delta_i Q_{ki}}{K(U_i)}$. Since $K(u)$ is unknown, it is usually estimated by the Kaplan-Meier estimator of the censoring survival curve $\hat{K}(U) = \prod_{u \le t}\{1 - \mathrm{d}N^c(u)/Y(u)\}$, with $N^c(u) = \sum_{i=1}^{n} I(U_i \le u, \Delta_i = 0)$ and $Y(u) = \sum_{i=1}^{n} I(U_i \ge u)$, resulting in an estimated weight function $\Delta_i Q_{ki}/\hat{K}(U_i)$. The estimator for the survival function $S_k(t)$ is then defined as:

$$\hat{S}_k^{IPMW}(t) = 1 - n^{-1} \sum_{i=1}^{n} \frac{\Delta_i Q_{ki}}{\hat{K}(U_i)} I(U_i \le t), \qquad k = 1, 2. \tag{3.2}$$

It was shown that if the true $K(\cdot)$ is substituted in the above equation, then $\hat{S}_k^{IPMW}(t)$ is unbiased for $S_k(t)$. $\hat{S}_k^{IPMW}(t)$ in equation (2) is an example of an inverse-probability-of-missing-weighted (IPMW) estimator (Horvitz-Thompson estimator, Horvitz (1952)). The second estimator was obtained by averaging using a probabilistically adjusted sample size, i.e.,

$$\hat{S}_k^{PA}(t) = 1 - \left\{ \sum_{i=1}^{n} \frac{\Delta_i Q_{ki}}{\hat{K}(U_i)} \right\}^{-1} \sum_{i=1}^{n} \frac{\Delta_i Q_{ki}}{\hat{K}(U_i)} I(U_i \le t), \qquad k = 1, 2 \tag{3.3}$$

Lunceford et al. [5] observed that both $\hat{S}_k^{IPMW}(t)$ and $\hat{S}_k^{PA}(t)$ are solutions of the equations of the form $\sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} \{Q_{ki} I(U_i \le t) + S_k(t) - 1 - \alpha_k (Q_{ki} - 1\} = 0$ with α_k set to 0 and $1 - S_k(t)$, respectively. Thus the third estimator was constructed by choosing the α_k that minimizes the variance among all solutions. To be specific, the third estimator has the form:

$$\hat{S}_k^{LDT}(t) = 1 - n^{-1} \sum_{i=1}^{n} \frac{\Delta_i Q_{ki}}{\hat{K}(U_i)} I(U_i \le t) + \hat{\alpha}_k n^{-1} \sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} (Q_{ki} - 1), \qquad k = 1, 2 \tag{3.4}$$

where

$$\hat{\alpha}_k = \left[n^{-1}\sum_{i=1}^{n}\Delta_i Q_{ki}(Q_{ki}-1)\frac{I(U_i \geq u)}{\hat{K}(V_i)} + \int_0^L \mathrm{d}N^c(u)\{\hat{K}(u)Y(u)\}^{-1}\hat{E}\{L_{1k}^{\alpha}(t,u)\}\right] \div \left[n^{-1}\sum_{i=1}^{n}(Q_{ki}-1)^2 + \int_0^L \mathrm{d}N^c(u)\{\hat{K}(u)Y(u)\}^{-1}\hat{E}\{G_k^{\alpha}(u)\}\right],$$

with

$$\hat{E}\{L_k^{\alpha}(t,u)\} = n^{-1}\sum_{i=1}^{n}\Delta_i\{Q_{ki}I(U_i \leq t) - \hat{G}_{1k}(t,u)\} \times \{Q_{ki} - 1 - \hat{G}_{Q_k}(u)\}\frac{I(U_i \geq u)}{\hat{K}(U_i)},$$

$$\hat{E}\{G_k^{\alpha}(u)\} = n^{-1}\sum_{i=1}^{n}\Delta_i\{Q_{ki} - 1 - \hat{G}_{Q_k}(u)\}^2\frac{I(U_i \geq u)}{\hat{K}(U_i)},$$

$$\hat{G}_{Q_k}(u) = \{n\hat{S}_k(u)\}^{-1}\sum_{i=1}^{n}\Delta_i(Q_{ki}-1)\frac{I(U_i \geq u)}{\hat{K}(U_i)},$$

$$\hat{G}_{1k}(u) = \{n\hat{S}_k(u)\}^{-1}\sum_{i=1}^{n}\Delta_i Q_{ki}I(U_i \leq t)\frac{I(U_i \geq u)}{\hat{K}(U_i)}.$$

The three estimators $\hat{S}_k^{IPMW}(t)$, $\hat{S}_k^{PA}(t)$ and $\hat{S}_k^{LDT}(t)$ are consistent and asymptotically normal. For details on the asymptotic property of these estimators we refer our readers to Lunceford et al. [5]. These estimators were defined on an ad hoc basis and the formal efficiency issue was not discussed.

3.2.3 SEMI-PARAMETRIC EFFICIENT ESTIMATOR

Wahed and Tsiatis [6] used the semi-parametric theory of missing data described in Robins, Rotnitzky and Zhao [7] to characterize the most efficient regular asymptotically linear (RAL) [17] estimator. They observed that any RAL estimator can be characterized by its influence function and their approach was to find the most efficient influence function for all RAL estimators of $S_k(t)$. However, the most efficient influence function for this problem contains a nuisance parameter in the form of the conditional expectation $\Pr(T_{ki} > t|T_i^R, V_i, R_i = 1, X_{ki} = 1)$. One way to construct useful estimators from the most efficient influence function is to approximate these conditional probability based on patient data history leading to locally efficient estimators. A natural way of estimating $\Pr(T_{ki} > t|T_i^R, V_i, R_i = 1, X_{ki} = 1)$

is to use a logistic regression of the binary outcome $I(T_i > t)$ on the covariates V_i and T_i^R within the subgroup of patients with $R = 1$ and $x_k = 1$. For instance, a logistic regression model

$$\Pr(T_{ki} > t | T_i^R, V_i, R_i = 1, X_{ki} = 1) = \frac{1}{1 + e^{-(\gamma_0 + \gamma_1 T_i^R + \gamma_2^T V_i)}} = g(T_i^R, V_i; \gamma)$$

will give rise to the locally efficient estimator:

$$\hat{S}_k^{LE} = \frac{1}{n} \sum_{i=1}^{n} [\{(1 - R_i) + \frac{R_i X_{ki}}{\pi_k}\} I(U_i > t) - R_i(\frac{X_{ki} - \pi_k}{\pi_k}) g(T_i^R, V_i; \hat{\gamma})] \tag{3.5}$$

for $k = 1, 2$. This estimator remains consistent even if the function form $g(\cdot)$ is not correctly specified, but if the regression relationship was incorrectly specified, then the gain of efficiency over the IPMW or LDT estimator could not be guaranteed. In the presence of right censoring, an inverse probability weighted version of the locally efficient estimator (3.5) is given by

$$\hat{S}_k^{IPCWLE} = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} \left[\left\{ (1 - R_i) + \frac{R_i X_{ki}}{\pi_k} \right\} I(U_i > t) - R_i(\frac{X_{ki} - \pi_k}{\pi_k}) g(T_i^R, V_i; \hat{\gamma}) \right] \tag{3.6}$$

for $k = 1, 2$. We will refer to it as the Inverse Probability of Censoring Weighted Local Efficient (IPCWLE) estimator. The properties of this estimator have not been investigated in previous studies. This estimator is asymptotically unbiased. In addition, in our simulation studies presented later, we find that the relative efficiency of this estimator over IPMW, PA or LDT estimator is close to unity. But this estimator also depends on the specification of the model g and therefore, is subjected to model mis-specification.

Wahed and Tsiatis [8] then extended the semi-parametric method to obtain the most efficient estimator in the presence of right censoring. In order to avoid cumbersome calculation in the construction of most efficient estimator, they restricted the search for the optimal estimator to a sub-class of the RAL estimators that contains the existing estimators. Letting $U_i^* = \min(C_i, T_i^R)$, $\Delta_i^* = I(C_i < T_i^R)$, $Y_i(u) = I(U_i \geq u)$, $\hat{E}_1(u) = \sum_{i=1}^n R_i I(U_i^* < u) X_{ki} Y_i(u)/Y(u)$, $\hat{E}_2(u) = Y^{-1}(u) \sum_{i=1}^n \{1 - R_i I(U_i^* < u)\} Y_i(u)$ and $L_{1i} = \{R_i I(U_i^* < u) X_{ki} - \hat{E}_1(u)\}/\pi_k$, $L_{2i} = 1 - R_i I(U_i^* < u) - \hat{E}_2(u)$, a simplified version of the regular asymptotic linear efficient (RALE) estimator is given by :

$$\hat{S}_k^{RALE}(t) = A_n / B_n \tag{3.7}$$

where

$$\begin{aligned} A_n &= \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i Q_{ki}}{\hat{K}(U_i)}I(U_i > t) - \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i^*(Q_{ki}-1)}{\hat{K}(U_i^*)}\hat{\gamma}^T W_i \\ &\quad + \sum_{j=1}^{2}\frac{1}{n}\sum_{i=1}^{n}\int\frac{\mathrm{d}N_i^c(u)}{\hat{K}(u)}\hat{\varphi}_j(u)L_{ji}(u), \end{aligned}$$

and

$$\begin{aligned} B_n &= \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i Q_{ki}}{\hat{K}(U_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i^*(Q_{ki}-1)}{\hat{K}(U_i^*)}\hat{\gamma}_\mu^T W_i \\ &\quad + \sum_{j=1}^{2}\frac{1}{n}\sum_{i=1}^{n}\int\frac{\mathrm{d}N_i^c(u)}{\hat{K}(u)}\hat{\varphi}_{j\mu}(u)L_{ji}(u), \end{aligned}$$

where $\hat{\boldsymbol{\gamma}} = \alpha^{-1}\beta$, $\hat{\boldsymbol{\gamma}}_\mu = -\alpha^{-1}\beta_\mu$, $\hat{\varphi}_1(u) = \zeta^{-1}(u)\eta(u)$, $\hat{\varphi}_{1\mu}(u) = 1$, $\hat{\varphi}_2(u) = \kappa^{-1}(u)\tau(u)$, $\hat{\varphi}_{2\mu}(u) = \kappa^{-1}(u) - \tau_\mu(u)$ where,

$$\alpha = n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{K}(U_i)}\left\{K^{-1}(U_i^*)(Q_{ki}-1)^2\boldsymbol{W}_i\boldsymbol{W}_i^T\right\},$$

$$\beta = n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{K}(U_i)}\left\{K^{-1}(U_i^*)Q_{ki}(Q_{ki}-1)I(U_i > t)\boldsymbol{W}_i\right\},$$

$$\beta_\mu = n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{K}(U_i)}\left\{K^{-1}(U_i^*)Q_{ki}(Q_{ki}-1)\boldsymbol{W}_i\right\},$$

$$\tau(u) = n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{K}(U_i)}\left[\left\{(1-R_i)I(U_i^* \geq u) + \frac{R_i X_{ki}}{\pi}\right\}I(U_i > t)\right],$$

$$\tau_\mu(u) = n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{K}(U_i)}\left\{(1-R_i)I(U_i^* \geq u) + \frac{R_i X_{ki}}{\pi}\right\},$$

$$\kappa(u) = n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{K}(U_i)}\left[I(U_i \geq u)\left\{1 - R_i I(T_i^R < u)\right\}\right],$$

$$\eta(u) = n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{K}(U_i)}\left\{I(U_i^* < u \leq U_i)R_i X_{ki} I(U_i > t)\right\},$$

$$\zeta(u) = n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{K}(U_i)}\left\{I(U_i^* < u \leq U_i)R_i X_{ki}\right\}.$$

The estimator $\hat{S}_k^{RALE}$ is consistent and asymptotically normaland is guaranteed to be asymptotically more efficient than the IPMW and LDT estimators since it is the most efficient estimator among a class of estimators including the IPMW and LDT estimators. For details on the proof of asymptotic properties, variance estimates, and the estimates of covariance between $\hat{S}_1^{RALE}(t)$ and $\hat{S}_2^{RALE}(t)$, we refer the readers to Wahed and Tsiatis [8].

3.2.4 WEIGHTED RISK SET ESTIMATOR

Guo and Tsiatis [10] derived the Weighted Risk Set Estimator (WRSE) using the concepts of counting process and risk sets, which is an extension of the Aalen-Nelson estimator. This estimator is more intuitive and easier to compute than the above ones. The intention was to use Aalen-Nelson estimator to estimate the cumulative hazard function, however, due to the property of two stage design, not all counting processes $N_i(u) = I(U_i \leq u, \Delta_i = 1)$ and at risk process $Y_i(u) = I(U_i \geq u)$ could be observed, because some of the patients who could have received treatment B_1 are instead randomized to receive B_2. Consequently, a time-varying weight function was defined for treatment strategy A_1B_1: $W_i(u) = 1 - R_i(u) + R_i(u)X_i/\pi_1$, where $R_i(u) = R_i I(T_i^R \leq u)$ is the indicator of response at time u for patient i. With this weight function, the extended Aalen-Nelson estimator for the cumulative hazard under policy A_1B_1 is defined as

$$\hat{\Delta}_1(t) = \int_0^t \frac{\sum_{i=1}^n W_i(u)\mathrm{d}N_i(u)}{\sum_{i=1}^n W_i(u)Y_i(u)} \tag{3.8}$$

and the corresponding estimator for the survival function follows as

$$\hat{S}_1^{WRSE}(t) = \exp\left\{-\int_0^t \frac{\sum_{i=1}^n W_i(u)\mathrm{d}N_i(u)}{\sum_{i=1}^n W_i(u)Y_i(u)}\right\} \tag{3.9}$$

It has been shown that WRSE is consistent and asymptotically normal. Detailed proof of the consistency and asymptotically normality of the WRSE is given by Guo and Tsiatis [10].

3.2.5 COX PROPORTIONAL HAZARD MODEL

Because of the wide use of Cox regression model in the analysis of survival data, Lokhnygina and Helterbrand [9] derived a consistent estimator for the log hazard ratio comparing strategies A_1B_1 and A_2B_1 in the Cox model. In addition to the sequential randomization assumption and the assumption of independent censoring, this construction like other applications using Cox model, requires the proportional hazard assumption between two treatment policies. As in a usual Cox proportional hazard model, consider the hazard corresponding to policy $A_{2-j}B_1$ be $\lambda(t|Z=j)$, $j=0,1$ where $\lambda(t|Z)=\lambda_0(t)\exp(Z\beta)$, The estimate of β can be obtained by solving the pseudo-score equation

$$U_{wn}(\beta)=\sum_{i=1}^{n}\int_0^\infty w_i\{Z_i-\bar{Z}_w(u,\beta)\}\mathrm{d}N_i(u)=0 \tag{3.10}$$

where $w_i=1-R_i+R_i(u)X_i/\pi_1$ acts as an inverse probability weight, and

$$\bar{Z}_w(u,\beta)=\frac{\sum_{i=1}^{n}w_iZ_iY_i(u)\exp(Z_i\beta)}{\sum_{i=1}^{n}w_iY_i(u)\exp(Z_i\beta)}.$$

Lokhnygina and Helterbrand [9] showed that the estimator of β is consistent and asymptotically normal. This estimator is easier to implement with available software and intuitively appealing.

3.3 SIMULATION STUDIES

To evaluate the performance of the methods reviewed in the previous section, several simulations were carried out following Lunceford et al. [5] strategy. We only simulated data for policy AB_1 and AB_2 since the data from A_1 and A_2 are independent. All simulations were based on a 2.5-year study for n=200 and 500 subjects. For each individual, censoring time C was generated as uniform(0,2.5) independent of all other variables. Remission/consent status R were sampled from Bernoulli(π_R). Two values of the response rate $\pi_R=0.4$ and $\pi_R=0.6$ were used in this simulation. The B treatment indicators were generated from Bernoulli(0.5) distribution. For non-responders ($R=0$), a survival time T^*_λ was generated from exponential(λ), where λ was taken to be 2.22 so that $E(T^*_\lambda)/L=0.3$, where $L=1.5$ was the upper limit of the restricted observed lifetime. For responders, a remission/consent time T^R was drawn from exponential(α). We take $T_1^{**}\sim EXP(e^{\beta_1})$, $T_2^{**}\sim EXP(e^{\beta_1+\beta_2T_1^{**}})$, where T_1^{**} and T_2^{**} are post-remission survival time under B_1 and B_2, respectively. The parameters α, β_1 and β_2 were chosen to be 6.67, 0.29 and -0.67, respectively, so that $E(T^R)/L=0.1$, $E(T_1^{**})/L=0.5$, and $E(T_2^{**})/L=1.0$. The potential restricted survival times were calculated as $T_1=\min\{(1-R)T^*_\lambda+R(T^R+T_1^{**}),L\}$ and $T_2=\min\{(1-R)T^*_\lambda+R(T^R+T_2^{**}),L\}$.

Table 3.1: Monte Carlo means, relative biases (bias as a percentage of the true value) and mean squared errors (MSE, expressed as multiples of 10^3) for estimation of survival probabilities based on 1000 data sets of sizes 200 each. The true values were $S_1(0.5) = 0.450$, $S_2(0.5) = 0.492$, $S_1(1.0) = 0.196$, $S_2(1.0) = 0.261$ for 40% response and $S_1(0.5) = 0.511$, $S_2(0.5) = 0.575$, $S_1(1.0) = 0.240$, $S_2(1.0) = 0.339$ for 60% response.

		$\pi_R = 0.4$						$\pi_R = 0.6$					
		Policy AB_1			Policy AB_2			Policy AB_1			Policy AB_2		
t(years)	Estimator	$\hat{S}(t)$	Bias(%)	MSE	$\hat{S}$(t)	Bias(%)	MSE	$\hat{S}$(t)	Bias(%)	MSE	$\hat{S}$(t)	Bias(%)	MSE
0.5	IPMW	0.452	0.4	4.28	0.493	0.2	4.42	0.511	0.0	5.48	0.578	0.5	5.93
	PA	0.450	0.0	2.44	0.492	0.0	2.38	0.511	0.0	2.73	0.575	0.0	2.56
	LDT	0.447	0.7(-)	2.11	0.489	0.6(-)	2.00	0.508	0.6(-)	2.41	0.571	0.7(-)	2.23
	IPCWLE	0.450	0.0	2.18	0.492	0.0	2.03	0.510	0.2(-)	2.51	0.574	0.2(-)	2.27
	WRSE	0.453	0.7	1.91	0.495	0.6	1.93	0.514	0.6	2.20	0.578	0.5	2.15
	RALE	0.446	0.9(-)	2.07	0.489	0.6(-)	1.98	0.508	0.6(-)	2.35	0.572	0.5(-)	2.18
1.0	IPMW	0.197	0.5	2.84	0.263	0.8	3.58	0.239	0.4	3.84	0.341	0.6	4.81
	PA	0.196	0.0	2.29	0.262	0.4	2.65	0.238	0.8(-)	2.93	0.338	0.2(-)	3.17
	LDT	0.193	1.5	2.03	0.259	0.8(-)	2.20	0.237	1.3(-)	2.62	0.335	1.2(-)	2.75
	IPCWLE	0.194	1.0(-)	2.13	0.261	0.0	2.36	0.238	0.8(-)	2.72	0.337	0.6(-)	2.92
	WRSE	0.200	2.0	1.71	0.267	1.5	2.00	0.243	1.3	2.25	0.343	1.2	2.53
	RALE	0.192	2.0(-)	1.87	0.259	0.8(-)	2.09	0.236	1.7(-)	2.41	0.336	0.9(-)	2.60

Table 3.2: Monte Carlo means, relative biases (bias as a percentage of the true value) and mean squared errors (MSE, expressed as multiples of 10^3) for estimation of survival probabilities based on 1000 data sets of sizes 500 each. The true values were $S_1(0.5) = 0.450$, $S_2(0.5) = 0.492$, $S_1(1.0) = 0.196$, $S_2(1.0) = 0.261$ for 40% response and $S_1(0.5) = 0.511$, $S_2(0.5) = 0.575$, $S_1(1.0) = 0.240$, $S_2(1.0) = 0.339$ for 60% response.

		$\pi_R = 0.4$						$\pi_R = 0.6$					
		Policy AB_1			Policy AB_2			Policy AB_1			Policy AB_2		
t(years)	Estimator	$\hat{S}(t)$	Bias(%)	MSE	$\hat{S}$(t)	Bias(%)	MSE	$\hat{S}$(t)	Bias(%)	MSE	$\hat{S}$(t)	Bias(%)	MSE
0.5	IPMW	0.451	0.2	1.54	0.494	0.4	1.71	0.512	0.2	2.10	0.576	0.2	2.23
	PA	0.451	0.2	0.95	0.493	0.2	0.94	0.512	0.2	1.05	0.575	0.0	0.97
	LDT	0.450	0.0	0.85	0.492	0.0	0.79	0.511	0.0	0.92	0.574	0.2(-)	0.85
	IPCWLE	0.450	0.0	0.85	0.493	0.2	0.80	0.511	0.0	0.97	0.576	0.2	0.85
	WRSE	0.452	0.4	0.77	0.494	0.4	0.79	0.513	0.4	0.85	0.576	0.2	0.81
	RALE	0.450	0.0	0.78	0.492	0.0	0.78	0.511	0.0	0.87	0.574	0.2(-)	0.82
1.0	IPMW	0.197	0.0	1.07	0.263	0.8	1.36	0.241	0.4	1.50	0.341	0.6	1.89
	PA	0.197	0.0	0.88	0.263	0.8	1.03	0.241	0.4	1.14	0.340	0.3	1.27
	LDT	0.196	0.0	0.80	0.262	0.4	0.87	0.240	0.0	1.02	0.339	0.0	1.11
	IPCWLE	0.197	0.5	0.87	0.263	0.8	0.89	0.241	0.4	1.09	0.340	0.3	1.14
	WRSE	0.199	1.5	0.68	0.264	1.2	0.82	0.243	1.3	0.88	0.342	0.9	1.01
	RALE	0.197	0.5	0.69	0.262	0.8	0.81	0.241	0.4	0.89	0.339	0.0	1.01

For each of 1000 Monte Carlo data sets, $P(T_k) > t, \quad k = 1, 2$ were estimated at time point 0.5 year and 1.0 year, reflecting early and late period of study. The mean squared errors were calculated from the bias of the estimated mean probability and the variance of the 1000 estimates. In calculating the IPCWLE and RALE estimators, the response time T_i^R was considered as the only auxiliary variable which the survival time could depend upon. For IPCWLE, to model the conditional expectation of survival probability among the responders who are consistent with the policy, logistic regression of survival probability on the response time was fitted. We did not include the Lokhnygina and Helterbrand's [9] Cox regression method in our simulation because its distinct property makes comparison less feasible.

Table 3.1 presents the mean, relative bias and mean squared errors for survival probability estimates based on 1000 samples of size 200 each. As shown in Table 1, almost all the relative biases, calculated as (bias/true value)×100, were less than 2%. By closely examining the table we notice that the relative biases were larger for $t = 1.0$ than $t = 0.5$, that is, the estimators were more biased for survival estimates at times towards the end of the study when there were more censoring present. In comparing the biases of different estimators in small samples, the PA estimator was generally the least biased, followed by the IPCWLE and IPMW estimators. LDT and RALE estimators always underestimated the true values whereas WRSE estimator overestimated them.

Comparing the MSE's, IPMW estimates were the least efficient as one would expect since no information from the censored patients or any auxiliary information is used in construction of such estimator. Among the IPMW, PA, LDT and RALE estimates whose influence functions belong to the same class, LDT estimates showed substantial gains in efficiency relative to both the first two, and RALE estimates are more efficient than LDT estimates in all scenarios, with the relative efficiency ranging from 1.01 to 1.18. The MSE of IPCWLE estimates were slightly larger than that of LDT estimates but substantially smaller than that of IPMW or PA estimates. In most instances WRSE estimator appeared to be the most efficient among all the estimates. The relative efficiencies of WRSE estimates with respect to LDT estimates ranged from 1.00 to 1.19 and the gain is bigger when more censoring is present. In general, the MSEs followed the pattern: IPMW $\geq$ PA $\geq$ IPCWLE $\geq$ LDT $\geq$ RALE $\geq$ WRSE.

Table 3.2 presents the mean, relative bias and mean squared errors for survival probability estimates based on 1000 samples of size 500 each. When the sample size was increased to 500, all the biases dropped to less than 1% except for the WRSE estimator. It was not surprising since the asymptotic unbiasness of WRSE estimator is achieved via the exponential functional of the cumulative hazard function. When the sample size was increased from 200 to 500, the efficiency of all the estimators improved, but the trend of relative efficiencies remained mostly unaffected.

4.0 INVERSE-PROBABILITY-WEIGHTING BASED SAMPLE SIZE FORMULA

4.1 BACKGROUND AND OBJECTIVE

Suppose the goal is to test the hypothesis that the probability of survival under a given strategy at a fixed time t does not differ from that under a different strategy sharing the same initial treatment. The goal can be accomplished by comparing the consistent estimators of survival probabilities under both strategies. Let us consider the IPMW estimator defined in 3.2.2 for this purpose. This inverse-probability weighting estimators are constructed based on the patients who actually received the combined treatments and those who did not respond to the initial treatment or those who refused further treatment. Thus when comparing the survival under two treatment strategies sharing the same initial treatment, the pair of estimators are not independent, since they are influenced by the same set of non-responders and non-consenters. In other words, the two groups of patients that are used to estimate survival are no longer random samples from those who could potentially follow respective treatment strategies. Consequently, one can not use the usual two-sample sample size formula for comparing the survival rates between two independent groups. In this section, the primary goal is to determine the required sample size for testing the hypothesis $H_0: \ F_1(t) = F_2(t)$ vs. $H_1: \ F_1(t) \neq F_2(t)$ where $F_k(t) = \Pr(T_k \leq t) = E\{I(T_k \leq t)\}$, denotes the probability of failure before or at time t for treatment strategy $A_1B_k, k = 1, 2$. We derive a sample size formula based on Lunceford et al.'s [5] estimator which is appropriate for censored data and sequential randomization. In other words, $\hat{F}_k(t) = 1 - \hat{S}_k(t)$, where $\hat{S}_k(t)$ was shown in equation (3.3).

Following the appendix in Lunceford et al. [5] , we have the large-sample property

$$n^{1/2}\{\hat{F}_k(t) - F_k(t)\} = n^{-1/2}\sum_{i=1}^{n}\psi_{ki} + o_p(1), \tag{4.1}$$

where ψ_{ki} is the influence function for $\hat{F}_k(t)$, given by

$$\psi_{ki} = Q_{ki}\{I(T_i \leq t) - F_k(t)\} - \int_0^L \frac{Q_{ki}\{I(T_i \leq t) - F_k(t)\} - G_{1k}(t,u)}{K(u)} dM_i^c(u), \tag{4.2}$$

where

$$G_k(t,u) = \frac{E[\{I(T_{ki} \leq t) - F_k(t)\}I(T_{ki} \geq u)]}{P(T_i > u)}, \tag{4.3}$$

$\lambda^c(u)$ is the hazard function for the censoring distribution, and $M_i^c(t)$ is the corresponding martingale process [11] $M_i^c(t) = N_i^c(t) - \int_0^t \lambda^c(u)Y_i(u)du$, where $N_i^c(t) = I(U_i \leq t, \Delta_i = 0)$ and $Y_i(u) = I(U_i \geq u)$. The variance of the influence function is given by

$$\sigma^2_{\psi_k} = E(\psi_{ki})^2 = E[Q_{ki}\{I(T_i \leq t) - F_k(t)\}]^2 + \int_0^L \frac{E\{L_{ki}(t,u)\}^2}{K(u)}\lambda^c(u)du, \tag{4.4}$$

where $L_{ki}(t,u) = [Q_{ki}\{I(T_i \leq t) - F_k(t)\} - G_k(t,u)]I(T_i \geq u)$. Similarly, the covariance of ψ_{1i} and ψ_{2i} is

$$\begin{aligned}\sigma_{\psi_1\psi_2} = E(\psi_{1i}\psi_{2i}) &= E[Q_{1i}Q_{2i}\{I(T_i \leq t) - F_1(t)\} \\ &\times \{I(T_i \leq t) - F_2(t)\}] + \int_0^L \frac{E\{L_{1i}(t,u)L_{2i}(t,u)\}}{K(u)}\lambda^c(u)du. \end{aligned} \tag{4.5}$$

For details on the derivation of (4.4) and (4.5) we refer the reader to Lunceford et al. [5].

4.2 A SAMPLE SIZE FORMULA FOR TESTING EQUALITY OF SURVIVAL PROBABILITIES

Denoting $F_1(t) - F_2(t)$ by D, our goal is to test the null hypothesis $H_0 : D = 0$ against $H_A : D \neq 0$. Utilizing the fact that the above estimators are consistent and asymptotically normal, the hypothesis testing could be performed using Wald's test with the test statistic being $Z = \hat{D}/\hat{\sigma}_D$, where $\hat{D} = \hat{F}_1(t) - \hat{F}_2(t)$ and $\hat{\sigma}^2_D$ is a consistent estimator of the variance of $\hat{D}$, σ^2_D. Since $\hat{F}_1(t)$ and $\hat{F}_2(t)$ are asymptotically normally distributed, Z is also asymptotically follow the standardized normal distribution under the null hypothesis. A consistent estimator of σ^2_D can be obtained by using the formula for variance and covariance estimates given in Lunceford et al. [5]. However, for power or sample size calculation in the absence of pilot data, one needs to have knowledge about the actual variance σ^2_D.

By (4.1), the asymptotic variance of $\hat{F}_k(t)$ is given by $\text{var}(\hat{F}_k(t)) = E(\psi^2_k)/n = \sigma^2_{\psi_k}/n$, and consequently, the asymptotic variance of $\hat{D}$ is

$$\sigma^2_D = \text{var}(\hat{D}) = \sigma^2_{\psi_1 - \psi_2}/n = \frac{\sigma^2_{\psi_1} + \sigma^2_{\psi_2} - 2\sigma_{\psi_1\psi_2}}{n}. \tag{4.6}$$

If the variabilities $\sigma^2_{\psi_1}$, $\sigma^2_{\psi_2}$ and $\sigma_{\psi_1\psi_2}$ were known with type I error set to α, the true difference in survival probabilities at time t, D, can be detected with pre-specified power $1 - \beta$, when the sample size is at least

$$n = \frac{\sigma^2_{\psi_1 - \psi_2} \cdot (z_{1-\alpha/2} + z_{1-\beta})^2}{D}, \tag{4.7}$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the $100(1-\alpha/2)^{th}$ and $100(1-\beta)^{th}$ percentile of the standard normal distribution, respectively. But $\sigma^2_{\psi_1}$, $\sigma^2_{\psi_2}$ and $\sigma_{\psi_1\psi_2}$ are unknown and hence educated guess should be made regarding them in order to use the above sample size formula. However, expressions (4.4) and (4.5) are too complicated. Our purpose is to express equations (4.4) and (4.5) in simplified forms in terms of the parameters of the survival distributions of sub-populations. Let us denote the cumulative distributions of the counterfactual variables T_0, T^*_{1i} and T^*_{2i} by F_0, F^*_1 and F^*_2, and the corresponding survival functions by S_0, S^*_1 and S^*_2, respectively. For $k = 1$, the first term on the right side of (4.4) can be rewritten as

$$\begin{aligned} E\{Q_{1i}^2[I(T_{1i} \leq t) - F_1(t)]^2\} &= E[\{I(T_{1i} \leq t) - F_1(t)\}^2 E(Q_{1i}^2|T_{1i})] \\ &= [Var(I(T_{1i} \leq t))] \cdot (1 - \pi_R + \pi_R/\pi_1) \\ &= F_1(t)(1 - F_1(t))(1 - \pi_R + \pi_R/\pi_1), \end{aligned} \tag{4.8}$$

where $F_1(t) = \mathrm{P}(T_{1i} \leq t) = (1 - \pi_R)F_0(t) + \pi_R F_1^*(t)$, and $\pi_R = \mathrm{P}(R_i = 1)$.

For the other term in (4.4), since $L_{1i}(t, u) = [Q_{1i}\{I(T_i \leq t) - F_1(t)\} - G_1(t, u)]I(T_i \geq u)$, we have

$$\begin{aligned} E\{L_{1i}(t, u)\}^2 = \quad & E\left[Q_{1i}^2\{I(T_i \leq t) - F_1(t)\}^2 I(T_i \geq u)\right. \\ & - 2Q_{1i}\{I(T_i \leq t) - F_1(t)\}G_1(t, u)I(T_i \geq u) \\ & \left. + G_1^2(t, u)I(T_i \geq u)\right]. \end{aligned} \tag{4.9}$$

Using the fact that $E(Q_{1i}^2|T_{1i}) = 1 - \pi_R + \pi_R/\pi_1$, the first part in equation (4.9) can be written as

$$\begin{aligned} & E\left[Q_{1i}^2\{I(T_{1i} \leq t) - F_1(t)\}^2 I(T_{1i} \geq u)\right] \\ = \; & E\left[Q_{1i}^2 I(T_{1i} \leq t)I(T_{1i} \geq u) - 2Q_{1i}^2 I(T_{1i} \leq t)F_1(t)I(T_{1i} \geq u) + Q_{1i}^2 F_1^2(t)I(T_{1i} \geq u)\right] \\ = \; & E\left[I(u \leq T_{1i} \leq t)E(Q_{1i}^2|T_{1i})\right] - 2F_1(t)E\left[I(u \leq T_{1i} \leq t)E(Q_{1i}^2|T_{1i})\right] \\ & + F_1^2(t)E\left[I(T_{1i} \geq u)E(Q_{1i}^2|T_{1i})\right] \\ = \; & (S_1(u) - S_1(t))(1 - \pi_R + \pi_R/\pi_1) - 2F_1(t)(S_1(u) - S_1(t))(1 - \pi_R + \pi_R/\pi_1) \\ & + F_1^2(t)S_1(u)(1 - \pi_R + \pi_R/\pi_1) \\ = \; & (1 - \pi_R + \pi_R/\pi_1)[F_1(t) - F_1(u) - 2F_1(t)(F_1(t) - F_1(u)) + F_1^2(t)S_1(u)]. \end{aligned} \tag{4.10}$$

Similarly, since $E(Q_{1i}|T_{1i}) = 1$, the second part in (4.9) is :

$$\begin{aligned} & -2G_1(t, u)E\left[\{I(u \leq T_{1i} \leq t) - I(T_{1i} \geq u)F_1(t)\}E(Q_{1i}|T_{1i})\right] \\ = \; & -2G_1(t, u)\left[S_1(u) - S_1(t) - (1 - S_1(t))S_1(u)\right] \\ = \; & 2G_1(t, u)S_1(t)F_1(u). \end{aligned} \tag{4.11}$$

The third part in (4.9) is :

$$\begin{aligned}
& G_1^2(t,u)E[I(T_i \geq u)] \\
= \; & G_1^2(t,u)\left[(1-\pi_R)S_0(u) + \pi_R\{\pi_1 S_1^*(u) + (1-\pi_1)S_2^*(u)\}\right]. \qquad (4.12)
\end{aligned}$$

Thus adding (4.10), (4.11), and (4.12)

$$\begin{aligned}
E\{L_{1i}(t,u)\}^2 \; = \; & (1-\pi_R+\pi_R/\pi_1)[F_1(t) - F_1(u) - 2F_1(t)(F_1(t) - F_1(u)) + F_1^2(t)S_1(u)] \\
& +2G_1(t,u)S_1(t)F_1(u) \\
& +G_1^2(t,u)\left[(1-\pi_R)S_0(u) + \pi_R\{\pi_1 S_1^*(u) + \pi_2 S_2^*(u)\}\right], \qquad (4.13)
\end{aligned}$$

where

$$\begin{aligned}
G_k(t,u) &= \frac{E[I(T_{ki} \leq t)I(T_{ki} \geq u) - F_k(t)I(T_{ki} \geq u)}{P(T_i \geq u)} \\
&= \frac{(S_k(u) - S_k(t)) - (1 - S_k(t))\, S_k(u)}{P(T_i \geq u)} = \frac{-S_k(t)F_k(u)}{P(T_i \geq u)},
\end{aligned}$$

$P(T_i \geq u) = (1-\pi_R)S_0(u) + \pi_R\left\{\pi_1 S_1^*(u) + \pi_2 S_2^*(u)\right\})$, and $S_1(t) = (1-\pi_R)S_0(t) + \pi_R S_1^*(t)$.

Thus, given the censoring distribution, the variance of ψ_{1i} can be explicitly expressed by

$$\sigma_{\psi_1}^2 = F_1(t)(1-F_1(t))(1-\pi_R+\pi_R/\pi_1) + \int_0^L \frac{(4.13) \times \lambda^c(u)}{K(u)} \mathrm{d}u. \qquad (4.14)$$

The variance of ψ_{2i}, $\sigma_{\psi_2}^2$ can be derived analogically. Now in order to obtain the variance of $\psi_{1i} - \psi_{2i}$, we need to simplify the covariance term in equation (4.5).

It can easily be shown that $E(Q_{1i}Q_{2i}|T_{1i}, T_{2i}) = 1 - \pi_R$, leading the first term in the right side of (4.5) to

$$\begin{aligned}
& E\left[Q_{1i}Q_{2i}\{I(T_i \leq t) - F_1(t)\}\{I(T_i \leq t) - F_2(t)\}\right] \\
= \; & E\left[Q_{1i}Q_{2i}\{I(T_{1i} \leq t) - F_1(t)\}\{I(T_{2i} \leq t) - F_2(t)\}\right] \\
= \; & E\left[\{I(T_{1i} \leq t)I(T_{2i} \leq t) - F_1(t)F_2(t)\}E(Q_{1i}Q_{2i}|T_{1i}, T_{2i})\right] \\
= \; & E\left[I(T_{1i} \leq t)I(T_{2i} \leq t) - F_1(t)F_2(t)\right](1-\pi_R). \qquad (4.15)
\end{aligned}$$

If we assume that $T_{1i}^* \perp T_{2i}^* | R=1$, then $E\left[I(T_{1i} \le t) I(T_{2i} \le t)\right]$ in (4.15) can be expressed as

$$\begin{aligned}
& E\left[I(T_{1i} \le t) I(T_{2i} \le t)\right] \\
= \; & P\left[I(T_{1i} \le t) I(T_{2i} \le t) | R=1\right] P(R=1) + P\left[I(T_{1i} \le t) I(T_{2i} \le t) | R=0\right] P(R=0) \\
= \; & \pi_R F_1^*(t) F_2^*(t) + (1-\pi_R) F_0(t). \qquad (4.16)
\end{aligned}$$

We substitute equation (4.16) into (4.15), then the first term in the right side of (4.5) becomes

$$\pi_R(1-\pi_R) F_1^*(t) F_2^*(t) + (1-\pi_R)^2 F_0(t) - (1-\pi_R) F_1(t) F_2(t) \qquad (4.17)$$

For the second part of (4.5), using derivations similar to (4.10), (4.11), and (4.12), we obtain

$$\begin{aligned}
E\{L_{1i}(t,u) L_{2i}(t,u)\} \; = \; & \{(1-F_1(t)-F_2(t))(F_0(t)-F_0(u)) + F_1(t)F_2(t)(1-F_0(u))\}(1-\pi_R) \\
& -G_2(t,u)\{-F_1(u)S_1(t)\} - G_1(t,u)\{-F_2(u)S_2(t)\} \\
& +G_1(t,u)G_2(t,u)\left[(1-\pi_R)S_0(u) + \pi_R(\pi_1 S_1^*(u) + \pi_2 S_2^*(u))\right]. \qquad (4.18)
\end{aligned}$$

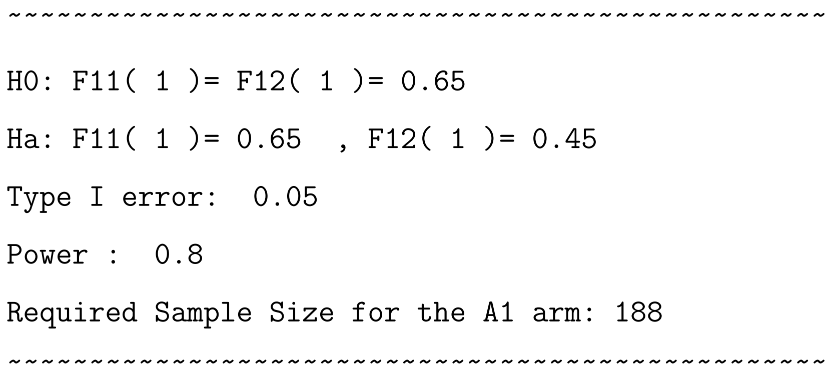
Thus the covariance $\sigma_{\psi_1\psi_2}$ between ψ_{1i} and ψ_{2i} in equation (4.5) is given by

$$\sigma_{\psi_1\psi_2} = (4.17) + \int_0^L \frac{(4.18)}{K(u)} \lambda^c(u) \mathrm{d}u. \qquad (4.19)$$

Finally, the variance of $\psi_{1i} - \psi_{2i}$ is

$$\sigma^2_{\psi_1-\psi_2} = \sigma^2_{\psi_{1i}} + \sigma^2_{\psi_{2i}} - 2\sigma_{\psi_1\psi_2}. \qquad (4.20)$$

Thus if we make working distributional assumptions for T_0, T_1^*, T_2^*, and make clinically meaningful estimates of the remission/consent rate π_R and the censoring distribution, given the randomization rate π_1, we would be able to calculate the variance $\sigma^2_{\psi_1-\psi_2}$ and hence determine the sample size using (4.7). If T_0, T_1^* and T_2^* are assumed to follow exponential distributions (irrespective of what the true distributions are), then the required distributional forms of T_0, T_1^* and T_2^* are identified by the means of T_0, T_1^* and T_2^* respectively, making computations simpler. We will compute the variance and the sample sizes based on the working assumption that the counterfactual survival times are exponentially distributed and will check the sensitivity of this assumption by generating samples from other distributions.



```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

HO: F11( 1 )= F12( 1 )= 0.65

Ha: F11( 1 )= 0.65  , F12( 1 )= 0.45

Type I error:  0.05

Power :  0.8

Required Sample Size for the A1 arm: 188

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

Figure 4.1: A snapshot of the output generated by the R routine for sample size.

Generally, the censoring time is assumed to follow a uniform distribution over some interval $[0, \tau]$, where τ is some time point beyond the length of the trial. We will adopt this assumption for the purpose of illustration. For this special case, $C \sim \text{UNIF}(0, \tau)$, $K(u) = 1 - u/\tau$ and $\lambda^c(u) = 1/(\tau - u)$. Other censoring distributions such as exponential censoring can also be implemented. We have developed an R [18] routine to calculate the sample size based on the proposed variance formula (see Appendix A.1). Figure 4.1 gives a snapshot of the output generated by the routine.

4.3 SIMULATION STUDIES

To evaluate the performance of the sample size formula proposed above, several simulations were carried out. We only simulated data for strategy A_1B_1 and A_1B_2. All simulations were based on a 2.5 year study, with the upper limit of the restricted observed lifetime being 1.5 years. For each individual, censoring time C was generated from a UNIF(0,3.5) independent of all other variables, resulting in 18% to 24% censoring at the end of one year. Remission/consent status R were sampled from and Bernoulli(0.5) in Table 4.1 and from Bernoulli(0.7) in Table 4.2. The B treatment indicators were generated from Bernoulli(0.5) distribution, i.e, for those who responded and agreed to further treatment, the randomization ratio between the two maintenance treatments was 1:1.

In the simulation scenarios described in Table 4.1 and Table 4.2, for nonresponders ($R = 0$), a survival time T_0 was drawn from EXP(α_0). For responders, we take $T_1^* \sim EXP(\alpha_1)$, $T_2^* \sim EXP(\alpha_2)$, where T_1^* and T_2^* are survival time for patients treated with B_1 and B_2, respectively. For each of the two response rate scenarios, specific values of α_0, α_1 and α_2 were chosen so that when $F_1(1.0) = 0.65$, $F_2(1.0)$ varies from 0.55 to 0.35, similarly, when $F_1(1.0) = 0.50$, $F_2(1.0)$ varies from 0.43 to 0.25. As a result, the difference in survival probability D to be detected ranged from 0.08 to 0.30. The potential restricted survival times were calculated as $T_1 = \min\{(1-R)T_0^* + RT_1^*, L\}$ and $T_2 = \min\{(1-R)T_0^* + RT_2^*, L\}$.

For the purpose of sample size determination, the variance $\sigma^2_{\psi_1-\psi_2}$ was calculated using equation (4.20) under the true assumption of exponential survival distributions (we check the sensitivity of this assumption in simulation scenarios presented later). The sample size n was then determined by the sample size formula (4.7) be setting $\alpha = 0.05$ and $\beta = 0.20$. We generated 2000 Monte-Carlo samples of size n from the true survival distributions. For each of these samples the test statistic $Z = \hat{D}/\hat{\sigma}_D$ was computed and compared to the null distribution. The observed power was calculated as the proportions of Monte Carlo data sets for which the null hypothesis was rejected at the .05 type I error level. For each scenario, data from null distributions, i.e, when the $F_2(1.0)$ was actually the same as $F_1(1.0)$, were also generated to assess the false rejection rates (type I error).

Results presented in Table 4.1 and Table 4.2 show that the sample sizes obtained using the proposed sample size formula provided observed power that are close to the expected power. In a few occasion, for smaller true differences in survival, average observed powers were smaller than the expected power. The test always maintained the nominated type I error level.

In practice, however, the distribution of true survival are unknown and thus the working assumption of the underlying survival distributions being exponential may not be valid. Simulations were also carried out where the true survival times were generated from log-normal, Weibull or a mixture of log-normal and Weibull distributions to assess the robustness of our sample size formula to the mis-specification of true distributions. The parameters of

Table 4.1: Sample size and achieved power for comparing survival probabilities at 1.0 year using inverse-probability-weighted estimator. Results are from simulation studies based on 2000 Monte Carlo data sets. Survival times for non-responders, responders in treatment strategy A_1B_1, and responders in A_1B_2 are generated from exponential distributions with means α_0,α_1 and α_2, respectively. Response rate is assumed to be 50%, randomization ratio is 1:1 for the B-treatment.

α_0	α_1	α_2	$F_1(1.0)$	$F_2(1.0)$	$\mathrm{Var}(\psi_1 - \psi_2)$	n	power	Type I error
0.825	1.10	1.64	0.65	0.58	0.7339	1166	0.780	0.056
0.825	1.10	1.98	0.65	0.55	0.7399	578	0.774	0.056
0.825	1.10	2.83	0.65	0.50	0.7438	261	0.775	0.051
0.925	0.98	3.66	0.65	0.45	0.7408	146	0.792	0.047
0.965	0.94	5.94	0.65	0.40	0.7321	92	0.805	0.052
1.054	2.04	3.52	0.50	0.43	0.7365	1177	0.817	0.050
1.054	2.04	4.83	0.50	0.40	0.7245	566	0.813	0.042
1.268	1.65	5.96	0.50	0.35	0.7107	248	0.849	0.046
2.262	0.97	3.59	0.50	0.30	0.7059	139	0.867	0.054
2.262	0.97	6.50	0.50	0.25	0.6768	85	0.887	0.056

Table 4.2: Sample size and achieved power for comparing survival probabilities at 1.0 year using inverse-probability-weighted estimator. Results are from simulation studies based on 2000 Monte Carlo data sets. Survival times for non-responders, responders in treatment strategy A_1B_1, and responders in A_1B_2 are generated from exponential distributions with means α_0,α_1 and α_2, respectively. Response rate is assumed to be 70%, randomization ratio is 1:1 for the B-treatment.

α_0	α_1	α_2	$F_1(1.0)$	$F_2(1.0)$	$\sigma^2_{\psi_1-\psi_2}$	n	power	Type I error
0.545	1.19	1.80	0.65	0.55	0.9591	760	0.720	0.045
0.545	1.19	2.29	0.65	0.50	0.9612	335	0.775	0.049
0.545	1.19	3.01	0.65	0.45	0.9540	188	0.771	0.051
0.545	1.19	4.21	0.65	0.40	0.9376	118	0.792	0.057
0.545	1.19	6.63	0.65	0.35	0.9122	80	0.772	0.051
0.855	1.84	2.61	0.50	0.43	0.9630	1513	0.805	0.052
0.855	1.84	3.10	0.50	0.40	0.9502	739	0.792	0.053
0.855	1.84	4.37	0.50	0.35	0.9220	321	0.790	0.052
0.855	1.84	7.01	0.50	0.30	08850	173	0.817	0.059
0.855	1.84	15.72	0.50	0.25	0.8397	106	0.844	0.061

Table 4.3: Achieved power when comparing survival probabilities at 1.0 year using inverse-probability-weighted estimator. Simulation studies were based on 2000 Monte Carlo data sets. Survival times for nonresponders, responders in treatment strategy A_1B_1, and responders in A_1B_2 were generated from log-normal distributions in Scenario 1, from Weibull distributions in Scenario 2. In Scenario 3, the distributions for the three survival times T_0, T_1^* and T_2^* are generated from exponential, log-normal and Weibull distribution, respectively. Sample sizes are based on the assumption that the survival times were from exponential distributions.

					Power(Type I error)	
π_R	$F_1(1.0)$	$F_2(1.0)$	n	Scenario 1	Scenario 2	Scenario 3
0.70	0.65	0.55	760	0.736(0.052)	0.793(0.050)	0.748(0.056)
	0.65	0.50	335	0.750(0.054)	0.771(0.048)	0.739(0.052)
	0.65	0.45	188	0.770(0.055)	0.795(0.053)	0.755(0.067)
	0.65	0.40	118	0.745(0.051)	0.772(0.051)	0.748(0.068)
	0.65	0.35	80	0.776(0.060)	0.789(0.046)	0.789(0.065)
0.50	0.50	0.43	1177	0.816(0.049)	0.814(0.046)	0.823(0.051)
	0.50	0.40	566	0.806(0.050)	0.821(0.050)	0.801(0.047)
	0.50	0.35	248	0.812(0.047)	0.824(0.051)	0.820(0.049)
	0.50	0.30	139	0.841(0.048)	0.832(0.043)	0.816(0.059)
	0.50	0.25	85	0.826(0.051)	0.841(0.057)	0.818(0.057)

these distributions were chosen such that the expected probabilities of survival at time 1.0 (year) is the same as those in tables 4.1 and 4.2. Table 4.3 shows that when we use the sample size formula as if the survival times were from exponential distributions, the observed power remains close to the expected power (0.80). For observed type I errors, they were again around 0.05, similar to the results presented in tables 4.1 and 4.2. In other words, the working assumption of exponential true survival played no major role in the variance and hence sample size calculation. Thus, as long as we know the means of T_0, T_1^* and T_2^*, we can ignore the true underlying distributions and calculate the variance as though the true counterfactuals followed exponential distributions. The implication of this finding is that when planning a two stage randomization clinical trial to test the equality of survival between two strategies sharing the same induction treatment, we only need to know the means of T_0, T_1^* and T_2^*, regardless of their actual distributions.

5.0 SUPREMUM WEIGHTED LOG-RANK TEST AND CORRESPONDING SAMPLE SIZE

5.1 INTRODUCTION

In the previous chapter we have considered testing the equality of two treatment strategies based on survival probabilities at fixed time t. However, survival probabilities at a fixed time t may not reflect the overall nature of the design. For the purpose of comparing overall survival pattern under various treatment strategies, Guo [12] proposed a weighted log-rank test, which was pointed out to have low power for detecting time-varying relative hazards[13].

In this chapter, we define a supremum weighted log-rank test for testing the equality of two adaptive treatment strategies sharing same maintenance treatment based on the overall survival distributions. Another important problem that we will address in this chapter is the determination of sample size to detect a particular class of alternatives. We derive a sample size formula based on the proposed supremum weighted log-rank statistic and conduct power analysis through simulation.

We will focus on comparing the survival distributions of treatment strategies A_1B_1 and A_2B_1, i.e., to compare the distributions of T_{11} and T_{21}. The method for comparing other pairs of treatment strategies follows analogously.

5.2 LOG-RANK TESTS

If everyone randomized to the initial treatment A_j remained on B_1 once they responded to A_j (that is, were there no second randomization), we could have observed the event times

$U_{j1i} = min(T_{j1i}, C_i)$ for the treatment strategy $A_jB_1, j = 1, 2$. A two sample standard log-rank statistic could then be used to test the null hypothesis of no difference between the two survival distributions related to the strategies A_1B_1 and A_2B_1. Using counting process representations [11], the standard log-rank test statistic could then be represented as

$$Z_n^{LR}(t) = \int_0^t \frac{Y_{11}(s)Y_{21}(s)}{Y_{11}(s) + Y_{21}(s)} \left\{ \frac{dN_{11}(s)}{Y_{11}(s)} - \frac{dN_{21}(s)}{Y_{21}(s)} \right\}, \tag{5.1}$$

where $N_{j1i}(s) = I(U_{j1i} \leq s, \Delta_i = 1)$, $Y_{j1i}(s) = I(U_{j1i} \geq s)$, $N_{j1}(s) = \sum_{i=1}^n N_{j1i}(s)$ and $Y_{j1}(s) = \sum_{i=1}^n Y_{j1i}(s)$ for $j = 1, 2$. Under the null hypothesis of no difference between two treatment strategies, $n^{-1/2}Z_n^{LR}(t)$ is asymptotically normally distributed with mean zero and a variance that is consistently estimated by $\sigma_n^{2\,LR}(t)$, where

$$\sigma_n^{2\,LR}(t) = n^{-1} \int_0^t \frac{Y_{11}(s)Y_{21}(s)}{Y_{11}(s) + Y_{21}(s)} \left\{ \frac{dN_{11}(s) + dN_{21}(s)}{Y_{11}(s) + Y_{21}(s)} \right\}.$$

For details on the properties of the log-rank statistic, we refer the readers to Fleming and Harrington [11].

In a TSRD study, however, U_{j1i} could not be observed for patients who are assigned to induction treatment A_j but randomized to maintenance treatment B_2 after responding to A_j. In order to account for the second randomization, Guo [12] modified the standard log-rank test by assigning a time-dependent weight to each observation, and using inverse weighting methods to derive a weighted log-rank test statistic. The weight function is defined as $W_i(s) = 1 - R_i(s) + R_i(s)X_i/\pi_z$, where $R_i(s) = R_iI(T_i^R \leq s)$ and π_z is the probability of a patient being assigned to maintenance treatment B_1 given that they responded and consented. In other words, if a patient has not responded at time s, they will have weight $W_i(s) = 1$; if the patient has responded/consented by time s and is assigned to B_1 in the second randomization, $W_i(s) = 1/\pi_z$; however, if the patient has responded/consented by time s and is assigned to B_2, which is not consistent with treatment strategy A_jB_1, then $W_i(s) = 0$. Define $N_{ji}(s) = I(U_i \leq s, \Delta_i = 1, Z_i = 2 - j)$, $Y_{ji}(s) = I(U_i \geq s, Z_i = 2 - j)$ and let $\overline{Y}_j(s) = \sum_{i=1}^n W_i(s)Y_{ji}(s)$, $\overline{N}_j(s) = \sum_{i=1}^n W_i(s)N_{ji}(s)$, be the weighted versions of at-risk and the death processes for the jth induction treatment. Based on these weighted processes

Guo [12] proposed the following inverse-probability-of-randomization-weighted (IPRW) log-rank statistic

$$Z_n(t) = \int_0^t \frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}_1(s) + \overline{Y}_2(s)} \left\{ \frac{d\overline{N}_1(s)}{\overline{Y}_1(s)} - \frac{d\overline{N}_2(s)}{\overline{Y}_2(s)} \right\}. \tag{5.2}$$

It has been shown that under the null hypothesis $n^{-1/2}Z_n(t)$ is asymptotically normally distributed with mean zero and variance

$$\sigma^2(t) = \sum_{j=1}^{2} E\left\{ \left[\int \frac{S_{3-j}^{(c)}(u)}{S_1^{(c)}(u) + S_2^{(c)}(u)} \sum_{i=1}^{n} W_i(u)\{dN_{ji}(u) - \lambda(u)Y_{ji}(u)du\} \right]^2 \right\}, \tag{5.3}$$

where $S_j^{(c)}(u) = P(U_{j1} \geq u)$ is the distribution of the overall survival time for treatment policy A_jB_1. Consequently, Guo [12] proposed to use the standardized IPRW log-rank statistic $T_n(L) = n^{-1/2}Z_n(L)/\sigma_n(L)$ to test the equality of survival curves between strategies A_1B_1 and A_2B_1, where $\sigma_n^2(t)$ is a consistent estimator of $\sigma^2(t)$.

Of importance to note here is that although the IPRW log-rank statistic accounts for the randomization in the second stage, it does not assign variable weights, for example, to early and late failures. To account for this, we define the following class of general IPRW log-rank statistics

$$Z_n^{\phi}(t) = \int_0^t \hat{\phi}_n(s) \frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}_1(s) + \overline{Y}_2(s)} \left(\frac{d\overline{N}_1(s)}{\overline{Y}_1(s)} - \frac{d\overline{N}_2(s)}{\overline{Y}_2(s)} \right), \tag{5.4}$$

where $\hat{\phi}_n(s)$ is a weight function uniformly consistent for some limiting function $\phi(s)$ over all closed subintervals of $[0, L)$ [13].

We show in section 5.3 that the variance of $n^{-1/2}Z_n^{\phi}(t)$ can be consistently estimated by

$$\sigma_n^{2^{\phi}}(t) = n^{-1} \int_0^t \left\{ \hat{\phi}_n(s) \right\}^2 \frac{\overline{Y}_2^2(s) \sum_{i=1}^n W_i^2(s)Y_{1i}(s) + \overline{Y}_1^2(s) \sum_{i=1}^n W_i^2(s)Y_{2i}(s)}{[\overline{Y}_1(s) + \overline{Y}_2(s)]^2} \left\{ \frac{d\overline{N}_1(s) + d\overline{N}_2(s)}{\overline{Y}_1(s) + \overline{Y}_2(s)} \right\}. \tag{5.5}$$

Corresponding standardized log-rank test statistics would be given by $T_n(L)$, where $T_n(t) = n^{-1/2}Z_n^{\phi}(t)/\sigma_n^{\phi}(L)$ and the supremum version by $sup_{t\in(0,L)} |T_n(t)|$. In the next section we will describe the limiting distribution of this statistic based on which a sample size formula will be derived.

5.3 LARGE SAMPLE PROPERTIES OF SUPREMUM LOG-RANK TEST

Eng and Kosorok introduced a sample size formula for the supremum log-rank statistic for two-sample censored data based on the standard contiguous time-varying proportional hazard alternative [13, 19], namely, the hazards of the two comparison groups are $\lambda_1^n(t) = \lambda_0(t)exp(\phi(t)\gamma^*/2n^{1/2})$ and $\lambda_2^n(t) = \lambda_0(t)exp(-\phi(t)\gamma^*/2n^{1/2})$, where $\lambda_0(t)$ is a continuous baseline hazard and γ^* is a scalar constant. We will use the same alternative for hazards $\lambda_1^n(t)$ and $\lambda_2^n(t)$ for strategies A_1B_1 and A_2B_1 respectively.

Let $\Lambda_j^n(t) = \int_0^t \lambda_j^n(s)ds$, $\Lambda^0(t) = \int_0^t \lambda^0(s)ds$, and

$$\overline{M}_j(t) = \overline{N}_j(t) - \int_0^t \overline{Y}_j(s)d\Lambda_j^n(s). \tag{5.6}$$

Since $N_{ji}(t)$ jumps only when the i^{th} individual from the j^{th} strategy fails, we can easily see that $W_i(t)N_{ji}(t) = W_i(t)\Delta_i I(U_i \leq t)I(X_i = 2-j)$, leading to $d(W_i(t)N_{ji}(t)) = W_i(U_i)\Delta_i I(U_i = t)I(X_i = 2-j) = W_i(t)dN_{ji}(t)$. Thus we can write,

$$\begin{aligned} d\overline{M}_j(t) &= d\overline{N}_j(t) - \overline{Y}_j(t)d\Lambda_j^n(t) \\ &= d\sum_{i=1}^n W_i(t)N_{ji}(t) - \sum_{i=1}^n W_i(t)Y_{ji}(t)d\Lambda_j^n(t) \\ &= \sum_{i=1}^n W_i(t)dM_{ji}(t), \end{aligned} \tag{5.7}$$

where $M_{ji}(t) = N_{ji}(t) - \int_0^t Y_{ji}(s)d\Lambda_j^n(s)$ is a continuous-time martingale. Now $n^{-1/2}Z_n^\phi(t)$ can be written as

$$n^{-1/2}Z_n^\phi(t) = G_n(t) + R_n(t), \tag{5.8}$$

where,

$$G_n(t) = n^{-1/2}\int_0^t \hat{\phi}_n(s)\frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}_1(s)+\overline{Y}_2(s)}\left\{\frac{d\overline{M}_1(s)}{\overline{Y}_1(s)} - \frac{d\overline{M}_2(s)}{\overline{Y}_2(s)}\right\}, \tag{5.9}$$

and

$$R_n(t) = n^{-1/2}\int_0^t \hat{\phi}_n(s)\frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}_1(s)+\overline{Y}_2(s)}(d\Lambda_1^n(s) - d\Lambda_2^n(s)). \tag{5.10}$$

Now, $G_n(t)$ can be expressed as

$$n^{-1/2}\left[\sum_{i=1}^n \int_0^t \hat{\phi}_n(s)\frac{\overline{Y}_2(s)W_i(s)}{\overline{Y}_1(s)+\overline{Y}_2(s)}dM_{1i}(s) - \sum_{i=1}^n \int_0^t \hat{\phi}_n(s)\frac{\overline{Y}_1(s)W_i(s)}{\overline{Y}_1(s)+\overline{Y}_2(s)}dM_{2i}(s)\right]. \tag{5.11}$$

By the martingale central limit theorem [11], $G_n(t)$ converges to a mean zero Gaussian process with variance equal to the limiting value of

$$\begin{aligned} & n^{-1}\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^t \left\{\hat{\phi}_n(s)\right\}^2 \frac{\overline{Y}_{3-j}^2(s)W_i^2(s)}{[\overline{Y}_1(s)+\overline{Y}_2(s)]^2} Y_{ji}(s)d\Lambda_j^n(s) \\ \approx\ & n^{-1}\int_0^t \phi^2(s)\frac{d\Lambda_0(s)}{[\overline{Y}_1(s)+\overline{Y}_2(s)]^2}\left\{\overline{Y}_2^2(s)\sum_{i=1}^{n}W_i^2(s)Y_{1i}(s)+\overline{Y}_1^2(s)\sum_{i=1}^{n}W_i^2(s)Y_{2i}(s)\right\}. \end{aligned} \tag{5.12}$$

The variance formula (5.12) leads to the consistent variance estimator (5.5).

5.4 A SAMPLE SIZE FORMULA FOR COMPARING SURVIVAL CURVES

Let n be the total number of patients and n_1 and n_2 be the number of patients assigned to A_1 and A_2, respectively. Assume that n_j/n converges to $a_j \in (0,1), j = 1,2$. Define $\pi_j^{NR}(s)$ to be the limiting distribution of $\sum_{i=1}^{n}(1-R_i(s))Y_{ji}(s)/n_j$, i.e., the proportion of patients who haven't responded to A_j at time s and are still at risk at time s among those who received A_j, and $\pi_{j1}^R(s)$ be the proportion of patients who have been assigned to A_jB_1 and are still at risk at time s among those who received A_j. Under these assumptions, we can write

$$\begin{aligned} \sum_{i=1}^{n}W_i^2(s)Y_{1i}(s) &= \sum(1-R_i(s))Y_{1i}(s)+\frac{1}{\pi_z^2}\sum R_i(s)Z_iY_{1i}(s) \\ &\approx na_1\pi_1^{NR}(s)+\frac{1}{\pi_z^2}na_1\pi_{11}^R(s), \end{aligned} \tag{5.13}$$

and similarly,

$$\sum_{i=1}^{n}W_i^2(s)Y_{2i}(s) \approx na_2\pi_2^{NR}(s)+\frac{1}{\pi_z^2}na_2\pi_{21}^R(s), \tag{5.14}$$

and

$$\overline{Y}_j(s) \approx na_j\pi_j^{NR}(s)+\frac{1}{\pi_z}na_j\pi_{j1}^R(s), j=1,2. \tag{5.15}$$

If we assume that the censoring and response rates are similar in the two induction treatment groups, then for example we can further write $\pi_1^{NR}(s) = \pi_2^{NR}(s) = \pi_0^{NR}(s)$ and $\pi_{11}^R(s) = \pi_{21}^R(s) = \pi_0^R(s)$. Consequently, equation (5.12) can be uniformly consistently approximated by

$$\int_0^t a_1a_2\phi^2(s)\left(\pi_0^{NR}(s)+\frac{\pi_0^R(s)}{\pi_z^2}\right)d\Lambda_0(s) = a_1a_2D_\phi(t) \tag{5.16}$$

where

$$\begin{aligned} D_\phi(t) &= \int_0^t \phi^2(s)\pi_0^{NR}(s)d\Lambda_0(s) + \frac{1}{\pi_z^2}\int_0^t \phi^2(s)\pi_0^R(s)d\Lambda_0(s) \\ &= \int_0^t \phi^2(s)dD^{NR}(t) + \frac{1}{\pi_z^2}\int_0^t \phi^2(s)dD^R(t), \qquad (5.17) \end{aligned}$$

where $D^{NR}(t)$ is the probability of observing an event by time t from the patients who are yet to respond by time t, $D^R(t)$ is the probability of observing an event by time t from the patients who have responded and received maintenance treatment B_1.

On the other hand, a Taylor series expansion of $\sqrt{n}(d\Lambda_1^n(s) - d\Lambda_2^n(s))$ shows that under the hypothesized contiguous alternative $R_n(t)$ can be written as

$$\begin{aligned} R_n(t) &= n^{-1}\int_0^t \hat{\phi}_n(s)\frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}_1(s)+\overline{Y}_2(s)}\sqrt{n}(d\Lambda_1^n(s) - d\Lambda_2^n(s)) \\ &\approx n^{-1}\int_0^t \phi(s)\frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}_1(s)+\overline{Y}_2(s)}\gamma^*\phi(s)\lambda_0(s)ds, \end{aligned}$$

which converges uniformly in probability to

$$\int_0^t a_1a_2\phi^2(s)\left(\pi_0^{NR}(s) + \frac{\pi_0^R(s)}{\pi_z}\right)\gamma^* d\Lambda_0(s) = \gamma^* a_1a_2 D'_\phi(t), \qquad (5.18)$$

where

$$D'_\phi(t) = \int_0^t \phi^2(s)dD^{NR}(t) + \frac{1}{\pi_z}\int_0^t \phi^2(s)dD^R(t) \qquad (5.19)$$

We can set $\gamma^* = n^{1/2}\gamma$ (for the purpose of determining the sample size), where γ refers to a fixed alternative, in which case $n^{-1/2}Z_n^\phi(t)$ converges weakly to the Gaussian process $W(a_1a_2D(t)) + n^{1/2}\gamma a_1a_2D'(t)$, where $W(\cdot)$ is a standard Brownian motion.

Using an entirely similar sets of arguments, it can be shown that $\sigma_n^{2\phi}(t)$ in Equation (5.5) converges uniformly to $a_1a_2D_\phi(t)$. Thus if we define $u(t) = D_\phi(t)/D_\phi(L)$ and assume that $\kappa = D'_\phi(t)/D_\phi(t)$ is a constant over time, then $T_n(t)$ converges weakly to

$$\begin{aligned} T(t) &= \frac{W(a_1a_2D_\phi(t))}{\sqrt{a_1a_2D_\phi(L)}} + \frac{\gamma a_1a_2\sqrt{n}D'_\phi(t)}{\sqrt{a_1a_2D_\phi(L)}} \\ &\sim W\left(\frac{D_\phi(t)}{D_\phi(L)}\right) + \gamma\kappa\sqrt{a_1a_2nD_\phi(L)}\frac{D_\phi(t)}{D_\phi(L)} \\ &= W(u(t)) + \mu u(t), \qquad (5.20) \end{aligned}$$

where $\mu = \gamma\kappa\sqrt{a_1a_2D}$ and $D = nD_\phi(L)$.

By definition, the assumption that κ is a constant follows when $D^R(t)$ and $D^{NR}(t)$ are proportional over time, i.e, the death rate for responders is proportional to that for the non-responders. This assumption also holds approximately if $D^{NR}(t)$ is large compared to $D^R(t)$, in which case (5.17) and (5.19) are dominated by the first term in their expressions. We notice that the expression (5.20) is a Brownian motion process with drift μ, denoted by $W_\mu(u)$. To compute the sample size required to achieve a power $1-\beta$ when the type I error is α in a two-sided hypothesis testing, we set $\phi(t) = 1$ (proportional hazard alternative). We then follow the procedure outlined in Eng and Kosorok [13] to obtain the critical value $S_{1-\alpha}$ for the supremum of standard Brownian motion such that

$$P(\sup_{u\in[0,1]} |W(u)| > S_{1-\alpha}) = \alpha. \tag{5.21}$$

We then solve for μ in the following expression

$$\overline{\Phi}(S_{1-\alpha} - \mu) + e^{2\mu S_{1-\alpha}}\overline{\Phi}(S_{1-\alpha} + \mu) = 1 - \beta, \tag{5.22}$$

where $\overline{\Phi} = 1-\Phi$ and Φ is the standard normal cumulative distribution. Finally, we compute

$$D = \frac{\mu^2}{a_1 a_2 \gamma^2 \kappa^2}, \tag{5.23}$$

and $n = D/D(L)$ is the required sample size. Routines for conducting the supremum/standard weighted log-rank test using software package R [18] is in Appendix A.2, and for calculating the sample size for supremum weighted log-rank test is in Appendix A.3. Notice that in the case where $\pi_z = 1$, that is, if there were no second randomization, then $D_\phi(t) = D'_\phi(t)$, leading to $\kappa = 1$, and the sample size formula coincides with the Eng and Kosorok's [13] sample size formula for the two-sample supremum log-rank test.

In the standard inverse-probability-of-randomization-weighted log-rank test, since the statistic $T_n(L)$ is asymptotically normally distributed with mean $\mu = \gamma\kappa\sqrt{a_1 a_2 D}$, the corresponding sample size can be calculated using Schoenfeld's (1983) Formula, i.e.,

$$D = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{a_1 a_2 \gamma^2 \kappa^2}, \qquad n = D/D(L). \tag{5.24}$$

where Z_q is the qth quantile of the standard normal distribution. We will compare the sample sizes resulting from the above two methods in our simulation study in section 5.6.

5.5 ANALYSIS OF CALGB 8932 DATA

We apply the supremum weighted log-rank test to a dataset from the CALGB 8923 trial reported by Stone et al. [3]. In the trial, 388 elderly patients with acute myelogenous leukaemia were randomized to two initial treatments following chemotherapy, 193 patients received infusions of granulocyte macrophage colony-stimulating factor, GM-CSF, and 195 patients received placebo. In the second stage, 37 out of the 79 patients who responded and consented in the GM-CSF group and 45 out of the 90 patients who responded and consented in the placebo group were randomized to intensification therapy I. The remaining patients who responded and consented were assigned to intensification therapy II. The goal was to compare survival distributions under different induction and maintenance combinations.

As a result of the thorough follow-up and the short survival times, 361 deaths were observed during the study length of over 3000 days. Since the responders responded at different times, each of them was assigned a set of time-varying weights depending on whether or not they had responded at that time and the status of their second-stage randomization.

Figure 5.1 compares the survival curves estimated using Guo & Tsiatis's weighted risk set estimates for two pairs of treatment strategies: GM-CSF/Intensification I vs. Placebo/ Intensification I, and GM-CSF/Intensification II vs. Placebo /Intensification II. Based on the standard weighted log-rank test, the first comparison yielded a p-value of 0.088 while the supremum weighted log-rank test described in section 5.2 produced a p-value of 0.178. For the second comparison the p-values were respectively 0.39 and 0.54 for weighted log-rank test and the supremum weighted log-rank test. Based on the results of supremum log-rank test, GM-CSF and Placebo had slightly different effects in treating acute myelogenous leukaemia patients when followed by intensification therapy I, but that, if they were followed by intensification therapy II, no significant difference was detected. The p-values from the supremum weighted log-rank test were larger than that from the standard weighted log-rank test. As shown in the hazard plots in Fig. 5.2, the hazard rates of the groups being compared cross over during the course of study, which indicates violation of the proportional hazard assumption. As a result, the supremum weighted log-rank test is more appropriate than the standard weighted log-rank test for this dataset.

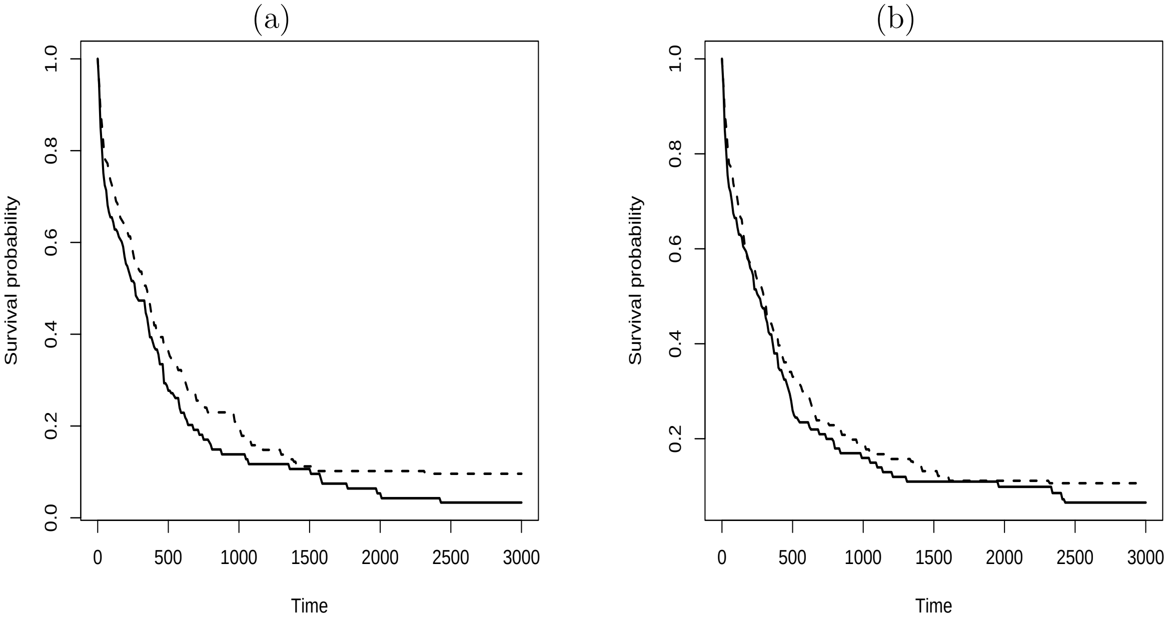


Figure 5.1: Leukaemia dataset. (a) Weighted risk set estimates of survival probability for strategies GM-CSF/Intensification I, solid line, and Placebo/Intensification I, dashed line, (b) for strategies GM-CSF/Intensification II, solid line, and Placebo/Intensification II, dashed line.

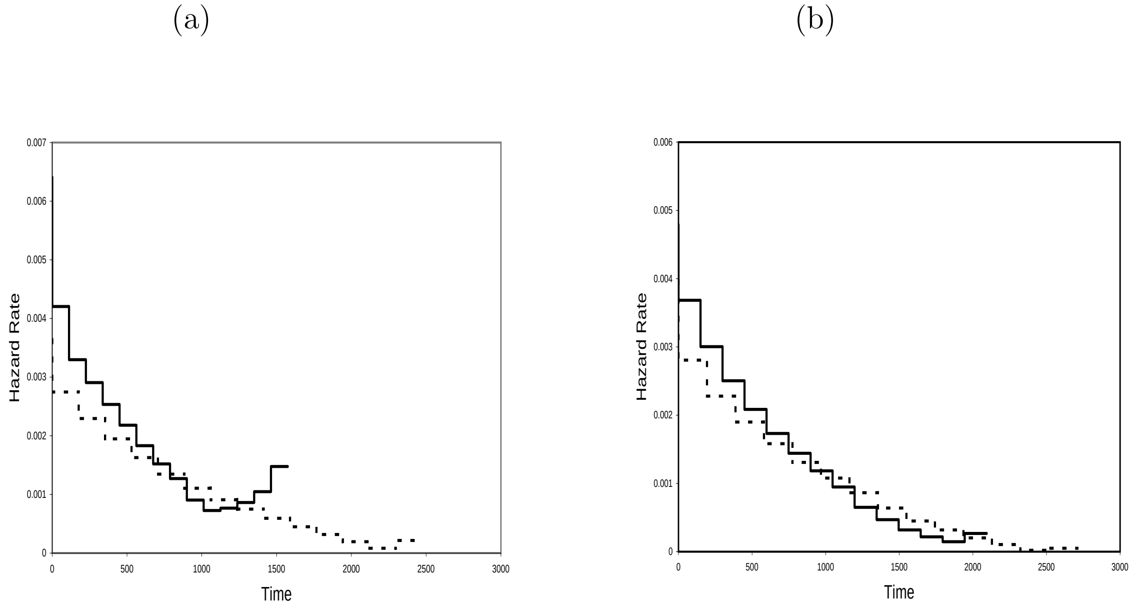


Figure 5.2: Leukaemia dataset. (a) Nonparametric hazard function estimates for strategies GM-CSF/Intensification I, solid line, and Placebo/Intensification I, dashed line, (b) for GM-CSF/Intensification II, solid line, and Placebo/Intensification II, dashed line.

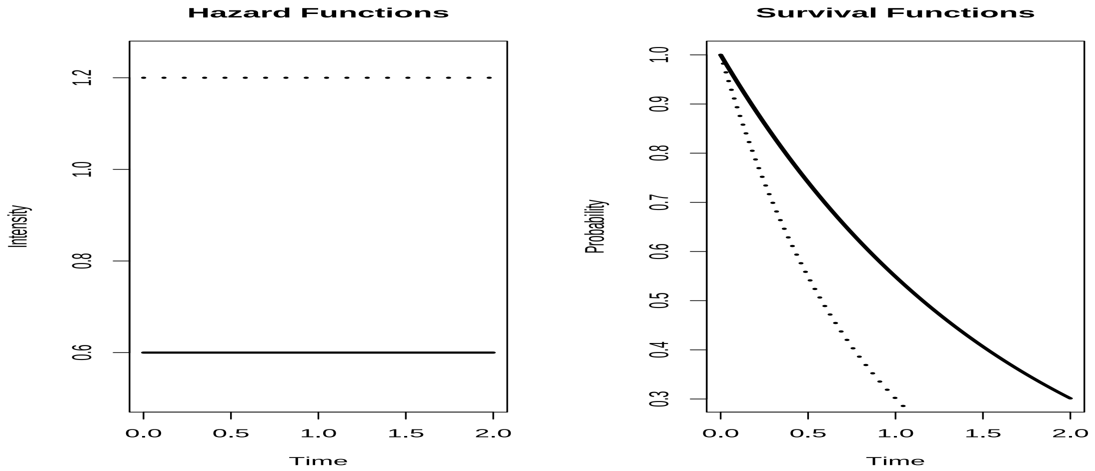


Figure 5.3: Hazard (left) and survival (right) functions when the hazards ratio between the two groups is 2. Dashed lines are for treatment strategy A_2B_1 and solid lines are for treatment strategy A_1B_1.

5.6 SIMULATION STUDIES

A series of Monte Carlo simulations were conducted to evaluate the proposed sample size formula for comparing treatment strategy A_1B_1 and A_2B_1. In the first simulation scenarios, following Lokhnygina and Helterbrand [9], to ensure a proportional hazard relationship between those two groups, we define the survival distribution for each strategy as a mixture of distributions for responders and nonresponders:

$$S_j(t) = \theta_j S_{R,j}(t) + (1-\theta_j) S_{NR,j}(t), \qquad j = 1, 2,$$

where

$$S_{R,j}(t) = \begin{cases} 1, & t \leq t^{resp}, \\ \theta_j^{-1}[exp\{-\lambda_0 exp((j-1)\gamma)(t)\} - (1-\theta_j)c_j exp(-\lambda_j^{NR}(t))], & t > t^{resp} \end{cases}, \tag{5.25}$$

$$S_{NR,j}(t) = \begin{cases} (1-\theta_j)^{-1}[exp\{-\lambda_0 exp((j-1)\gamma)t\} - \theta_j], & t \leq t^{resp}, \\ c_j exp(-\lambda_j^{NR} t), & t > t^{resp}, \end{cases} \tag{5.26}$$

t^{resp} is the time of the response assessment, θ_j is the proportion of responders in the induction treatment group A_j, e^{γ} is the hazard ratio for group A_2B_1 relative to A_1B_1, and c_j is the normalizing constant.

Since we are only interested in comparing A_1B_1 and A_2B_1 in this demonstration, the survival distribution for responders assigned to B_2 was simulated to be the same as those assigned to B_1. The baseline hazard λ_0 was set to be 0.8. The proportion of responders in treatment arms A_1 and A_2 are 0.5 and 0.4, respectively, and $\lambda_1^{NR} = 0.85$, $\lambda_2^{NR} = 0.88$, and $t^{resp} = 0.35(years)$. All patients who were still at risk at time $L = 2.0(years)$ are censored at that time. Figure 5.3 illustrates the survival functions for the two comparison groups when the hazard ratio is 2.0.

We considered two values of type I error, $\alpha = 0.01$ and $\alpha = 0.05$, two target powers, 0.80 and 0.90, and two values of the hazard ratio, 2.0 and 1.6. Also, censoring times were assumed to be distributed uniformly on the interval (0, 3.0) or (0, 2.05), yielding respectively 30% and 40% of observations censored. For each of the above 2^4 scenarios, 5000 Monte Carlo datasets were generated for the sample size calculated from the formula proposed for the supremum weighted log-rank test in section 5.4. Then both standard and supremum log-rank tests were performed on each dataset and observed powers were calculated as the proportion of times on which the null hypothesis was rejected. To see how the false rejection rate of the test matched up to the nominal value, we also generated data from the null hypothesis, $\gamma = 0$, and obtained the rejection rates.

The results are shown in Table 5.1 for 30% censoring, when censoring time was distributed as Un(0, 3.0) and 40% censoring, when censoring time was distributed as Un(0, 2.05). The target powers were achieved in almost all scenarios, and the type I errors were close to the nominal level. The power of the supremum weighted log-rank test were very close to that of Guo's standard test [12] in most scenarios, and the sample sizes required by the two tests did not differ by more than 6%.

To assess the sensitivity of the sample size and power to the assumption of constant κ, we considered other sets of simulations. Table 5.2 shows the results when κ was defined by $\kappa = D'(L/2)/D(L/2)$, instead of $\kappa = D'(L)/D(L)$ in Table 5.1, in other words, it was evaluated at the middle of the following-up time interval instead of at the last time point.

Table 5.1: Sample size, achieved power and type I error when comparing the survival curves under treatment strategies A_1B_1 and A_2B_1 using supremum IPRW log-rank test and standard IPRW log-rank test based on 5000 Monte Carlo data sets. The results from regular IPRW log rank test are given in parentheses. κ in Equation (5.23) is taken as $D'(L)/D(L)$. The rejection rates for IPRW log-rank test are based on the sample size from the Supremum IPRW log-rank test.

	Hazard	Target	Target type I	Sample		
% Censored	ratio(e^γ)	power	error(α)	size	Observed power	False rejection rate
30 %	2.0	0.80	0.05	136(128)	0.790(0.845)	0.051(0.062)
			0.01	198(191)	0.815(0.826)	0.011(0.011)
		0.90	0.05	181(172)	0.896(0.920)	0.047(0.056)
			0.01	252(243)	0.911(0.920)	0.008(0.011)
	1.6	0.80	0.05	306(289)	0.818(0.829)	0.053(0.062)
			0.01	447(430)	0.825(0.849)	0.012(0.019)
		0.90	0.05	408(387)	0.918(0.921)	0.056(0.062)
			0.01	568(548)	0.913(0.925)	0.014(0.015)
40 %	2.0	0.80	0.05	153(145)	0.812(0.843)	0.048(0.067)
			0.01	224(215)	0.829(0.848)	0.010(0.012)
		0.90	0.05	204(194)	0.919(0.936)	0.048(0.061)
			0.01	284(274)	0.921(0.930)	0.009(0.010)
	1.6	0.80	0.05	344(326)	0.830(0.837)	0.056(0.062)
			0.01	504(485)	0.829(0.861)	0.013(0.018)
		0.90	0.05	460(436)	0.920(0.936)	0.054(0.060)
			0.01	641(618)	0.913(0.924)	0.012(0.015)

Table 5.2: Sample sizes, achieved powers and type I errors when comparing the survival curves under treatment strategies A_1B_1 and A_2B_1 using the supremum inverse-probability-of-randomization-weighted log-rank test and the standard inverse-probability-of-randomization-weighted log-rank test based on 5000 Monte Carlo datasets. The results from the standard log-rank test are given in parentheses. The value of κ in equation (5.23) is taken as $\kappa = D^{`}(L/2)/D((L/2)$. The rejection rates for the standard log-rank test are based on the sample size from the supremum test.

	Hazard	Target	Target type I	Sample		
% Censored	ratio	power	error	size	Observed power	False rejection rate
30 %	2.0	0.80	0.05	126(119)	0.769(0.779)	0.041(0.052)
			0.01	184(177)	0.764(0.811)	0.008(0.011)
		0.90	0.05	168(160)	0.877(0.889)	0.043(0.055)
			0.01	234(226)	0.880(0.896)	0.010(0.012)
	1.6	0.80	0.05	282(267)	0.804(0.829)	0.054(0.057)
			0.01	412(396)	0.811(0.823)	0.012(0.018)
		0.90	0.05	376(357)	0.908(0.919)	0.052(0.060)
			0.01	524(505)	0.904(0.913)	0.013(0.017)
40%	2.0	0.80	0.05	145(137)	0.801(0.833)	0.050(0.051)
			0.01	212(204)	0.820(0.821)	0.009(0.012)
		0.90	0.05	193(183)	0.908(0.922)	0.050(0.052)
			0.01	269(259)	0.910(0.929)	0.012(0.012)
	1.6	0.80	0.05	324(306)	0.795(0.821)	0.051(0.060)
			0.01	474(456)	0.805(0.823)	0.011(0.014)
		0.90	0.05	432(410)	0.916(0.919)	0.054(0.059)
			0.01	602(581)	0.904(0.914)	0.013(0.016)

Comparing to Table 5.1, both the sample size and observed power decreased, but the changes were not substantial and the expected powers were still achieved: when the hazard ratio was 1.6, the observed power was a little lower than expected, while, when it was 2.0, the observed power was still higher than expected. Other choices of κ between $D'(L/2)/D(L/2)$ and $D'(L)/D(L)$ resulted in sample sizes and powers comparable to those presented in Table 5.2. Thus the choice of κ had little impact on the sample size based on the supremum test.

Next we consider scenarios in which the hazard ratio between the two comparison groups is not constant, i.e., when the proportional hazard assumption does not hold. In the scenarios presented in Table 5.3, we generated data by setting the value of γ in equations (5.25) and (5.26) to be $\log(1.86)$ over the time interval $[0, 1.0]$, $\log(0.90)$ over the interval $[1.0, 1.8]$ and $\log(1.10)$ thereafter. Figure 5.4 shows the hazard and survival functions of the two groups for the simulated data. It is clear that, although the hazards cross over at certain time points, the survival curve for treatment strategy A_2B_1 is always below that for A_1B_1. The goal is to see if the tests can still detect the difference in survival when the hazards are not proportional. The sample sizes are computed based on supremum weighted log-rank test with a proportional hazards assumption, where the log hazard ratio γ was set to $\log(1.6)$.

Table 5.3 shows that the powers observed for the supremum weighted log-rank test achieved the desired levels in almost all scenarios, while the standard log-rank test, which required proportional hazards assumption, failed to do so. The results provide strong evidence that the supremum weighted log-rank test is more powerful than standard weighted log-rank test in comparing strategies from TSRD when the alternative is not proportional hazards.

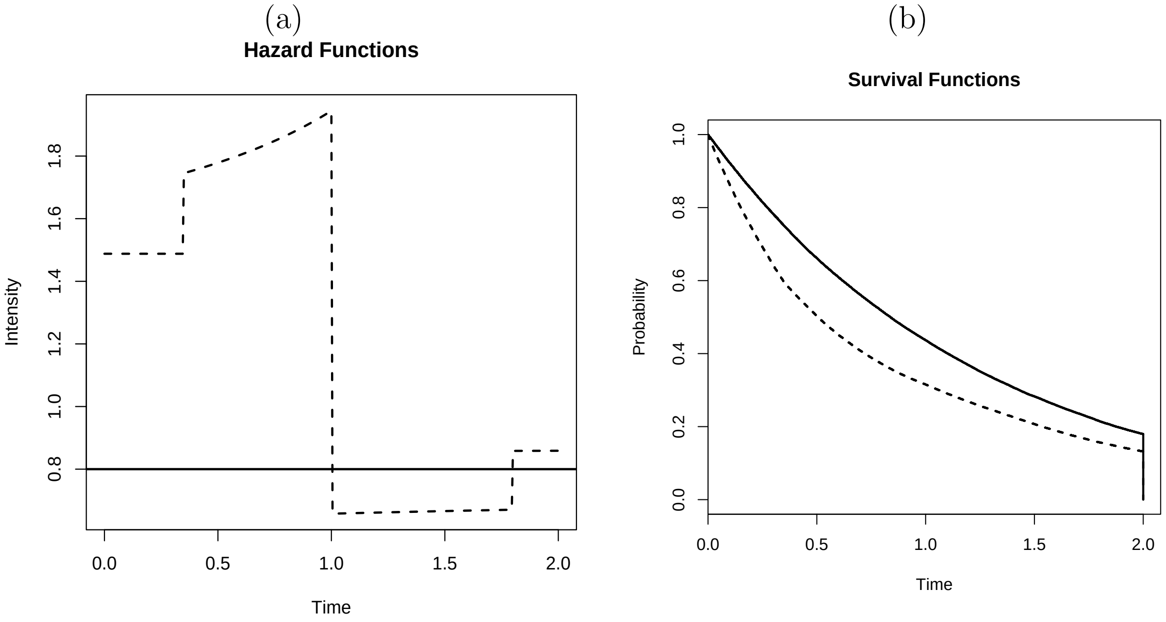


Figure 5.4: (a) Hazard and (b) survival functions when the hazards of the two groups are not proportional. Dashed lines are for treatment strategy A_2B_1 and solid lines are for treatment strategy A_1B_1.

Table 5.3: Achieved powers when comparing the survival curves under treatment strategies A_1B_1 and A_2B_1 using supremum inverse-probability-of-randomization-weighted log-rank test and Guo's inverse-probability-of-randomization-weighted log rank test, from simulation studies based on 5000 Monte Carlo datasets, when the alternative is non-proportional hazards.

				Observed power	
	Target	Target type I	Sample	Supremum	Regular
% Censored	power	error(α)	size	IPRW LR test	IPRW LR test
30 %	0.80	0.05	306	0.784	0.656
		0.01	447	0.762	0.643
	0.90	0.05	408	0.906	0.785
		0.01	568	0.879	0.731
40 %	0.80	0.05	344	0.828	0.742
		0.01	504	0.823	0.730
	0.90	0.05	460	0.919	0.843
		0.01	641	0.926	0.849

6.0 DISCUSSION AND FUTURE WORK

Sequential randomization designs are being broadly accepted in clinical trials for the purpose of conducting adaptive treatment strategies. While traditional methods of data analysis can not make efficient use of all the information obtained from such trials, recent methodologies have shown considerable advancement in this area. Lunceford et al. [5] first proposed methods for estimating survival distribution and mean restricted survival time for treatment strategies from two-stage randomization designs. The inverse-probability-weighted estimators proposed by them are consistent and asymptotically normal. However, these estimators are not asymptotically efficient, mainly because they fail to take into account the information from censored observations. Nevertheless, their method was the first valid approach toward statistical inference from two-stage designs. The estimators developed by Wahed and Tsiatis [6, 8] improves efficiency over Lunceford et al. [5] estimators by taking into account auxiliary covariates, which provides additional gain in efficiency when the covariates are prognostic of the survival time among responders. These estimators are not as simple or intuitive as the inverse-probability-weighted estimator[5] or the weighted risk set estimator defined by Guo and Tsiatis [10].

The weighted risk set estimator defined as a natural extension of the Aalen-Nelson estimator is more intuitive and easier to implement than other estimators such as inverse-probability-weighted estimator or the regular asymptotically linear efficient estimator. Our simulation study shows that weighted risk set estimator is the most efficient among the ones discussed in chapter 3, however, the estimate of survival probability shows some bias in small samples possibly due to its non-linear functional dependence on the cumulative hazard function. The small-sample bias of this estimator is larger than other estimates in most cases. The regular asymptotically linear estimator proposed by Wahed and Tsiatis [8] is the most efficient in its class, i.e, the class of regular asymptotically linear estimators, although the idea is not as intuitive and the implementation is more computationally involved.

For the purpose of comparing survival probabilities between different treatment strategies, we presented a sample size formula based on an inverse-probability-weighted consistent and asymptotically normal estimator. In determining the variance of the estimated difference between survival rates, we made the working assumption that the survival times follow exponential distributions. The simulation results show that the sample size formula achieves the desired power even when the true survival distributions are not exponential. This gives our sample size formula a broader applicability. Possible future work includes the consideration of informative censoring and the comparison among more than two strategies in similar or more complex designs.

In comparing survival curves for different treatment strategies in two stage randomization trials with censored data, we have presented a weighted supremum weighted log-rank test. This approach takes into account the second randomization, which makes use of the information for the non-responding patients as well as the patients assigned to other treatment strategies, enhancing the efficiency of the test. The supremum weighted log-rank test is more powerful than the usual log-rank test in the case of non-proportional hazard alternative. The sample size formula provided in our study is based on the limiting distribution of the test statistic and the contiguous time-varying proportional hazard alternative, and has been shown to provide desired power and nominal type I errors. As two-stage randomization is being used in many clinical trials in recent times, there is a growing need for a sample size formula for the purpose of designing such trials. The sample size formula developed in this article will serve that need, while the supremum IPRW log rank test can serve as an efficient tool to analyze such data.

APPENDIX

PROGRAMS WRITTEN IN R$^{©}$

A.1 SAMPLE SIZE FOR INVERSE PROBABILITY WEIGHTED WALD'S TEST FOR ADAPTIVE TREATMENT STRATEGIES

```
##########################################
# Name:
#    Sample.Size.tsrd.walds
# Purpose:
#    To calculate sample Size for comparing two two-stage adaptive treatment
#    strategies sharing common induction treatment.
# Arguments:
#    m0: mean survival time for the non-responders;
#    m1: mean survival time for those who responded and received the
#        maintenance treatment I;
#    m2: mean survival time for those who responded and received the
#        maintenance treatment II;
#    pir: the response/consent rate;
#    pi: the probability for a respondent to be randomized to the
#         maintenance treatment I;
#    t: the time at which survival probability to be compared;
#    LL: the restiction of lifetime (the uplimit of the survival time);
#    c: the uplimit of the uniform distribution of censoring time,
#       i.e, Censoring time ~ UNIF(0,c);
#    alpha: pre-specified Type I error;
#    beta: pre-specified Type II error.
##########################################

SampleSize<-function(m0,m1,m2,pir,pi,t,LL,c,alpha,beta)
{
 alpha0<-1/m0;alpha1<-1/m1;alpha2<-1/m2;
 S0.t<-exp(-t/alpha0);S11.t.star<-exp(-t/alpha1);S12.t.star<-exp(-t/alpha2);
 F0.t<-1-S0.t;F11.t.star<-1-S11.t.star;F12.t.star<-1-S12.t.star;
 F11.t<-(1-pir)*F0.t+pir*F11.t.star;S11.t<-1-F11.t;
 F12.t<-(1-pir)*F0.t+pir*F12.t.star;S12.t<-1-F12.t;
 cons<-1-pir+(pir/pi);

 S0.u<-function(u){exp(-u/alpha0)};F0.u<-function(u){1-S0.u(u)};
 S11.u.star<-function(u){exp(-u/alpha1)};
 S12.u.star<-function(u){exp(-u/alpha2)};
 F11.u.star<-function(u){1-S11.u.star(u)};
 F12.u.star<-function(u){1-S12.u.star(u)};
 F11.u<-function(u){(1-pir)*F0.u(u)+pir*F11.u.star(u)};
 S11.u<-function(u){1-F11.u(u)}
 F12.u<-function(u){(1-pir)*F0.u(u)+pir*F12.u.star(u)};
 S12.u<-function(u){1-F12.u(u)}

 Pr.T.u<-function(u){(1-pir)*S0.u(u)+pir*(pi*S11.u.star(u)
                     +(1-pi)*S12.u.star(u))};
 Gprim.1<-function(u){-S11.t*F11.u(u)/Pr.T.u(u)};
 Gprim.2<-function(u){-S12.t*F12.u(u)/Pr.T.u(u)};
 E.L2.1<-function(u){cons*(S11.u(u)-S11.t)-2*F11.t*cons*(S11.u(u)-S11.t)
        +cons*(F11.t^2)*S11.u(u)+2*Gprim.1(u)*S11.t*F11.u(u)
        +((Gprim.1(u))^2)*Pr.T.u(u)};   #E(L_11^2)
 E.L2.2<-function(u){cons*(S12.u(u)-S12.t)-2*F12.t*cons*(S12.u(u)-S12.t)
        +cons*(F12.t^2)*S12.u(u)+2*Gprim.2(u)*S12.t*F12.u(u)
        +((Gprim.2(u))^2)*Pr.T.u(u)};   #E(L_12^2)

 int.1<-function(u){E.L2.1(u)*(c)/((c-u)^2)};
 int.2<-function(u){E.L2.2(u)*(c)/((c-u)^2)};
 var.F11<-F11.t*S11.t*cons+integrate(int.1,0,LL)$value; # variance of phi_1
 var.F12<-F12.t*S12.t*cons+integrate(int.2,0,LL)$value; # variance of phi_2

 cov.fst<-pir*(1-pir)*F11.L.star*F12.L.star+(1-pir)^2*F0.L-(1-pir)*F11.L*F12.L
 # first term in the covariance expression
 E.L1L2<-function(u){
              (1-pir)*((1-F11.t-F12.t)*(F0.t-F0.u(u))+F11.t*F12.t*(1-F0.u(u)))
              +Gprim.2(u)*F11.u(u)*S11.t+Gprim.1(u)*F12.u(u)*S12.t
              +Gprim.1(u)*Gprim.2(u)*Pr.T.u(u)
                    }; # E(L11*L12)
 inter<-function(u){E.L1L2(u)*(c)/((c-u)^2)};
 cov.snd<-integrate(inter,0,LL)$value;
 # second term in the covariance expression
 var.F11.F12<-var.F11+var.F12-2*(cov.fst+cov.snd)    #variance of phi_1-phi_2

 Delta=F11.u(t)-F12.u(t);
 n<-((qnorm(1-alpha/2)+qnorm(1-beta))^2)*var.F11.F12/(Delta^2)
 cat(' ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~','\n',
' H0: F11(',t,')= F12(',t,')=',round(F11.t,2),'\n',
' Ha: F11(',t,')=',round(F11.t,2),' ,', 'F12(',t,')=',round(F12.t,2),'\n','
```

```
  Type I error: ', alpha,'\n',
' Power : ', 1-beta,'\n',' Requried Sample Size for the A1 arm:',ceiling(n),'\n',
'~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~','\n' )
return(invisible(list("response rate"=pir,"randomization rate"=pi,
     Delta=round(Delta,2),alpha=alpha,beta=beta,"sample size"=ceiling(n))))
}
```

A.2 COMPARING ADAPTIVE TREATMENT STRATEGIES USING SUPREMUM/STANDARD WEIGHTED LOG-RANK TEST

```
##########################################################################
# Name:
#      sup.log.rank.two.stage
# Purpose:
#     returns the p-value for the supremum weighted
#      log-rank test proposed in chapter 5
# Arguments:
#    time: the observed event time;
#    delta: the censoring indicator, 1 if death, 0 if censoring;
#    group: group indicator, 1 for A1, 2 for A2;
#    R: response indicator, 1 for responders, 0 for non-responders;
#    t.r: the time for assessing the response status;
#    pi.z: the probability of responders to be randomized to second treatment B1;
#    error: the tolerance.
# Acknowledgment:
#   Functions associated with Brownian motion are in courtesy of Professor
#   Kosorok of UNC Chapel Hill
##########################################################################

sup.log.rank.two.stage<-function (time, delta, group, R,Z,t.r, pi.z,error=1.0e-8)
{
  sup.G<-function(x,m=10)
    # This is to calculate the CDF of supremum Brownian motion
    {
     k<-m:0
     (4/pi)*sum(((-1)^k)/(2*k+1)*exp(-(pi^2)*((2*k+1)^2)/(8*x^2)))
    }
```

```
cnorm<-function(z,thresh=3.6,delta=0.6,kk=4){
   # This is to calculate the 1 - CDF of standard normal distribution
 check<-F
  if(z<0){
   z<-(-1)*z
   check<-T
         }
  if(z<thresh){
   out<-1-pnorm(z)
   }
      else{
   term<-1
   tally<-term
   if(kk>1){
        for(k in 1:(kk-1)){
             term<-(-1)*term*(2*k-1)/z^2
             tally<-tally+term
        }
   }
   out<-tally*dnorm(z)/z
   if(z<thresh+delta){
        x<-1-pnorm(z)
        out<-x+(z-thresh)*(out-x)/delta
   }
 }
 if(check){out<-1-out}
 out
 }


   n<-length(time)
   weight<-rep(1,n)
   X <- group - 1
   n2 <- sum(X)
   n1<- n-n2
   Wb<-rep(1,n)  # Weight Before time of response
   Wa<-1-R+R*Z/pi.z  # Weight After time of response

 Iu1<-matrix(0,n1,n1);Iu2<-matrix(0,n2,n2)
 y1.temp<-rep(0,n1);y2.temp<-rep(0,n2);d1.temp<-rep(0,n1);d2.temp<-rep(0,n2);
 N.u1<-matrix(0,n1,n1);N.u2<-matrix(0,n2,n2);
 y1.new.temp<-rep(0,n1);y2.new.temp<-rep(0,n2);
 V1<-time[X==0]
 V2<-time[X==1]
 delta1<-delta[X==0]
 delta2<-delta[X==1]
 Wb1<-Wb[X==0];Wa1<-Wa[X==0]
 Wb2<-Wb[X==1];Wa2<-Wa[X==1]

 for (i in 1:n1)
   { Iu1[i,]<-ifelse(V1>=V1[i],1,0)
     N.u1[i,]<-ifelse(V1<=V1[i],1,0)
     y1.temp[i]<-sum((Wb1*(V1[i]<t.r[i])+Wa1*(V1[i]>=t.r[i]))*Iu1[i,])
     y1.new.temp[i]<-sum((Wb1*(V1[i]<t.r[i])+Wa1*(V1[i]>=t.r[i]))^2*Iu1[i,])
     d1.temp[i]<-(Wb1[i]*(V1[i]<t.r[i])+Wa1[i]*(V1[i]>=t.r[i]))*delta1[i]
   }

 for (i in 1:n2)
   { Iu2[i,]<-ifelse(V2>=V2[i],1,0)
     N.u2[i,]<-ifelse(V2<=V2[i],1,0)
     y2.temp[i]<-sum((Wb2*(V2[i]<t.r[i])+Wa2*(V2[i]>=t.r[i]))*Iu2[i,])
    y2.new.temp[i]<-sum((Wb2*(V2[i]<t.r[i])+Wa2*(V2[i]>=t.r[i]))^2*Iu2[i,])
    d2.temp[i]<-(Wb2[i]*(V2[i]<t.r[i])+Wa2[i]*(V2[i]>=t.r[i]))*delta2[i]
  }

new.v<-append(V1,V2)
mis<-rep(-1,n2)
y1.temp<-append(y1.temp,mis)
y1.new.temp<-append(y1.new.temp,mis)
mis<-rep(-1,n1)
y2.temp<-append(mis,y2.temp)
y2.new.temp<-append(mis,y2.new.temp)
mis.d<-rep(-1,n2)
d1.temp<-append(d1.temp,mis.d)
mis.d<-rep(-1,n1)
d2.temp<-append(mis.d,d2.temp)
otime<-order(new.v)
new.v<-new.v[otime]
y1<-y1.temp[otime];y1.new<-y1.new.temp[otime];
y2<-y2.temp[otime];y2.new<-y2.new.temp[otime];
d1<-d1.temp[otime];
d2<-d2.temp[otime];

for (i in (n-1):1){
  if (y1[i]<0) y1[i]<-y1[i+1]
  if (y2[i]<0) y2[i]<-y2[i+1]
  if (y1.new[i]<0) y1.new[i]<-y1.new[i+1]
  if (y2.new[i]<0) y2.new[i]<-y2.new[i+1]
                 }
y1.new.temptemp<-y1.new[y1.new>0]
y2.new.temptemp<-y2.new[y2.new>0]
```

```
 for (i in 1:n){
   if (y1.new[i]<=0) y1.new[i]<-y1.new[length(y1.new.temptemp)]
   if (y2.new[i]<=0) y2.new[i]<-y2.new[length(y2.new.temptemp)]
               }
 y1.temptemp<-y1[y1>0]
 y2.temptemp<-y2[y2>0]
 for (i in 1:n){
   if (y1[i]<=0) y1[i]<-y1[length(y1.temptemp)]
   if (y2[i]<=0) y2[i]<-y2[length(y2.temptemp)]
               }

for (i in 1:n) {
   if (d1[i]<0) d1[i]<-0
   if (d2[i]<0) d2[i]<-0
                }

    #weight<-tapply(weight,time,"max")
    w <- (y1 * y2)/(y1 + y2)
    w.new<-(y1^2*y2.new+y2^2*y1.new)/(y1+y2)^2
    terms <- (d1/y1 - d2/y2)[w > 0]
    terms<-terms[!is.na(terms)]
    temp<-y1+y2-1
    temp<-ifelse(temp<1,1,temp)
    cc<-1-(d1+d2-1)/temp
    cc<-1
    vterms <- (cc*(d1 + d2)/(y1 + y2))[w > 0]
    weight<-weight[w > 0]
    w <- w[w > 0]
    w.new<-w.new[w.new>0]
    #terms <- ( w * terms)/sqrt(sum( w * vterms))
```

```
     terms.new <- ( w * terms)/sqrt(sum( w.new * vterms))
     temp<-c(0,cumsum(terms))
     temp.new<-c(0,cumsum(terms.new))
     xs<-max(temp.new)
     xi<-min(temp.new)
     if(abs(xs)>abs(xi)){test<-xs} else test<-xi
     x <- abs(test)
     m<-ceiling(max(c(1,(x*sqrt(2)/pi)*sqrt(max(c(1,log(1/(pi*error)))))-0.5)))
     p<-1-sup.G(x,m=m)
     out <- NULL
     out$test <- test
     out$p <- p

          x.logrank<-temp.new[length(temp.new)]
          out$test.logrank<-x.logrank
          out$p.logrank<-2*cnorm(abs(x.logrank))

     cat(' ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~','\n',
' Test statistics for Supremum weighted log rank test :',test,';','\n',
' P-value for Supremum weighted log rank test :',p,';','\n',
' Test statistics for Regular weighted log rank test :',x.logrank,';','\n',
' P-value for Regular weighted log rank test :',out$p.logrank,';','\n',
'~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~','\n' )

}
```

A.3 SAMPLE SIZE FOR COMPARING ADAPTIVE TREATMENT STRATEGIES USING SUPREMUM WEIGHTED LOG-RANK TEST

```
##########################################################################
# Name:
#   sample.size.sup.log.rank.two.stage
# Purpose:
#    Returns required sample size for comparing
#     the two strategies A1B1 vs. A2B1;
# Arguments:
#    alpha: Type I error rate ;
#    power: desired power ;
#    pi.z: proportion of responders to be randomized to second treatment B1;
#    gamma: the hazard ratio of A1B1 vs A2B1 in the alternative hypothesis;
#    D.NR.tau: expected proportion of death among the non-responders at the
#    end of the study;
#    D.R.tau: expected proportion of death among the responders at the end
#    of the study;
# Acknowledgment:
#   Functions associated with Brownian motion are in courtesy of Professor
#   Kosorok of UNC Chapel Hill
##########################################################################

sample.size.sup.log.rank.two.stage<-
    function(alpha,power,pi.z,gamma,D.NR.tau,D.R.tau)
{
sup.G<-function(x,m=10)
## This is to calculate the CDF of supremum brownian motion
{
     k<-m:0
     (4/pi)*sum(((-1)^k)/(2*k+1)*exp(-(pi^2)*((2*k+1)^2)/(8*x^2)))
}
sup.g<-function(x,m=10)
## This is to calculate the PDF of supremum brownian motion
{
     k<-m:0
     (pi/x^3)*sum(((-1)^k)*(2*k+1)*exp(-(pi^2)*((2*k+1)^2)/(8*x^2)))
}
cnorm<-function(z,thresh=3.6,delta=0.6,kk=4){
## This is to calculate the 1-CDF of standard normal distribution
check<-F
if(z<0){
     z<-(-1)*z
     check<-T
}
if(z<thresh){
     out<-1-pnorm(z)
}
else{
     term<-1
     tally<-term
     if(kk>1){
          for(k in 1:(kk-1)){
               term<-(-1)*term*(2*k-1)/z^2
               tally<-tally+term
          }
     }
     out<-tally*dnorm(z)/z
     if(z<thresh+delta){
          x<-1-pnorm(z)
          out<-x+(z-thresh)*(out-x)/delta
     }
}
if(check){out<-1-out}
out
}
sup.inverse<-function(alpha,error=1e-8)
# This is to calculate the critical value of
#supremum brownian motion: S_{1-alpha}
{
     x<-qnorm(1-alpha/4)
     temp<-max(1,2/x)
     m<-ceiling((x/pi)*sqrt(2*log(temp/(pi*error)))-0.5)
     if(m<0){m<-0}
     interror<-1
     while(interror>error)
     {
          yx<-sup.G(x,m=m)
          dg<-sup.g(x,m=m)
          delta<-(1-alpha-yx)/dg
          x<-x+delta
          interror<-sup.G(x)-(1-alpha)
     }
     x
}
sup.mu<-function(alpha, beta, error=1e-8)
# This is to calculate R (
#the ratio of sample size between supremum and regular)
{
     u<-sup.inverse(alpha,error=error)
     y<-1-beta
     ml<-qnorm(1-alpha/2)+qnorm(1-beta)
     x<-ml
     delta<-1
     while(delta>error)
     {
          yx<-cnorm(u-x)+exp(2*x*u)*cnorm(u+x)
          dp<-dnorm(u-x)-exp(2*u*x)*dnorm(u+x)+2*u*exp(2*u*x)*cnorm(u+x)
          delta<-(y-yx)/dp
          x<-x+delta
     }
     x
}

D.tau<-D.NR.tau+(1/pi.z^2)*D.R.tau

D.tau.prime<-D.NR.tau+(1/pi.z)*D.R.tau
kappa<-D.tau.prime/D.tau

mu.star<-sup.mu(alpha,1-power)
beta<-log(gamma)
D<-mu.star^2/(pi.z*(1-pi.z)*(beta)^2*kappa^2)
# size is the sample size using the supremum weighted log rank test #
size<-D/D.tau
size #146 when alhpa=0.05,beta=0.2, 195 when alhpa=0.05,beta=0.1
# size.wlr is the sample size using the regular weighted log rank test #
size.wlr<-(qnorm(1-alpha/2)
          +qnorm(power))^2*D.tau/(pi.z*(1-pi.z)*(beta)^2*D.tau.prime^2)
size.wlr
```

```
cat(' ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~','\n',
' Type I error :',alpha,';','Power :',power,';','\n',
' Probability of being randomized to B1 :',pi.z,';','\n',
' Hazard ratio : ', gamma,';','\n',
' Requried Sample Size Using Regular Weighted Log Rank Test:',
ceiling(size.wlr),'\n',
' Requried Sample Size Using Supremum Weighted Log Rank Test:',
 ceiling(size),'\n',
'~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~','\n' )
return(invisible(list("randomization rate"=pi.z,alpha=alpha,power=power,
               "hazard ratio"=gamma,"sample size.supremum"=ceiling(size),
                  "sample size.regular"=ceiling(size.wlr))))

}
```

BIBLIOGRAPHY

[1] A. John Rush, Maurizio Fava, Stephen R. Wisniewski, Philip W. Lavori, Madhukar H. Trivedi, Harold A. Sackeim, Michael E. Thase, Andrew A. Nierenberg, Frederic M. Quitkin, and T. Michael Kashner. Sequenced treatment alternatives to relieve depression (star*d): rationale and design. *Controlled Clinical Trials*, 25:119–142, 2004.

[2] Susan A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24:1455–1481, 2005.

[3] Richard M. Stone, Deborah T. Berg, Stephen L. George, Richard K. Dodge, Paolo A. Paciucci, Philip P. Schulman, Edward J. Lee, Joseph O. Moore, Bayard L. Powell, Maria R. Baer, Clara D. Bloomfield, , and Charles A. Schiffer. Postremission therapy in older patients with de novo acute myeloid leukemia: a randomized trial comparing mitoxantrone and intermediate-dose cytarabine with standard-dose cytarabine . *Blood*, 98:548–553, 2001.

[4] Jane N. Winter, Edie A. Weller, Sandra J. Horning, Maryla Krajewska, Daina Variakojis, Thomas M. Habermann, Richard I. Fisher, Paul J. Kurtin, William R. Macon, Mukesh Chhanabhai, Raymond E. Felgar, Eric D. Hsi, L. Jeffrey Medeiros, James K. Weick, John C. Reed, and Randy D. Gascoyne. Prognostic significance of bcl-6 protein expression in dlbcl treated with chop or r-chop: a prospective correlative study . *Blood*, 107:4207–4213, 2006.

[5] Jared K. Lunceford, Marie Davidian, and Anastasios A. Tsiatis. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 58:48–57, 2002.

[6] Abdus S. Wahed and Anastasios A. Tsiatis. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 60:124–133, 2004.

[7] James M. Robins, Andrea Rotnitsky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994.

[8] Abdus S. Wahed and Anastasios A. Tsiatis. Semiparametric efficient estimation of survival distributions in two-stage randomization designs in clinical trials with censored data. *Biometrika*, 93:163–177, 2006.

[9] Yuliya Lokhnygina and Jeffrey D. Helterbrand. Cox regression methods for two-stage randomization designs. *Biometrics*, 63:422–428, 2007.

[10] Xiang Guo and Anastasios Tsiatis. A weighted risk set estimator for survival distributions in two-stage randomization designs with censored survival data. *The International Journal of Biostatistics*, 1:1–15, 2005.

[11] Thomas R. Fleming and David P. Harrington. Counting processes and survival analysis. *New Youk: Wiley*, 1991.

[12] Xiang Guo. Statistical analysis in two stage randomization designs in clinical trials. (unpublished Ph.D. thesis, Department of Statistics, North Carolina State University). *http://www.lib.ncsu.edu/theses/available/etd-06232005-143538/unrestricted/etd.pdf*, 2005.

[13] Kevin Hasegawa Eng and Michael R. Kosorok. A sample size formula for the supremum log-rank statistic. *Biometrics*, 61:86–91, 2005.

[14] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[15] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

[16] Donald B. Rubin. Discussion of "'randomization analysis of experimental data in the fisher randomization test"' by d. basu. *Journal of the American Statistical Association*, 75:591–593, 1980.

[17] Whitney K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5:99–135, 1990.

[18] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[19] Michael R. Kosorok and C. Y. Lin. The versatility of function-indexed weighted log-rank statistics. *Journal of the American Statistical Association*, 94:320–332, 1999.

INDEX