Abstract

LOKHNYGINA, YULIYA. Topics in Design and Analysis of Clinical Trials. (Under the direction of Anastasios A. Tsiatis.)

In the first part of this dissertation we derive optimal two-stage adaptive group-sequential designs for normally distributed data which achieve the minimum of a mixture of expected sample sizes at the range of plausible values of a normal mean. Unlike standard group-sequential tests, our method is adaptive in that it allows the group size at the second look to be a function of the observed test statistic at the first look. Using optimality criteria, we construct two-stage designs which we show have advantage over other popular adaptive methods. The employed computational method is a modification of the backward induction algorithm applied to a Bayesian decision problem.

Two-stage randomization designs (TSRD) are becoming increasingly common in oncology and AIDS clinical trials as they make more efficient use of study participants to examine therapeutic regimens. In these designs patients are initially randomized to an induction treatment, followed by randomization to a maintenance treatment conditional on their induction response and consent to further study treatment. Broader acceptance of TSRDs in drug development may hinge on the ability to make appropriate intent-to-treat type inference as to whether an experimental induction regimen is better than a standard regimen in the absence of maintenance treatment within this design framework. Lunceford, Davidian, and Tsiatis (2002, Biometrics 58, 48-57)

introduced an inverse-probability-weighting based analytical framework for estimating survival distributions and mean restricted survival times, as well as for comparing treatment policies in the TSRD setting. In practice Cox regression is widely used, and in the second part of this dissertation we extend the analytical framework of Lunceford et. al. to derive a consistent estimator for the log hazard in the Cox model and a robust score test to compare treatment policies. Large sample properties of these methods are derived and illustrated via a simulation study. Considerations regarding the application of TSRDs compared to single randomization designs are discussed.

Topics in Design and Analysis of Clinical Trials

by

Yuliya Lokhnygina

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

STATISTICS

in the

GRADUATE SCHOOL
at
NC STATE UNIVERSITY
2004

Anastasios A. Tsiatis Chair of Advisory Committee	Dennis Boos
Marie Davidian	Daowen Zhang

Biography

Yuliya Lokhnygina was born and raised in Dolgoprudny, Moscow region, Russian Federation. She earned a Bachelor of Science (in 1997) and a Master of Science (in 1999) degrees in Operations Research from Moscow Institute of Physics and Technology. She also earned a Master of Science degree in Operations Research from Cornell University in 2001 before entering the Ph.D. program in Statistics at North Carolina State University in the same year. After earning her PhD, the author will join the Statistics faculty at Duke Clinical Research Institute and the Department of Biostatistics and Bioinformatics at Duke University.

Acknowledgements

I would like to express my deepest gratitude and appreciation to my academic advisor, Dr. Anastasios Tsiatis, for his constant support, guidance and inspiration.

I would also like to thank Jeff Helterbrand for the fruitful collaboration on the topic of the analysis of two-stage randomization designs, as well as for introducing me to the world of statistical research in industry.

Finally, I want to thank my parents for their endless support, love and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Preface

Advances in the field of the design and analysis of clinical trials are often motivated by practical problems suggested by the drug development industry. In this dissertation we consider two such problems.

Most phase III clinical trials are extremely expensive and involve a large number of patients. It may take many years for the potential cures of serious diseases to reach the general patient population. Consequently, investigators are often interested in designing the trials in a way that would require the least amount of resources while still guarantee the specified significance level and power. Group-sequential designs offer a reduction of the average sample size of a trial by monitoring the results periodically to allow the possibility of stopping the study early. In traditional clinical trial designs determination of the sample size requires specification of a "clinically important treatment difference", which is often an artificial and ambiguous construct. Recently, various "adaptive" designs have been proposed in the literature, that use

estimated treatment differences at interim analyses to adaptively modify the trial design and sample size. However, the rationale for these designs has been ad hoc. In this dissertation we pursued the goal of developing optimal two-stage sequential designs for normally distributed data which achieve the minimum of the expected sample sizes at the range of plausible values of a normal mean.

The problem of estimating optimal dynamic treatment regimens has received considerable attention in the last few years. During the course of treating such complex diseases as HIV and cancer, physicians have to make multiple treatment decisions along the way. Dynamic treatment regimen is a set of decision rules, one per time interval, for how the treatment type and the dosage should vary. Naturally, the challenge is to find the treatment regimen which results in the best response. For example, in the case of two time intervals and when considering combining a novel treatment to existing standard of care treatment, it is often unclear whether the best strategy is to give the novel treatment concurrently with current treatments as an induction strategy, sequentially after current treatments as a possible maintenance strategy, or both concurrently and sequentially. Such multiple questions can be assessed with a single two-stage randomization design, where patients are initially randomized to an induction treatment, followed by randomization to a maintenance treatment conditional on their induction response and consent to further participation in the study. Even though two-stage randomization designs are becoming increasingly popular, little has been done to develop adequate statistical analysis methods for such designs. In this dissertation we consider a Cox regression approach to the analysis of two-

stage randomization designs. This topic was suggested during the author's summer internship at Genentech, Inc., and was motivated by a E4494 clinical trial.

This dissertation is organized as follows. In Chapter 1 we formulate the statistical testing problem and the concept of optimal two-stage sequential designs, introduce our method of finding optimal designs and present the numerical results. Chapter 2 describes the Cox regression approach to the problem of analysis of two-stage randomization designs and presents the results of the study of performance of the proposed method via a series of simulation experiments. Some computational details for the construction of optimal two-stage designs and the proof of large sample properties of the proposed Cox regression analysis of two-stage randomization designs are given in the Appendix.

Chapter 2

Optimal Two-stage Designs

2.1 Introduction

Most large scale phase III clinical trials are monitored periodically using group-sequential tests that allow the possibility of stopping the study early if sufficiently large or small treatment differences are observed at an interim analysis while preserving the operating characteristics of the test. Recently, there has been a great deal of interest in what are termed 'adaptive sequential designs'. Traditionally, in the design of a clinical trial, the sample size is computed so that a 'clinically important treatment difference' can be detected with some specified power. Often, the criterion for the choice of such a clinically important difference is not straightforward. The appeal of the adaptive design is that it uses estimated treatment differences at interim analyses to adaptively modify the design and sample size. Roughly speaking, rather than fixing the sample size in advance, the sample size may be modified based

on the observed treatment difference according to some power or conditional power considerations. For example, Proschan and Hunsberger (1995) proposed a procedure that guarantees that a two-stage test has a desired conditional power to detect the treatment effect equal to the estimated treatment effect at the first stage, given the observed test statistic at the first stage. Another adaptive strategy, suggested by Cui, Hung and Wang (1999), is a modification of a standard two-stage group-sequential test where an initial sample size was computed based on some best guess for a clinically important alternative. The sample size, to be collected at the second stage, would be increased proportionally by the squared ratio of the treatment difference which the study was originally powered to detect and the estimated treatment difference at the first stage, if this ratio is greater than one. Because of the adaptive nature of these designs, the test statistic is modified to preserve the overall type I error. For instance, the test statistic used by Cui et al. adaptively weights the increments of the commonly used test statistic. Other examples of such adaptive designs are given by Posch and Bauer (1999) and Lehmacher and Wassmer (1999).

Because of administrative constraints, some clinical trials may not be able to conduct many interim analyses. It may be that a two-stage sequential procedure is the only logistically feasible design. In such cases, the issue then becomes how to best design two-stage sequential procedures. Rather than using standard group-sequential methods, we consider optimal two-stage designs. That is, we derive the two-stage design which minimizes the expected sample size for some fixed alternative or minimizes a weighted average of expected sample sizes across a range of alternatives

among all two-stage designs with specified level and power. We show that such an optimal test is an adaptive test where the sample size at the second stage depends on the data at the first stage. However, rather than using an ad-hoc adaptive design, we construct tests based on a rationale of optimality.

2.2 Method

We focus our attention on the problem where the data $\mathbf{X} = X_1, X_2, ...$, observed sequentially, are independent and normally distributed with mean μ and unit variance, and we wish to test the null hypothesis $H_0 : \mu \leq 0$ against the alternative $H_1 : \mu > 0$ with type I error α and type II error β at $\mu = 0$ and $\mu = \delta$ respectively. For example, in a clinical trial, the parameter μ can measure the effectiveness of the experimental treatment compared to control, with positive values of μ indicating the superiority of the experimental treatment. The data X_i may represent one observation from the i-th individual or a statistic computed from the i-th group of observations obtained between interim analyses. Although this example seems oversimplified, most test statistics used to test treatment differences, whether the outcomes be continuous, discrete, survival or longitudinal, will have, asymptotically, the same distributional structure as above; that is, most test statistics, properly normalized, computed sequentially over time, will have a joint distribution which is asymptotically normal with independent increments and variance proportional to the Fisher information; see Scharfstein, Tsiatis and Robins (1997). Consequently, for the general problem, the Fisher information would play the same role as sample size in the scenario above.

The sufficient statistic for μ after k observations is given by the partial sum $S_k = X_1 + \ldots + X_k$ and any optimal procedure will always be a function of the data through the sufficient statistic. Here we focus on two-stage group-sequential designs where a decision to either stop and reject or accept the null hypothesis or to continue to sample can be made after n_1 observations (the first stage), and the number $n_2(S_{n_1})$ of additional observations to be collected at the next stage (possibly zero) can be determined based on the observed data S_{n_1}. For simplicity, we denote the total sample size by $N = n_1 + n_2(S_{n_1})$. Note that N defined this way is a random variable. To fully describe the optimal design, we must find the first stage sample size n_1, the rule to determine the number of additional observations $n_2(S_{n_1})$ to be collected at the second stage as a function of the observed data at the first stage, and a critical value a_N where the null hypothesis is rejected if $S_N \geq a_N$ and accepted otherwise. Consequently, a two-stage design can be described by the triplet$\{n_1, n_2(S_{n_1}), a_N$; for all possible $S_{n_1}\}$.

Among two-stage designs with type I error α and type II error β at $\mu = 0$ and $\mu = \delta$ respectively, we define the optimal two-stage design as the one which minimizes the weighted average of the expected sample sizes across a range of plausible values of μ. Specifically, we want to minimize

$$\int E(N|\mu) f(\mu) dh(\mu),$$

where $\int f(\mu) dh(\mu) = 1$.

Remark 1. The specific value of δ where the power $1 - \beta$ is desired is not important in the subsequent development but serves as an anchor for comparative purposes.

That is, δ could be increased or decreased and this only changes the problem by proportionately decreasing or increasing the sample sizes at the two stages.

Remark 2. We allow $h(\mu)$ to be either Lebesgue measure to reflect continuous densities or counting measure to allow for mass at only a fixed number of μ.

Remark 3. A standard two-stage group-sequential test is one which either rejects or accepts the null hypothesis at the first stage or continues to a second stage at a fixed value $n_1 + n_2$. This test which can be represented by $n_2(S_{n_1}) = 0$ if S_{n_1} is in the first stage rejection or acceptance region or $n_2(S_{n_1}) = n_2$ if S_{n_1} is in the first stage continuation region is just a special case of the two-stage designs considered above.

Following the general approach of Lai (1973), Eales and Jennison (1992), Chang (1996) and Barber and Jennison (2002), a convenient way to devise an optimal group-sequential test is to find a solution to a corresponding Bayes sequential decision problem where we must choose between decisions $D_0 : \mu = 0$ and $D_\delta : \mu = \delta$ with associated loss function $C(D, \mu)$ taking values $C(D_0, \delta) = d_\delta$, $C(D_\delta, 0) = d_0$ and zero otherwise. For the two-stage design $D_\delta(\mathbf{X}) = (S_N \geq a_N)$, corresponding to the rejection of the null hypothesis and $D_0(\mathbf{X}) = (S_N < a_N)$ if we accept the null hypothesis. The description of the problem should also include the prior distribution of the parameter of interest $\pi(\mu)$ and the sampling cost per observation $c(\mu)$. We take the sampling cost to equal one for values of μ along the support $h(\mu)$ used to compute the average expected sample size; i.e. $c(\mu) = 1$ whenever $dh(\mu) > 0$. The prior distribution is chosen as follows: we put probability mass of $1/3$ at $\mu = 0$ and $\mu = \delta$; that is $\pi(0) = 1/3$, $\pi(\delta) = 1/3$, and spread the remaining probability of

1/3 with density $\pi(\mu) = f(\mu)/3$ with respect to the dominating measure $h(\mu)$. The solution to this Bayes problem minimizes the expected total cost consisting of the sum of the expected cost of sampling and the expected cost of the decision which equals

$$E(L) = \int E(N|\mu)\frac{f(\mu)}{3}dh(\mu) + \frac{1}{3}d_0 Pr(S_N \geq a_N|\mu = 0) + \frac{1}{3}d_\delta Pr(S_N < a_N|\mu = \delta). \quad (2.1)$$

Remark 4. The conversion of the original problem of minimizing the functional

$$\int E(N|\mu)f(\mu)dh(\mu)$$

into a Bayesian decision problem can be viewed as a version of a well-known conversion of the problem of optimizing a function $H(x)$ subject to the constraints $\phi_1(x) = 0, \phi_2(x) = 0$ to that of optimizing $g(x) = H(x) - d_0\phi_1(x) - d_1\phi_2(x)$ via using Lagrange multipliers d_0, d_1.

Remark 5. The choice of one-third prior mass at $\mu = 0$ and $\mu = \delta$ is not important. Any prior mass could be used and the choice of d_0 and d_δ can be adjusted to lead to the same $E(L)$ up to a proportionality constant.

Two classes of objective functions are of particular interest. In the first, we minimize the expected sample size at a fixed value $\mu = \phi$. For this case, we took $h(\mu)$ to be point mass at $\mu = \phi$ and $f(\phi) = 1$. We will denote functions of this class by $G_{1,\phi}$. For the second class of objective functions, denoted by $G_{2,\phi}$, we minimize the weighted average of the expected sample size over a range of μ. For this second class, we took $h(\mu)$ to be Lebesgue measure. We consider the density $f(\mu)$ to be normally

distributed with mean ϕ and standard deviation $\delta/2$. The standard deviation was chosen as a multiple of δ to reflect the range of interest when specifying the type I and type II errors. The specific choice of $\delta/2$ was arbitrary, but, experimenting with other values for the standard deviation, we found the results to be insensitive to this choice.

Thus, the expected total loss for the first class of objective functions is given by

$$E(L)_{1,\phi} = \frac{1}{3}E_{\phi}(N) + \frac{1}{3}\{d_0 Pr_0(D_\delta) + d_\delta Pr_\delta(D_0)\}, \tag{2.2}$$

and for the second class of objective functions, by

$$E(L)_{2,\phi} = \frac{1}{3}\int E(N|\mu)\frac{2}{\delta}\phi\left(\frac{\mu-\phi}{\delta/2}\right)d\mu + \frac{1}{3}\{d_0 Pr_0(D_\delta) + d_\delta Pr_\delta(D_0)\}. \tag{2.3}$$

The optimal Bayes rule should also satisfy the condition

$$Pr_0(D_\delta) = \alpha, Pr_\delta(D_0) = \beta, \tag{2.4}$$

which can be achieved by finding the appropriate values for d_0 and d_δ. As the next equation shows, the critical value a_N at the final stage, and hence the error probabilities in (2.4), depend on the ratio of d_0 and d_δ. The solution minimizing the expected total cost can be found for any values of costs d_0 and d_δ. Thus, if d_0 and d_δ are chosen to satisfy (2.4), the optimal Bayes rule will also minimize the objective function among all decision rules with error probabilities α and β, yielding the desired optimal decision rule.

The optimal solution can be computed using the dynamic programming algorithm, also known as the backward induction algorithm as it proceeds by finding the stopping

boundaries of a group-sequential design going from the final stage to the first stage. Note that the first stage can also be the final stage if $n_2(S_{n_1}) = 0$. The critical value a_N, at the final stage, is the value such that the expected loss from rejecting the null hypothesis is equal to the expected loss from accepting the null hypothesis:

$$d_0 Pr(\mu = 0|N, S_N = a_N) = d_\delta Pr(\mu = \delta|N, S_N = a_N) \tag{2.5}$$

For any value s_1 of the test statistic S_{n_1} at the first stage, the optimal additional number of observations $n_2(s_1)$ can be found by minimizing the additional conditional expected loss $E(L|S_{n_1} = s_1)$ incurred by sampling $n_2(s_1)$ observations at the second stage and proceeding optimally. When the optimal $n_2(s_1) = 0$, this corresponds to stopping the study at the first stage after n_1 observations, in which case, we reject or accept the null hypothesis according to whether S_{n_1} is greater than or smaller than a_{n_1} respectively, where a_{n_1} is defined by (2.5). By searching over the costs d_0 and d_δ we find the values for which condition (2.4) is satisfied. Finally, to minimize the expected sample size, we search over a range of possible values for n_1.

The computational details of the algorithm for the first class of objective functions are given in the Appendix.

2.3 Results

Investigators are often interested in the stopping properties of the sequential procedure at or near the clinically important alternative of interest. Consequently, we will focus attention on the performance of optimal tests for values of μ in a range

around δ. We investigated the performance of our optimal tests for the objective function $G_{1,\phi}$ at a range of values of ϕ in the interval $\{\delta/2, 3\delta/2\}$, and for the objective function $G_{2,\phi}$ where ϕ takes values $\delta/2, 3\delta/2$. We considered tests with error probabilities $\alpha = 0.05, \beta = 0.1$. We define the sample size n_{fix} to be that necessary for a fixed-sample level α test to have power $1-\beta$ to detect the alternative δ. For $\alpha = 0.05$ and $\beta = 0.1$, $n_{fix} = \{(1.64 + 1.28)/\delta\}^2$. For ease of comparison, we will present sample sizes and expected sample sizes as a percentage of n_{fix}. For each optimal decision rule, the expected sample size was evaluated at $\mu = 0, \frac{\delta}{8}, \frac{\delta}{4}, ..., \frac{11\delta}{8}, \frac{3\delta}{2}$. Figure 2.1 plots, for each optimal test, the ratio of the expected sample size to n_{fix} for the range of the parameter values at which the expected sample size was evaluated.

By construction, at a given value of $\mu = \phi$, the minimum expected sample size among all two-stage level α tests with power $1-\beta$ for $\mu = \delta$ is given by the optimal test minimizing the objective function $G_{1,\phi}$. Therefore, the expected sample size for these optimal tests as a function of ϕ in the range $\{\delta/2, 3\delta/2\}$ serves as an minimal envelope for comparing the global behavior (in terms of expected sample size across this range of alternatives) of any two-stage design. As we can see, optimal rules that minimize the objective function $G_{1,\phi}$ with the values of ϕ at the lower end of the range of possible alternatives are not very efficient at the higher end of the range of possible alternatives, with ϕ in the range $\{\delta, 3\delta/2\}$. On the other hand, optimal rules that minimize the objective function $G_{1,\phi}$ with the values of ϕ at the higher end of the range of possible alternatives are inefficient for ϕ in the range $\{0, \delta/2\}$, with expected sample size evaluated at the null hypothesis achieving values as high as $1.3n_{fix}$.

In contrast, the optimal tests based on the objective function $G_{2,\phi}$, although not best for any fixed value $\mu = \phi$, showed excellent global performance across a wide range of μ. We also notice that the performance of $G_{2,\delta/2}$ was similar to that of $G_{2,3\delta/2}$. Based on these results, we recommend using the integral objective function $G_{2,\phi}$ with ϕ chosen to emphasize the part of the parameter space the investigator wants to focus on. For example, if the investigator wants good early stopping properties for values of $\mu \geq \delta$, then a good choice is the two-stage design which minimizes the objective function $G_{2,3\delta/2}$.

For ease of implementation, these optimal two-stage designs can be described in terms of standardized statistics, relative alternative ϕ/δ and sample sizes expressed as percentages of the fixed sample size n_{fix}. Consequently, we define the sample size at the first stage as a percentage of n_{fix}, n_1/n_{fix}, and the standardized test statistic at the first stage by $Z_1 = S_{n_1}/\sqrt{n_1}$.

We denote the upper and lower standardized boundaries at the first stage as u_1 and l_1, where

$$u_1 = \min\{s_1/\sqrt{n_1} \text{ such that } n_2(s_1) = 0 \text{ and } s_1 \geq a_{n_1}\}$$

and

$$l_1 = \max\{s_1/\sqrt{n_1} \text{ such that } n_2(s_1) = 0 \text{ and } s_1 \leq a_{n_1}\}.$$

If the value of the standardized statistic $Z_1 = z_1$ at the first stage exceeds u_1, we stop and reject the null hypothesis, if $z_1 \leq l_1$, then we stop and accept the null hypothesis. If $l_1 < z_1 < u_1$ then we increase the sample size to a total sample size as a percentage

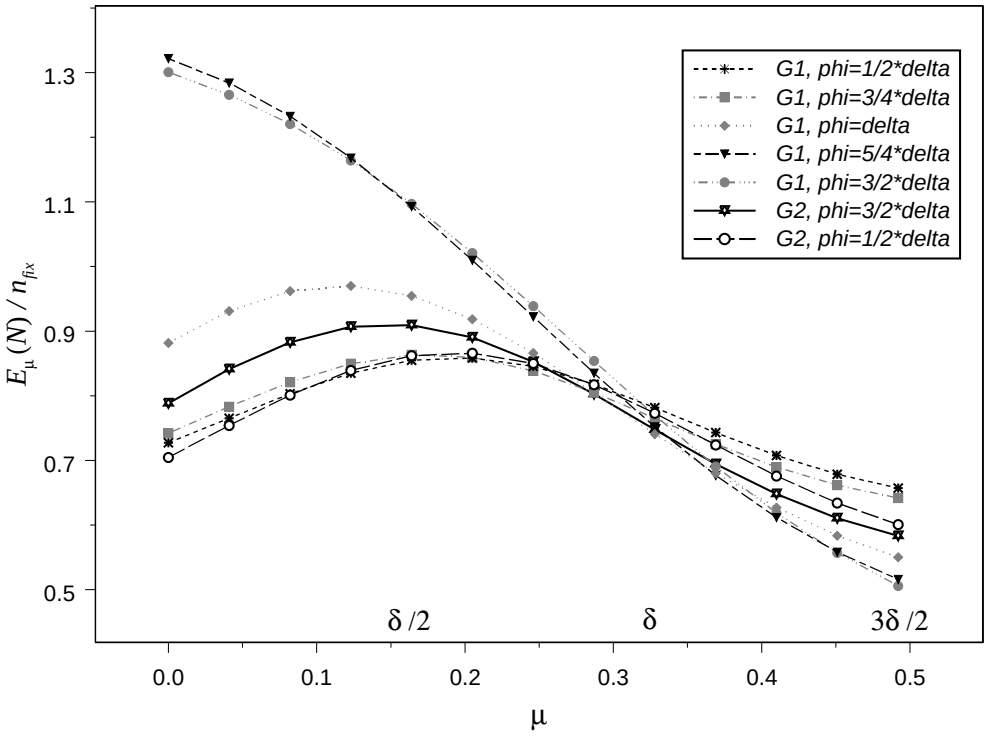

Figure 2.1: Expected sample size for optimal designs, as a proportion of n_{fix}.

of n_{fix}, N/n_{fix}. We define the standardized test statistic at the second stage by $Z_2(z_1) = S_N/\sqrt{N}$ and the boundary at the second stage by $u_{2N} = a_N/\sqrt{N}$. The test would reject the null hypothesis at the second stage if $l_1 < Z_1 < u_1, Z_2(Z_1) \geq u_{2N}$ and accept the null hypothesis if $l_1 < Z_1 < u_1, Z_2(Z_1) < u_{2N}$.

The methods described above can be used to find optimal two-stage designs for a given $\alpha, \beta, \delta, \phi$. Because the partial sum of iid normal random variables approximates Brownian motion, the design, presented in terms of standardized statistics and ratios of n_{fix}, will depend on α, β and the ratio ϕ/δ and will be insensitive to the actual value of δ as long as the sample size is sufficiently large for a good approximation to Brownian motion. Thus, for various error probabilities, tables or graphs with specifications of the optimal designs can be constructed to be used by practicing statisticians. As an example, here we consider an optimal design which minimizes the objective function $G_{2,3\delta/2} = \int E(N|\mu)\frac{2}{\delta}\phi(\frac{\mu-3\delta/2}{\delta/2})d\mu$ with error probabilities $\alpha = 0.05, \beta = 0.1$. Figure 2.2 shows the total sample size N expressed as a percentage of n_{fix}, as a function of the standardized statistic z_1 where z_1 is in the continuation region, and the boundary value at the second stage as a function of z_1.

To use this design we would proceed as follows. First we compute n_{fix}, i.e. the sample size necessary to detect the alternative of interest δ with 90% power using a test at the .05 level of significance. In accordance with the left hand panel of Figure 2.2, the first stage would be conducted after $.49 \times n_{fix}$ observations. A standardized test statistic Z_1 is then computed. Using the right hand panel of Figure 2.2, if Z_1 exceeds 2.0 then we stop the study and reject the null hypothesis. If Z_1 is less than

.40 then we stop the study and accept the null hypothesis. Otherwise we continue to a second stage with a total sample size, which is given as a function of Z_1 in the left hand panel of Figure 2.2. For example, if $Z_1 = 1.0$, then the total sample size at stage 2 is $1.29 \times n_{fix}$. Finally, we reject or accept the null hypothesis if the standardized statistic Z_2, computed using all the data, is greater than or less than the boundary value u_2, given as a function of Z_1 on the right hand panel of Figure 2.2, respectively. Again, if $Z_1 = 1.0$ we would reject the null hypothesis if $Z_2 \geq 1.85$ and accept otherwise.

To further investigate the performance of the optimal tests relative to some common adaptive designs, we compared the conditional power adaptive design by (Proschan & Hunsberger 1995) with our two optimal adaptive designs which minimize the objective functions $G_{2,\delta/2}$ and $G_{2,3\delta/2}$. All three tests have 90% power to detect the same alternative δ at the .05 level of significance. Figure 3 shows a substantial decrease in the expected sample size of either of the two optimal designs compared to conditional power adaptive test, uniformly over the range of alternatives $0 \leq \mu \leq 3\delta/2$. We observed the same effect when we compared our optimal tests with the design of (Cui, Hung & Wang 1999). In this case, we used tests at the .025 level of significance with 90% power to detect the same alternative δ. These results are depicted in Figure 4.

2.4 Discussion

In this dissertation we presented two-stage tests for normally distributed data with specified power to detect the minimally accepted treatment difference and the

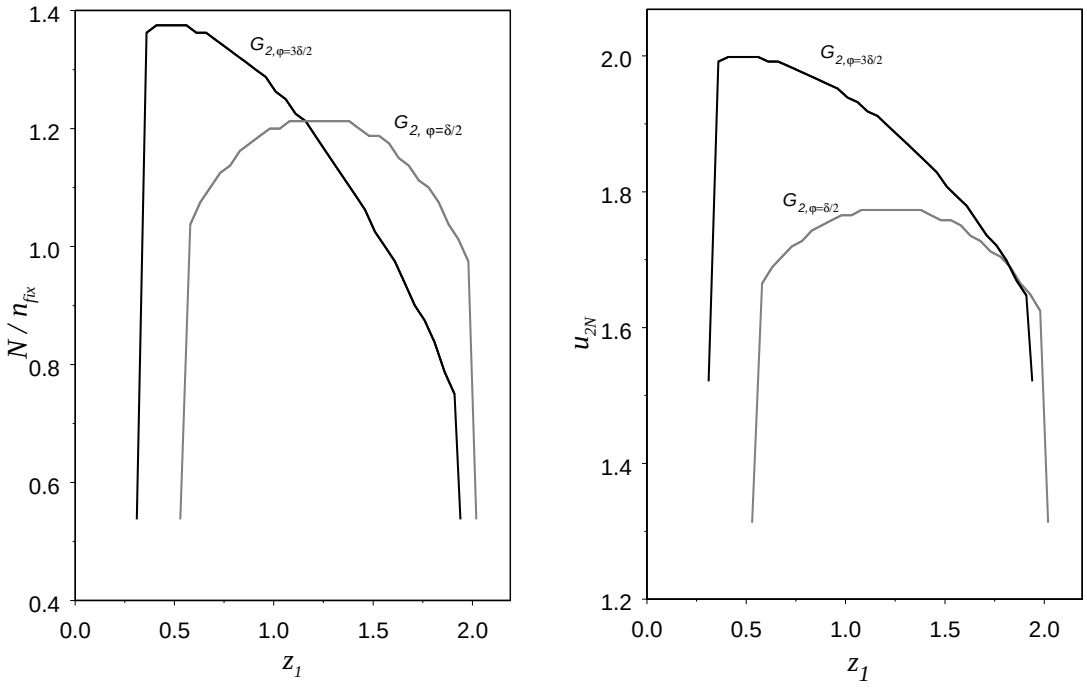

Figure 2.2: Total sample size N/n_{fix} and the standardized stopping boundary u_{2N} for objective functions $G_{2,3\delta/2}$ and $G_{2,\delta/2}$

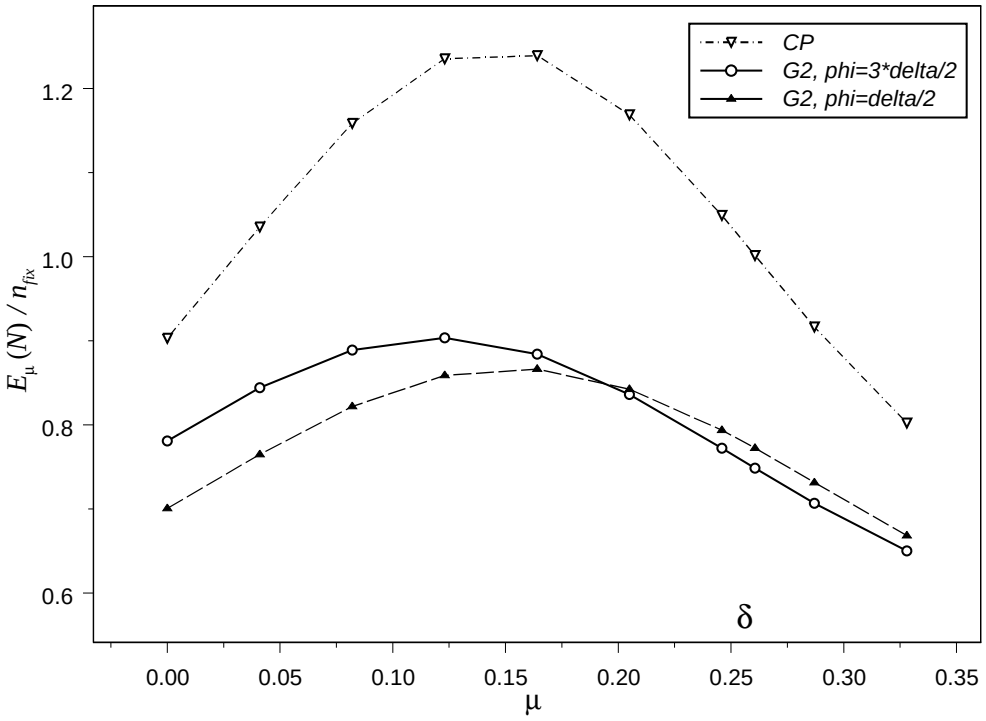

Figure 2.3: Comparison to the conditional power (CP) adaptive design.

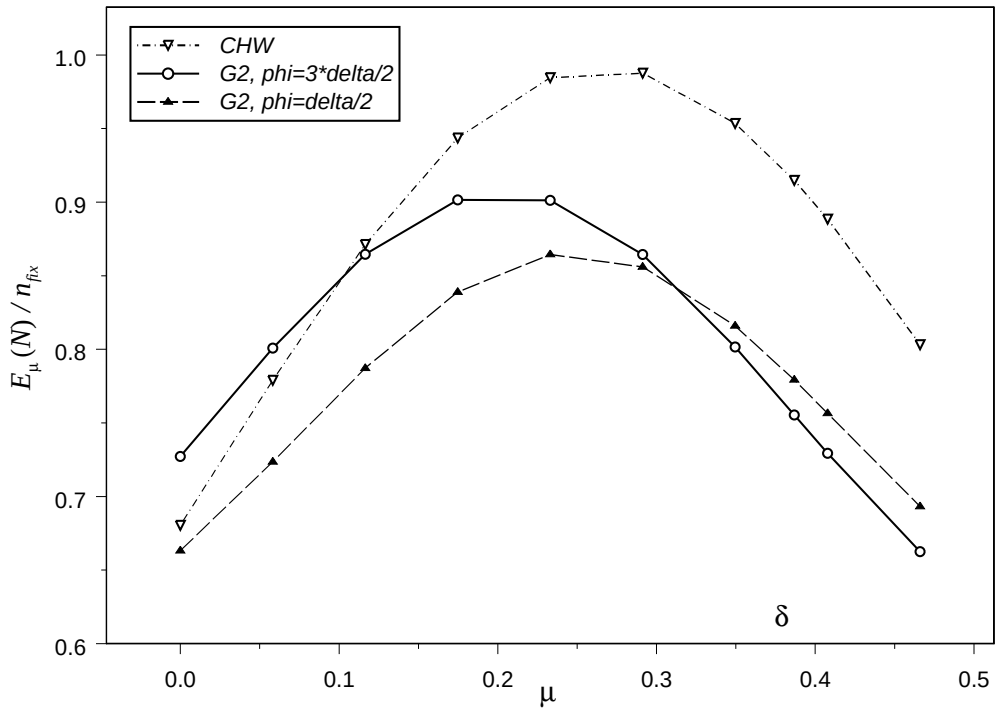

Figure 2.4: Comparison to the adaptive design of Cui et al.

property of stopping early on the average for the larger plausible treatment differences. We have shown that the design that minimizes the weighted average of the expected sample size over a range of the values of the normal mean is fairly robust under a range of possible alternatives. The proposed designs outperform some popular adaptive tests.

Chapter 3

Cox Regression Methods for Two-stage Randomization Designs

3.1 Introduction

Two-stage randomization designs (TSRDs) are becoming increasingly popular in cancer and AIDS clinical trials. In such designs patients are randomly assigned initially to an induction treatment, followed by a second randomization to a maintenance treatment if the patient responds to the induction treatment and consents to further study treatment. Many times the primary objective is to compare different combinations of treatment regimens, referred as treatment policies, to identify the best such policy with respect to a time-to-event endpoint such as survival. Recent examples of such trials are given in Thall, Sung and Estey (2002), Tummarello et al. (1994) and

Joss et al. (1994).

In drug development when considering combining a novel treatment to existing standard of care treatment, it is often unclear whether the best strategy is to give the novel treatment concurrently with current treatments as an induction strategy, sequentially after current treatments as a possible maintenance strategy, or both concurrently and sequentially. Such questions can be assessed with a single TSRD. However, broad acceptance of TSRDs will require the ability to assess in an appropriate intent-to-treat manner whether the experimental concurrent induction regimen including the novel treatment is better than the standard induction regimen.

Consider for example the E4494 clinical trial conducted by the Eastern Cooperative Oncology Group (ECOG), the Cancer and Leukemia Group B (CALGB) and the Southwest Oncology Group (SWOG) where a TSRD was employed (Habermann et al., 2003). For this Phase III clinical trial, patients were initially randomized to either chemotherapy alone (CHOP) or a combination of chemotherapy and Rituxan (R-CHOP). After receiving 6-8 courses of CHOP (approximately 6 months from the initial randomization), patients' response was assessed and responders who consented to the second randomization were assigned with equal probability to either observation or treatment with maintenance Rituxan. As the superiority of R-CHOP vs. CHOP as an induction strategy had not been demonstrated in a US-based clinical trial, a main objective of this TSRD study was to assess whether R-CHOP (followed by observation) should be the preferred induction treatment relative to CHOP (followed by observation) in terms of providing a survival benefit.

Applying standard analytic approaches in a TSRD framework cannot directly answer the question of which induction strategy is preferable with a plausible intent-to-treat interpretation and often may lead to biased inference. Recently Lunceford, Davidian and Tsiatis (2002) (subsequently referred to as LDT) and Wahed and Tsiatis (2004) have developed an analytical framework for policy inference in the TSRD setting. In these papers, estimators for the mean restricted survival time and the survival distribution for treatment policies were developed and Wald-based tests specified. Cox regression is widely used by practicing statisticians to perform inference and estimate treatment effects. In this dissertation we extend the LDT framework to the analysis of TSRDs based on the Cox proportional hazards model for the potential survival times associated with treatment combinations and illustrate its relevance for comparing induction strategies for a fixed maintenance regimen (e.g., observation).

For concreteness and following LDT, we will consider the TSRD trial where patients are initially randomized to one of two induction treatments, say A_1 or A_2, upon entry into the trial. Among those eligible for maintenance therapy, a second randomization is offered to one of the maintenance therapies B_1 or B_2. Our objective is to use the Cox model to compare the survival distributions associated with the treatment policies $A_jB_k, j, k = 1, 2$, where A_jB_k represents the policy "treat with A_j followed by B_k if the patient responds and consents to subsequent maintenance therapy."

This chapter is organized as follows. In Section 2 we introduce the model framework and the methods for estimation and testing. In Section 3 we study the performance of the method relative to alternative approaches via a series of Monte Carlo

simulation studies. In Section 4 we discuss the factors one must weigh when considering a TSRD relative to more traditional single randomization designs.

3.2 Model framework and proposed method of analysis

Let $X_i = 0(1)$ if a patient was randomized to $A_1(A_2)$, $R_i = 1$ if a patient went into remission and consented to further participation in the trial, and 0 otherwise, $Z_i = 0(1)$ if a patient was randomized to $B_1(B_2)$, defined only if $R_i = 1$. Denote $\pi_X = Pr[X = 1]$, $\pi_Z = Pr[Z = 1|R = 1]$. The problem of comparison of different treatment policies can be conveniently conceptualized through potential outcomes, or counterfactuals (Holland 1986). Let T_{jki} denote the potential survival time of the ith patient, were this patient assigned to treatment $A_jB_k, j, k = 1, 2$, and $R_{1i}(R_{2i})$ denote the potential remission/consent status the ith patient would achieve if assigned to one of the treatment policies A_1B_k (A_2B_k). We make a reasonable assumption that the survival time of the patients who did not respond or did not consent to the second randomization would be the same under treatment policies A_jB_1 and A_jB_2: $T_{j1} = T_{j2}, j = 1, 2$. We also make a standard assumption of no unmeasured confounders, also known as "strongly ignorable treatment assignment":

$$(T_{11i}, T_{12i}, T_{21i}, T_{22i}) \perp\!\!\!\perp X_i,$$

$$(T_{11i}, T_{12i}, T_{21i}, T_{22i}) \perp\!\!\!\perp Z_i \mid R_i = 1.$$

Let us consider two induction therapies A_1 and A_2 in the absence of maintenance treatment, i.e. policies A_1B_1 and A_2B_1 in the framework of LDT. Note that the framework described below can be modified to compare any pair of treatment policies. The actual observed data for the ith patient are (X_i, R_i, R_iZ_i, U_i), where $U_i = min(T_i, C_i)$, T_i is the underlying survival time and C_i is the time to right censoring. We assume that for the patients assigned to the induction treatment A_k, potential survival times are related to the observed data as

$$T_i = (1 - R_i)T_{k1i} + R_i(1 - Z_i)T_{k1i} + R_iZ_iT_{k2i}, \tag{3.1}$$

and that potential and observed consent/remission statuses are related as

$$R_i = (1 - X_i)R_{1i} + X_iR_{2i}.$$

We assume the distribution of time to censoring to be the same for all patients in each induction treatment arm, but we allow it to be different between the two induction treatment arms. Formally, this means that C_i is conditionally independent of $(R_i, R_iZ_i, T_{j1i}, T_{j2i})$, given $X_i = j - 1, j = 1, 2$.

Following a standard approach to modeling survival data, we consider Cox proportional hazards model

$$\lambda(t|X) = \lambda_0(t)exp(X\beta),$$

where $\lambda(t|X = j)$ is the hazard corresponding to therapy $A_{j+1}B_1,\ j = 0, 1$. Usual Cox regression analysis would then estimate the log-hazard ratio β from a score equation and test the hypothesis $H_0 : \beta = 0$ using a score statistic. We propose to modify the

standard as follows. The estimate of β can be obtained from solving the pseudo-score equation

$$U_{wn}(\beta) = \sum_{i=1}^{n} \int_0^{\infty} w_i[X_i - \overline{X}_w(u,\beta)]dN_i(u) = 0, \tag{3.2}$$

where

$$w_i = I[U_i \le t^{resp}] + I[U_i > t^{resp}]\{1 - R_i + \frac{R_i(1-Z_i)}{1-\pi_Z}\},$$
$$\overline{X}_w(u,\beta) = \frac{\sum_{i=1}^{n} w_j X_j Y_j(u) exp(X_j\beta)}{\sum_{i=1}^{n} w_j Y_j(u) exp(X_j\beta)},$$

$N_i(t) = I[T_i \le t, \Delta_i = 1], Y_i(t) = I[U_i \ge t], \Delta_i = I[T_i \le C_i]$, and t^{resp} is the time of the response assessment.

Theorem 3.2.1 *$n^{1/2}(\hat{\beta}_n - \beta_0)$, where $\hat{\beta}_n$ is a solution of the equation (3.2), is asymptotically normal with mean zero and covariance matrix $\Sigma = A^{-1}(\beta_0)B(\beta_0)A^{-1}(\beta_0)$, which can be consistently estimated by $\hat{\Sigma} = \hat{A}_n^{-1}\hat{B}_n\hat{A}_n^{-1}$, where*

$$\begin{aligned}
\hat{A}_n &= -n^{-1}\frac{dU_{wn}(\beta)}{d\beta}\Big|_{\beta=\hat{\beta}_n}, \\
\hat{B}_n &= n^{-1}\sum_{i=1}^{n} G_i^2(\hat{\beta}_n), \\
G_i(\beta) &= \int_0^{\infty}\{X_i - \frac{S_w^{(1)}(u,\beta)}{S_w^{(0)}(u,\beta)}\}w_i dN_i(u) \\
&\quad - \int_0^{\infty}\frac{w_i Y_i(u) exp(X_i\beta)}{S_w^{(0)}(u,\beta)}\{X_i - \frac{S_w^{(1)}(u,\beta)}{S_w^{(0)}(u,\beta)}\}d\frac{\overline{N}_w(u)}{n}, \\
\overline{N}_w(u) &= \sum_{i=1}^{n} w_i N_i(u) \quad \textit{and} \\
S_w^{(k)}(u,\beta) &= n^{-1}\sum_{i=1}^{n} w_i X_i^k Y_i(u) exp(X_i\beta), k = 0, 1.
\end{aligned}$$

The detailed proof of Theorem 3.2.1 is given in the Appendix.

The hypothesis $H_0 : \beta = 0$ can be tested with the pseudo-score statistic $\frac{U_{wn}(0)}{\sqrt{\hat{B}_n}}$, where variance of $U_{wn}(0)$ can be consistently estimated by $\hat{B}_n$.

Note that in this instance, when we are comparing A_1B_1 relative to A_2B_1, w_i is equal to 1 if the patient does not respond or consent of if the patient dies or is lost to follow-up before second randomization, $\frac{1}{\pi_Z}$ if the patient responds and is assigned B_1 at the second randomization, and equal to 0 if the patient responds and is assigned B_2. Thus, w_i acts as an inverse probability weight in the Cox regression, where patients receiving the second randomization treatment of interest represent themselves as well as the response of $(\frac{1}{\pi_Z} - 1)$ similar individuals included in the second randomization who have "missing data" with respect to A_jB_1 since they were randomized to the other maintenance treatment.

The idea of inverse probability weighting to remove bias that can result from analyzing only a sub-cohort of patients is not new. Horvitz and Thompson (1952) considered the inverse probability weighting approach to a missing data problem. Prentice (1986) proposed a pseudo-score function as an estimate of the usual partial likelihood based on complete cohort information for a problem of fitting Cox regression when covariate histories are obtained only for individuals who fail. Lin and Ying (1993) extended the pseudo-score approach to the general missing data problem under Cox regression model. Pugh et al. (1993) and Wang and Chen (2001) have used a pseudo-score approach in the problem with missing covariates. However, the application of the pseudo-score methodology to the TSRD setting has not been

established.

As shown in Therneau and Grambsch (2000) (Section 7.3), the S-Plus function $coxph()$ with option $weights$ and term $cluster(id)$ in the model statement can be used to implement the pseudo-score function approach. It employs the robust jackknife estimate of the variance of the pseudo-score statistic, which is algebraically equivalent to $\hat{B}_n$, and produces a robust score test and correct standard errors for $\hat{\beta}_n$.

3.3 Simulation study

To study the large sample properties of the proposed method to compare A_1B_1 to A_2B_1, we carried out a series of Monte Carlo simulations based on the sample size of 630 patients targeted for the E4494 study. The survival distribution for each policy was generated as a mixture of distributions for responders and non-responders:

$$S_K(t) = \theta_K S_{R,K}(t) + (1-\theta_K) S_{NR,K}(t), K = 0, 1,$$

where

$$S_{R,K}(t) = \begin{cases} 1, t \leq t^{resp}, \\ \theta_K^{-1}\{exp[-\lambda_0 exp(K\beta)t] - (1-\theta_K)c_K exp(-\lambda_K^{NR}t)\}, & t > t^{resp}, \end{cases}$$

$$S_{NR,K}(t) = \begin{cases} (1-\theta_K)^{-1}\{exp[-\lambda_0 exp(K\beta)t] - \theta_K\}, & t \leq t^{resp}, \\ c_K exp(-\lambda_K^{NR}t), & t > t^{resp}, \end{cases}$$

t^{resp} is the time of the response assessment, θ_K is the proportion of responders in the induction treatment arm A_{K+1}, and c_K is the normalizing constant.

Modeling the survival distributions in this manner ensures a proportional hazards relationship for A_1B_1 relative to A_2B_1. The choice of conditional distribution for responders makes sure that all responders survive past the response assessment time t^{resp}. In the simulation, the baseline hazard λ_0 was set equal to 0.04. The survival distribution for responders assigned to active maintenance (B_2) was simulated using the hazard function proportional to that of responders assigned to observation: $\lambda^M_{R,K}(t) = A_K\lambda_{R,K}(t)$. We considered scenarios with censoring times distributed uniformly on the interval $[0, L]$. All patients who did not fail or were censored before time L, are censored at L. With these specifications, 350-450 overall death events were observed in each simulation.

We were interested in the performance of the proposed method for the following situations: 1) no induction or maintenance effect; 2) induction effect but no maintenance effect; 3) induction effect with the same maintenance treatment effect in both induction subgroups; 4) induction effect with the maintenance treatment effect confined to the A_1 induction subgroup.

For scenarios 1-3 and 7-9, we assume the same probability of response and consent in the two induction groups. For Scenarios 4-6 and 10-12 we assume the induction response rate is greater for induction group A_2 than A_1. In scenarios 2 and 3 for completeness, we assume no induction effect, with a similar maintenance effect in both induction subgroups in scenario 2 and a maintenance effect limited to the A_1 induction subgroup for scenario 3. We assessed the performance of the proposed inverse probability weighting method (IPW) with three alternative approaches one may

consider to compare A_1B_1 to A_2B_1: 1) include all induction patients in the analysis as assigned with all patients getting equal weight (ALL); 2) include those patients who were not randomized to drug maintenance (B_2) and give these patients equal weight (UWM); 3) include all induction patients with equal weight and censor patients randomized to drug maintenance (B_2) at the time of the second randomization (CEN).

The results of the simulation experiment are presented in Tables 3.1 (with no induction effect, $\beta = 0$) and 3.2 (with induction effect, $\beta = log(0.7)$). In the tables, $Bias$ means the average of the estimated log-hazard ratio values minus the true value, SD denoted the square root of the sample variance of the estimates, ASE is the average of the standard error estimates, α denotes the average of the type I errors, Cov is the average of the sample coverages of the 95% Wald CIs and MSE denotes the mean squared error of the estimates. For all four estimators, standard error estimates SD and ASE were remarkably close. As expected by the theory, for all scenarios the IPW estimator was approximately unbiased and did maintain the targeted type 1 error. As expected, the IPW estimator with its robust variance did have greater standard errors than the alternative estimators.

As it uses all available death information, the ALL estimator consistently had the lowest standard errors. However, in two types of instances the ALL estimator exhibited considerable bias and did not maintain type I error: 1) if there was maintenance treatment effect on survival, along with different induction response rates or induction treatment effect (scenarios 5, 8, 11); or 2) if the maintenance treatment

Table 3.1: No induction effect ($\beta = 0$); N=630, $\lambda_0 = 0.04, t^{resp} = 6, L = 60$, MC = 5000 runs. Average standard errors: 0.002($Bias$), 0.002(SD), 0.002(α), 0.004(Cov)

	$(\theta_0, \theta_1, \lambda_0^{NR}, \lambda_1^{NR}, A_0, A_1)$		IPW	ALL	UWM	CEN
1	(.6,.6,.07,.07,1,1)	$Bias$	-0.004	-0.002	-0.004	-0.004
		SD	0.13	0.10	0.11	0.11
		ASE	0.13	0.10	0.12	0.12
		MSE	0.016	0.010	0.013	0.013
		α	0.026	0.028	0.025	0.025
		Cov	0.953	0.950	0.953	0.954
2	(.6,.6,.07,.07,.5,.5)	$Bias$	0.004	0.004	0.005	0.005
		SD	0.13	0.11	0.12	0.12
		ASE	0.13	0.11	0.12	0.12
		MSE	0.017	0.011	0.013	0.013
		α	0.024	0.021	0.021	0.021
		Cov	0.949	0.952	0.951	0.952
3	(.6,.6,.07,.07,.5,1)	$Bias$	0.002	0.16	0.002	0.002
		SD	0.13	0.10	0.12	0.12
		ASE	0.13	0.10	0.12	0.12
		MSE	0.017	0.035	0.014	0.014
		α	0.023	0.0002	0.024	0.021
		Cov	0.949	0.673	0.948	0.952
4	(.4,.7,.07,.07,1,1)	$Bias$	0.000	0.001	0.032	-0.071
		SD	0.13	0.10	0.12	0.12
		ASE	0.13	0.10	0.12	0.12
		MSE	0.016	0.010	0.014	0.018
		α	0.027	0.024	0.014	0.090
		Cov	0.948	0.949	0.937	0.905
5	(.4,.7,.07,.07,.5,.5)	$Bias$	-0.003	-0.10	0.029	-0.073
		SD	0.13	0.11	0.12	0.12
		ASE	0.13	0.11	0.12	0.12
		MSE	0.017	0.022	0.014	0.019
		α	0.026	0.16	0.013	0.093
		Cov	0.949	0.839	0.941	0.902
6	(.4,.7,.07,.07,.5,1)	$Bias$	0.000	0.088	0.032	-0.071
		SD	0.13	0.10	0.12	0.11
		ASE	0.13	0.10	0.12	0.11
		MSE	0.016	0.018	0.014	0.018
		α	0.024	0.003	0.011	0.084
		Cov	0.955	0.867	0.944	0.912

Table 3.2: Induction effect ($\beta = log(0.7)$); N=630, $\lambda_0 = 0.04, t^{resp} = 6, L = 60$, $MC = 5000$ runs. Average standard errors: 0.002($Bias$), 0.002(SD), 0.004(Cov)

	$(\theta_0, \theta_1, \lambda_0^{NR}, \lambda_1^{NR}, A_0, A_1)$		IPW	ALL	UWM	CEN
7	(.6,.6,.07,.07,1,1)	$Bias$	0.000	0.000	0.057	0.064
		SD	0.14	0.11	0.12	0.12
		ASE	0.14	0.11	0.12	0.12
		MSE	0.019	0.012	0.018	0.019
		Cov	0.951	0.949	0.925	0.913
8	(.6,.6,.07,.07,.5,.5)	$Bias$	0.002	0.045	0.059	0.066
		SD	0.14	0.11	0.12	0.12
		ASE	0.14	0.11	0.12	0.12
		MSE	0.019	0.014	0.018	0.019
		Cov	0.949	0.930	0.920	0.913
9	(.6,.6,.07,.07,.5,1)	$Bias$	-0.003	0.16	0.055	0.062
		SD	0.14	0.11	0.12	0.12
		ASE	0.14	0.11	0.12	0.12
		MSE	0.018	0.038	0.017	0.018
		Cov	0.951	0.674	0.926	0.919
10	(.4,.7,.07,.07,1,1)	$Bias$	0.000	0.002	0.062	-0.023
		SD	0.14	0.11	0.12	0.12
		ASE	0.14	0.11	0.12	0.12
		MSE	0.019	0.011	0.019	0.015
		Cov	0.948	0.950	0.913	0.945
11	(.4,.7,.07,.07,.5,.5)	$Bias$	0.000	-0.065	0.061	-0.023
		SD	0.13	0.11	0.12	0.12
		ASE	0.13	0.11	0.12	0.12
		MSE	0.018	0.017	0.018	0.015
		Cov	0.950	0.911	0.917	0.949
12	(.4,.7,.07,.07,.5,1)	$Bias$	-0.001	0.093	0.061	-0.024
		SD	0.14	0.11	0.12	0.12
		ASE	0.14	0.11	0.12	0.12
		MSE	0.019	0.020	0.019	0.015
		Cov	0.946	0.858	0.917	0.947

effect was different based on the induction treatment received (scenarios 3, 6, 9, 12). Note that the latter case of a maintenance by induction interaction (e.g., scenario 9) would be likely in practice if the novel treatment exhibited a benefit when combined with standard care either concurrently or sequentially, but garnished little additional benefit as maintenance therapy if received as part of the induction regimen.

The UWM and CEN estimators exhibited bias in every scenario with an induction treatment effect (scenarios 7-12) or when there was a difference in induction response rates (scenarios 4-6). This is not surprising as responding patients in these scenarios are differentially under-represented in the assessments of the two individual induction survival distributions and this therefore translates to the estimate of the log hazard.

We have also considered using time-dependent weights

$$w_i^*(t) = I[t \leq t^{resp}] + I[t > t^{resp}]\{1 - R_i + \frac{R_i(1 - Z_i)}{1 - \pi_Z}\},$$

which potentially can lead to efficiency gain as more subjects are used to estimate the survival distributions in the time interval $[0, t^{resp}]$. However, our simulation experiments have shown that this efficiency gain is small (less than 5%) and may be outweighed by the convenience of implementing IPW analysis using standard software.

3.4 Considerations for two-stage randomization designs

Prior to its use, one should compare the implications of conducting a TSRD clinical trial relative to two kinds of single randomization designs (SRD): 1) a SRD where patients are randomized upfront to one of the available four induction/maintenance combinations, and 2) a SRD that compares the two induction treatments with no maintenance randomization.

In the first instance and as discussed in LDT, the analytical framework presented here for TSRDs allows a similar intent-to-treat type inference of the four policies to be made while making more efficient use of the information from patients who do not respond or do not consent. Patients randomized to induction A_1 who do not respond or do not consent are used when estimating survival distributions for both A_1B_1 and A_1B_2.

Additionally, the ability to compare maintenance regimens in an intent-to-treat manner is compromised by the upfront randomization, as one is required to compare all patients randomized to B_1 and B_2 regardless of whether the patient actually responded to induction treatment and therefore received the maintenance treatment assigned, introducing further variability. This reflects the idea that it is best to compare any two maintenance treatments by randomizing patients as closely as possible to the time of maintenance treatment start.

In the second instance, TSRDs have the obvious advantage that a maintenance

hypothesis can be assessed in the same study where an induction hypothesis is tested, and the analysis of maintenance strategies is straightforward. However, non-standard analysis methods as introduced in LDT and here must be employed in the TSRD setting to compare induction strategies for a given maintenance treatment in an intent-to-treat manner. Although badly biased in some instances as shown above, one may be tempted to compare induction strategies using the ALL method. However, the intent-to-treat interpretation of such an analysis is non-plausible: "treat with A_j followed by the physician tossing a two-sided coin to determine next treatment if the patient responds and consents to the coin flip."

A SRD is preferable to a TSRD if one is truly only interested in the induction comparison as all patients can be used in the intent-to-treat analysis yielding more precise estimates of effect and greater power. However, it is often questioned whether a novel treatment should given concurrently, subsequently, or both concurrently and subsequently relative to current treatments, and this can be effectively assessed in one TSRD.

The additional assumption we have to impose when considering a TSRD vs. a SRD is that within each induction subgroup, responding patients randomized to B_1 should be suitably similar to those assigned to B_2 with respect to factors at baseline and prior to the second randomization. In addition to the second randomization process itself, stratification at the second randomization for important factors can be used to promote this similarity. Additionally, the potential impact of any lack of similarity should be put into the perspective of the other inverse weighting that already occurs

in survival analysis, namely, due to censoring where randomization and stratification are not available to promote similarity. Fortunately, any potential imbalances can be assessed directly and then addressed via standard sensitivity analyses.

3.5 Discussion

In this chapter we extend the analytical framework of LDT and propose re-weighted versions of the usual score estimating equation and the score test in the Cox model to be used for the analysis of TSRD clinical trials. These methods yield relevant intent-to-treat interpretations. The simulation results demonstrate that the IPW estimator is unbiased and maintains type I error while alternative estimators, while yielding small standard errors, can be badly biased in some instances. In particular, when the effect of maintenance treatment differs by the induction treatment received, the alternative estimators will perform poorly.

The Cox model analytical framework described here can be easily extended to incorporate baseline covariates, as shown in Binder (1992). One open question is whether inverse probability weights as used here should be favored over empirical weights that are determined by the actual number of patients randomized to the two maintenance strategies. Both approaches will yield consistent estimators and tests. Additionally, methods for sample size determination in TSRDs need development, as one now may need to ensure enough power to test both an induction and a maintenance hypothesis in the same study.

The TSRD framework offers several advantages in terms of efficiency. The broader

acceptance of TSRDs in drug development will likely hinge on the ability to make appropriate intent-to-treat type inference as to whether an experimental induction regimen is better than a standard regimen within this design framework. The methods described in this dissertation create such an analytical framework, with mild additional assumptions relative to single randomization designs that can be assessed and effectively dealt with if violated.

The E4494 study illustrates the need for considering the methods developed in this dissertation. At the 2003 American Society of Hematology meeting, study authors reported a considerable maintenance Rituxan by induction treatment interaction, with maintenance effects predominantly confined to the CHOP induction subgroup (Habermann, Weller, Morrison, Cassileth, Cohn, Dakil, Gascoyne, Woda, Fisher, Peterson & Horning 2003). The authors proceeded to conduct two induction strategy comparisons, effectively the ALL and IPW analyses for survival. For the ALL analysis, the authors reported a p-value of 0.29, while for the IPW analysis they reported a p-value of 0.03 and a hazard ratio of 0.69 in favor of R-CHOP over CHOP. Hence, induction strategy conclusions related to survival in this instance would be largely dependent on the analytical method assessed as most relevant.

Bibliography

Andersen, P. & Gill, R. (1982), 'Cox's regression model for counting processes: a large sample study', *The Annals of Statistics* **10**, 1100–1120.

Barber, S. & Jennison, C. (2002), 'Optimal asymmetric one-sided group sequential tests', *Biometrika* **89**, 49–60.

Binder, D. (1992), 'Fitting cox's proportional hazards models from survey data', *Biometrika* **79**, 139–147.

Chang, M. N. (1996), 'Optimal designs for group sequential clinical trials', *Communications in statistics A* **25**, 361–379.

Cui, L., Hung, H. M. J. & Wang, S.-J. (1999), 'Modification of sample size in group sequential clinical trials', *Biometrics* **55**, 853–857.

Eales, J. D. & Jennison, C. (1992), 'An improved method for deriving optimal one-sided group sequential tests', *Biometrika* **79**, 13–24.

Eales, J. D. & Jennison, C. (1995), 'Optimal two-sided group sequential tests', *Sequential analysis* **14**, 273–286.

Habermann, T., Weller, E., Morrison, V., Cassileth, P., Cohn, J., Dakil, S., Gascoyne, R., Woda, B., Fisher, R., Peterson, B. & Horning, S. (2003), *Rituxan-CHOP vs. CHOP with a 2nd randomization to maintenance Rituximab or Observation in patients = 60 years with diffuse large B-cell lymphoma (DLBCL)*, American Society of Hematology (ASH), San Diego, CA, USA.

Holland, P. (1986), 'Statistics and causal inference', *J. Amer. Statist. Assoc.* **81**, 945–960.

Horvitz, D. & Thompson, D. (1952), 'A generalization of sampling without replacement from a finite universe', *J. Amer. Statist. Assoc.* **47**, 663–685.

Joss, R., Alberto, P., Bleher, E., Ludwig, C., Siegenthaler, P., Martinelli, G., Sauter, C., Schatzmann, E. & Senn, H. (1994), 'Combined-modality treatment of small-cell lung cancer: randomized comparison of three induction chemotherapies followed by maintenance chemotherapy with or without radiotherapy to the chest. swiss group for clinical cancer research', *Annals of Oncology* **5**, 921–928.

Lai, T. L. (1973), 'Optimal stopping and sequential tests which minimize the maximum expected sample size', *Annals of statistics* **1**, 659–673.

Lehmacher, W. & Wassmer, G. (1999), 'Adaptive sample size calculations in group sequential trials', *Biometrics* **55**, 1286–1290.

Lin, D. & Wei, L. (1989), 'The robust inference for the cox proportional hazards model', *J. Amer. Statist. Assoc.* **84**, 1074–1078.

Lin, D. & Ying, Z. (1993), 'Cox regression with incomplete covariate measurements', *J. Amer. Statist. Assoc.* **88**, 1341–1349.

Lunceford, J., Davidian, M. & Tsiatis, A. A. (2002), 'Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials', *Biometrics* **58**, 48–57.

Posch, M. & Bauer, P. (1999), 'Adaptive two stage designs and the conditional error function', *Biometrical Journal* **41**, 689–696.

Prentice, R. (1986), 'A case-cohort design for epidemiologic cohort studies and disease prevention trials', *Biometrika* **73**, 1–11.

Proschan, M. A. & Hunsberger, S. A. (1995), 'Designed extension of studies based on conditional power', *Biometrics* **51**, 1315–1324.

Pugh, M., Robins, J., Lipsitz, S. & Harrington, D. (1993), Inference in the cox proportional hazards model with missing covariate data, Technical report, Harvard School of Public Health, Department of Biostatistics.

Scharfstein, D. O., Tsiatis, A. A. & Robins, J. M. (1997), 'Semiparametric efficiency and its implications on the design and analysis of group-sequential studies', *J. Amer. Statist. Assoc.* **92**, 1342–1350.

Thall, P., Sung, H.-G. & Estey, E. (2002), 'Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials', *J. Amer. Statist. Assoc.* **97**, 29–39.

Therneau, T. & Grambsch, P. (2000), *Modeling survival data: Extending the Cox model*, Springer, New York, NY, USA.

Tummarello, D., Mari, D., Graziano, F., Isidori, P., Cetto, G., Pasini, F., Santo, A. & Cellerino, R. (1994), 'A randomized, controlled phase iii study of cyclophosphamide, doxorubicin and vincristine with etposide (cav-e) or teniposide (cav-t), followed by recombinant a-interferon maintenance therapy or observation, in small cell lung carcinoma patients with complete responses', *Anticancer Research* **14**, 2221–2228.

Wahed, A. & Tsiatis, A. (2004), 'Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials', *Biometrics* **60**, 124–133.

Wang, C. & Chen, H. (2001), 'Augmented inverse probability weighted estimator for cox missing covariate regression', *Biometrics* **57**, 414–419.

Appendix A

Application of the backward induction algorithm for the objective function $G_{1,\phi} = E(N|\phi)$

Suppose we want to find the two-stage optimal test of $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ which minimizes the objective function $G_{1,\phi} = E(N|\phi)$ and attains error probabilities α and β at $\mu = 0$ and $\mu = \delta$ respectively. To achieve this, we consider a Bayes decision problem introduced in the Section 2, where we must choose between decisions $D_0 : \mu = 0$ and $D_\delta : \mu = \delta$ with associated loss function $C(D, \mu)$ taking values $C(D_0, \delta) = d_\delta$, $C(D_\delta, 0) = d_0$ and zero otherwise; we set cost function $c(\mu)$ to be one at $\mu = \phi$ and zero otherwise and we place a three point prior distribution $\pi(0) = \pi(\delta) = \pi(\phi) = \frac{1}{3}$. The optimal solution should minimize the expected total

loss

$$E(L)_{1,\phi} = \frac{1}{3}E_{\phi}(N) + \frac{1}{3}(d_0 Pr_0(D_\delta) + d_\delta Pr_\delta(D_0)),$$

and satisfy the condition (2.4) for the error probabilities.

As we explained above, the critical value a_N must satisfy condition (2.5). If we solve this equation, we find that for any given values of n_1 and n_2,

$$a_N = \frac{1}{\delta} log \frac{d_0}{d_\delta} + \frac{n_1 + n_2}{2}\delta \tag{A.1}$$

For a given value s_1 of the observed test statistic at the first look, the additional expected loss incurred by sampling $n_2 = n_2(s_1)$ observations at the second stage is

$$\begin{aligned} E\{L(s_1)\}_{1,\phi} = n_2 Pr(\mu = \phi | S_1 = s_1) + d_0 Pr(\mu = 0, S_1 + S_2 > a_N | S_1 = s_1) \\ + d_\delta Pr(\mu = \delta, S_1 + S_2 \leq a_N | S_1 = s_1) \end{aligned} \tag{A.2}$$

The probabilities involved in these calculations are

$$\begin{aligned} Pr(\mu = \phi | S_1 = s_1) &= \frac{\pi(\phi) Pr(S_1 = s_1 | \mu = \phi)}{\sum_{\mu' \in \{0,\delta,\phi\}} \pi(\mu') Pr(S_1 = s_1 | \mu = \mu')} \\ &= \frac{\phi(\frac{s_1 - n_1\phi}{\sqrt{n_1}})}{\phi(\frac{s_1}{\sqrt{n_1}}) + \phi(\frac{s_1 - n_1\delta}{\sqrt{n_1}}) + \phi(\frac{s_1 - n_1\phi}{\sqrt{n_1}})}, \end{aligned}$$

$$\begin{aligned} Pr(\mu = 0, S_1 + S_2 > a_N | S_1 = s_1) &= Pr(S_2 > a_N - s_1 | S_1 = s_1, \mu = 0) \\ &\times Pr(\mu = 0 | S_1 = s_1) = \Phi\left(\frac{-a_N + s_1}{\sqrt{n_2}}\right) Pr(\mu = 0 | S_1 = s_1), \end{aligned}$$

$$Pr(\mu = \delta, S_1 + S_2 \leq a_N | S_1 = s_1) = Pr(S_2 \leq a_N - s_1 | S_1 = s_1, \mu = \delta)$$
$$\times Pr(\mu = \delta | S_1 = s_1) = \Phi\left(\frac{a_N - s_1 - n_2\delta}{\sqrt{n_2}}\right) Pr(\mu = \delta | S_1 = s_1),$$

where $\phi(x)$ and $\Phi(x)$ are respectively density and cdf of the standard normal distribution.

Then the backward induction algorithm proceeds as follows. We choose some initial values for n_1 and costs d_0, d_δ. For each s_1, optimal additional sample size $n_2(s_1)$ can be found by minimizing (A.2). We search over the values of s_1 until we find standardized first stage boundaries l_1, u_1 such that

$$u_1 = \min\{s_1/\sqrt{n_1} \text{ such that } n_2(s_1) = 0 \text{ and } s_1 \geq a_{n_1}\},$$
$$l_1 = \max\{s_1/\sqrt{n_1} \text{ such that } n_2(s_1) = 0 \text{ and } s_1 \leq a_{n_1}\},$$

where a_{n_1} is defined by (2.5). The error probabilities α^* and β^* of the two-stage test with the first stage sample size n_1, additional sample size at the second stage defined by the function $n_2(s_1)$, and boundaries $l_1, u_1, u_{2N} = a_N/\sqrt{N}$ can be computed as

$$\alpha^* = \Phi(-u_1) + \int_{l_1}^{u_1} \Phi\left(\frac{-a_N + z_1\sqrt{n_1}}{\sqrt{n_2}}\right) \phi(z_1)\, dz_1,$$

$$\beta^* = \Phi(l_1 - \delta\sqrt{n_1}) + \int_{l_1}^{u_1} \Phi\left(\frac{a_N - z_1\sqrt{n_1} - n_2\delta}{\sqrt{n_2}}\right) \phi(z_1 - \delta\sqrt{n_1})\, dz_1.$$

We search over a two-dimensional range of values (d_0, d_δ) until the error probabilities of the test satisfy

$$\sqrt{\left(1-\frac{\alpha^*}{\alpha}\right)^2+\left(1-\frac{\beta^*}{\beta}\right)^2}\leq\epsilon,$$

for some small ϵ (we took $\epsilon = 0.005$). The distance in the left part of this inequality is a measure of the proximity of the test error probabilities to the target values α and β. After the appropriate costs (d_0, d_δ) are determined for a given n_1, we compute the objective function

$$G_{1,\phi} = E(N|\phi) = n_1 + \int_{l_1\sqrt{n_1}}^{u_1\sqrt{n_1}} n_2(s_1)\frac{1}{\sqrt{n_1}}\phi\left(\frac{s_1 - n_1\phi}{\sqrt{n_1}}\right) ds_1$$

By searching over the range of possible values for n_1, we find the one that minimizes the objective function $G_{1,\phi}$.

Appendix B

Proof of Theorem 3.2.1

The following proof uses the fact that under the assumptions of Section 2 there exists a neighborhood B of β_0 such that, for each $\tau < \infty$ and $k = 0, 1$

$$\sup_{t\in[0,\tau],\beta\in B} ||S_w^{(k)}(\beta,t) - s_w^{(k)}(\beta,t)|| \to 0$$

in probability as $n \to \infty$ and $\frac{s_w^{(1)}(\beta,\tau)}{s_w^{(0)}(\beta,\tau)}$ is bounded on $B \times [0,\tau]$.

First, we will show that $n^{-1/2}U_{wn}(\beta_0)$ can be expressed as a sum of n i.i.d. terms and a remainder term that converges in probability to zero, by applying the techniques used in the proof of theorem 2.1 of Lin and Wei (1989). Notice that the pseudo-score function can be expressed as

$$U_{wn}(\beta) = \sum_{i=1}^{n} \int_0^\infty w_i X_i dN_i(u) - \int_0^\infty \frac{S_w^{(1)}(u,\beta)}{S_w^{(0)}(u,\beta)} d\overline{N}_w(u),$$

where $\overline{N}_w(u) = \sum_{i=1}^n w_i N_i(u)$ and $S_w^{(k)}(u,\beta) = n^{-1}\sum_{i=1}^n w_i X_i^k Y_i(u) exp(X_i\beta), k = 0, 1$.

Taylor expansion of $U_{wn}(\hat{\beta}_n)$ around β_0 results in

$$n^{1/2}(\hat{\beta}_n - \beta_0) = \hat{A}^{-1}(\beta^*)n^{-1/2}U_{wn}(\beta_0),$$

where $\hat{A}(\beta) = -n^{-1}\frac{dU_{wn}(\beta)}{d\beta}$ and β^* lies between $\hat{\beta}_n$ and β_0. Assuming consistency of $\hat{\beta}_n$, consistency of $\hat{A}(\beta^*)$ for $A(\beta_0)$ can be shown using the techniques from the proofs of theorems 3.2 and 4.2 of Andersen & Gill (1982).

Further, $n^{-1/2}U_{wn}(\beta_0)$ can be expressed as

$$\begin{aligned} n^{-1/2}U_{wn}(\beta_0) &= n^{-1/2}\sum_{i=1}^{n}\int_0^\infty w_i X_i dN_i(u) - n^{1/2}\int_0^\infty \frac{S_w^{(1)}(u,\beta_0)}{S_w^{(0)}(u,\beta_0)} d\left\{E\left(\frac{\overline{N}_w(u)}{n}\right)\right\} \\ &\quad -n^{1/2}\int_0^\infty \frac{s_w^{(1)}(u,\beta_0)}{s_w^{(0)}(u,\beta_0)} d\left\{\frac{\overline{N}_w(u)}{n} - E\left(\frac{\overline{N}_w(u)}{n}\right)\right\} \qquad \text{(B.1)} \\ &\quad -n^{1/2}\int_0^\infty \left\{\frac{S_w^{(1)}(u,\beta_0)}{S_w^{(0)}(u,\beta_0)} - \frac{s_w^{(1)}(u,\beta_0)}{s_w^{(0)}(u,\beta_0)}\right\} d\left\{\frac{\overline{N}_w(u)}{n} - E\left(\frac{\overline{N}_w(u)}{n}\right)\right\}, \end{aligned}$$

where $s_w^{(k)}(u,\beta) = E\{S_w^{(k)}(u,\beta)\}$, $k = 0, 1$. Since $n^{-1}\{\overline{N}_w(u) - E(\overline{N}_w(u))\}$ converges in distribution to a mean zero Gaussian process, the last term in (B.1) is $o_p(1)$. The second term in (B.1) can be shown to be equal to

$$\begin{aligned} & n^{1/2}\int_0^\infty s_w^{(0)}(u,\beta_0)^{-1}\left\{S_w^{(1)}(u,\beta_0) - \frac{s_w^{(1)}(u,\beta_0)}{s_w^{(0)}(u,\beta_0)} \times\right. \\ \times \quad & \left.[S_w^{(0)}(u,\beta_0) - s_w^{(0)}(u,\beta_0)]\right\} d\left\{E\left(\frac{\overline{N}_w(u)}{n}\right)\right\} + o_p(1) \end{aligned}$$

Combining the terms, we observe that

$$n^{-1/2}U_{wn}(\beta_0) = n^{-1/2}\sum_{i=1}^{n} g_i(\beta_0) + o_p(1),$$

where $g_i(\beta)$, $i = 1, \ldots, n$, are i.i.d. terms,

$$\begin{aligned} g_i(\beta) &= \int_0^\infty \left\{ X_i - \frac{s_w^{(1)}(u,\beta)}{s_w^{(0)}(u,\beta)} \right\} w_i dN_i(u) \\ &- \int_0^\infty \frac{w_i Y_i(u) exp(X_i\beta)}{s_w^{(0)}(u,\beta)} \left\{ X_i - \frac{s_w^{(1)}(u,\beta)}{s_w^{(0)}(u,\beta)} \right\} d\left\{ E\left(\frac{\overline{N}_w(u)}{n} \right) \right\}. \end{aligned}$$

To show the asymptotic unbiasedness of $n^{-1/2} \sum_{i=1}^n g_i(\beta_0)$, we first observe that

$$\begin{aligned} &E\left[\int_0^\infty \left\{ X_i - \frac{s_w^{(1)}(u,\beta_0)}{s_w^{(0)}(u,\beta_0)} \right\} \frac{w_i Y_i(u) exp(X_i\beta)}{s_w^{(0)}(u,\beta_0)} d\left\{ E\left(\frac{\overline{N}_w(u)}{n} \right) \right\} \right] = \\ &\int_0^\infty E\left[\left\{ X_i - \frac{s_w^{(1)}(u,\beta_0)}{s_w^{(0)}(u,\beta_0)} \right\} \frac{w_i Y_i(u) exp(X_i\beta)}{s_w^{(0)}(u,\beta_0)} \right] d\left\{ E\left(\frac{\overline{N}_w(u)}{n} \right) \right\} = \\ &\int_0^\infty \left[\frac{s_w^{(1)}(u,\beta_0)}{s_w^{(0)}(u,\beta_0)} - \frac{s_w^{(1)}(u,\beta_0)}{s_w^{(0)}(u,\beta_0)} \frac{s_w^{(0)}(u,\beta_0)}{s_w^{(0)}(u,\beta_0)} \right] d\left\{ E\left(\frac{\overline{N}_w(u)}{n} \right) \right\} = 0. \end{aligned}$$

Let us consider a representation for the weighted counting processes $w_i N_i(u)$ and $w_i Y_i(u), u \in (0, \infty), i = 1, \ldots, n$, in terms of counterfactuals using (3.1). Since

$$\begin{aligned} I[U_i \le t^{resp}] N_i(u) &= \begin{cases} N_i(u), & u \le t^{resp}, \\ N_i(t^{resp}), & u > t^{resp}, \end{cases} \\ I[U_i > t^{resp}] N_i(u) &= \begin{cases} 0, & u \le t^{resp}, \\ I[C_i > t^{resp}]\{N_i(u) - N_i(t^{resp})\}, & u > t^{resp}, \end{cases} \end{aligned}$$

the weighted counting processes $w_i N_i(u)$ and $w_i Y_i(u)$ can be expressed as

$$
\begin{aligned}
w_i N_i(u) &= \begin{cases} N_i(u), & u \le t^{resp}, \\ N_i(t^{resp}) + \{1 - R_i + \frac{R_i(1-Z_i)}{1-\pi_Z}\} \times & \\ \quad \times I[C_i > t^{resp}]\{N_i(u) - N_i(t^{resp})\}, & u > t^{resp}, \end{cases} \\
&= \begin{cases} (1-X_i)N_{11i}(u) + X_i N_{21i}(u), & u \le t^{resp}, \\ (1-X_i)N_{11i}(t^{resp}) + X_i N_{21i}(t^{resp}) + I[C_i > t^{resp}] \times & \\ \times \{1 - R_i + \frac{R_i(1-Z_i)}{1-\pi_Z}\}\big[(1-X_i)N_{11i}(u) + X_i N_{21i}(u) - & u > t^{resp}, \\ -(1-X_i)N_{11i}(t^{resp}) - X_i N_{21i}(t^{resp})\big], & \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
w_i Y_i(u) &= \begin{cases} Y_i(u) - Y_i(t^{resp}) + \{1 - R_i + \frac{R_i(1-Z_i)}{1-\pi_Z}\} Y_i(t^{resp}), & u \le t^{resp}, \\ \{1 - R_i + \frac{R_i(1-Z_i)}{1-\pi_Z}\} Y_i(u), & u > t^{resp}, \end{cases} \\
&= \begin{cases} (1-X_i)Y_{11i}(u) + X_i Y_{21i}(u) + & \\ + \{\frac{R_i(1-Z_i)}{1-\pi_Z} - R_i\}\big\{(1-X_i)Y_{11i}(t^{resp}) + X_i Y_{21i}(t^{resp})\big\}, & u \le t^{resp}, \\ \{1 - R_i + \frac{R_i(1-Z_i)}{1-\pi_Z}\}\big\{(1-X_i)Y_{11i}(u) + X_i Y_{21i}(u)\big\}, & u > t^{resp}, \end{cases}
\end{aligned}
$$

Using a conditional expectation argument, we establish that

$$
E\Big[\{\frac{R_i(1-Z_i)}{1-\pi_Z} - R_i\}\big\{(1-X_i)Y_{11i}(u) + X_i Y_{21i}(u)\big\}\Big] =
$$

$$
E\bigg[E\Big\{\frac{R_i(1-Z_i)}{1-\pi_Z} - R_i | X_i\Big\} E\Big\{(1-X_i)Y_{11i}(u) + X_i Y_{21i}(u) | X_i, R_i\Big\}\bigg] = 0
$$

The last assertion is true because

$$
E\Big\{\frac{R_i(1-Z_i)}{1-\pi_Z} - R_i | X_i, R_i = 0\Big\} = E\Big\{\frac{R_i(1-Z_i)}{1-\pi_Z} - R_i | X_i, R_i = 1\Big\} = 0
$$

Therefore,

$$
E\{w_i Y_i(u) | X_i\} = E\Big\{(1-X_i)Y_{11i}(u) + X_i Y_{21i}(u) | X_i\Big\},
$$

which results in $s_w^{(k)}(u,\beta) = s^{(k)}(u,\beta),\ k = 0, 1$, where

$$s^{(k)}(u,\beta) = E\{n^{-1}\sum_{i=1}^{n} X_i^k \widetilde{Y}_i(u) exp(X_i\beta)\},$$

and $\widetilde{Y}_i(u)$ corresponds to the survival time we would observe if the patients were only randomized to one of the induction therapies in a SRD.

Using repeatedly a conditional expectation argument and observing that $I[C_i > t^{resp}]dN_{jki}(u) = dN_{jki}(u), j, k \in \{1, 2\}, u \geq t^{resp}$, it is easy to show that

$$E\big\{w_i dN_i(u)|X_i\big\} = E\big\{(1 - X_i)dN_{11i}(u) + X_i dN_{21i}(u)|X_i\big\}$$

We have shown that

$$E\bigg[\int_0^\infty \bigg\{X_i - \frac{s_w^{(1)}(u,\beta)}{s_w^{(0)}(u,\beta)}\bigg\} w_i dN_i(u)\bigg] = E\bigg[\int_0^\infty \bigg\{X_i - \frac{s^{(1)}(u,\beta)}{s^{(0)}(u,\beta)}\bigg\} d\widetilde{N}_i(u)\bigg]$$

where $S(\beta) = \int_0^\infty \bigg\{X_i - \frac{s^{(1)}(u,\beta)}{s^{(0)}(u,\beta)}\bigg\} d\widetilde{N}_i(u)$ is the usual score vector that is routinely used for estimating β in a SRD, with the property $E\big\{S(\beta_0)\big\} = 0$. Therefore, $E\big\{n^{-1/2}\sum_{i=1}^n g_i(\beta_0)\big\} = 0$ and $\hat{\beta}_n$ is an unbiased estimate for β_0. This concludes the proof of the Theorem 3.2.1.