2014

Practical Methods for Equivalence and Non-Inferiority Studies with Survival Response

Elvis Englebert Martinez

FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

PRACTICAL METHODS FOR EQUIVALENCE AND NON-INFERIORITY STUDIES WITH SURVIVAL RESPONSE

By

ELVIS ENGLEBERT MARTINEZ

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Fall Semester, 2014

Elvis Englebert Martinez defended this dissertation on November 07, 2014.
The members of the supervisory committee were:

Debajyoti Sinha
Professor Directing Dissertation

Cathy Levenson
University Representative

Eric Chicken
Committee Member

Stuart Lipsitz
Committee Member

Dan McGee
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

For my older brothers Roberto and Felix and my younger sister Jones. Thank you for always believing in me.

ACKNOWLEDGMENTS

I would like to thank my lead adviser, Dr. Debajyoti Sinha, for always encouraging me, never letting me settle, and for always being there when I needed his advice. In addition, I would like to thank the faculty and staff at Florida State Departments of Statistics, Mathematics, and College of Medicine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Determining the equivalence or non-inferiority of a new drug (test drug) with a existing treatment (reference drug) is an important topic of statistical interest. Wellek (1993) pioneered the way for log-rank based equivalence and non-inferiority testing by formulating a testing procedure using proportional hazards model (PHM) of Cox (1972). In many equivalence and non-inferiority trials, two hazards functions may converge to one rather than being proportional for all time-points. In this case, the proportional odds survival model (POSM) of Bennett (1983) will be more sufficient than a Cox's PHM assumption. We show in both cases, when the wrong modeling assumption is made and Cox's PH assumption is violated, the popular procedure of Wellek (1993) has an inflated type I error. On the contrary, our proposed POSM based equivalence and non-inferiority tests maintains the practitioners desired 5% level of significance regardless of the underlying modeling assumption (e.g. Cox, 1972; Wellek, 1993). Furthermore for non-inferiority trials, we introduce a method to determine the optimal sample size required when a desired power and type I error is specified and the data follows the POSM of Bennett (1983). For both of the above trials, we present simulation studies showing the finite approximation of powers and type I error rates, when the underlying modeling assumption are correctly specified and when the assumptions are misspecified.

CHAPTER 1

BACKGROUND

1.1 Survival Data

Time-to-event data (i.e. survival data) occurs very frequently in nature and appears in many fields including oncology (cancer trials), automated systems, and any process that requires waiting for an outcome (event) to occur. Some examples of survival data are, time to a drug relapse, parole violation, system breakdown, cancer recurrence, and death. A vital part of survival analysis, is the ability to calculate the probability of an event materializing at some given time-point. In addition, some additional quantities survival analysis aims to measure, are mean-lifetime of a process, and/or testing for statistical differences in distributions (survival responses) of two competing drugs administered to there respective reference and control groups.

When dealing with survival data a major drawback that may arise, especially when human subjects are involved, is the presence of censored observations. Censoring is very common with survival analysis with human participants. There are three typical censoring schemes that can transpire. Interval censoring, this is the case when we do not know the exact time the event of interest transpired, but rather a time interval the event of interest occurred in. For the case of left-censoring, all that is known is that the event of interest happened at some time before the outset of the trial. Finally right-censored data, this is the most common form of censoring, occurs when the only knowledge is a patient did not have the event of interest by some time-point. Furthermore when dealing with presence of right-censored observations, there are two additional types of right-censoring. Type I right-

censoring appears, when the event is consider part of the study if it occurs before a fixed time-point, then all other observations are censored. Rather, with Type II right-censoring, a clinical trial proceeds until a pregiven number of patients experience the event, the at the conclusion of the trial all other observation are deemed right-censored.

In addition to censoring of survival data, another problematic issue is truncation. When truncation occurs, set restrictions are placed on patients either prior to the analysis or during the onset of the trial. Only appropriate subjects that satisfy predetermined criteria(s) are consider acceptable for the study. In the case of left truncation, patients must first survive (i.e. not have the event) up to a predetermined time-point before their survival time is considered part of the study. Whereas, right truncation restricts participants in the study to those persons who attain the event of interest before a particular time.

Define $T \geq 0$ as a random continuous variable that denotes the survival time (i.e. duration time until the event). T has a probability density function $f(t)$ and associated cumulative distribution function $F(t) = P(T \leq t)$. Since we are concerned with the probability of surviving (i.e. not having the event) up to a time-point t, we introduce the survival function $S(t) = P(T > t) = 1 - F(t)$ and its corresponding hazard function $\lambda(t) = \lim_{h \to 0} P[t \leq T < t+h | T \geq t]/h = f(t)/S(t)$. The hazard function (rate) $\lambda(t)$ is a pivotal measure when dealing with survival analysis. This quantity may be interpreted as the instantaneous rate of having the event at a given time-point. Hazard rates may be increasing, decreasing or both, with $\lambda(t) \geq 0$ being the only binding condition. When a process ages naturally or worsens over time, an increasing hazard function should appear. While, decreasing hazard functions, not as common as increasing hazard functions, corresponds to instances in which the rate of the event of interest materializes is initially high but then decreases rapidly after-wards. The life-span of certain electronics and digital devices may follow decreasing hazard rates. A hazard rate is regarded as 'hump-shaped' when it increases sharply at first and then declines as times goes. The hazard rates of many surgical procedures follow this trend. Where after

surgery, the risk of infection or death may be high leading to an increase in hazard rates with time, but eventually this risk abates. In addition, we define the cumulative hazard function as $\Lambda(t) = \int_0^t \lambda(u)du$. With the introduction of the cumulative hazard function $\Lambda(t)$, we can now define the survival function $S(t)$ as $S(t) = \exp\{-\Lambda(t)\}$ where $S(0) = 1$ and the $\lim_{t\to\infty} S(t) = 0$. Similarly, the hazard function can also be represented in terms of the survival function $S(t)$, where $\lambda(t) = \frac{d}{dt}(-\ln\{S(t)\})$. Even though for most aspects of a survival analysis the hazard and survival functions can be used interchangeably, many practitioners and statisticians are more at ease interpreting results using the hazard rate. Thus, many statistical tests are conducted using the hazard rate or the hazard ratio (i.e. the ratio of two hazard rates).

A non-parametric approach to modeling survival data, consist of the product limit estimator of Kaplan EL (1958). Using the product limit estimator of Kaplan EL (1958) the survival function (curve) of a given survival data set can now be attained. The Kaplan EL (1958) estimator is explained as follows, given n independent time points (measurements) $t_{(i)},\ i = 1, ..., n$, where $t_{(i)}$ corresponds to event time of a subject in the study and censoring indicator $c_{(i)},\ i = 1, ..., n$, where $c_{(i)} = 1$ when the event occurs and $c_{(i)} = 0$ if the survival time is censored. Let C represent the sum total of actual events, where $C \leq n$ and $C = \sum_{i=1}^{n} c_{(i)}$. Additionally, $t_{(k)},\ k = 1, ..., c$, express the ascending ordered event times. Further, let the number of subjects at danger of experience an event at $t_{(i)}$ as $Y_{(i)}$. Thus $Y_{(i)}$ is the count of subjects yet to experience the event of interest by time $t_{(i)}$. Now let $d_{(i)}$ equal the number of subjects who experience the event of interest at $t_{(i)}$. We can now formulate the conditional probability that an individual will have an event at $t_{(i)}$ given he/she has survived just before $t_{(i)}$ as $\frac{d_{(i)}}{Y_{(i)}}$. Thus for a given survival data set, we can now estimate $S(t)$

by its Kaplan-Meier estimates given by

$$\tilde{S}(t) = \left\{ \begin{array}{ll} 1 & , t < t_{(1)} \\ \prod\limits_{t_{(i)} \leq t} 1 - \dfrac{\hat{d}_{(i)}}{Y_{(i)}} & , t \geq t_{(1)} \end{array} \right\}, \tag{1.1}$$

for $i = 1, ..., n$.

Figure 1.1 shows an example of three Kaplan-Meier curves estimated using (1.1). Figure 1.1 was created using SAS 9.4 and the SAS data set *sashelp.BMT*.

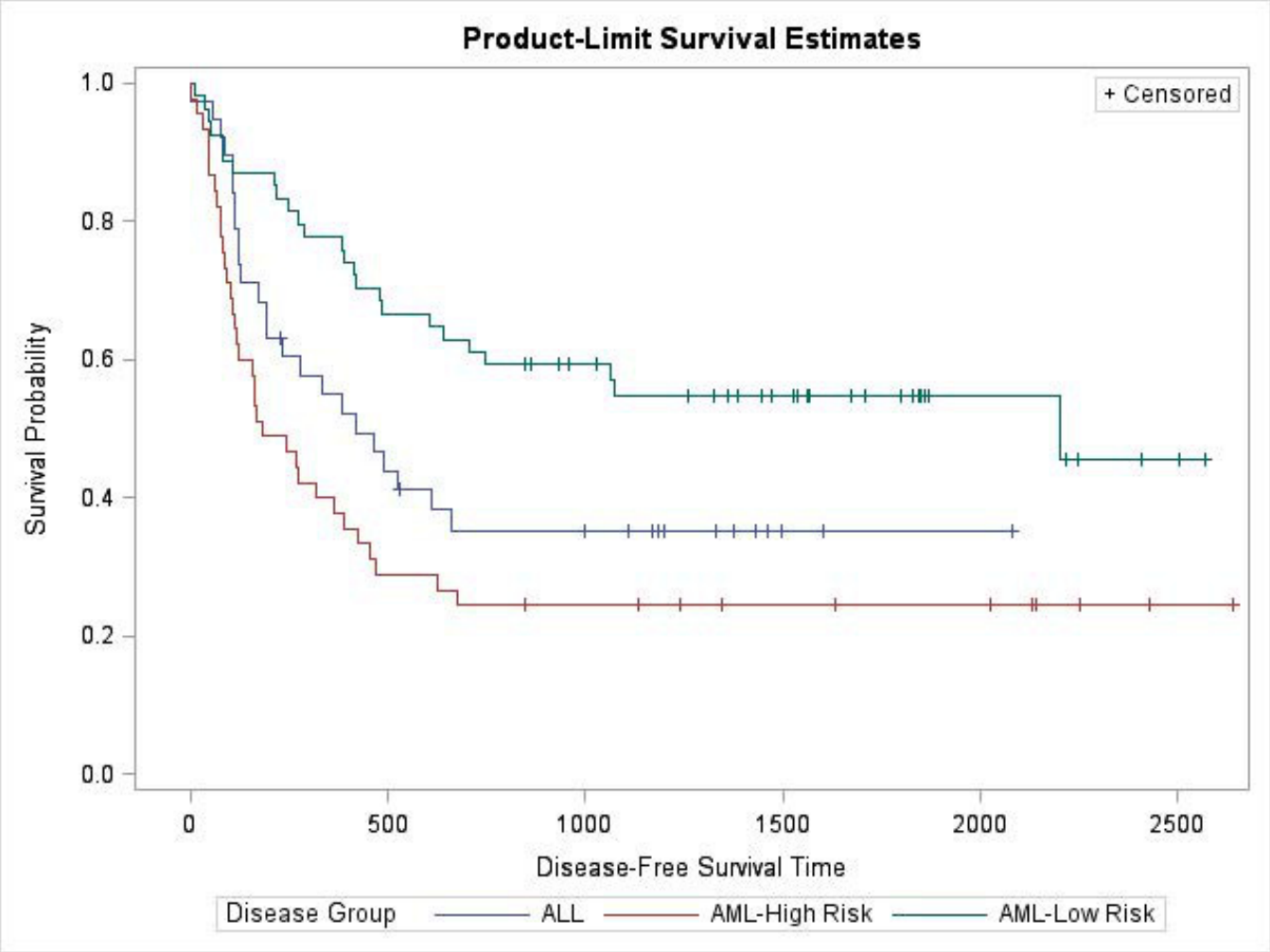

Figure 1.1: Example of Survival curves estimated using (1.1)

1.2 Survival Models

There are four frequently used models for analyzing (modeling) the covariate (treatment) effect on a patients survival:

• Parametric Families

• Non-Parametric

• Semi-Parametric: Proportional Hazards Model

• Semi-Parametric: Proportional Odds Survival Model.

This paper will focus on the latter two, Proportional Hazards Model (PHM) and Proportional Odds Survival Model (POSM) with one covariate of treatment arms.

1.2.1 Cox's Proportional Hazards

Currently, the most common modeling assumption when dealing with survival functions is Cox's PH model (1972)

$$h_1(t)/h_0(t) = exp(\eta), \tag{1.2}$$

where $h_0(t)$ is the baseline hazard and η is the regression parameter of Cox's model relating to the difference in log hazards between the two treatment arms $S_1(t)$ and $S_0(t)$. In this case, the statistician assumes the treatment effect is proportional, i.e., one hazards function dominates the other by the same proportion, for all time points t.

1.2.2 Bennett's Proportional Odds

We now introduce Bennett's POSM (1983),

$$\frac{1 - S_1(t)}{S_1(t)} = \theta \left[\frac{1 - S_0(t)}{S_0(t)} \right], \tag{1.3}$$

for all $t > 0$. Here θ is the time-constant odds ratio between new and standard treatment. An advantage of this model selection over (1.2), is that it allows the hazards of two survival functions $S_1(t)$ and $S_0(t)$ to converge to 1 as the effects of the treatments diminishes over time.

1.3 Clinical Trials of Two Survival Functions

1.3.1 Superiority

When conducting a clinical trial with the desire to test whether a treatment, $S_1(t)$, is superior to a placebo or an existing treatment, $S_0(t)$, the clinical scientific hypothesis is

$$H_0 : S_1(t) = S_0(t) \text{ for all t vs } H_a : S_1(t) - S_0(t) > \delta \text{ for all t.} \tag{1.4}$$

In this case, the null hypothesis, H_0, means that there is no significant difference in the two survival functions $\{S_1(t)\}$ and $\{S_0(t)\}$. Only if this new treatment is greater than the placebo or existing treatment should we reject the H_0 and conclude $S_1(t)$ is a superior treatment. In addition, the clinicians and experimenters have some knowledge through a *Phase I* trial that $S_1(t)$ is effective, before developing a statistical hypothesis of interest. The type I error of the hypothesis of (1.4) is the probability of declaring a new treatment arm superior when it is not. For the case of a new treatment vs placebo superiority trial, a type I error will still lead to some treatment, granted it may not be more effective that a placebo. However, in the case of testing for superiority of new treatment vs an active control and death as a survivor outcome, committing a type I error may result in a shortened life-span. In this case, the H_0 does not depend on the modeling assumption. If the modeling assumption is violated the type I error will still be below the nominal rate. However, the H_a does depend on the model assumption being true. For the case of the H_a, a wrong modeling assumption

may lead to an under-powered test.

1.3.2 Equivalence

The aim of a clinical equivalence trial is to test the clinical scientific hypothesis:

$$H_0 : |S_1(t) - S_0(t)| \geq \delta \text{ for any t vs } H_a : |S_1(t) - S_0(t)| < \delta \text{ for all t}, \tag{1.5}$$

where δ is now our cutoff value for equivalence between two survival functions $S_1(t)$ and $S_0(t)$. The construction of our H_0 and H_a test inequalities in (1.5) are in reverse of the customary view of testing (1.4). This is to ensure that the patient's health is the main concern when considering type I error. With the hypothesis setting of (1.5), the type I error is the probability of declaring two treatment arms equivalent when in fact they are not.

1.3.3 Non-Inferiority

Non-inferiority trials differ from equivalence trial in the sense you are assuming the new treatment arm, $S_1(t)$, is less effective than the current treatment arm $S_0(t)$. Thus the clinical scientific hypothesis of non-inferiority is stated as:

$$H_0 : S_0(t) - S_1(t) \geq \delta \text{ for some t vs } H_a : S_0(t) - S_1(t) < \delta \text{ for all t}, \tag{1.6}$$

where δ is our margin for non-inferiority between two survival functions $S_1(t)$ and $S_0(t)$. Similar to the equivalence trial of two survival curves, the H_0 and H_a test inequalities in (1.6) are again opposite of the superiority hypothesis of (1.4). This guarantees that the non-inferiority trial will account for a patient's safety first with respect to type I error rate. In this sense, the type I error associated with the hypothesis of (1.6) is the probability of declaring the new treatment arm non-inferior to the existing treatment arm when in fact the new

treatment arm was statistically inferior. For both the equivalence and non-inferiority studies the H_0 does depend on the modeling assumption. For both of these cases, a misspecified modeling assumption may result in a inflated type I error.

CHAPTER 2

TESTS FOR EQUIVALENCE OF TWO SURVIVAL FUNCTIONS

2.1 Introduction

For the equivalence trial with survivor outcomes from two treatment groups, the most popular testing procedure is the extension (e.g. Wellek, 1993) of log-rank based test under proportional hazards model (PHM). We show that the actual type I error rate for the popular procedure of Wellek (1993) is higher than the intended nominal rate when survival responses from two treatment arms satisfy the proportional odds survival model (POSM). When the true model is POSM, we show that the hypothesis of equivalence of two survival functions can be formulated as a statistical hypothesis involving only the survival odds-ratio parameter. We further show that our new equivalence test, formulation, and related procedures are applicable even in the presence of additional covariates beyond treatment arms, and the associated equivalence test procedures have correct type I error rates under the PHM as well as the POSM. These results show that use of our test will be a safer statistical practice for equivalence trials of survival responses than the commonly used log-rank based tests.

Clinical trials for determining equivalence of a new treatment with a standard treatment of proven efficacy have become increasingly commonplace in recent years. With growing financial and ethical pressures (e.g. hel, 2000) to switch from an expensive and invasive standard treatment/procedure to a cheaper and less-invasive treatment, we can expect an increasingly higher number of equivalence trials to be conducted in future years. Our pa-

per deals with the important concern about the validity of the conclusions from equivalence studies when the key modeling assumptions of the test is violated. Statistical methods used for equivalence trial for survival response are often based on methods of Wellek (1993) using the proportional hazards model (PHM) of Cox (1972). The reason behind the popularity of this method for equivalence trial is given below. One main challenge for developing a convenient hypothesis testing method for an equivalence trial is the formulation of the statistical hypothesis using only the parameter of the treatment effect. For a two-arm (placebo versus treatment) superiority trial under any semi-parametric model (e.g. Cox, 1972), it is straightforward to make a statistical/mathematical formulation of the alternative hypothesis H_a (clinically important difference) of scientific interest. Any difference in the regression parameter η of the treatment arm implies some difference $S_1(t) \neq S_0(t)$ in survival curves S_1 and S_0 from two different arms at least at one time point t, and the converse is also true. For example, when two treatment arms follow proportional hazards model (PHM) of Cox (1972) with hazard ratio η, the alternative hypothesis H_a: $S_1(t) \neq S_0(t)$ for some t, implies H_a^*: $\eta \neq 1$ and vice versa. However, for an equivalence trial, when the alternative H_a is $|S_1(t) - S_0(t)|$ being within the prespecified range of equivalence for every time-point t (to be explained later), it is not straightforward to express this H_a as a statistical hypothesis H_a^* involving only the regression parameter η (which is free of time t). For example, in Cox's PHM, it is not obvious that $|S_1(t) - S_0(t)|$ less than a small known constant for all t does imply that η is within a known interval. Wellek (1993) paved the way for a convenient log-rank based equivalence test by deriving this result for the PHM, and only for the case of no covariates beyond treatment arms. Our result, an extension of the result of Wellek (1993) to the case of POSM, allows us to formulate an equivalence test for the POSM based on a rejection region which only involves the estimate and the corresponding standard error of the treatment effect parameter. Please see Wellek (2010) (Section 6.7) for a thorough review of the justifications behind formulating statistical hypothesis of equivalence based on

the treatment effect parameter.

Due to the results of Wellek (1993), the existing literature on equivalence trials for survival responses is dominated by the log-rank test based on the assumption of a PHM for the two treatment arms, without any consideration for alternative semi-parametric models and the presence of other covariates. Non parametric procedures (e.g. Com-Nougue *et al.*, 1993) and others often require much higher sample sizes than tests based on semi-parametric models. In practice, often the hazard functions of two treatment arms are not proportional over time and there may be other covariates in addition to treatment arms. We show that a log-rank based test of equivalence has a higher than intended type I error rate when treatment arms do not follow the PHM. This points to the practical need to consider new equivalence tests based on other semi-parametric models. For example, the ratio of two hazards may converge towards one over time when the initial benefit of one treatment arm over the other treatment arm diminishes over time. In this situation, the proportional odds survival model (POSM) of Bennett (1983) will be more appropriate than a PHM. In this paper, we also show that a POSM based equivalence test has correct type I error even when the true model is either POSM or PHM. This shows that the POSM based equivalence test is a safer option in practice compared to the log-rank based test, especially when the underlying modeling assumption is under suspicion.

We place high emphasis on controlling the type I error rate for an equivalence trial because, unlike a superiority trial, an effective standard treatment already exists for an equivalence trial. Wrongly accepting the alternative H_a of equivalence can potentially replace an effective standard treatment with an ineffective treatment in the market. Whereas, even if we wrongly accept the null of non equivalence, i.e, do not accept the new treatment as equivalent, we will still have the effective standard treatment available in the market. In this case, wrongly rejecting the null is a more serious mistake than wrongly accepting the null. However, we first deal with a major impediment for developing an equivalence test for a

POSM. The clinicians and other non statisticians have understandable difficulty in defining the clinically important difference between the two treatment arms in terms of the ratio of two survival odds. On the contrary, most clinical experts and researchers are comparatively more at ease to express the clinical equivalence of two treatment arms in terms of a clinically important difference between two survival functions. The development of equivalence trial methodology for POSM depends on whether the alternative hypothesis of the equivalence of two survival curves (or two hazard curves) can be properly expressed as an alternative hypothesis in the regression parameter of the POSM.

In Section 2.2, we first derive the formulation of the alternative statistical hypothesis, H_a^*, that only uses the odds ratio of the POSM, such that H_a^* also corresponds to the scientific (clinical) hypothesis related to the "equivalence" of the survival functions of two treatment arms. In Section 2.3, we describe the statistical methods including rejection regions for two-sample and one-sample equivalence studies under POSM. In Section 2.4, we show that even in the presence of additional covariates, testing equivalence of the survival functions for two treatment arms is the same as statistically testing the survival odds ratio parameter to be within a small interval. This result allows us to develop the statistical test of equivalence of two treatments under POSM, even in the presence of additional covariates. In Section 2.5, we study the relationship between sample size and intended type I error rates with tests based on Cox's model and our new POSM based tests. Our theoretical and simulation studies show that when the POSM assumption is true for the trial in question, log-rank based equivalence test of Wellek (1993) tends to reject the correct null hypothesis more often than the desired level of significance. On the contrary, our POSM based equivalence tests achieve desired type I error rates and power when the true model is either POSM or Cox's model.

2.2 Formulation of Hypothesis Under POSM

For the time being, we consider no covariate other than treatment arm. We later extend our methods to include other covariates. The POSM of Bennett (1983) assumes

$$\frac{1-S_1(t)}{S_1(t)} = \theta \left[\frac{1-S_0(t)}{S_0(t)}\right], \tag{2.1}$$

for all time points $t > 0$, where θ is the time-constant survival odds ratio between new treatment and standard treatment. With corresponding survival functions $S_1(t)$ and $S_0(t)$ respectively. For example, one may consider two treatments are clinically equivalent if $|S_1(t) - S_0(t)|$, the difference between two survival functions, is smaller than a predetermined equivalence level δ over time. Thus two treatment arms are equivalent only when $|S_1(t) - S_0(t)| < \delta$ for all t. Here, the additional quantity $\delta > 0$ indicates the maximum clinical difference allowed between the standard therapy and a therapeutically equivalent experimental therapy. The value of δ is usually determined by clinical experts and regulatory agencies involved in determining the practical definition of the equivalence of two treatments under consideration. However, in order to implement a statistical test for the equivalence of two treatments under POSM of (2.1), the alternative statistical hypothesis H_a^* must be based on a range (interval) of θ, where the interval depends on the practical (clinical) meaning of the equivalence of two survival curves $S_1(t)$ and $S_0(t)$. Furthermore, it is difficult for clinicians and non statisticians to express the therapeutic equivalence in terms of a prespecified range of θ, because θ is a ratio of odds, unlike difference in probabilities of any observable event under two treatment arms. To facilitate the formulation of a statistical hypothesis testing procedure for evaluating the clinical (scientific) alternative hypothesis H_a: $|S_1(t) - S_0(t)| < \delta$ for all t, under POSM of (2.1) we develop the following theorem.

Theorem 1 *Under POSM of (2.1) with continuous $S_0(t)$, testing H_a: $|S_1(t) - S_0(t)|$ ¡ δ for all t¿0, is the same as testing H_a^*: $(1+\epsilon)^{-1} < \theta < 1+\epsilon$, where $\epsilon = (4\delta)/(1-\delta)^2$ is a*

known function of δ.

Theorem 1 (proof in the Appendix (A.1)) shows that under the POSM of (2.1), if the clinicians and practitioners can specify the maximum allowable difference δ between two survival functions $S_1(t)$ and $S_0(t)$ of two equivalent treatment arms, we can derive the corresponding statistical alternative hypothesis H_a^* based on the time-constant survival odds ratio θ. This H_a^* can now be tested using statistical hypothesis testing tools.

Many authors (e.g. Rothmann *et al.*, 2012) advocated testing the equivalence of two treatments using the hazard ratio, because the hazard ratio of Cox's model does not depend on the baseline population. The hazard ratio is also the popular parameter for comparing treatments in efficacy trials (at least in the field of oncology). One may specify the alternative (scientific) hypothesis H_a of equivalence of the two treatment arms via H_a: $|h_1(t)/h_0(t)| < \rho$ for all time points $t > 0$, where $h_1(t)$ and $h_0(t)$ are hazard functions for new and standard treatments respectively. Similar to δ for Theorem 1, the maximum allowable hazard ratio $\rho > 1$ for two clinically equivalent treatments is determined from a clinical perspective. To expedite the equivalence trial under POSM of (2.1) for H_a based on hazards ratio, we have the following theorem (proof is again in the Appendix (A.2)).

Theorem 2 *Under the POSM assumption of (2.1), the alternative hypothesis of interest* H_a: $|h_1(t)/h_0(t)| < \rho$ *for all* t, *is the same as testing* H_a^*: $\rho^{-1} < \theta < \rho$.

We note that the H_a^* here is identical to the H_a^* of Theorem 1 with $(1+\epsilon)$ replaced by ρ. This indicates that for POSM of (2.1), the formulation of the statistical hypothesis H_a^* is the same while testing the equivalence of two treatment arms based on either the maximum hazards ratio over time or the maximum difference of the survival functions over time. Both of these alternative hypothesis can be reduced to testing the statistical hypothesis H_a^* involving only time constant parameter θ in (2.1). In the next section, we present the statistical tests and corresponding critical regions for this hypothesis H_a^* for two cases – the two-sample

case when the baseline survival function $S_0(t)$ of standard treatment is unknown and the one-sample case when $S_0(t)$ is known from historical data.

2.3 Implementation of Equivalence Tests

First we discuss the statistical tests for the equivalence of two treatment arms under the POSM of (2.1) when n patients are randomized to two treatment arms with $z_i = 1$ when patient i receives the new treatment, and $z_i = 0$ when she/he receives the standard treatment. We denote the observed right-censored data as $(\underset{\sim}{\boldsymbol{Y}}, \underset{\sim}{\boldsymbol{d}}, \underset{\sim}{\boldsymbol{z}})$, where $\underset{\sim}{\boldsymbol{Y}} = (Y_1, \ldots, Y_n)$ and observed censoring indicators $\underset{\sim}{\boldsymbol{d}} = (d_1, \ldots, d_n)$ where Y_i is the observed survival when $d_i = 1$ and Y_i is the right-censoring time when $d_i = 0$. Survival time T_i is at risk of non informative random right censoring. In practice, the decision about therapeutic equivalence of two treatment arms will be based on testing H_0: $|S_1(t) - S_0(t)| \geq \delta$ for some time-point t, versus H_a: $|S_1(t) - S_0(t)| < \delta$ for all $t > 0$. From Theorem 1 and (2.1), we know that testing this hypothesis is equivalent to testing

$$H_0^*\text{: } |\beta| \geq \log(1+\epsilon) \text{ versus } H_a^*\text{: } |\beta| < \log(1+\epsilon), \tag{2.2}$$

where $\epsilon = (4\delta)/(1-\delta)^2$ and $\beta = \log(\theta)$ in (2.1). Due to the formulation of this statistical equivalence test based solely on parameter β of POSM, we can use the test statistic of a superiority test under POSM as the test statistic for testing (2.2). However, the new test for (2.2) has a different rejection region.

One option is to use the semi-parametric Maximum likelihood estimator (SPMLE) $(\hat{\beta}, \hat{B})$ of Murphy SA (1997) obtained via maximizing the following semi-parametric likelihood

$$L(\beta, B_0 | \underset{\sim}{\boldsymbol{Y}}, \underset{\sim}{\boldsymbol{d}}) \propto \prod_{i=1}^{n} \left(\frac{exp(z_i\beta)}{B_0(Y_i) + exp(z_i\beta)} \right) \left(\frac{\Delta B_0(Y_i)}{B_0(Y_i-) + exp(z_i\beta)} \right)^{d_i},$$

where the baseline odds function $B_0(t) = S_0(t)/\{1 - S_0(t)\}$ is a non decreasing, right continuous function and with jumps $\Delta B_0(t) = B_0(t) - B_0(t-)$ at the observed failure times. The rejection region of the large sample based asymptotically most powerful test for (2.2) is given as:

$$\left\{ |\hat{\beta}| \sqrt{\hat{I}_\beta} < C_\alpha \left(\sqrt{\hat{I}_\beta} \log(1+\epsilon) \right) \right\}, \tag{2.3}$$

where the $C^2_\alpha(\psi)$ is the αth quantile of a χ^2 distribution with $df = 1$ and non centrality parameter ψ^2. Numerical differentiation of the profile likelihood $\text{prlik}_n = \log\{L(\beta, \hat{B}_0 | \boldsymbol{Y}, \boldsymbol{\beta})\}$ is used to obtain

$$\hat{I}_\beta \approx -\frac{1}{nh^2}\{\text{prlik}_n(\hat{\beta}+h) - 2\,\text{prlik}_n(\hat{\beta}) + \text{prlik}_n(\hat{\beta}-h)\},$$

for some small enough h (see Murphy SA, 1997).

An alternative semi-parametric approach for testing (2.2) is to use the test statistic of Chen *et al.* (2002) based on the estimator $\tilde{\beta}$ obtained via iteratively solving a set of estimating equations. The iterative steps are outlined in Appendix (A.3). Using the test statistic of Chen *et al.* (2002) we can similarly derive the rejection region for testing (2.2) as

$$\left\{ \frac{|\tilde{\beta}|}{\nu(\tilde{\beta})} < C_\alpha \left(\frac{\log(1+\epsilon)}{\nu(\tilde{\beta})} \right) \right\}, \tag{2.4}$$

where $C^2_\alpha(\psi)$ is the αth quantile of a χ^2 with df=1 and non centrality parameter ψ^2. This approach avoids the high dimensional numerical maximization and the estimator $\nu^2(\tilde{\beta})$ of the asymptotic variance of $\tilde{\beta}$ has a closed-form expression. We omit the closed form expression of $\nu(\tilde{\beta})$ (see Chen *et al.*, 2002) for the sake of brevity. Although the estimator $\tilde{\beta}$ is not the most efficient estimator, the efficiency loss is typically small.

In many equivalence trials, particularly in oncology, for all practical purposes, we may know the baseline survival $S_0(t)$ of the standard treatment. In particular, there often exists

a considerable amount of historical data on the survival function $S_0(t)$ of the standard treatment because its efficacy has been already studied. In this situation, every patient with observed survival data Y_i and censoring indicator d_i for $i = 1, \ldots, n$ receives the new treatment. We note that the logic and the result of Theorem 1 still apply here and the hypothesis of equivalence of two treatment is again reduced to the hypothesis of (2.2). Since $S_0(t)$ is known, we can find the MLE ($\hat{\beta}$) of β by solving the score equation:

$$n - \sum_{i=1}^{n}(1 + d_i)\frac{exp(\beta)}{exp\{-B(y_i)\} + exp(\beta)} = 0,$$

where $B(t) = S_0(t)/\{1 - S_0(t)\}$ is known. Using the usual asymptotic theory, the large-sample rejection region is $\frac{|\hat{\beta}|}{\nu(\hat{\beta})} < C_\alpha\left(\frac{\log(1+\epsilon)}{\nu(\hat{\beta})}\right)$, where $C_\alpha(\psi)$ is the same as in (2.4) and the estimated variance $\nu^2(\hat{\beta})$ has the closed form expression

$$\nu^2(\hat{\beta}) = \sum_{i=1}^{n}(1 + d_i)\frac{B_0(y_i)exp(\hat{\beta})}{n\{B_0(y_i) + exp(\hat{\beta})\}^2}.$$

The computer codes for computing the test statistic of (2.3) and corresponding critical region of (2.4) are available from the authors upon request. The authors also have codes for the competing test statistic and critical region of Wellek (1993).

2.4 Extension to Include Other Covariates

We now extend our previously described procedure of equivalence tests to accommodate even other covariates $\boldsymbol{x}_i$, in addition to treatment arm indicator z_i. Even though it is very much conceivable to have additional covariates in practice, we have not yet come across any previous research on equivalence tests to accommodate additional covariates. We assume that the underlying model with additional covariate $\boldsymbol{x}$ is a natural extension of the POSM

of 2.1 with

$$\frac{1-S_1(t|\boldsymbol{x})}{S_1(t|\boldsymbol{x})} = \theta\left[\frac{1-S_0(t|\boldsymbol{x})}{S_0(t|\boldsymbol{x})}\right] = \theta e^{\boldsymbol{\gamma x}}\left[\frac{1-S_0(t)}{S_0(t)}\right], \tag{2.5}$$

where $\boldsymbol{\gamma}$ is the regression parameter of $\boldsymbol{x}$ and θ is again the treatment effect of interest. For this situation, the relevant clinical hypothesis of interest is H_a: $|S_1(t|\boldsymbol{x}) - S_0(t|\boldsymbol{x})| < \delta$ for all covariates $\boldsymbol{x}$ and for all $t > 0$. Similar to the statement of Theorem 1, we can show that H_a for this case is equivalent to testing the statistical hypothesis H_a^*: $(1+\epsilon)^{-1} < \theta < (1+\epsilon)$, where $\epsilon = (4\delta)/(1-\delta)^2$ (proof omitted). It is important to note that H_a^* does not depend on either γ or $\boldsymbol{x}$. This result shows that for survival response with the POSM assumption, the hypothesis of equivalence of two patients with the same covariate $\boldsymbol{x}$ but from different treatment arms is the same as testing the statistical hypothesis H_a^*. This result allows us to extend the formulation of the statistical hypothesis of equivalence in Theorem 1 to the equivalence studies under POSM with additional covariates $\boldsymbol{x}$. However, the test statistic and corresponding critical region are now different from those used for equivalence tests with no covariates. The new test statistic, its corresponding critical region, and associated computational steps are given in the Appendix (A.4).

2.5 Error Rates of Tests

Since the properties of our equivalence testing procedures do not depend on additional covariates $\boldsymbol{x}$, for the sake of simplicity, we do not include covariate $\boldsymbol{x}$ for our theoretical and simulation studies to compare the error rates of competing procedures. In this section, we first theoretically show inflation of type I error rate of the PHM based test when true model is POSM. After that, we also perform simulation studies to study the finite sample properties (type I error and power) of both the POSM-based tests and the log-rank based tests under correctly and incorrectly specified models.

In practice, the most frequently used semi-parametric procedure for testing the equiv-

alence (e.g. Wellek, 1993) is via a log-rank based statistic under the assumption of the proportional hazards model (Cox, 1972):

$$h_1(t)/h_0(t) = exp(\eta), \tag{2.6}$$

where $h_0(t)$ is the baseline hazard and $exp(\eta)$ is the hazards ratio of the two treatment arms under the PHM. In spite of substantial literature on the robustness of a log-rank statistic based on the PHM of (2.6) for superiority tests, there is not much research studying the effect of wrongly using a log-rank based test statistic for an equivalence hypothesis when the true underlying model is not of (2.6). We examine the type I error rate for wrongly using a log-rank based equivalence test when the true underlying model is the POSM of (2.1) with true value of β as $\beta_0 = 2\log\{(1+\delta)/(1-\delta)\}$. This implies that two treatment arms following the POSM of (2.1) have the maximum difference of δ between their survival curves. If we wrongly use a log-rank based equivalence test with the same δ, we actually use a test based on the partial likelihood estimate $\hat{\eta}$ of Cox (1972). In this case, the asymptotic density of $\hat{\eta}$ is not centered around true parameter value β_0 of model (2.1). Instead, Lin & Wei (1989) showed that $n^{1/2}(\hat{\eta}-\eta^*)$ follows an asymptotic normal distribution with mean 0 and variance $v^2(\eta)$, where η^* is the unique solution of the equation

$$n_1 - \int_0^{+\infty} \frac{n_1 e^{\eta} S_0(t)}{n_1 e^{\eta} S_0(t) + n_0 S_1(t)} dt = 0, \tag{2.7}$$

and where n_0 and n_1 are the sample sizes for the standard treatment and new treatment respectively. Here, $v(\eta)$ is the estimated standard error of $\hat{\eta}$ obtained from Cox (1972). When the sample sizes n_0 and n_1 in the two treatment arms increase to $+\infty$, we can show that the center of the asymptotic distribution of η is $|\eta| < \log(1+\epsilon_h)$, where ϵ_h satisfies $(1+\epsilon_h)^{-1/\epsilon_h} - (1+\epsilon_h)^{-(1+\epsilon_h)/\epsilon_h} = \delta$ (the proof is in the Appendix (A.4)). Since the rejection

region for the log-rank based test is

$$\left\{ \frac{|\eta|}{v(\eta)} < C_\alpha \left(\frac{\log(1+\epsilon_h)}{v(\eta)} \right) \right\},$$

the necessary condition for controlling the type I error rate within 0.05 for large sample size is $|\eta^*| = \log(1+\epsilon_h)$. Under the null hypothesis H_0, as sample sizes become sufficiently large and $|\eta^*|$ goes below $\log(1+\epsilon_h)$, the type I error rate for a log-rank based test becomes greater than 0.05, the intended type I error rate of the test. Below, we also show, via simulation studies, the approximate levels of inflation of the type I error rate for finite sample sizes if we wrongly use a log-rank based test when the true model is POSM of (2.1) with true regression parameter β_0.

Our simulation studies with underlying POSM use a log normal baseline survival function $S_0(t) = \Phi(2 - \log(t))$ with mean = 2 and variance = 1, and an exponential censoring distribution with mean 50. The test-statistics for the log-rank and POSM based tests were calculated in Matlab. We take the maximum allowable difference in survival curves between two equivalent treatments as $\delta = 0.15$ (see Wellek, 1993). Using Theorem 1, we get the corresponding $\epsilon = 0.8304$, the cut-off for the equivalence test based on POSM. Each entry gives the fraction of times out of $1,000$ replications of simulated data sets for which the test statistic falls in the critical region of (2.3) with $\delta = 0.15$ (that is $\epsilon = 0.8304$). The columns for $m = max|S_1(t) - S_0(t)| = 0$ and 0.10 represent the approximate powers of the tests. The rest of the columns represent the type I error rates (sizes) of the tests at different $m \geq 0.15$. Table 2.1 shows the approximate powers and sizes using the POSM test, it appears to be below the nominal significance level of 0.05.

Table 2.2 summarizes the approximate powers and sizes for (wrongly) using the log-rank test of Com-Nougue *et al.* (1993) and Wellek (1993) for equivalence using the same $1,000$ replicate data sets simulated from the POSM models. We use the rejection region of Wellek

(1993), with intended test size 0.05 and the maximum difference in survival curves $\delta = 0.15$ as the margin of equivalence (same as Table 2.1). Each entry gives the fraction of replications for which the test statistic falls in the critical region of Wellek (1993) for $\delta = 0.15$. The simulation results show that the type I error rates at the boundary of the null H_0 of the log-rank based tests are greater than 0.05 when the true model is POSM. The difference between the actual (estimated) size type I error rate and the intended probability of type I error (5%) increases as the sample size increases. This indicates that when we wrongly use a log-rank based test, the probability of accepting the alternative that the two treatments are equivalent even when they are actually different from each other (null is true) is higher than the intended level of significance of the test.

Our next simulation study use data sets generated from the proportional hazards model (PHM) with baseline survival $S_0(t) = \Phi(2-\log(t))$ and random censoring density of exponential with mean 50. Both $S_0(t)$ and censoring density are same as the one used in previous simulation study. We again compare the type I error rates as well as the power of the log-rank based test with those of our POSM based tests. The rejection regions for both tests are determined using the equivalence margin $\delta = 0.15$ and an intended level of significance of 0.05. The values in Table 2.3 are the results using the log-rank based test under PHM. Table 2.4 values represent the POSM based test when the simulation model is PHM. Although we have only limited amount of loss of power for wrongly assuming the POSM compared to the powers of the log-rank based test, the type I error rates (sizes) of POSM based test remain below and close to the intended 5% level. This shows that the test based on a POSM assumption is a more conservative and robust approach, when compared to the log-rank based test, even when the true underlying model has proportional hazards.

For the sake of brevity, we skip the results of the simulation study of the one sample case comparing the PHM based test and the POSM based test. Similar to the two-sample case, the size of our POSM based one-sample test has type I error lower than intended significance

level test even when the sample size is small. The power of the one-sample test is almost double compared to the corresponding power of the two-sample test, indicating that we need a smaller number of patients compared to the two sample case when $S_0(t)$ is known.

2.6 Data Example & Conclusion

The main goal of the pediatric oncology trial of mo Nam *et al.* (2005) was to evaluate whether a seven months long maintenance treatment (the standard) is equivalent to a shorter (less toxic) four months long treatment (new treatment) for non-Hodgkin's malignant type B lymphoma. It is necessary to accept a "small" decrease in survival rate as a trade-off for the better tolerability and less toxicity of a new shorter treatment. For this study, the investigators decided that $\delta = 0.09$ would be the "threshold of equivalence region" - the maximum difference between the survival rates of two equivalent treatment arms (mo Nam *et al.*, 2005). Using this $\delta = 0.09$, the p-value of the log-rank based equivalence test was 0.024 based on the observed data with 11 and nine failures out of sample sizes of 84 and 82 respectively from standard long and new shorter treatment regimens. This p-value ($0.024 < 0.05$) is a highly significant evidence in favor of the alternative hypothesis of equivalence of two treatments. We cannot re-analyze the clinical trial because the original data is proprietary. Instead, we would like to demonstrate using a simulation study the comparison between a log-rank based test and a POSM based test when we follow the design and the censoring mechanism similar to this trial and the true difference between treatment arms is at the margin of equivalence ($m = max|S_1(t) - S_0(t)| = 0.09$). We simulate 1000 data sets with 11 out of $n_1 = 84$ and nine failures out of $n_2 = 82$ for two treatment arms following POSM. We use the censoring scheme and monitoring length (18 months) similar to mo Nam *et al.* (2005). The proportion of log-rank based test statistics with p-values more extreme than 0.024 is 0.038. When using the POSM test on the same set of data, we find the corresponding proportion to be 0.012.

This shows that when the true model is POSM, the probability that a log-rank test will have a p-value less than 0.024 (same or more significant than the p-value obtained by mo Nam *et al.* (2005)) is almost three times more than the probability of having such a significance level with the POSM based test. This further demonstrates that a log-rank based test has a high probability to give a highly-significant (very small) p-value, even though we do not want to reject the null hypothesis H_0 when the actual $\delta = 0.09$.

Unlike Cox's model where one hazard function dominates the other over time, a POSM can allow two hazard functions to merge over time. This may be a possible explanation for POSM based equivalence test being more conservative than the log-rank based test. Several authors (e.g. Li Y, 2003; Betensky *et al.*, 2002) have argued that even for superiority trials, the efficiency and validity of a log-rank based test likely is questionable for solid tumor oncology studies in which there is heterogeneity of tumors due to existence of various unidentified genetic subtypes. For these studies, the hazard functions from two treatment arms may merge over time. A POSM assumption with POSM based equivalence tests would be a wise choice in this situation.

In this paper, we have presented the test statistics, critical region and robustness and other related properties of POSM based tests restricted only to right-censored survival data. We would like point out that the statements of Theorem 1 and Theorem 2 are valid irrespective of the censoring mechanism. Arguably, for any type of censoring, it is possible to develop an equivalence test based on POSM if one can determine the appropriate statistic (preferably based on the SPML estimate of β of POSM from interval-censored data) and the corresponding critical region. However, equivalence trials with different types of censoring such as interval-censoring are beyond the scope of this paper. These are important topics of future research.

Table 2.1: For different values of maximum difference in survival curves $m = max|S_1(t) - S_0(t)|$, the Pr(Rejecting H_0) for using POSM based test when the true model is POSM (sample size = $n_1 + n_2$ for $n_1 = n_2$).

Sample	Power		Type I error rate		
size	$m = 0$	$m = 0.10$	$m = 0.15$	$m = 0.20$	$m = 0.30$
50	0.114	0.072	0.049	0.030	0.006
100	0.210	0.115	0.050	0.010	0.000
150	0.378	0.154	0.050	0.012	0.000
200	0.598	0.200	0.055	0.007	0.000
400	0.930	0.308	0.044	0.004	0.000

Table 2.2: For different values of maximum difference in survival curves $m = max|S_1(t) - S_0(t)|$, Pr(Rejecting H_0) for using log-rank based test when the true model is POSM (sample size = $n_1 + n_2$ for $n_1 = n_2$).

Sample	Power		Type I error rate		
size	$m = 0$	$m = 0.10$	$m = 0.15$	$m = 0.20$	$m = 0.30$
50	0.120	0.085	0.069	0.032	0.007
100	0.286	0.155	0.085	0.038	0.006
150	0.497	0.235	0.111	0.034	0.004
200	0.685	0.335	0.130	0.033	0.000
400	0.964	0.539	0.180	0.033	0.000

Table 2.3: For different values of maximum difference in survival curves $m = max|S_1(t) - S_0(t)|$, Pr(Rejecting H_0) for using log-rank based test when the true model is PHM (sample size = $n_1 + n_2$ for $n_1 = n_2$).

Sample	Power		Type I error rate		
size	$m = 0$	$m = 0.10$	$m = 0.15$	$m = 0.20$	$m = 0.30$
50	0.127	0.087	0.045	0.012	0.005
100	0.268	0.131	0.052	0.016	0.000
150	0.510	0.156	0.047	0.010	0.000
200	0.676	0.201	0.049	0.009	0.000
400	0.966	0.355	0.052	0.002	0.000

Table 2.4: For different values of maximum difference in survival curves $m = max|S_1(t) - S_0(t)|$, Pr(Rejecting H_0) for using POSM based test when the true model is PHM (sample size = $n_1 + n_2$ for $n_1 = n_2$).

Sample	Power		Type I error rate		
size	$m = 0$	$m = 0.10$	$m = 0.15$	$m = 0.20$	$m = 0.30$
50	0.104	0.076	0.042	0.026	0.008
100	0.211	0.116	0.046	0.015	0.001
150	0.381	0.143	0.044	0.009	0.000
200	0.594	0.207	0.049	0.003	0.000
400	0.920	0.302	0.044	0.003	0.000

CHAPTER 3

TESTS FOR NON-INFERIORITY OF TWO SURVIVAL FUNCTIONS WITH SAMPLE SIZE APPROXIMATIONS

3.1 Introduction

Recently, there is a substantial interest in the bio-pharmaceutical industry to develop new treatments that may have crucial advantages such as easier administration, lower cost, and increased tolerability compared to a prevailing treatment of proven efficacy. An essential step for approving such a new treatment is to conduct a non-inferiority trial to evaluate whether the new treatment is at least as efficacious as the existing treatment in terms of the main survival outcome. Here, we first investigate the serious consequences of the usual log-rank based non-inferiority tests when the proportional hazard model (PHM) assumption fails. Then we introduce a practical formulation of the non-inferiority hypothesis as well as the corresponding statistical test based on proportional odds survival model (POSM). Unlike the commonly used procedures, our test ensures a proper control of the type I error rate when two treatment arms satisfy either the POSM or PHM. We also present a practical method for determining sample sizes to guarantee a desired power.

In contrast, to a superiority trial to identify which treatment is superior to another, a non-inferiority trial aims to evaluate whether the new treatment arm is acceptably as efficacious as the current treatment arm within a pre-specified margin of non-inferiority (e.g.

Rothmann *et al.*, 2003). Non-inferiority trial is useful when the new treatment arm has additional benefits such as less toxicity, or collateral effects, easier administration, lower cost, or protection from drug resistance. In addition, non-inferiority trials have been crucial in assessing most modern-day antibiotics (e.g. Plikaytis & Carlone, 2005; Chan *et al.*, 2003). Such trials also have potential to play prominent roles in the evaluation of immunotherapy agents including cancer vaccines (Rothmann *et al.*, 2012) for treating adjuvant melanoma and other cancers.

Even though the existing statistical methods for non-inferiority trials for survival response often use a log-rank based test statistic under the proportional hazards model (PHM) assumption of Cox (1972), in practice, the hazard functions of two treatment arms are often not proportional over time (e.g. Betensky *et al.*, 2002; Cooner *et al.*, 2007). Indeed, numerous articles and exceptionally credible bio-statistical research have clearly indicated that it is nearly impossible to interpret any PHM based statistic as a treatment contrast under non-proportional hazards, especially in presence of censoring (e.g. Lin & Wei, 1989; Kalbfleisch & Prentice, 2002; Rudser *et al.*, 2012). In this paper, we demonstrate, both theoretically (in Section 3.2) and via simulation study, a severe consequence of the log-rank based non-inferiority test of, say, Wellek (1993), is the inflation of the type I error rate when the true model has non-proportional hazards. Controlling the type I error rate of a non-inferiority trial is particularly important because a serious adverse consequence of erroneously accepting the alternative hypothesis of non-inferiority is that an ineffective treatment can potentially replace a current treatment of proven efficacy.

One strategy for dealing with the inflated type I error for wrong modeling assumption is to use non-parametric tests proposed by many authors including Rothmann *et al.* (2003), Freitag *et al.* (2006) and Zhang *et al.* (2011). These non-inferiority tests are in general based on one of the following quantities, (1) difference in restricted mean survival times (Royston & Parmar, 2011), (2) average hazards ratio (appropriate mainly for PHM), (3) integrated

difference in survival functions (Zhao *et al.*, 2012), and (4) difference in medians (Koti, 2013). In Section 3.2, we discuss the formulation and benefits of using our semi parametric model based non-inferiority test over the existing non-parametric tests. Furthermore, required sample sizes for the non-parametric non-inferiority tests (e.g. Com-Nougue *et al.*, 1993) are substantially higher than those for tests under semi-parametric models (see Wellek, 2010).

We develop a proportional odds survival model (Bennett, 1983) based non-inferiority hypothesis formulation and the corresponding statistical test in Section 3.3. Unlike the commonly used log-rank based procedure (e.g. Wellek, 1993), our test ensures a type I error rate lower than the desired nominal rate when two treatment arms satisfy either the POSM or Cox's (1972) PHM. In practice, this implies that our test is a safer statistical procedure than the current method. We present a serviceable method for determining sample sizes to guarantee a desired power, at different non-inferiority margin values between two survival curves. We also introduce a simple formula for obtaining sample sizes when the data are without censored observations. Moreover, we show our sample size approximation works reasonably well under various censoring schemes.

In Section 3.4, we conduct simulation studies to demonstrate the potential effect of wrong modeling assumptions on the type I error rate and power for both PHM and POSM based tests. Section 3.5 present's conclusions and some discussions.

3.2 Formulation of Non-Inferiority Hypothesis

We consider survival time response from two competing treatment arms, standard and new, with corresponding survival functions $S_0(t)$ and $S_1(t)$, and hazards functions $h_0(t)$ and $h_1(t)$. Statistical tests for evaluating differences between survival responses from two treatment arms are primarily based on comparing the following, (1) hazard functions, (2) survival functions, (3) means and (4) medians. Without loss of generality, we assume that

the event of interest is unfavorable to the patient (e.g., relapse, death). For a superiority trial, the logical clinical/scientific hypothesis of interest H_a (i.e. alternative hypothesis) is that the new treatment is superior to existing one. In this case a logical formulation of H_a is H_a: $h_1(t)/h_0(t) < 1$ for all t. The principal reason for using an alternative hypothesis H_a based on the hazard ratio (or difference) is that $h_1(t)/h_0(t) < 1$ for all $t \Rightarrow S_1(t) > S_0(t)$ for all t, however, the reverse is not true. Also, when dealing with the corresponding ordering of the associated means (μ_0, μ_1), H_a: $h_1(t)/h_0(t) < 1$ for all $t \Rightarrow \mu_1 > \mu_0$. Notably, ordering of means, medians and particular quantiles do not imply any ordering of corresponding hazards. Unlike superiority trials, a non-inferiority trial aims to test the alternative H_a that new treatment is clinically non-inferior to the standard, however importantly, only within a level of non-inferiority margin.This non-inferiority margin δ is the maximum measured value a new treatment is allowed to be inferior to while still being considered efficacious. For example using this δ, a formulation of non-inferiority alternative hypothesis H_a based on difference between two survival functions is H_a: $S_0(t) - S_1(t) < \delta$ for all t, versus the null hypothesis that H_0: $S_0(t) - S_1(t) > \delta$ at **some** time t, where δ is the predetermined margin of non-inferiority (Rothmann *et al.*, 2012; Tsiatis & Mehta, 2003) when H_a is formulated based on survival difference.

Unlike the superiority trial, it is not obvious that the alternative hypothesis H_a of non-inferiority should be based on hazard functions $h_1(t)$ and $h_0(t)$. There does not exist any reasonable margin, say, $\delta^* \in (0, 1)$, such that H_a: $h_1(t) - h_0(t) < \delta^*$ for all t can ensure $S_0(t) - S_1(t) < \delta$ for all t for a chosen margin $0 < \delta < 1$. In this case, we argue that it is logical to formulate H_a based on the maximum difference $S_0(t) - S_1(t)$ of survival functions. Initially, it is convenient to define non-inferiority margin δ on difference in two survival probabilities. Also, H_a based on margin of $S_0(t) - S_1(t)$ implies a stricter definition of non-inferiority compared to any other non-inferiority hypothesis based on differences of either medians or any other quantiles. Another existing method for a non-inferiority test compares

the restricted mean survival times (RMST) $\mu_{k\tau} = \int_0^\tau S_k(t)dt$, of two groups (Royston & Parmar, 2011; Zhao *et al.*, 2012; Tian *et al.*, 2013), where $(0, \tau)$ is the pre-specified time-interval of clinical interest. A potentially controversial aspect of the RMST method, is the requirement of a pre-specified interval of interest. We note that H_a: $S_0(t) - S_1(t) < \delta$ for all $t \Rightarrow \mu_{0\tau} - \mu_{1\tau} < \delta\tau$, however, the reverse is again not true. The above finding confirms, using an H_a based on survival differences corresponds to a stricter and more encompassing definition of non-inferiority than that based on RMST.

The required sample sizes for RMST and other non-parametric non-inferiority tests (e.g. Com-Nougue *et al.*, 1993) are usually higher than those for tests under semi-parametric models (see Wellek, 2010). A currently popular choice for a semi-parametric model based non-inferiority test is the log-rank based test of Wellek (2010) using the partial likelihood estimate ($\hat{\eta}$) of the PHM of Cox (1972) given by

$$h_1(t) = \eta h_0(t), \tag{3.1}$$

where η is the hazard ratio. For this non-inferiority test, the null hypothesis is H_0: $\eta \geq \rho$, where $\rho > 1$ is the non-inferiority margin of the hazards-ratio. Using results of Struthers & Kalbfleisch (1986) under assumption of no censoring, we can show that $n^{1/2}(\hat{\beta} - \beta^*)$ has an asymptotic normal distribution with mean zero and finite variance, where $\eta^* = e^{\beta^*}$ is the unique solution of the equation

$$\pi + \int_0^{+\infty} \frac{\eta\pi S_1(t)}{\eta\pi S_1(t) + (1-\pi)S_0(t)} \left[\pi \frac{d}{dt} S_1(t) + (1-\pi)\frac{d}{dt} S_0(t)\right] dt = 0, \tag{3.2}$$

$\hat{\beta}$ is the PHM based estimator of $\beta = \log \eta$ and π is the proportion of subjects in new treatment arm out of n total subjects. This η^* is the true hazard ratio η_0 when assumption of (3.1) is correct. However, when the assumption of (3.1) fails, this η^* generally under represents the true difference $\max_t\{S_0(t) - S_1(t)\}$. Theorem 3 gives the exact statistical

definition and it's corresponding proof is given in the Appendix B.1.

Theorem 3 *When $S_0(t)$ and $S_1(t)$ follow the proportional odds survival model (POSM) of Bennett (1983)*

$$\frac{1-S_1(t)}{S_1(t)} = \theta \left[\frac{1-S_0(t)}{S_0(t)}\right], \tag{3.3}$$

where θ is the time-constant survival odds ratio, then $\max_t\{S_0(t) - S_1(t)\} > \max_t\{S_0(t) - S_1^*(t)\}$, *where $S_1^*(t) = S_0(t)^{\eta^*}$ and η^* is the solution of equation (3.2) for any $0 < \pi < 1$ and $S_0(t)$.*

As a consequence, when the targeted type I error rate α for log-rank based non-inferiority test is aimed at the margin $\eta = \rho$, based on $\hat{\beta}$ is centered at $\eta^* < \rho$ even when $n \to \infty$ (see Appendix B.1). Hence, the type I error becomes greater than the intended level, even for large sample sizes. We later illustrate the actual inflation of type I error rates for finite sample sizes via simulation studies.

Figure 3.1 shows the survivals curves $S_0(t)$ under an exponential with mean 1 (solid line) and $S_1(t)$ (dotted line) with the maximum difference of $S_0(t) - S_1(t)$ being 0.20, when $S_0(t)$ and $S_1(t)$ follow the POSM of (3.3). The dashed line shows the survival curve $S_1^*(t)$ when $S_1^*(t)$ and $S_0(t)$ satisfy PHM assumption with hazard ratio η^*, where η^* is the solution of (3.2) for $\pi = 1/2$. The maximum difference between the $S_1^*(t)$ and $S_0(t)$ is less than 0.15. This figure demonstrates that even when the true maximum difference $S_0(t) - S_1(t)$ is 0.20, if we have $S_0(t)$ and $S_1(t)$ satisfying POSM, the survival curve $S_1^*(t)$ (fitted curve using PHM as $n \to \infty$) for new treatment has the maximum difference of $S_0(t) - S_1(t)$ as 0.15 that is less than the true value of 0.20.

This figure as well as our theoretical result demonstrate that a PHM based non-inferiority test inflates the type I error when the true model is not PHM, and the amount of inflation does not decrease with an increase in sample size. This motivates us to look for an alternative non-inferiority test based on a semi-parametric model different from the PHM. For example

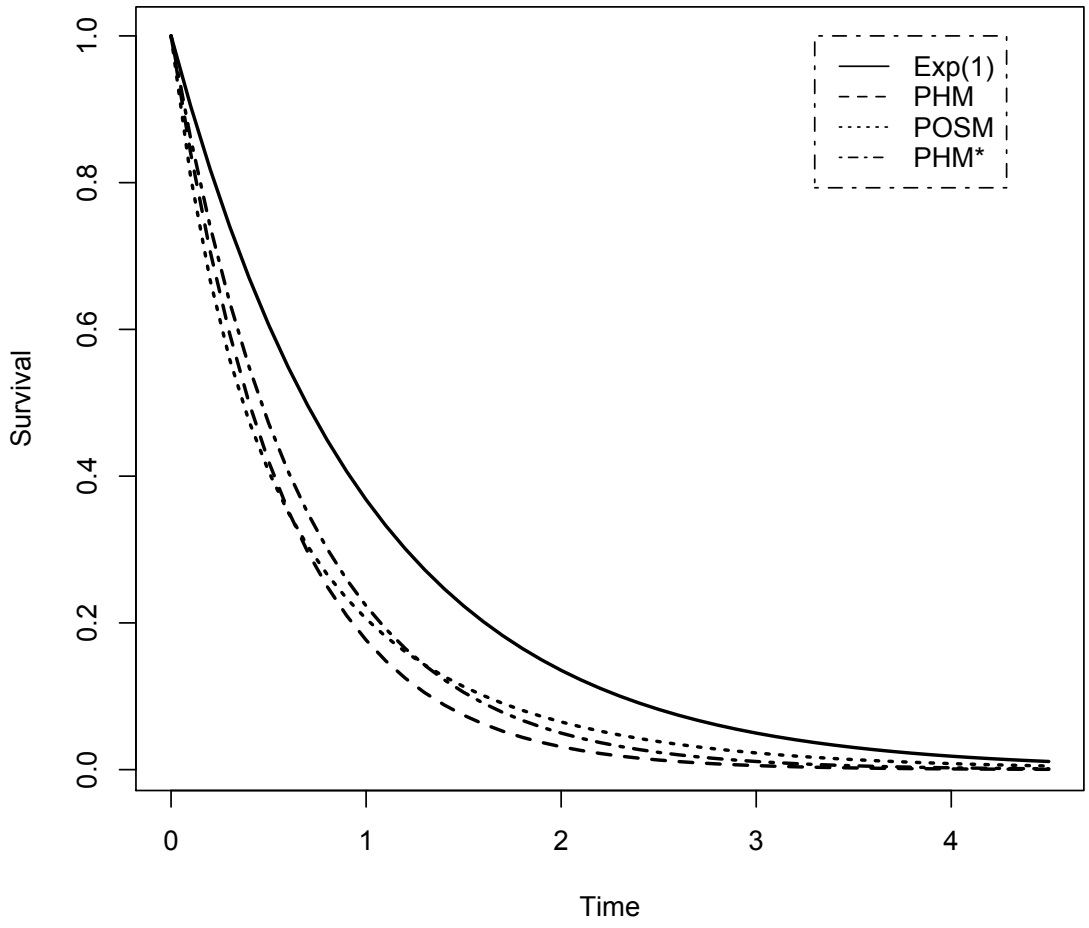

Figure 3.1: Survival curves for the standard treatment, true PHM, true POSM and PH convergence model of POSM (with $*$) when the sample size goes to infinity.

an alternative of Cox's model of (3.1) is the POSM of Bennett (1983), with $S_0(t)$ and $S_1(t)$ following (3.3). However, in order to achieve an efficient and practical non-inferiority test under a semi-parametric model such as (3.3), we must first formulate the statistical hypothesis of non-inferiority consisting of only the treatment effect parameter (Wellek, 2010). Therefore, the null and alternative statistical hypotheses based on POSM of (3.3) should be based on ranges of the scalar odds ratio parameter θ. Still, the margin of non-inferiority δ is often easier to describe as a limit on the difference in survival probabilities. Thus Theorem 4 formulates a null and corresponding alternative statistical hypotheses for non-inferiority testing established on ranges of θ from the non-inferiority hypotheses based on the difference in survival functions.

Theorem 4 *Under POSM of (3.3) with continuous* $S_0(t)$*, testing* $\mathrm{H}_0 : \ S_0(t) - S_1(t) > \delta$ *for some* $t > 0$ *versus* $\mathrm{H}_a : \ S_0(t) - S_1(t) < \delta$*, for all* $t > 0$ *is equivalent to testing* $\mathrm{H}_0^* : \ \theta < \Delta$ versus $\mathrm{H}_a^* : \ \theta > \Delta$ *where* $\Delta = [(1+\delta)/(1-\delta)]^2$.

The proof of Theorem 4 is in the Appendix B.2.

Many practitioners consider the hazard ratio as a very logical quantity for comparing treatments. The suitable alternative hypothesis of non-inferiority is $\mathrm{H}_{a2} : \ h_1(t)/h_0(t) < \rho$ for all $t > 0$ for this case. Theorem 5 shows that under the POSM of (3.3), the H_{a_2} based on the hazards ratio is also equivalent to our hypothesis based on range of θ.

Theorem 5 *The POSM assumption of (3.3), the alternative hypothesis of interest* H_{a2}*:* $h_1(t)/h_0(t) < \rho$ *for all* t*, is the same as testing* H_a^**:* $\theta < \rho$.

The proof is provided in the Appendix B.3.

Comparing the results of Theorems 4 and 5, we conclude that the corresponding alternative statistical hypotheses (H_a^*) are the same. This shows that for POSM, both formulations of the null and alternative statistical hypotheses can be reduced to testing the statistical hypothesis H_0^* versus H_a^* involving only ranges of θ.

3.3 Non-Inferiority Tests and Sample-Size

Following the formulation of our statistical hypotheses, we now explain the method for obtaining the respective rejection region. Theorem 4 can be rewritten as

$$\mathrm{H}_0^* : \beta \geq \log \Delta \text{ versus } \mathrm{H}_a^* : \beta < \log \Delta, \tag{3.4}$$

where $\beta = \log \theta$. A rejection region for testing (3.4) with level significance α is

$$R_\alpha = \left\{ \hat{\beta} : \frac{\hat{\beta} - \log \Delta}{v(\hat{\beta})/\sqrt{n}} < z_\alpha^*, \right\}, \tag{3.5}$$

where z_α^* is the α quantile of a standard normal distribution, $\hat{\beta}$ is any consistent and efficient estimator of β, and $v^2(\hat{\beta})$ is a consistent estimator of the asymptotic variance of $\hat{\beta}$. We will use estimators based on the methods proposed by Chen *et al.* (2002), these methods consist of using an iterative approach to procure $\hat{\beta}$ and its associated variance, $v^2(\hat{\beta})$. Even though, $\hat{\beta}$ is not the most efficient estimate of β, the amount of efficiency lost is small and the estimated variance $v^2(\hat{\beta})$ has a closed form expression. The iterative steps to find $\hat{\beta}$ and $v^2(\hat{\beta})$ from possibly right-censored data from a two-arm randomized trial are in the Appendix (B.4). Since the test statistic in (3.5) follows an approximate normal distribution, we can now develop an accurate method (asymptotically correct) to obtain the optimal sample size n_{opt} to guarantee a desired power and a required type I error given the non-inferiority margin δ and under POSM.

Determining sample size for a non-inferiority test is a crucial part of the clinical trial. A wrongly sized non-inferiority trial may lead to inadequacies in power and optimal number of patients. It may be unreasonable (e.g. high cost) to conduct the clinical trial when the required sample size to guarantee a particular power is too large.

For a targeted type I error rate α_1, power $(1 - \alpha_2)$ at pre-specified $\beta_0 < \log \Delta$, and

pre-specified baseline function $S_0(t)$, the optimal sample size n_{opt} for testing (3.4) can be evaluated solving the equation $P[\hat{\beta}_n \in R_{\alpha_1} | S_0, \beta = \beta_{\alpha_2}] = 1 - \alpha_2$, where $\hat{\beta}_n$ is the estimate of β with sample size n, and R_{α_1} is the rejection region from (3.5) with level of significance α_1. When no censoring occurs, the equation has a closed form solution (under asymptotic distribution of $\hat{\beta}$)

$$n_{\text{opt}} = \frac{\sigma^2 (z^*_{\alpha_1} + z^*_{\alpha_2})^2}{\pi(1-\pi)(\beta_{\alpha_2} - \log \Delta)^2}, \tag{3.6}$$

where z^*_α is the upper 100α percentile of a standard normal distribution, π is the proportion of subjects in treatment arm, and σ^2 is the inverse of the expected information matrix of β given pre-specified $S_0(t)$. For POSM with uncensored observations, the corresponding expected information matrix of β is $I_\beta = \frac{n_1}{3\theta^2}$ (proof in Appendix (B.5)). For the case with censoring, the optimal sample sizes for our proposed non-inferiority test will be approximated using

$$n_{\text{opt}_c} = \frac{n_{\text{opt}}}{1 - \pi_c}, \tag{3.7}$$

where π_c is the expected proportionality of censoring.

3.4 Simulation Studies

We now conduct simulation studies to show the behavior of the type I error and power rates for the POSM as well as the log-rank PHM based tests under both accurately and inaccurately stated models. We generate 1,000 data sets assuming that the relationship between the standard and the new treatments follow a POSM. We use both the log-rank test (Wellek, 1993; Com-Nougue *et al.*, 1993) and the POSM based test proposed here to obtain the power and the type I error rates considering different maximum distance (m) between the survival curves, where $m = 0$ and $m = 0.10$ represent the power analysis and $m = 0.15, m = 20$ and $m = 0.30$ represent the type I error rates for our simulation studies.

All data were generated with the standard treatment following a log-normal distribution with mean equal to two and variance equal to one. Following Wellek (1993), we use the cut-off margin of non-inferiority $\delta = 0.15$ and significance level $\alpha = 0.05$. Tables 3.1 and 3.2 show the proportions of samples for which the null hypothesis were rejected when we use the POSM based test and the log-rank test, respectively. Unlike our POSM based tests, the required type I error rate of 5% cannot be attained using the PHM based log-rank test when the data simulated follows the POSM. In this case, the type I error rate increases with sample size, this means the probability of approving an inferior treatment is actually increasing with sample size.

Using the same simulation procedure described above, we now generate 1,000 data sets such that the standard and the new treatments follow a PHM. The null hypothesis rejection rates given in Table 3.3 are calculated when we use a log-rank test and Table 3.4 values are acquired when we use our POSM based test. Comparing the results from Tables 3.3 and 3.4 we conclude that both tests have similar power and type I error rates when the true model follows a PHM, in other words, our POSM based test maintains the required 5% level of significance even if the true model is not a POSM. Hence, our POSM based approach to non-inferiority testing is more conservative and robust than the current log-rank based test, regardless of whether the underlying modeling assumption follows the PHM.

In addition, we determine the optimal sample size to conduct a clinical trial that assumes a POSM (e.g, non-inferiority), using the expression given by (3.6). The results shown in Table 3.5 allow us to conclude that expression (3.6) is a reliable approximation. Since our simulation studies in Table 3.1 have sample sizes with similar powers ($m = 0$) to those of Table 3.5 when $\delta = 0.15$.

The computational codes for both POSM and PHM based non-inferiority tests were implemented using SAS 9.4 and they can be request directly from the authors.

3.5 Conclusion

The non-inferiority tests between two survival treatments usually are conducted under the assumption that the true model is the PHM, without worrying about the consequences of the PH assumption being wrong. We prove that an incorrect PHM assumption inflates the type I error of non-inferiority tests, and this inflation increases with sample size. To circumvent this serious problem we propose a non-inferiority test for two survival curves based on the POSM of Bennett (1983). In addition to taking into account two hazards functions merging over time, another major advantage of our new non-inferiority test is the control of type I error even if the true model is the PHM of Cox (1972). We have shown this inflation of type I error theoretically and through simulation studies. In this paper, we are emphasizing the important public health issue of ensuring the strict control of the type I error rate of a non-inferiority trial even when the modeling assumption is under suspicion.

Even though additional covariates is not a primordial issue besides treatment arm indicator, covariates can be incorporated in the POSM of (3.3) considering a multiplicative effect on the odds ratio as shown in (3.8),

$$\frac{1 - S_1(t|\boldsymbol{x})}{S_1(t|\boldsymbol{x})} = \theta \left[\frac{1 - S_0(t|\boldsymbol{x})}{S_0(t|\boldsymbol{x})}\right] = \theta e^{\boldsymbol{\gamma}\boldsymbol{x}} \left[\frac{1 - S_0(t)}{S_0(t)}\right], \tag{3.8}$$

where $\boldsymbol{\gamma}$ is the regression parameter of $\boldsymbol{x}$ and θ remains the treatment effect of importance. Then, the corresponding alternative hypothesis for the non-inferiority test is H_a: $S_0(t|\boldsymbol{x}) - S_1(t|\boldsymbol{x}) < \delta$ for all $\boldsymbol{x}$ and $t > 0$. Analogous to what we demonstrated in Theorem 4, we also have that the alternative hypothesis is equivalent to testing the statistical hypothesis $\mathrm{H}_a^* : \theta < [(1+\delta)/(1-\delta)]^2$. The proof is similar to the proof presented in Appendix B.2. It is worth acknowledging that the H_a^* does not rely on either $\boldsymbol{\gamma}$ or $\boldsymbol{x}$, this means when testing for non-inferiority of two patients receiving different treatments, but with same covariates and survival responses following the POSM assumption, it is again enough to analyze the survival

odds ratio. A direct consequence of this result is that we now can incorporate covariates in our methodology for non-inferiority studies under the POSM.

Non-inferiority trials that deal with interval and other kinds of censoring besides right-censoring stand beyond the scope of this paper. Nevertheless, Theorems 4 and 5 are justified regardless of the censoring scheme. In order to extend our non-inferiority test formulation based on POSM to all cases of censored data, we first need to find the correct test statistic and critical region for each proposed censoring mechanism. These are explicitly important topics of future research and consideration.

Table 3.1: Pr(Rejecting H_0) for using POSM based test when the true model is POSM, where $m = \max\{S_0(t) - S_1(t)\}$ and sample size $= n_1 + n_2$ for $n_1 = n_2$.

Sample	Power		Type I error rate		
size	$m = 0$	$m = 0.10$	$m = 0.15$	$m = 0.20$	$m = 0.30$
50	0.340	0.105	0.042	0.019	0.000
100	0.525	0.144	0.046	0.018	0.001
150	0.695	0.186	0.050	0.009	0.000
200	0.791	0.216	0.048	0.007	0.000
300	0.919	0.286	0.051	0.001	0.000

Table 3.2: Pr(Rejecting H_0) for using log-rank based test when the true model is POSM, where $m = \max\{S_0(t) - S_1(t)\}$ and sample size $= n_1 + n_2$ for $n_1 = n_2$.

Sample	Power		Type I error rate		
size	$m = 0$	$m = 0.10$	$m = 0.15$	$m = 0.20$	$m = 0.30$
50	0.413	0.172	0.088	0.031	0.003
100	0.637	0.254	0.117	0.059	0.004
150	0.799	0.360	0.146	0.043	0.000
200	0.882	0.438	0.184	0.041	0.001
300	0.981	0.553	0.225	0.046	0.000

Table 3.3: Pr(Rejecting H_0) for using log-rank based test when the true model is PHM, where $m = \max\{S_0(t) - S_1(t)\}$ and sample size $= n_1 + n_2$ for $n_1 = n_2$.

Sample	Power		Type I error rate		
size	$m = 0$	$m = 0.10$	$m = 0.15$	$m = 0.20$	$m = 0.30$
50	0.418	0.122	0.044	0.020	0.007
100	0.650	0.160	0.056	0.016	0.003
150	0.792	0.202	0.049	0.004	0.000
200	0.876	0.242	0.054	0.001	0.000
300	0.963	0.334	0.042	0.004	0.000

Table 3.4: Pr(Rejecting H_0) for using POSM based test when the true model is PHM, where $m = \max\{S_0(t) - S_1(t)\}$ and sample size $= n_1 + n_2$ for $n_1 = n_2$.

Sample	Power		Type I error rate		
size	$m = 0$	$m = 0.10$	$m = 0.15$	$m = 0.20$	$m = 0.30$
50	0.323	0.109	0.045	0.019	0.001
100	0.565	0.135	0.050	0.013	0.000
150	0.683	0.168	0.050	0.007	0.000
200	0.786	0.212	0.053	0.001	0.000
300	0.909	0.256	0.045	0.003	0.000

Table 3.5: Optimal Sample Size Approximations for Non-inferiority trials without Censoring.

δ	Power	n_{opt}
0.10	0.70	220
0.10	0.80	288
0.10	0.90	400
0.15	0.70	82
0.15	0.80	108
0.15	0.90	150
0.20	0.70	37
0.20	0.80	48
0.20	0.90	66

CHAPTER 4

FUTURE WORK

We would like to further extend equivalence and non-inferiority testing for two survival curves $S_0(t)$ and $S_1(t)$ using a Bayesian approach through the use of a utility function for the PHM of Cox (1972) and POSM Bennett (1983). The aim of a utility function, is to give a numerical value to the expected gain or loss by choosing one action (a) over the other. For our cases, it is the difference between the utility of rejection and the utility of acceptance of equivalence or non-inferiority of two survival functions. For example using the PHM of Cox (1972), with a prior distribution, we can the attain the posterior density $P(\eta|D_n)$, where η is the hazard ratio of Cox (1972) and D_n is the given data. We would like to then incorporate a utility function, $U(\eta, a)$ to capture the reward for making the correct decision. In our cases, actions a_1, rejecting the H_0 when we should reject the H_0, and a_0 accepting the H_0 when we should accept the H_0. We then would find the corresponding $E\{U(\eta, a_0|D_n)\}$ and $E\{U(\eta, a_1|D_n)\}$, where $E\{U(\eta, a|D_n)\} = \int U(\eta, a)P(\eta|D_n)d\eta$. For this case, $E\{U(\eta, a|D_n)\}$ is a univariate density and can be solved analytically using common integration techniques. The only thing left to do is determine the criteria for choosing an appropriate utility function.

We would like to choose $U(\eta, a)$ with two primary restrictions. First, we would like to ensure a Bayesian type I error of 5%. By Bayesian, we mean only 5% of the time would we like to falsely conclude an action a. Secondly, we would like to maximize the power of our Bayesian utility driven test.

Using this approach, we can now quantify the expected gain or loss from choosing one action over the other. This allows the practitioner a way to measure how badly the result of committing a type I error would be. This is a crucial advantage gained by using the aid of

a utility function. We can now, through the use of a utility function, measure the affect of committing a type I error for different clinical trials. This is future work that we would like to justify via simulation studies and theoretical proofs.

APPENDIX A

PROOFS FROM CHAPTER 2

A.1 Proof of Theorem 1

For the POSM of (2.1), $|S_0(t) - S_1(t)| < \delta$ for all t,
$\Leftrightarrow |\{1 + B_0(t)\}^{-1} - \{1 + B_0(t)\theta\}^{-1}| < \delta$ for all t, where $B_0(t) = \{1 - S_0(t)\}/S_0(t)$. Using standard calculus, we can show that

$$\max_{t>0} = \left|\{1 + B_0(t)\}^{-1} - \{1 + B_0(t)\theta\}^{-1}\right| = \left|\frac{\theta^{\frac{1}{2}} - \theta^{-\frac{1}{2}}}{(1 + \theta^{\frac{1}{2}})(1 + \theta^{-\frac{1}{2}})}\right| = M(\theta),$$

because $B_0(t)$ is continuous. Please note that $M(\theta)$ is a decreasing (increasing) function when $\theta \in (0, 1)$ (when $\theta \in (1, +\infty)$). Therefore, the condition $|S_0(t) - S_1(t)| < \delta$ for all t, is equivalent to $M(\theta) < \delta \Leftrightarrow (1+\epsilon)^{-1} < \theta < 1+\epsilon$, where ϵ should satisfy $\frac{(1+\epsilon)^{\frac{1}{2}} - (1+\epsilon)^{-\frac{1}{2}}}{\{1+(1+\epsilon)^{\frac{1}{2}}\}\{1+(1+\epsilon)^{-\frac{1}{2}}\}} = \delta \;\Rightarrow\; \epsilon = (4\delta)/(1-\delta)^2$.

A.2 Proof of Theorem 2

Let $h_0(t)$ and $h_1(t)$ be the hazard function for $S_0(t)$ and $S_1(t)$, respectively. POSM of (2.1) implies $h_1(t)/h_0(t) = [1 + (\theta - 1)S_0(t)]^{-1}$, which is an increasing (a decreasing) function of t converging to 1 when $\theta > 1$ (when $\theta < 1$). This implies that, the maximum of $|\log\{h_1(t)/h_0(t)\}|$ for all $t > 0$, is equal to $|\log\{1 + (\theta - 1)S_0(0)\}| = |\log(\theta)|$ Therefore, testing $|\log\{h_1(t)/h_0(t)\}| < \log(\rho)$ for all $t > 0$, is equivalent to testing $|\log(\theta)| < \log(\rho) \Leftrightarrow \rho^{-1} < \theta < \rho$.

A.3 Iterative Steps to Estimate Parameters

Chen *et al.* (2002) estimator $(\tilde{\beta}, \tilde{B})$ of (β, B) is obtained via iteratively solving the following set of estimating equations, where $B(t)$ is a non decreasing and non negative function with jumps only at observed failure times $t_{(1)} < \ldots < t_{(k)}$.

Step 0. Choose an initial value $\beta^{(0)}$ for β.

Step 1. Obtain $B^{(0)}(t_{(1)})$ by solving

$$\sum_{i=1}^{n} \frac{Y_i(t_{(1)}) \exp(z_i\beta^{(0)})}{\exp(z_i\beta^{(0)}) + \exp\{-B(t_{(1)})\}} = 1,$$

where $Y_i(t)$ is the indicator that subject i is under observation at time t. Then obtain $B^{(0)}(t_{(k)})$, for $k = 2, \ldots, K$, sequentially by solving:

$$\sum_{i=1}^{n} \frac{Y_i(t_{(k)}) \exp(z_i\beta^{(0)})}{\exp(z_i\beta^{(0)}) + \exp\{-B(t_{(k)})\}} = 1 + \sum_{i=1}^{n} \frac{Y_i(t_{(k)}) \exp(z_i\beta^{(0)})}{\exp(z_\beta^{(0)}) + \exp\{-B(t_{(k-)})\}}.$$

Step 2. Obtain new estimate $\beta^{(1)}$ of β by solving

$$U(\beta, B) \equiv \sum_{i=1}^{n} z_i[d_i - \Lambda\{B(t) + z_i\beta\}] = 0,$$

$$\sum_{i=1}^{n} [d_i - \Lambda\{B(t) + z_i\beta\}] = 0,$$

where $\Lambda(u) = \{1 + \exp(-u)\}^{-1}$.

Step 3. Repeat Step 1 and Step 2 until the convergence.

A.4 Estimating Equations with Covariates

The estimates $(\tilde{\beta}, \tilde{\boldsymbol{\gamma}}, \tilde{B})$ of $(\beta, \boldsymbol{\gamma}, B)$ of (2.5) (based on Chen *et al.* (2002)) obtained via solving the estimating equations:

$$U(\beta, \boldsymbol{\gamma}, B) \equiv \sum_{i=1}^{n} \int_0^{\infty} [z_i; \boldsymbol{x}_i][dN_i(t) - \tilde{Y}_i(t) d\Lambda\{B(t) + z_i\beta + \boldsymbol{x}_i'\boldsymbol{\gamma}\}] = 0,$$

$$\sum_{i=1}^{n} [dN_i(t) - \tilde{Y}_i(t) d\Lambda\{B(t) + z_i\beta + \boldsymbol{x}_i'\boldsymbol{\gamma}\}] = 0.$$

The detailed iterative steps have been omitted for the sake of brevity.

Proof for $|\eta^*| < \log(1 + \epsilon_h)$ when $n_0, n_1 \to +\infty$

For brevity, we only show the proof when $n_0 = n_1$. The equation of (2.7) converges to

$$1 - \int_0^{+\infty} \frac{e^{\eta^*} S_0(t)}{S_1(t) + e^{\eta^*} S_0(t)} \left\{ -S_1'(t) - S_0'(t) \right\} dt = 0, \tag{A.1}$$

where $S_1(t) = S(t|z = 1)$, $S_0(t) = S(t|z = 0)$ follow POSM of (2.1), and $S_1'(t)$ and $S_0'(t)$ are the derivatives of $S_1(t)$ and $S_0(t)$ respectively.

Using $b_0(t) = dB_0(t)/dt$, in (2.7), we get $S_0(t) = \{1 + e^{\beta_0} B_0(t)\}^{-1}$, $S_1(t) = \{1 + B_0(t)\}^{-1}$ and $-S_1'(t) = \frac{e^{\beta_0} b_0(t)}{\{1+e^{\beta_0} B_0(t)\}^2}$, $-S_0'(t) = \frac{b_0(t)}{\{1+B_0(t)\}^2}$. Using these $S_j(t)$ and $-S_j'(t)$ for $j = 1, 2$ in (A.1), we can get the equation for η^* as

$$1 - \int_0^{+\infty} \frac{\{1 + B_0(t)\} e^{\eta^*}}{\{1 + B_0(t) e^{\beta_0}\} + \{1 + B_0(t)\} e^{\eta^*}} \times \left\{ \frac{b_0(t)}{\{1 + B_0(t)\}^2} + \frac{b_0(t) e^{\beta_0}}{\{1 + e^{\beta_0} B_0(t)\}^2} \right\} dt = 0.$$

With further Calculus manipulation, we show η^* to be the unique solution of

$$U(\eta) = e^{\eta} \log \left\{ \frac{1 + e^{\eta}}{e^{\beta_0} + e^{\eta}} \right\} - e^{\beta_0 - \eta} \log \left\{ \frac{e^{\beta_0} + e^{\eta}}{(1 + e^{\eta}) e^{\beta_0}} \right\} = 0,$$

where $U(\eta)$ is an decreasing (increasing) function for any fixed $\beta_0 > 0$ ($\beta_0 < 0$). Now recall that ϵ_h must satisfy $\delta = (1+\epsilon_h)^{-1/\epsilon_h} - (1+\epsilon_h)^{-(1+\epsilon_h)/\epsilon_h}$, and our true model is POSM of (2.1) with $\beta = \beta_0$ such that
$\delta = \frac{|e^{\beta_0/2} - e^{-\beta_0/2}|}{(1+e^{\beta_0/2})(1+e^{-\beta_0/2})}$ (margin for the true maximum of $|S_1(t) - S_0(t)|$). We can numerically show that $U(\eta) < 0$ for $\beta_0 \in (-\infty, 0) \cup (0, +\infty)$. Therefore, we can conclude that when $\beta_0 < 0$, $\eta^* > -\log(1+\epsilon_h)$ and when $\beta_0 > 0$, $\eta^* < \log(1+\epsilon_h)$. This is equivalent to $|\eta^*| < \log(1+\epsilon_h)$ when the true model is POSM with $\beta_0 \neq 0$.

APPENDIX B

PROOFS FROM CHAPTER 3

B.1 Proof of Theorem 3

Assuming that the true model is a POSM and considering the results from Struthers & Kalbfleisch (1986) with $z = 0$ and $z = 1$, the covariates values from standard treatment and new treatment, respectively, we obtain that η^* satisfies the equation (3.2).

Let $G_0(t) = [1 - S_0(t)]/S_0(t)$ and $\frac{d}{dt}G_0(t) = G_0'(t)$. By making the substitution $u = G_0(t)$ and considering $n_0 = n_1$ we can rewrite the left side of the equation (3.2) as

$$1 - \int_0^\infty \frac{\eta}{(1+u)[\eta(1+u)+\theta u+1]}du - \int_0^\infty \frac{\theta\eta(1+u)}{(1+\theta u)^2[\eta(1+u)+\theta u+1]}du. \tag{B.1}$$

The indefinite integrals for the integrals in (B.1) are

$$\frac{\eta}{1-\theta}\log\left[\frac{1+u}{\eta(1+u)+\theta u+1}\right] \quad \text{and} \quad \frac{\theta}{\eta(1-\theta)}\log\left[\frac{\theta u+1}{\eta(1+u)+\theta u+1}\right] - \frac{1}{\theta u+1}. \tag{B.2}$$

Finally the solution of (B.1) is

$$\eta\log\frac{1+\eta}{\theta+\eta} - \frac{\theta}{\eta}\log\frac{\theta+\eta}{(1+\eta)\theta} \doteq U(\theta,\eta). \tag{B.3}$$

Then η^* is a function of θ defined by $U(\theta,\eta^*) = 0$. For every θ and η^* we have that $\max_t\{S_0(t) - S_1(t)\} = (1+\theta^{-1/2})^{-1} - (1+\theta^{1/2})^{-1}$ and $\max_t\{S_0(t) - S_1(t)^{\eta^*}\} = \eta^{*\frac{1}{1-\eta^*}} - \eta^{*\frac{\eta^*}{1-\eta^*}}$. We can show numerically that $\max_t\{S_0(t) - S_1(t)\} - \max_t\{S_0(t) - S_1(t)^{\eta^*}\} > 0$, for all $\theta > 1$ concluding the proof.

Proof of $\eta^* < \rho$

For each θ it is possible to show that $U(\theta, \eta)$ given in (B.3) is a decreasing function, since $\theta > 1$. Taking $\rho = \Delta$, we can numerically show that $U(\Delta, e^{\hat{\beta}}) < 0$, for all $\Delta > 1$. Therefore, assuming a type I error η^* must satisfies $\eta^* < \rho$.

B.2 Proof of Theorem 4

The function $\max_t\{S_0(t) - S_1(t)\} = (1 + \theta^{-1/2})^{-1} - (1 + \theta^{1/2})^{-1}$ is a increasing function in θ. Then $S_0(t) - S_1(t) < \delta, \forall t \Leftrightarrow \max_t S_0(t) - S_1(t) < \delta \Leftrightarrow \theta < \left(\frac{1+\delta}{1-\delta}\right)^2$.

B.3 Proof of Theorem 5

Following the POSM parametrization given in (3.3) we have that $h_1(t)/h_0(t) = \theta[\theta - \theta S_0(t) + S_0(t)]^{-1} \doteq \Lambda(t, \theta)$. For fixed θ, $\max_t \Lambda(t, \theta) = 1$ if $0 \leq \theta < 1$ or θ if $\theta \geq 1$. Since $\rho \geq 1$ for non-inferiority tests, $h_1(t)/h_0(t) < \rho,\ \forall t$, is satisfied if $\theta < \rho$.

B.4 Iterative Steps for Parameter Estimation

Chen *et al.* (2002) estimator To estimate $(\hat{\beta}, \hat{B})$ of (β, B), iteratively solve the given set of estimating equations, where $B(t)$ is a non decreasing and non negative function with jumps only at observed failure times $t_{(1)} < \ldots < t_{(k)}$.

Step 0. Initiate a given value $\beta^{(0)}$ for β.

Step 1. Approximate $B^{(0)}(t_{(1)})$ by attaining

$$\sum_{i=1}^{n} \frac{Y_i(t_{(1)}) \exp(z_i\beta^{(0)})}{\exp(z_i\beta^{(0)}) + \exp\{-B(t_{(1)})\}} = 1,$$

where $Y_i(t)$ is the indicator that subject i is being observed at time t. After, obtain $B^{(0)}(t_{(k)})$, for $k = 2, \ldots, K$, by successively solving:

$$\sum_{i=1}^{n} \frac{Y_i(t_{(k)}) \exp(z_i\beta^{(0)})}{\exp(z_i\beta^{(0)}) + \exp\{-B(t_{(k)})\}} = 1 + \sum_{i=1}^{n} \frac{Y_i(t_{(k)}) \exp(z_i\beta^{(0)})}{\exp(z_\beta^{(0)}) + \exp\{-B(t_{(k-)})\}}.$$

Step 2. Then update estimate $\beta^{(1)}$ of β by solving

$$U(\beta, B) \equiv \sum_{i=1}^{n} z_i[d_i - \Lambda\{B(t) + z_i\beta\}] = 0,$$

$$\sum_{i=1}^{n} [d_i - \Lambda\{B(t) + z_i\beta\}] = 0,$$

where $\Lambda(u) = \{1 + \exp(-u)\}^{-1}$.

Step 3. Reiterate Step 1 and Step 2 until desired convergence criteria is met.

B.5 Proof of Fisher Information

Considering the POSM given by (3.3) we have that $S_1(t) = \frac{S_0(t)\theta}{1-S_0(t)+S_0(t)\theta}$ and $f_1(t) = \frac{f_0(t)\theta}{[F_0(t)+\theta-\theta F_0(t)]^2}$, where $F_0(t) = 1-S_0(t)$ and $f_0(t) = F_0'(t)$. Supposing that all observations are independent and that are valid some commonly regularity conditions, the Fisher information of θ, $I(\theta)$, is given by $I(\theta) = E([\frac{d}{d\theta}\sum_{i=1}^{n} \log f_i(T_i)]^2) = E(-\frac{d^2}{d\theta^2}\sum_{i=1}^{n} \log f_i(T_i))$, where $f_i(t_i)$ is the density function for the ith observation. Define $L(\theta) = \sum_{i=1}^{n} \log f_i(t_i)$, then $L(\theta) = \sum_{i=1}^{n_0} \log f_0(t_i) + \sum_{i=1}^{n_1}\{\log\theta - 2\log[F_0(t_i) + \theta - \theta F_0(t_i)]\}$, $\frac{d}{d\theta}L(\theta) = \frac{n_1}{\theta} - \sum_{i=1}^{n_1} \frac{2[1-F_0(t_i)]}{F_0(t_i)+\theta-\theta F_0(t_i)}$ and $\frac{d^2}{d\theta^2}L(\theta) = -\frac{n_1}{\theta^2} + 2\sum_{i=1}^{n_1} \frac{[1-F_0(t_i)]^2}{[F_0(t_i)-\theta-\theta F_0(t_i)]^2}$. Therefore

$$\begin{aligned} I(\theta) &= E\left(\tfrac{n_1}{\theta^2} + 2\textstyle\sum_{i=1}^{n_1} \frac{[1-F_0(t_i)]^2}{[F_0(t_i)-\theta-\theta F_0(t_i)]^2}\right) \\ &= \tfrac{n_1}{\theta^2} - \tfrac{2}{\theta^2}\textstyle\sum_{i=1}^{n_1} E\left(\left[\frac{\theta(1-F_0(t_i))}{F_0(t_i)-\theta-\theta F_0(t_i)}\right]^2\right) \\ &= \tfrac{n_1}{\theta^2} - \tfrac{2}{\theta^2}\textstyle\sum_{i=1}^{n_1} E(U_i^2), \end{aligned} \tag{B.4}$$

where $U_i = \frac{\theta(1-F_0(t_i))}{F_0(t_i)+\theta-\theta F_0(t_i)} = \frac{\theta S_0(t_i)}{1-S_0(t_i)+\theta S_0(t_i)}$. Since $U_i \sim \text{Uniform}(0,1)$, $E(U_i) = 1/3$, thus we conclude that $I(\theta) = \frac{n_1}{3\theta^2}$.

BIBLIOGRAPHY

(2000). World medical association declaration of helsinki. ethical principles for medical research involving human subjects. *JAMA*, (284), 3043–3045. 9

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, **2**, 273–277. ix, 11, 13, 28, 31, 33, 37, 41

Betensky, R. A., Louis, D. N. & Cairncross, J. G. (2002). Influence of unrecognized molecular heterogeneity on randomized clinical trials. *Journal of Clinical Oncology*, **20**(10), 2495–2499. 23, 27

Chan, I. S., Wang, W. & Heyse, J. (2003). Vaccine clinical trials. *Encyclopedia of Biopharmaceutical Statistics*, **2**, 1005–1022. 27

Chen, K., Jin, Z. & Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, **89**(3), 659–668. 16, 34, 44, 45, 48

Com-Nougue, C., Rodary, C. & Patte, C. (1993). How to establish equivalence when data are censored: A randomized trial of treatments for b non-hodgkin lymphoma. *Statistics in Medicine*, **12**(14), 1353–1364. 11, 20, 28, 30, 35

Cooner, F., Banerjee, S., Carlin, B. P. & Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association*, **102**(478), 560–572. 27

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187–220. ix, 10, 19, 27, 30, 37, 41

Freitag, G., Lange, S. & Munk, A. (2006). Non-parametric assessment of non-inferiority with censored data. *Statistics in Medicine*, **25**, 1201–1217. 27

Kalbfleisch, J. & Prentice, R. (2002). *The statistical analysis of failure time data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. J. Wiley. ISBN 9780471363576. 27

Kaplan EL, M. P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), 968–976. 3

Koti, K. M. (2013). New tests for assessing non-inferiority and equivalence from survival data. *Open Journal of Statistics*, **3**, 55–64. 28

Li Y, Adelstein DJ, e. a. (2003). An intergroup phase iii comparison of standard radiation therapy and two schedules of concurrent chemoradiotherapy in patients with unresectable squamous cell head and neck cancer. *J Clin Oncol*, (21), 9298. 23

Lin, D. Y. & Wei, L.-J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, **84**(408), 1074–1078. 19, 27

mo Nam, J., Kim, J. & Lee, S. (2005). Equivalence of two treatments and sample size determination under exponential survival model with censoring. *Computational Statistics and Data Analysis*, **49**(1), 217–226. 22, 23

Murphy SA, Rossini AJ, v. d. V. A. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, **92**(439), 968–976. 15, 16

Plikaytis, B. D. & Carlone, G. M. (2005). Statistical considerations for vaccine immunogenicity trials: Part 1: introduction and bioassay design and analysis. *Vaccine*, **23**(13), 1596–1605. 27

Rothmann, M., Li, N., Chen, G., Chi, G. Y. H., Temple, R. & Tsou, H.-H. (2003). Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine*, **22**, 239–264. 27

Rothmann, M. D., Wiens, B. L. & Chan, I. S. F. (2012). *Design and analysis of non-inferiority trials*. Chapman & Hall/CRC Biostatistics Series. CRC Press, Boca Raton, FL. ISBN 978-1-58488-804-8. 14, 27, 29

Royston, P. & Parmar, M. K. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*, **30**(19), 2409–2421. 27, 30

Rudser, K. D., LeBlanc, M. L. & Emerson, S. S. (2012). Distribution-free inference on contrasts of arbitrary summary measures of survival. *Statistics in Medicine*, **31**(16), 1722–1737. 27

Struthers, C. A. & Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, **73**(2), 363–369. 30, 47

Tian, L., Zhao, L. & Wei, L. J. (2013). On the restricted mean event time in survival analysis. *Havard University Biostatistics Working Paper Series*, **Working paper 156**. 30

Tsiatis, A. A. & Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, **90**(2), 367–378. 29

Wellek, S. (1993). A log-rank test for the equivalence of two survival functions. *Biometrics*, **49**, 877–881. ix, 9, 10, 11, 12, 17, 19, 20, 21, 27, 28, 35, 36

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. CRC Press, Boca Raton, FL, second edition. ISBN 978-1-4398-0818-4. 10, 28, 30

Zhang, X., Xu, J. & He, J. (2011). Assessing non-inferiority with time-to-event data via the method of non-parametric covariance. *Statistical Methods in Medical Research*, **22**, 346–360. 27

Zhao, L., Tian, L., Uno, H., Solomon, S. D., Pfeffer, M. A., Schindler, J. S. & Wei, L. J. (2012). Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical Trials*, **9**(5), 570–577. 28, 30

BIOGRAPHICAL SKETCH

Elvis Englebert Martinez was born November 10, 1977 in New York, New York. His family moved to Florida in 1993 and earned his GED in 1996. Elvis started Broward Community College in 1997. In 2005 Elvis started his education process at Tallahassee Community College where he graduated with honors in 2006. The same year Elvis enrolled at Florida State University. Elvis then went on to double major in Applied Mathematics and Statistics. With the encouragement of Dr. Steven Ramsier, he then started the graduate program at the Department of Statistics at Florida State University. He received his Master of Science degree in Biostatistics in August 2011. He defended his Ph.D. dissertation in November 2014. Elvis is the first person in his family to attend college.