NONPARAMETRIC AND SEMIPARAMETRIC INFERENCE ON QUANTILE LOST LIFESPAN

by

Lauren C. Balmert

B.S., Fairfield University, 2012

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Lauren C. Balmert

It was defended on

April 5, 2017

and approved by

Dissertation Advisor:

Jong H. Jeong, Ph.D.

Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Committee Members:

Jeanine M. Buchanich, M.Ed., Ph.D.

Research Associate Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Ying Ding, Ph.D.

Assistant Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Yu Cheng, Ph.D.
Associate Professor
Department of Statistics
Kenneth P. Dietrich School of Arts and Sciences
University of Pittsburgh

NONPARAMETRIC AND SEMIPARAMETRIC INFERENCE ON QUANTILE LOST LIFESPAN

Lauren C. Balmert, PhD

University of Pittsburgh, 2017

ABSTRACT

A new summary measure for time-to-event data, termed lost lifespan, is proposed in which the existing concept of reversed percentile residual life, or percentile inactivity time, is recast to show that it can be used for routine analysis to summarize life lost. The lost lifespan infers the distribution of time lost due to experiencing an event of interest before some specified time point. An estimating equation approach is adopted to avoid estimation of the probability density function of the underlying time-to-event distribution to estimate the variance of the quantile estimator. A K-sample test statistic is proposed to test the ratio of quantile lost lifespans. Simulation studies are performed to assess finite properties of the proposed statistic in terms of coverage probability and power. The concept of life lost is then extended to a regression setting to analyze covariate effects on the quantiles of the distribution of the lost lifespan under right censoring. An estimating equation, variance estimator, and minimum dispersion statistic for testing the significance of regression parameters are proposed and evaluated via simulation studies. The proposed approach reveals several advantages over existing methods for analyzing time-to-event data, which is illustrated with a breast cancer dataset from a Phase III clinical trial conducted by the National Surgical Adjuvant Breast and Bowel Project.

Public Health Significance: The analysis of time-to-event data can provide important information about new treatments and therapies, particularly in clinical trial settings. The methods provided in this dissertation will allow public health researchers to analyze effective-

ness of new treatments in terms of a new summary measure, life loss. In addition to providing statistical advantages over existing methods, analyzing time-to-event data in terms of the lost lifespan provides a more straightforward interpretation beneficial to clinicians, patients, and other stakeholders.

Keywords: Lost lifespan; residual life; survival analysis; time-to-event; right censoring.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to express my tremendous gratitude to my dissertation advisor, Dr. Jong-Hyeon Jeong. His guidance throughout this process has been invaluable. I am truly grateful for his support and very thankful for the opportunity to work with him. I would also like to thank my dissertation committee: Dr. Yu Cheng, Dr. Jeanine Buchanich, and Dr. Ying Ding, for their time and valuable feedback. I would especially like to thank Dr. Buchanich, as my graduate student research advisor, for the opportunities to be involved in the research process of a variety of studies.

I am also very grateful to the Department of Biostatistics for the opportunity to learn and grow as a student, researcher, and teaching fellow. I have had the pleasure of working with several faculty and staff who have aided in my professional development. Specifically, I would like to extend a sincere thank you to my former academic advisor, Dr. Sally Morton, for her mentorship throughout my education.

Finally, I would like to thank my family and friends for their continuous love and support.

1.0 INTRODUCTION

Time-to-event data, which focuses on the time until occurrence of an event of interest, can be encountered in many research areas such as engineering, economics, medicine, and social sciences. Statistical methods for analyzing this type of data have mainly considered cumulative information up to the time of analysis or residual information beyond the time of analysis. To our best knowledge, there exist no methods in the literature to analyze censored time-to-event data in terms of quantiles of time loss. Here, a summary measure termed *lost lifespan* is proposed to consider the time lost due to experiencing an event of interest before some specified time point. The contribution of the proposed methods is several-fold. First, the interpretation of the lost lifespan at a specific follow-up time point is more straightforward and more informative than that of existing methods, including the hazard function-based results. For example, a physician can explain an intervention effect as "Taking this drug is expected to reduce your life loss by 50% on average at 5 years after beginning treatment". In comparison, the interpretation of the hazard function-based results could be less transparent to the laymen because of the definition of the hazard function as the conditional limiting probability. The proposed method also offers important advantages over residual life based methods, which can be heavily influenced by censored observations. In the lost lifespan analysis, the observations beyond the fixed time point t_0 are excluded, so that it would be substantially less affected by heavy censoring at the tail of the distribution.

We propose here a nonparametric quantile-based method, which is more robust than the mean-based method for often asymmetric time-to-event data, yet our inference procedure does not require estimation of the probability density function of the true time-to-event distribution under censoring to evaluate the variance of the quantile. We then extend the methods to a quantile regression setting to allow for examination of covariate effects on the

lost lifespan. This extension is important for considering the relationship between the lost lifespan and a covariate of interest while adjusting for potential confounding factors.

Chapter 2 will provide a literature review of the existing methods related to nonparametric inference and quantile regression. Chapter 3 will introduce the quantile lost lifespan and provide a non-parametric estimation and inference procedure. Simulation studies will assess the proposed methods, and an application to a real data set will illustrate their use. The research presented in this Chapter has been previously published in Biometrics[1], a journal of the International Biometric Society, by Wiley-Blackwell Publishing (Balmert and Jeong, 2017). A regression model and method for testing significance of covariates will be presented in Chapter 4. The model will be evaluated via simulations studies and applied to real data. Chapter 5 will conclude with a brief discussion and future directions of the research.

[1]Available at http://onlinelibrary.wiley.com/doi/10.1111/biom.12555/full

2.0 LITERATURE REVIEW

2.1 NONPARAMETRIC INFERENCE LITERATURE REVIEW

The third chapter focuses on recasting the concept of reversed percentile residual life, or reversed inactivity time, to show how it can be used for routine analysis of time to event data to summarize 'life lost'. Time-to-event analysis can be based on cumulative information up to a given time point or residual information after the given time point. Popular summary measures for time-to-event outcomes from medical or reliability studies have been the hazard function and associated survival probability or quantile survival time. The adoption of residual life as a summary measure for time to event data stemmed from the interest of knowing the remaining life expectancy at a specified time point beyond the initial diagnosis or start of treatment. The mean residual life function, defined as $E(T-t|T \geq t)$, has been studied extensively in the literature (Berger et al., 1988; Chen et al., 1983; Chiang, 1960; Oakes and Dasu, 1990; Maguluri and Zhang, 1994; Chen et al., 2005). Here, T represents the event time, and t is the specified time point. More recently, there has been examination of the median life function, defined as $median(T-t|T \geq t)$ (Schmittlein and Morrison, 1981; Fligner and Rust, 1982; Wang and Hettmansperger, 1990; Su and Wei, 1993). The median is particularly important in analysis of skewed data, as it is more robust to outliers than the mean. An important nonparametric method was proposed for median residual life estimation using martingale increments, and an asymptotically chi-square test statistic was derived for comparing the ratio of median residual lifetimes between groups (Jeong et al., 2008). These methods were also extended to summarize other quantiles that may be of interest to investigators. The methods proposed here will be extended to the lost lifespan in Chapter 3. Further detail for the inference and regression procedures for quantile residual life has been documented in the literature (Jeong, 2014).

The hazard function, or conditional failure rate, has also been studied extensively. The traditional hazard function is defined as the instantaneous failure rate given that a subject did not experience an event of interest previously, or the ratio of the probability density function to the survival function (Klein and Moeschberger, 2003), ie for a non-negative random variable T,

$$h(t) = \lim_{\Delta t \to 0} \Pr(t - \Delta t \leq T \leq t | T \geq t) = \frac{f_T(t)}{S_T(t)}.$$

On the other hand, the *reversed* hazard function (Block et al., 2009; Chandra and Roy, 2001) specifies the instantaneous failure rate given that a subject did experience an event previously, i.e for a non-negative random variable T,

$$h_R(t) = \lim_{\Delta t \to 0} \Pr(t - \Delta t \leq T \leq t | T \leq t) = \frac{f_T(t)}{F_T(t)}, \tag{2.1.1}$$

where $f_T(\cdot)$ and $F_T(\cdot)$ are the probability density function and the cumulative distribution function of T, respectively. The importance of the reversed hazard function particularly in the presence of left censoring has also been discussed (Andersen et al., 1993; Kalbfleisch and Lawless, 1989). The reversed hazard function will be used to show in Chapter 3 that it characterizes the distribution of the proposed lost lifespan, also known as the inactivity time (Ruiz and Navarro, 1996; Li and Lu, 2003).

The concept of restricted mean lifetime has also been briefly addressed in the literature to identify the average time lived before some specified time point (Irwin, 1949; Karrison, 1987; Andersen et al., 2004; Royston and Parmar, 2011). Andersen, 2013 (Andersen, 2013) then defined "years lost" in terms of the restricted mean lifetime at a specified time point under competing risks, using the difference between the specified time point and the mean restricted lifetime. Specifically, the expected number of years lost before time τ can be described as,

$$L(0, \tau) = \tau - \int_0^\tau S(t)dt.$$

Despite this recent development, there have been no estimation or inference methods proposed for the quantile lost lifespan under the scenario of right-censored data.

2.2 QUANTILE REGRESSION LITERATURE REVIEW

The fourth chapter focuses on extending the lost lifespan to a quantile regression setting. Quantile regression, originally developed by Koenker and Basset (Koenker and Bassett, 1978), is a well studied extension of linear regression (Jung, 1996; Portnoy and Koenker, 1997). As defined in the literature (Koenker and Bassett, 1978), the θ-regression quantile is a solution to the minimization of

$$min_{b\in\mathbb{R}}\left[\sum_{t\in\{t:y_t\geq x_t b\}}\theta|y_t-x_t b|+\sum_{t\in\{t:y_t<x_t b\}}(1-\theta)|y_t-x_t b|\right]$$

of the regression process $\mu_t = y_t - x_t\beta$. Methods have also been established for analyzing survival data in the presence of censoring (Lindgren, 1997; McKeague et al., 2001; Yin and Cai, 2005; Peng and Huang, 2008). More recently, the concept of residual life has seen extensions to regression analysis. Covariate effects on residual life were examined under the proportional hazards and accelerated life models (Raja rao et al., 1992), regression models on the median failure time were proposed (Ying et al., 1995), and bayesian modeling was considered on median residual life (Gelfand and Kottas, 2003).

A method for quantile regression on the residual life function was recently developed, in which covariate effects on the quantile failure time among individuals surviving beyond a specified time point can be estimated (Jung et al., 2009). The proposed method is useful for situations where the researcher is interested in examining the residual life while controlling for important demographic, environmental, or medical factors. A log-linear regression method was considered for the ϵ-quantile, such that

$$\epsilon - quantile\{\ln(T_i - t_0)|T_i \geq t_0, Z_i\} = \beta'_{\epsilon|t_0} Z_i.$$

An asymptotically chi-square test statistic was proposed to test the significance of covariates within the model utilizing the minimum dispersion statistic of Basawa and Koul (Basawa and Koul, 1988). The methods proposed here will be extended to the lost lifespan in the fourth chapter. Further detail, including parametric and semi-parametric estimation of regression parameters, can be found in the literature (Jeong, 2014). This method has also been extended

to cause-specific quantile residual life regression (Lim and Jeong, 2015), and more recently to methods allowing for dynamic predictions (Li et al., 2016).

3.0 NONPARAMETRIC INFERENCE ON QUANTILE LOST LIFESPAN

3.1 INTRODUCTION

In Chapter 3, we introduce the lost lifespan distribution, providing a derivation and constructing an estimating equation to estimate the median lost lifespan. We propose an inference procedure through the confidence interval approach. A two-sample statistic is derived to compare the median lost lifespans between groups, utilizing the minimum dispersion statistic. The two-sample test statistic is also extended to K samples to compare multiple groups to a baseline group. Simulation studies are performed to demonstrate finite properties of the proposed two-sample statistic in terms of coverage probability and power. Finally, the proposed method is exemplified with a real data example from a breast cancer study. The benefits of the proposed method over traditional methods are discussed in the conclusion.

3.2 DERIVATION OF LOST LIFESPAN DISTRIBUTION

The reversed hazard function (Block et al., 2009; Chandra and Roy, 2001), as previously mentioned, specifies the instantaneous failure rate given that a subject did experience an event previously. It can be shown that the this function characterizes the distribution of the lost lifespan. Specifically define $T^* = t_0 - T$ to be a time loss due to an event occurence prior to t_0. Here t_0 can be any time point during an observation period whose maximum might be determined by administrative censoring. Therefore when T is defined as time to death, the lost lifespan can be interpreted as the time period being dead, in contrast to being alive, or life lost due to death prior to t_0. In general, it can be viewed as the time

span after occurrence of an event of interest, so that the longer it is, the less favorable it is to a patient in a disease setting. Figure 3.2 compares the traditional residual lifetime and the lost lifespan at a fixed time point t_0. At t_0, we can either consider the remaining lifetime given that an individual has not yet experienced an event (residual life), or the life lost given that an individual has already experienced an event (lost lifespan). As t_0 is shifted to the right, more information in included in the lost lifespan analysis.

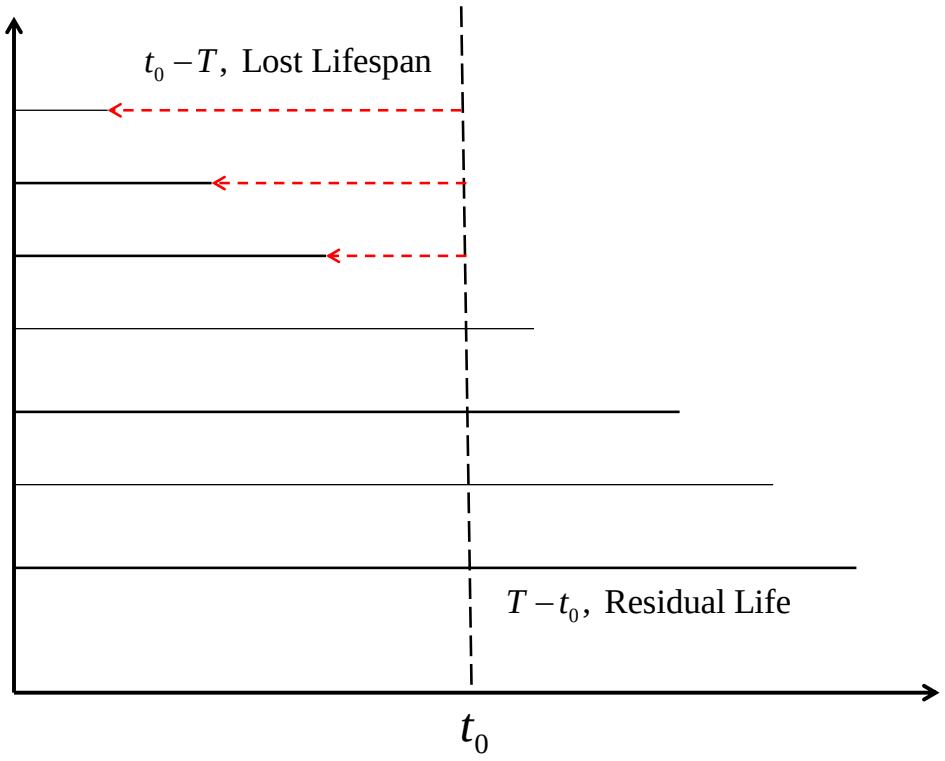

Figure 3.2.1: *Comparison of Residual Life and Lost Lifespan at* t_0

Now we show that the hazard function of the lost lifespan distribution is characterized by the reversed hazard function defined in Equation (2.1.1). The cumulative distribution function of T^* can be defined as

$$\begin{aligned} F_{T^*}(s) &= \Pr(t_0 - T \leq s | T \leq t_0) \\ &= \frac{F_T(t_0) - F_T(t_0 - s)}{F_T(t_0)}. \end{aligned} \tag{3.2.1}$$

Since the probability density function of T^* is

$$f_{T*}(s) = \frac{dF_{T^*}(s)}{ds} = \frac{f_T(t_0 - s)}{F_T(t_0)},$$

its hazard function is given by

$$h_{T^*}(s) = \frac{f_{T^*}(s)}{1 - F_{T^*}(t_0 - s)} = \frac{f_T(t_0 - s)}{F_T(t_0 - s)},$$

which is the reversed hazard function of the original random variable T at a backward time point $t_0 - s$.

From Equation (3.2.1), the λ-percentile, $\theta_{t_0}^{(\lambda)}$, of the lost lifespan distribution can be defined by solving

$$\frac{F_T(t_0) - F_T\left(t_0 - \theta_{t_0}^{(\lambda)}\right)}{F_T(t_0)} = \frac{S_T\left(t_0 - \theta_{t_0}^{(\lambda)}\right) - S_T\left(t_0\right)}{1 - S_T(t_0)} = \lambda. \tag{3.2.2}$$

With a continuous survival function $S_T(\cdot) = 1 - F_T(\cdot)$, Equation (3.2.2) provides the solution

$$\theta_{t_0}^{(\lambda)} = t_0 - S_T^{-1}\left[\lambda + (1 - \lambda) S_T(t_0)\right]. \tag{3.2.3}$$

For notational simplicity and without loss of generality, we will denote the median lost lifespan throughout the paper as

$$\theta_{t_0} = \text{median}(t_0 - T | T \leq t_0),$$

which can be interpreted as the median time loss among individuals who experienced an event before time t_0. While the estimation and inference procedures in the following sections will focus on the median, the results can be easily adapted for other quantiles of interest by adjusting λ.

3.3 INFERENCE ON QUANTILE LOST LIFESPAN: ONE SAMPLE CASE

3.3.1 Notation

Here, notation used throughout the third chapter will be defined. Let T_{ik} denote the event time of the ith patient in group k ($k = 1, 2$) and n_k the the number of patients in the kth group. In the situation where observations are censored prior to experiencing the event, C_{ik} will denote the censoring time. Thus the observed information for each patient will be $X_{ik} = \min(C_{ik}, T_{ik})$ with an event indicator $\Delta_{ik} = I(T_{ik} < C_{ik})$. We will make the assumption that event times are independent of censoring times. $S_k(t)$ will be the survival function of T_{ik} for the kth group and $\hat{S}_k(t)$ the corresponding Kaplan-Meier estimator. Let $Y_{ki}(t) = I(X_{ki} \geq t)$ and $N_{ki}(t) = \Delta_{ki} I(X_{ki} \leq t)$ be the at-risk and event processes, respectively, for patient i in group k. Also, we define $Y_k = \sum_{i=1}^{n_k} Y_{ki}$ and $N_k = \sum_{i=1}^{n_k} N_{ki}$.

3.3.2 Estimation

Suppressing the subscript k, from Equation (3.2.2) the median of the lost lifespan distribution at t_0 can be estimated for each group by solving $\hat{u}(\theta_{t_0}) = 0$ for θ_{t_0} where

$$\hat{u}(\theta_{t_0}) = \hat{S}(t_0 - \theta_{t_0}) - \frac{1}{2}\hat{S}(t_0) - \frac{1}{2}, \tag{3.3.1}$$

with $\hat{S}(\cdot)$ being the Kaplan-Meier estimate of the true event time distribution. Thus, the median lost lifespan can be nonparametrically estimated by

$$\hat{\theta}_{t_0} = t_0 - \hat{S}^{-1}[(1/2)\{1 + \hat{S}(t_0)\}].$$

Let $\theta_{t_0,0}$ denote the true median lost lifespan at time t_0 and $S^*(\cdot)$ the true survival function, so that

$$u(\theta_{t_0,0}) = S^*(t_0 - \theta_{t_0,0}) - \frac{1}{2}S^*(t_0) - \frac{1}{2} = 0.$$

By rearranging (3.3.1) at $\theta_{t_0,0}$, we have

$$\begin{aligned}
\hat{u}(\theta_{t_0,0}) &= \hat{S}(t_0 - \theta_{t_0,0}) - S^*(t_0 - \theta_{t_0,0}) - \frac{1}{2}[\hat{S}(t_0) - S^*(t_0)] + S^*(t_0 - \theta_{t_0,0}) - \frac{1}{2}S^*(t_0) - \frac{1}{2} \\
&= \hat{S}(t_0 - \theta_{t_0,0}) - S^*(t_0 - \theta_{t_0,0}) - \frac{1}{2}[\hat{S}(t_0) - S^*(t_0)],
\end{aligned}$$

which can be represented as the sum of independent martingales by Corollary 3.2.1 of Fleming and Harrington (1991)

$$\hat{u}(\theta_{t_0,0}) = -\sum_{i=1}^{n} S^*(t_0 - \theta_{t_0,0}) \int_0^{t_0-\theta_{t_0,0}} \frac{dM_i(s)}{Y(s)} + \frac{1}{2}\sum_{i=1}^{n} S^*(t_0) \int_0^{t_0} \frac{dM_i(s)}{Y(s)} + o_p(n^{-\frac{1}{2}}),$$

where $M_i(t) = N_i(t) - \int_0^t Y_i(s)d\Lambda(s)$ is a martingale with the cumulative hazard function $\Lambda(s)$, so that $E[dM_i(t)|F_{t^-}] = 0$ for filtration $\{F_t : t \geq 0\}$. The $o_p(n^{-\frac{1}{2}})$ term indicates that remainder terms, multiplied by $n^{\frac{1}{2}}$, will converge in probability to 0. Also, $Y(t)/n$ uniformly converges to $y(t)$ over $[0, \xi]$, where ξ is the maximum follow-up time, so we have

$$\hat{u}(\theta_{t_0,0}) = \sum_{i=1}^{n} \epsilon_i + o_p(n^{-\frac{1}{2}}),$$

where

$$\epsilon_i = -S^*(t_0 - \theta_{t_0,0}) \int_0^{t_0-\theta_{t_0,0}} \frac{dM_i(s)}{ny(s)} + \frac{1}{2}S^*(t_0) \int_0^{t_0} \frac{dM_i(s)}{ny(s)}.$$

Again since $u(\theta_{t_0,0}) = 0$ at the true value $\theta_{t_0,0}$, substituting in $\frac{1}{2}S^*(t_0) + \frac{1}{2}$ for $S^*(t_0 - \theta_{t_0,0})$ alternatively gives

$$\epsilon_i = \frac{1}{2}S^*(t_0) \int_{t_0-\theta_{t_0,0}}^{t_0} \frac{dM_i(s)}{ny(s)} - \frac{1}{2}\int_0^{t_0-\theta_{t_0,0}} \frac{dM_i(s)}{ny(s)}.$$

Note that ϵ_1,...,ϵ_n are independent random variables with mean 0 and variance $\sigma^2_{t_0}$. This follows from martingale theory where for filtration $\{F_t : t \geq 0\}$,

$$E(dM_i(t)) = E\{E(dM_i(t)|F_t)\} = 0.$$

The variance can then be estimated by $\hat{\sigma}^2_{t_0} = \sum_{i=1}^n \hat{\epsilon}_i^2$ with

$$\hat{\epsilon}_i = \frac{1}{2}\hat{S}(t_0) \int_{t_0-\hat{\theta}_{t_0}}^{t_0} \frac{d\hat{M}_i(s)}{Y(s)} - \frac{1}{2}\int_0^{t_0-\hat{\theta}_{t_0}} \frac{d\hat{M}_i(s)}{Y(s)},$$

where

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s)d\hat{\Lambda}(s),$$

and $\hat{\Lambda}(s) = \int_0^t Y^{-1}(s)dN(s)$ is the Nelson-Aalen Estimator (Nelson, 1972; Aalen, 1978). By applying the ordinary Central Limit Theorem, a one-sample test statistic $\hat{u}(\theta_{t_0})^2/\hat{\sigma}^2_{t_0}$ follows a χ^2-distribution with 1 degree of freedom. A large value of the proposed statistic would

reject the null hypothesis, $H_0 : \theta_{t_0} = \theta_{t_0,0}$. The test statistic has been developed based on a martingale representation of the entire estimating equation, instead of relying on the underlying event distribution. Finally, a 100(1-α)% confidence interval can be constructed by inverting

$$\left\{\theta_{t_0} : \frac{\hat{u}(\theta_{t_0})^2}{\hat{\sigma}^2_{t_0}} < \chi^2_{1,1-\alpha}\right\}. \tag{3.3.2}$$

3.4 TWO SAMPLE TEST STATISTIC AND CONFIDENCE INTERVAL

In order to compare two groups, we could consider the ratio of the median lost lifespans at t_0. If we are interested in the case of an equal median lost lifespans, then the hypotheses can be specified as H_0: $\tau_{t_0} = 1$ versus H_1: $\tau_{t_0} \neq 1$, where $\tau_{t_0} = \theta_{2,t_0}/\theta_{1,t_0}$, with θ_{k,t_0} representing the median lost lifespan for the kth group ($k = 1, 2$). The estimating function for group k is defined as

$$\hat{u}_k(\theta_{k,t_0}) = \hat{S}_k(t_0 - \theta_{k,t_0}) - \frac{1}{2}\hat{S}_k(t_0) - \frac{1}{2},$$

so that under the null hypothesis of $\theta_{2,t_0} = \tau_{t_0}\ \theta_{1,t_0}$ we have the two-sample test statistic

$$W_{t_0}(\tau_{t_0}, \theta_{1,t_0}) = \frac{\hat{u}^2_1(\theta_{1,t_0})}{\hat{\sigma}^2_{1,t_0}} + \frac{\hat{u}^2_2(\tau_{t_0}\theta_{1,t_0})}{\hat{\sigma}^2_{2,t_0}}, \tag{3.4.1}$$

where $\hat{\sigma}^2_{k,t_0}$ $(k = 1, 2)$ is the variance estimate of $\hat{u}_k(\theta_{k,t_0})$ for group k, and θ_{2,t_0} is replaced in the second term with $\tau_{t_0}\ \theta_{1,t_0}$ from the null hypothesis.

A well-known approach to eliminating the unknown nuisance parameter θ_{1,t_0} from the test statistic (3.4.1) would be to minimimize the statistic over the parameter θ_{1,t_0}, resulting in the minimum dispersion statistic (Basawa and Koul, 1988). For any given time t_0, under the null hypothesis of $H_0 : \tau_{t_0} = \tau_{t_0,0}$ it can be shown (Jeong et al., 2008) that

$$Q_{t_0}(\tau_{t_0,0}) = \inf_{\theta_{1,t_0}} W_{t_0}(\tau_{t_0,0}, \theta_{1,t_0}) \tag{3.4.2}$$

follows asymptotically a χ^2_1-distribution. Specifically, for group k $(= 1, 2)$, let n_k denote the sample size and $\hat{\theta}_k$ denote the estimator of the median residual lifetime, θ_{k0}, at time t. For $n = n_1+n_2$, $n_k/n \to p_k \in (0, 1)$ as $n \to \infty$. Suppose that $t+\theta_{k0} < \xi = \sup\{t : S(t)G(t) > 0\}$.

Following similar arguments as in the Appendix of Su and Wei (Su and Wei, 1993), we have

$$\widehat{S}_1(t-\theta_{10}) - \frac{1}{2}\widehat{S}_1(t) - \frac{1}{2} = f_1(t-\theta_{10})(\theta_{10}-\hat{\theta}_1) + o_p(n^{-1/2}) \tag{3.4.3}$$

using a Taylor series expansion, where $f_k(\cdot)$ is the probability density function of $S_k(\cdot)$. Let $\tau_0 = \theta_{20}/\theta_{10}$ and $\hat{\tau} = \hat{\theta}_2/\hat{\theta}_1$. Then, similarly for group 2,

$$\widehat{S}_2(t-\theta_{20}) - \frac{1}{2}\widehat{S}_2(t) - \frac{1}{2} = f_2(t-\theta_{20})(\theta_{20}-\hat{\theta}_2) + o_p(n^{-1/2})$$

$$= f_2(t-\theta_{20})\{(\theta_{10}-\hat{\theta}_1)(\tau_0-\hat{\tau}) + \theta_{10}(\tau_0-\hat{\tau}) + \tau_0(\theta_{10}-\hat{\theta}_1)\} + o_p(n^{-1/2}).$$

By the consistency of $\hat{\theta}_k$, $(\theta_{10}-\hat{\theta}_1)(\tau_0-\hat{\tau}) = O_p(n^{-1})$, so that

$$\widehat{S}_2(t-\theta_{20}) - \frac{1}{2}\widehat{S}_2(t) - \frac{1}{2} = f_2(t-\theta_{20})\{\theta_{10}(\tau_0-\hat{\tau}) + \tau_0(\theta_{10}-\hat{\theta}_1)\} + o_p(n^{-1/2}). \tag{3.4.4}$$

From 3.4.3 and 3.4.4, we have $W_t(\tau_0,\theta_{10}) = V_t(\tau_0,\theta_{10}) + o_p(n^{-1})$, where

$$V_t(\tau_0,\theta_{10}) = (\tau_0-\hat{\tau},\theta_{10}-\hat{\theta}_1)\Gamma_t' \begin{pmatrix} \sigma_1^2(\theta_{10}) & 0 \\ 0 & \sigma_2^2(\theta_{20}) \end{pmatrix} \Gamma_t \begin{pmatrix} \tau_0-\hat{\tau} \\ \theta_{10}-\hat{\theta}_1 \end{pmatrix}$$

and

$$\Gamma_t = \begin{pmatrix} 0 & f_1(t+\theta_{10}) \\ \theta_{10}f_2(t+\theta_{20}) & \tau_0 f_2(t+\theta_{20}) \end{pmatrix}.$$

Therefore, the minimum of $W_t(\tau_0,\theta_1)$ with respect to θ_1 is asymptotically equivalent to $\inf_{\theta_1} V_t(\tau_0,\theta_1)$, which is $(\tau_0-\hat{\tau})^2/\mathrm{var}(\hat{\tau})$ and is asymptotically distributed as χ_1^2 by the delta-method. Thus, we reject the null hypothesis with type 1 error probability of α if $Q_{t_0}(\tau_{t_0,0})$ $> \chi^2_{1,1-\alpha}$. An important advantage of using this type of statistic is that there is no need for estimating the underlying probability density function of event times under censoring to make inference about the ratio of the two median lost lifespans. From (3.4.2), a 100(1-α)% confidence interval for τ_{t_0} can be obtained from

$$\{\tau_{t_0} : Q_{t_0}(\tau_{t_0}) < \chi^2_{1,1-\alpha}\}. \tag{3.4.5}$$

Note that to achieve a confidence interval from (3.4.2), the statistic $W_{t_0}(\tau_{t_0},\theta_{1,t_0})$ first needs to be minimized over θ_{1,t_0} for each fixed value of τ_{t_0}, and then the values of τ_{t_0} corresponding to the range where the value of $\chi^2_{1,1-\alpha}$ exceeds the minimum dispersion statistic $Q_{t_0}(\tau_{t_0,0})$ will form the desired confidence interval.

3.4.1 Extension to K-Sample Case

In this section, the proposed two-sample statistic is extended to K samples for comparisons among multiple groups. We are interested in testing the null hypothesis of $H_0 : \frac{\theta_2}{\theta_1} = \frac{\theta_3}{\theta_1} = ... = \frac{\theta_k}{\theta_1} = 1$, the equality of the median lost lifespans from $K-1$ groups (θ_2, θ_3, ... , θ_K) being compared simultaneously to one from a reference group (θ_1). The alternative hypothesis would then be H_1: at least one of the median lost lifespans is different from the reference group. This would be analogous to a regression model including one covariate with K categories that requires creating $K-1$ dummy variables where each pairwise comparison is performed relative to the reference group. Specifically, let us consider a log-linear regression model in the median lost lifespan at time t_0,

$$\text{med}\{\log(t_0 - T_i)|T_i \leq t_0, x_{1i}\} = \beta_{t_0}^{(0)} + \beta_{t_0}^{(1)} x_{1i}, \tag{3.4.6}$$

where x_{1i} is a binary covariate, say, 0 for control group and 1 for treatment group. Because the natural logarithm is a monotone transformation and by the invariance property of the median with respect to monotone transformations, the model (3.4.6) is equivalent to

$$\text{med}(t_0 - T_i|T_i \leq t_0, x_{1i}) = \exp(\beta_{t_0}^{(0)} + \beta_{t_0}^{(1)} x_{1i}). \tag{3.4.7}$$

Here $\theta_0 = \exp(\beta_{t_0}^{(0)})$ and $\theta_1 = \exp(\beta_{t_0}^{(0)} + \beta_{t_0}^{(1)})$ can be interpreted as the median lost lifespans for the control and treatment groups at time t_0, respectively, so that the slope parameter $\beta_{t_0}^{(1)}$ can be interpreted as the logarithm of the ratio of the two median lost lifespans at time point t_0. Therefore testing the null hypothesis of $H_0 : \beta_{t_0}^{(1)} = 0$ would be equivalent to testing $H_0 : \theta_1/\theta_0 = 1$. If x_{1i} has K categories, then $K-1$ dummy variables need to be created and each summary variable compares the indicated category to the reference group.

Therefore under the null hypothesis of $H_0 : \frac{\theta_2}{\theta_1} = \frac{\theta_3}{\theta_1} = ... = \frac{\theta_k}{\theta_1} = \tau_{t_0,0}$ in general, we propose a nonparametric K-sample test statistic

$$W_{t_0}(\tau_{t_0,0}, \theta_{1,t_0}) = \frac{\hat{u}_1^2(\theta_{1,t_0})}{\hat{\sigma}_{1,t_0}^2} + \sum_{k=2}^{K} \frac{\hat{u}_k^2(\tau_{t_0,0}\theta_{1,t_0})}{\hat{\sigma}_{k,t_0}^2}. \tag{3.4.8}$$

Again, for any given time t_0 it can be shown (Jeong et al., 2008) that

$$Q_{t_0}(\tau_{t_0,0}) = \inf_{\theta_{1,t_0}} W_{t_0}(\tau_{t_0,0}, \theta_{1,t_0}) \tag{3.4.9}$$

asymptotically follows a χ^2-distribution with $K-1$ degrees of freedom. We reject the null hypothesis with type 1 error probability of α if $Q_{t_0}(\tau_{t_0,0}) \geq \chi^2_{K-1,1-\alpha}$.

3.5 SIMULATION STUDIES

3.5.1 Type I Errors

First a simulation study was performed to assess the proposed two-sample test statistic under the null hypothesis $H_0 : \tau_{t_0} = 1$, or equal median lost lifespans. Empirical coverage probabilities at the significance level of 5% and the median lengths of 95% confidence intervals for the ratio (τ_{t_0}) of two median lost lifespans were considered. The empirical coverage probability is how often the estimated confidence intervals from the two-sample test statistic include the null value of the lost lifespan ratio. Event times were simulated from the Weibull distribution with survival function

$$S(t) = \exp\{-(\rho t)^{\eta}\},$$

where ρ and η are the scale and shape parameters set equal to .2 and 2, respectively. Specifically, event times were generated through the probability integral transformation,

$$T_i = \frac{1}{\rho}(-\log(1-V_i))^{\frac{1}{\eta}},$$

where V_i is a uniform random variable over the interval (a, b), with a and b determining the desired censoring proportion. For example, to generate samples with 10% censoring proportion, a and b were set to 4 and 15, respectively. One thousand (1,000) samples were simulated with 4 different scenarios of censoring proportions (0%, 10%, 20%, 30%) and 3 different sample sizes (n = 50, 100, 200) at 4 different time points (t_0 = 5, 6, 7, 8). The results are displayed in Table 3.5.1.

Table 3.5.1: *Empirical 95% coverage probabilities and median lengths of empirical 95% confidence intervals of the two-sample test statistic for comparing the median lost lifespans*

		Censoring Proportion			
t_0	Obs.	0%(ML)	10%(ML)	20%(ML)	30%(ML)
5	50	0.969(1.18)	0.971(1.20)	0.972(1.23)	0.980(1.29)
	100	0.971(0.79)	0.970(0.79)	0.970(0.82)	0.977(0.85)
	200	0.969(0.52)	0.968(0.53)	0.969(0.55)	0.969(0.58)
6	50	0.969(0.90)	0.968(0.92)	0.974(0.96)	0.975(1.02)
	100	0.969(0.64)	0.971(0.64)	0.977(0.66)	0.978(0.70)
	200	0.969(0.43)	0.968(0.43)	0.971(0.45)	0.972(0.48)
7	50	0.974(0.74)	0.971(0.75)	0.978(0.79)	0.984(0.82)
	100	0.972(0.51)	0.972(0.51)	0.973(0.54)	0.973(0.58)
	200	0.962(0.35)	0.960(0.35)	0.971(0.37)	0.972(0.39)
8	50	0.970(0.60)	0.974(0.62)	0.977(0.65)	0.969(0.68)
	100	0.972(0.42)	0.969(0.42)	0.967(0.45)	0.972(0.47)
	200	0.964(0.29)	0.957(0.29)	0.964(0.31)	0.967(0.32)

The coverage probabilities increase slightly as the censoring proportion increases, but no consistent pattern exists as t_0 changes. Although coverage probabilities often decrease towards 95% as sample size increases, they generally remain conservative. The median lengths of empirical 95% confidence intervals become narrower as sample size increases and as t_0 increases. The latter phenomenon is due to the fact that more lost lifespan observations are included in the analysis as t_0 increases.

Additionally, a simulation study was performed to assess the proposed K-sample test statistic to test the null hypothesis $H_0 : \frac{\theta_2}{\theta_1} = \frac{\theta_3}{\theta_1} = 1$, or to compare the median lost lifespans among 3 groups. Empirical coverage probabilities at the significance level of 5% and the median lengths of 95% confidence intervals are summarized in Table 3.5.2, where the observations represent the sample size in each of the three groups.

Similarly to the 2-sample case, coverage probabilities generally decrease towards 95% as sample size increases, but they remain conservative. Median lengths of 95% confidence intervals become narrower as sample size increases and t_0 increases. Compared to the two-sample case, the median lengths are generally narrower in the three-sample simulations, due to the increase in total sample size with the addition of the third group.

Table 3.5.2: *Empirical 95% coverage probabilities and median lengths of empirical 95% confidence intervals of the three-sample test statistic for comparing the median lost lifespans*

		Censoring Proportion			
t_0	Obs.	0%(ML)	10%(ML)	20%(ML)	30%(ML)
5	50	0.973(0.87)	0.975(0.88)	0.976(0.93)	0.972(0.97)
	100	0.979(0.59)	0.976(0.59)	0.978(0.61)	0.976(0.64)
	200	0.964(0.40)	0.964(0.40)	0.965(0.42)	0.968(0.44)
6	50	0.973(0.67)	0.972(0.68)	0.973(0.73)	0.973(0.76)
	100	0.967(0.47)	0.963(0.47)	0.968(0.50)	0.975(0.53)
	200	0.966(0.32)	0.969(0.33)	0.969(0.35)	0.968(0.36)
7	50	0.973(0.55)	0.966(0.55)	0.974(0.58)	0.976(0.62)
	100	0.967(0.39)	0.966(0.39)	0.969(0.41)	0.967(0.43)
	200	0.969(0.26)	0.968(0.27)	0.965(0.28)	0.964(0.30)
8	50	0.997(0.45)	0.974(0.46)	0.972(0.48)	0.986(0.52)
	100	0.962(0.32)	0.959(0.32)	0.968(0.34)	0.965(0.36)
	200	0.968(0.21)	0.967(0.22)	0.966(0.23)	0.967(0.24)

3.5.2 Power Analysis

The parametric proportional hazards model (Cox, 1972), with a single covariate, was used to perform power analyses. Event times for the control and intervention groups were generated as previously described, but β was introduced to simulate differences in survival distributions between the two groups. Thus, given a binary covariate z as an indicator for the intervention group ($z = 0$ for control; $z = 1$ for intervention), the survival function was specified as

$$S(t; z) = \exp\{-(\rho t)^{\eta} \exp(\beta z)\},$$

where ρ and η are the Weibull parameters, which were set equal to .2 and 2, and β is a regression coefficient associated with z. The median lost lifespan function under this model is given by

$$\theta_{t_0}(z) = t_0 - \frac{1}{\rho}[\exp(-\beta z)\{\log(2) - \log(1 + \exp(-(\rho t_0)^{\eta} \exp(\beta z))\}]^{\eta}. \tag{3.5.1}$$

Table 3.5.3 shows the values of $\theta_{t_0}(z)$ at different combinations of t_0 and β obtained from (3.5.1). Various scenarios were considered, with β ranging from 0 to -3.0. Under the null hypothesis of $H_0 : \beta = 0$, the median lost lifespan increases as t_0 increases, as expected. On the other hand, as β decreases, the distribution of event times shifts to the right in the intervention group, and hence the median lost lifespan decreases. Thus, the ratio of the median lost lifespans between the intervention group and the control group decreases as β decreases. For example, at $t_0 = 5$, the median lost lifespans corresponding to $\beta = 0$, -0.9, -1.2, and -1.5 are 1.92, 1.65, 1.60, and 1.56, respectively. Therefore, the true values of the lost lifespan ratio corresponding to β = -0.9, -1.2, and -1.5 are 0.86, 0.83, and 0.81, respectively.

Table 3.5.3: *True median lost lifespans at different combinations of β and t_0*

	β						
t_0	0	-0.5	-0.9	-1.2	-1.5	-2.0	-3.0
5	1.92	1.74	1.65	1.60	1.56	1.52	1.49
6	2.53	2.23	2.08	1.99	1.93	1.86	1.80
7	3.25	2.80	2.56	2.42	2.33	2.22	2.11
8	4.07	3.46	3.10	2.90	2.76	2.59	2.43
9	4.95	4.19	3.71	3.43	3.23	2.99	2.77
10	5.89	4.99	4.38	4.02	3.74	3.42	3.11
11	6.86	5.86	5.12	4.66	4.30	3.88	3.46
12	7.85	6.77	5.92	5.36	4.91	4.36	3.82

The results of the power analyses under 3 different sample sizes (n = 50, 100, 200) and 4 different time points (t_0 = 5, 6, 7, 8) are displayed for 10% censoring in Table 3.5.4, 20% censoring in Table 3.5.5, and 30% censoring in Table 3.5.6. Power generally increased as β decreased, indicating higher power to detect a larger difference between the control and intervention groups. Additionally, power and hence efficiency generally increased with larger t_0, as more information was gained by shifting t_0 to the right. Increasing the censoring proportion, as seen in the different tables, slightly reduced the power possibly due to a decrease in the number of events. Finally, increasing sample size was obviously associated with higher power.

Table 3.5.4: *Empirical powers of the two-sample test statistic for comparing the median lost lifespans at a 5% significance level with 10% censoring*

t_0	n	β = -0.9	β = -1.2	β = -1.5
5	50	0.031	0.046	0.043
	100	0.068	0.065	0.066
	200	0.111	0.112	0.120
	500	0.268	0.306	0.306
6	50	0.058	0.071	0.081
	100	0.124	0.157	0.153
	200	0.256	0.304	0.320
	500	0.602	0.682	0.697
7	50	0.118	0.158	0.150
	100	0.253	0.288	0.313
	200	0.507	0.594	0.646
	500	0.903	0.951	0.954
8	50	0.209	0.265	0.295
	100	0.434	0.546	0.576
	200	0.771	0.870	0.892
	500	0.990	0.994	1.000

Table 3.5.5: *Empirical powers of the two-sample test statistic for comparing the median lost lifespans at a 5% significance level with 20% censoring*

t_0	n	β = -0.9	β = -1.2	β = -1.5
5	50	0.030	0.041	0.041
	100	0.064	0.064	0.066
	200	0.104	0.114	0.083
	500	0.260	0.310	0.117
6	50	0.063	0.063	0.085
	100	0.119	0.149	0.145
	200	0.253	0.288	0.304
	500	0.582	0.667	0.686
7	50	0.112	0.149	0.153
	100	0.245	0.284	0.315
	200	0.491	0.578	0.603
	500	0.902	0.941	0.952
8	50	0.207	0.264	0.302
	100	0.419	0.539	0.547
	200	0.745	0.856	0.887
	500	0.990	0.996	1.000

Table 3.5.6: *Empirical powers of the two-sample test statistic for comparing the median lost lifespans at a 5% significance level with 30% censoring*

t_0	n	β = -0.9	β = -1.2	β = -1.5
5	50	0.029	0.044	0.042
	100	0.069	0.065	0.055
	200	0.097	0.110	0.114
	500	0.251	0.292	0.294
6	50	0.059	0.060	0.072
	100	0.113	0.139	0.143
	200	0.225	0.283	0.268
	500	0.542	0.649	0.680
7	50	0.095	0.141	0.149
	100	0.217	0.262	0.309
	200	0.446	0.566	0.596
	500	0.870	0.932	0.943
8	50	0.194	0.269	0.294
	100	0.381	0.495	0.534
	200	0.697	0.834	0.872
	500	0.983	0.993	0.999

3.6 APPLICATION TO NSABP B-04 DATA

In this section, we apply the proposed method to a real data set from a phase III clinical study on breast cancer, conducted by the National Surgical Adjuvant Breast and Bowel Project (NSABP). The data set, referred to as the NSABP B-04 data, includes 1,665 women with over 25 years of follow-up (Fisher et al., 2002). The original study was designed to compare a radical mastectomy (RM) with a less intensive total mastectomy (TM). As shown in Figure 3.6, the 1,079 node-negative patients were randomly assigned to either a radical mastectomy, total mastectomy, or total mastectomy with regional irradiation (TMR). The 586 node-positive patients were randomly assigned to either a radical mastectomy or total mastectomy with regional irradiation.

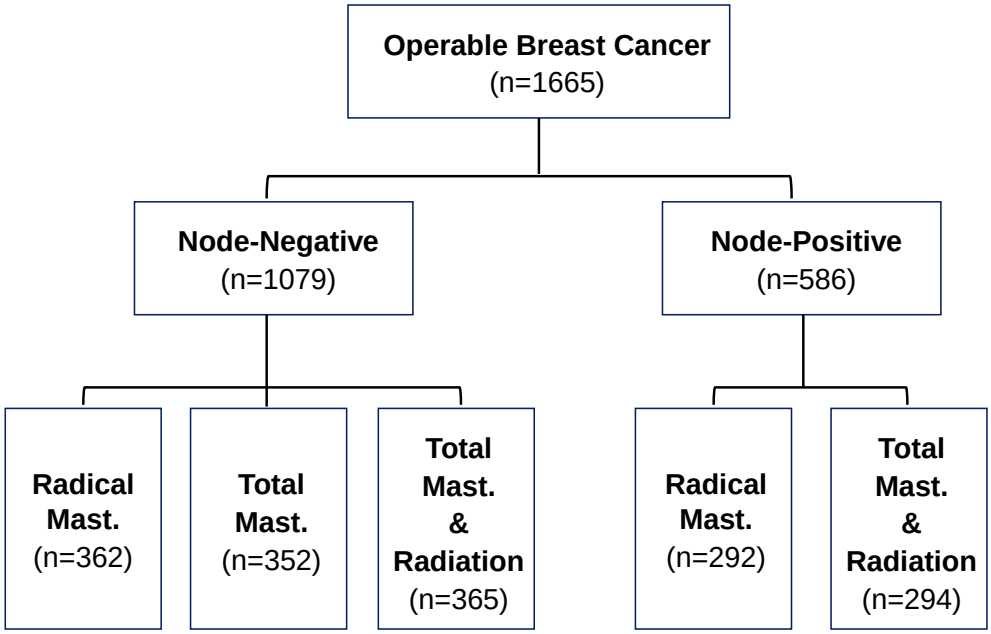

Figure 3.6.1: *NSABP B-04 Study Design*

(Adapted from *Fisher et al. (1981)*)

In the original study, the data were analyzed with Kaplan Meier curves and Cox proportional hazards models (Fisher et al., 2002). End points considered were disease-free survival, relapse-free survival, distant-disease-free survival, and overall survival. Follow-up time began from the date of mastectomy. The following two figures replicate the Kaplan-Meier curves for overall survival by treatment for the node-negative group (Figure 3.6) and node-positive group (Figure 3.6), adapted from Fisher et al. (2002) with full follow-up time. No significant differences among treatments were found within either nodal status group based on log-rank tests (Fisher et al., 2002).

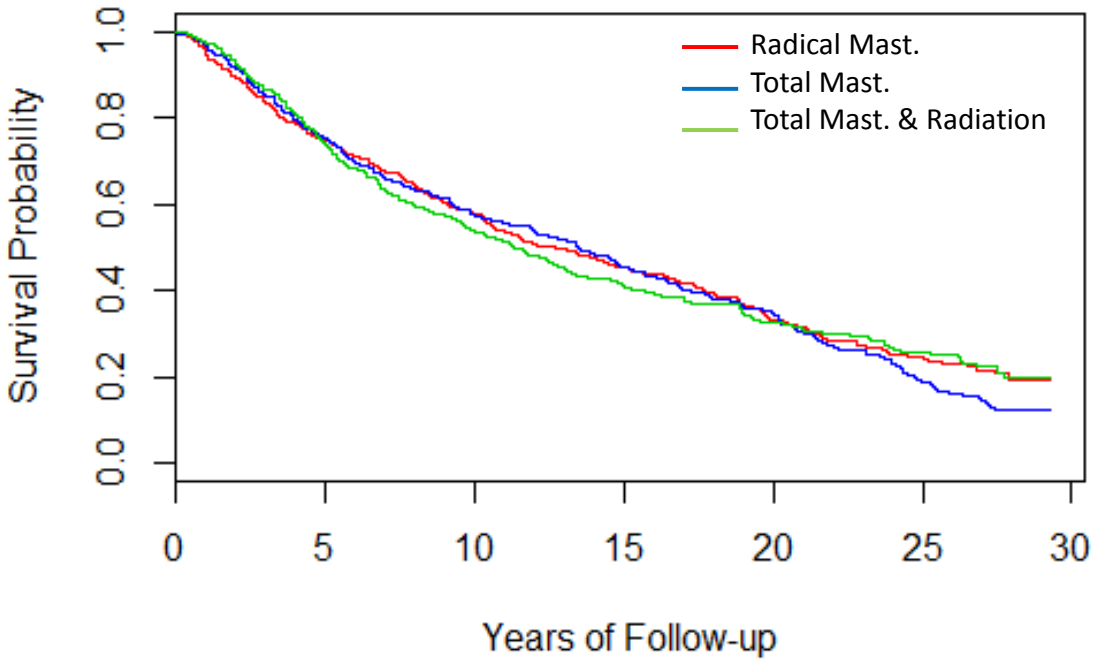

Figure 3.6.2: *Kaplan-Meier curves for overall survival by treatment in node-negative patients*

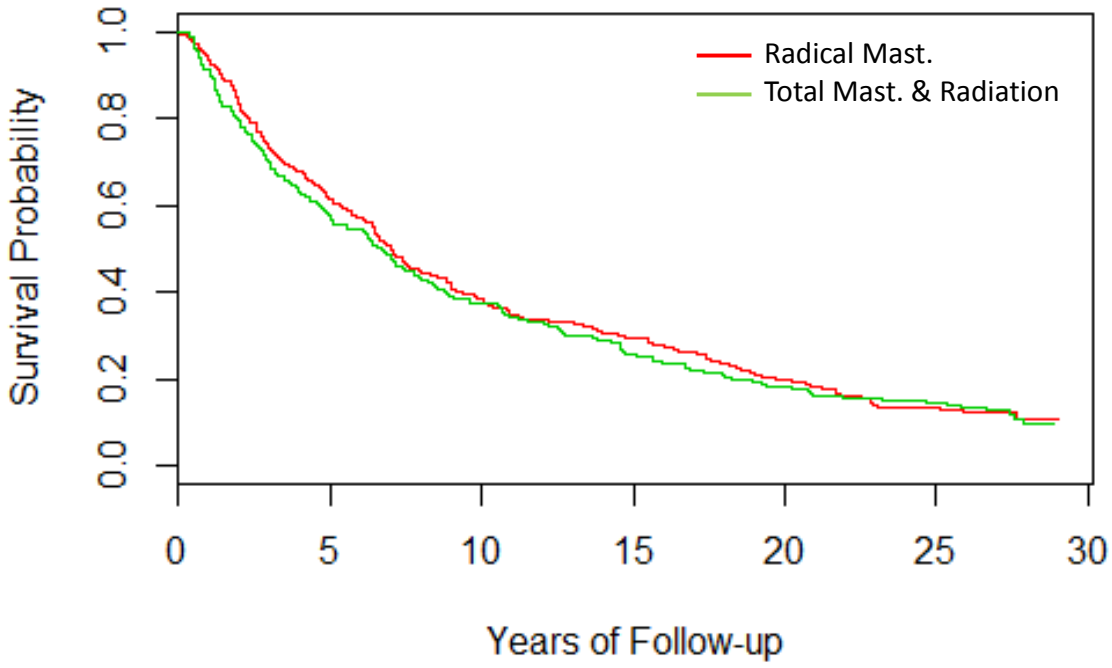

Figure 3.6.3: *Kaplan-Meier curves for overall survival by treatment in node-positive patients*

Additionally, hazard ratios from Cox proportional-hazards models further indicated no significant differences among groups of node-negative women or node-positive women when considering overall survival, disease-free survival, relapse-free survival, or distant-disease-free survival (Table 3.6.1). The overall conclusions were no significant advantage from a radical mastectomy, consistent with previous results. There was also no significant improvement in survival from radiation therapy after total mastectomy among node-negative women (Fisher et al., 2002).

Table 3.6.1: *Hazard Ratios (HR) comparing treatment groups by nodal status and endpoint of interest from Fisher et al. (2002)*

Endpoint	Nodal Status	Comparison	HR	95% CI	P-Value
Overall	Negative	RM vs TMR	1.08	(0.91, 1.28)	0.38
		RM vs TM	1.03	(0.87, 1.23)	0.72
		TM vs TMR	0.96	(0.81, 1.13)	0.60
	Positive	RM vs TMR	1.06	(0.89, 1.27)	0.49
Distant-Disease-free	Negative	RM vs TMR	1.08	(0.88, 1.34)	0.44
		RM vs TM	1.10	(0.89, 1.35)	0.39
		TM vs TMR	1.02	(0.83, 1.25)	0.85
	Positive	RM vs TMR	1.07	(0.87, 1.32)	0.51
Disease-free	Negative	RM vs TMR	1.06	(0.90, 1.25)	0.49
		RM vs TM	1.07	(0.91, 1.27)	0.39
		TM vs TMR	1.02	(0.87, 1.21)	0.78
	Positive	RM vs TMR	1.12	(0.94, 1.33)	0.20
Relapse-free	Negative	RM vs TMR	0.96	(0.76, 1.21)	0.74
		RM vs TM	1.14	(0.91, 1.42)	0.27
		TM vs TMR	1.18	(0.94, 1.48)	0.15
	Positive	RM vs TMR	1.09	(0.89, 1.35)	0.40

Thus, to illustrate the use of our proposed method for identifying differences in lost lifespans between groups, we will focus here on comparing by nodal status, an important prognostic factor. First, Kaplan-Meier curves were replicated for overall survival by nodal status, similar to analyses in the original study (Fisher et al., 2002), but with full follow-up time. Figure 3.6 illustrates the results by nodal status with corresponding 95% confidence intervals. Survival probability is shown to be consistently higher in the node-negative group compared to node-positive group.

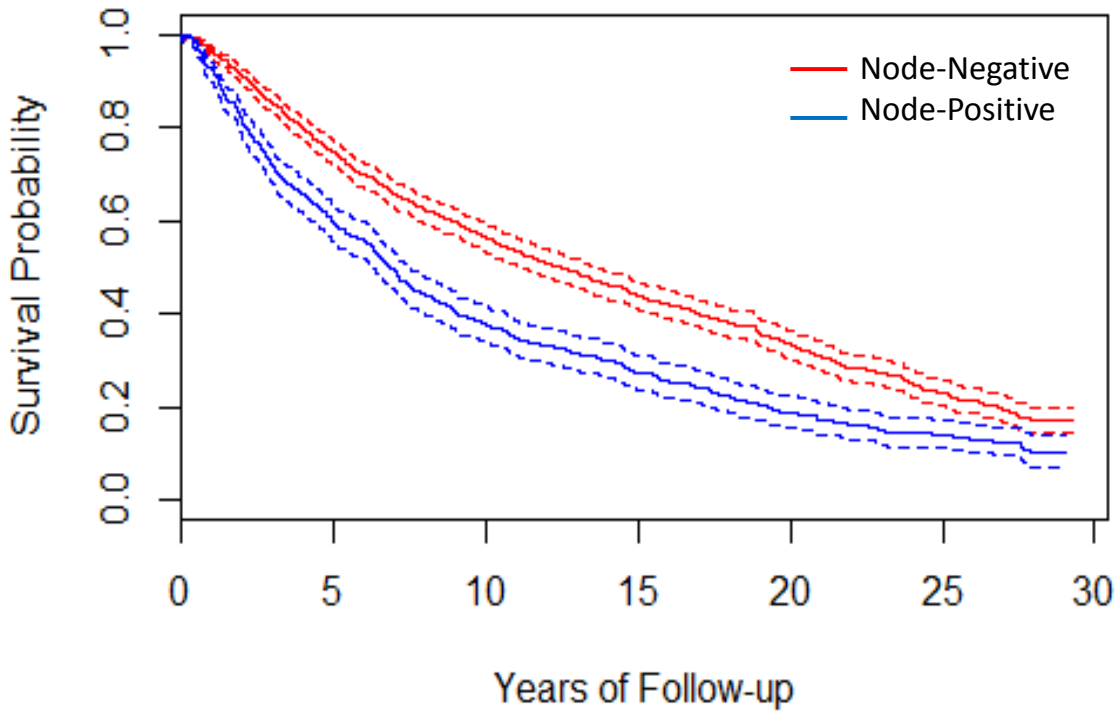

Figure 3.6.4: *Kaplan-Meier curves for overall survival by nodal status*

Then, the same data were analyzed with our proposed method to compare lost lifespans by nodal status. Specifically, lost lifespans were estimated in each nodal status group using the proposed nonparametric estimator (3.3.1). The two-sample test statistic (3.4.1) and associated confidence intervals (3.4.2) were used to compare lost-lifespans between node-negative and node-positive groups. Time-to-event was defined as time to any death, so that the lost lifespan was estimated as the years lost in each nodal group. Here the two-sample test statistic was evaluated at 5 different time points (t_0 = 13, 15, 20, 24, and 26 years post mastectomy) and for three different percentiles ($25th$, $50th$, and $75th$). Regardless of the time points specified, the median lost lifespan was significantly different between the two groups. The median lost lifespan was consistently longer in the node-postive group indicating greater life loss, as shown in Table 3.6.2. The ratio of median lost lifespans can be interpreted as a percent increase or deficit in life loss between groups. For example, at 13 years post treatment, there is an expected 14% increase in median life loss for the node-positive group compared to the node-negative group. The test statistic was also applied to the $25th$ and $75th$ lost lifespan percentiles to assess differences for the 'better' and 'worse' case scenarios. This was accomplished by replacing λ in (3.2.2) with 1/4 and 3/4, respectively. The $25th$ lost lifespan percentile represents patients with longer survival, as can be seen by the smaller lost lifespans. The ratios between node-positive and node-negative groups are larger compared to those of the $50th$ percentile, indicating that among patients with longer survival, the difference in lost lifespan between two nodal status groups is greater. Conversely, the 75th lost lifespan percentile represents patients with shorter survival, which can be seen by the larger lost lifespans. The smaller ratios indicate that among patients with shorter survival, the difference in lost lifespan between two nodal status groups is smaller, although still statistically significant.

Table 3.6.2: *Estimated median ($\lambda = 0.5$), 25th ($\lambda = 0.25$) and 75th ($\lambda = 0.75$) percentile lost lifespans in node-negative and node-positive groups, with ratios and 95% confidence intervals*

λ	t_0	Node-Negative	Node-Postive	Ratio	95% CI
0.5	13	7.91	9.02	1.14	(1.04, 1.25)
	15	9.49	10.68	1.12	(1.05, 1.21)
	20	13.20	14.98	1.13	(1.06, 1.21)
	24	16.04	18.45	1.15	(1.07, 1.25)
	26	17.40	20.27	1.16	(1.09, 1.25)
0.25	13	4.68	5.96	1.27	(1.13, 1.40)
	15	5.55	7.48	1.35	(1.19, 1.51)
	20	7.60	10.63	1.40	(1.21, 1.59)
	24	8.90	13.21	1.48	(1.29, 1.72)
	26	9.26	15.06	1.63	(1.34, 1.80)
0.75	13	10.38	11.11	1.07	(1.03, 1.14)
	15	12.13	13.00	1.07	(1.03, 1.12)
	20	16.63	17.88	1.08	(1.03, 1.11)
	24	20.30	21.75	1.07	(1.04, 1.10)
	26	22.11	23.71	1.07	(1.04, 1.10)

The proposed method was then compared to the existing methods based on the median residual life and hazard functions. Table 3.6.3 shows the median residual lifetimes in node-negative and node-positive groups at given time points t_0 together with their ratios and associated 95% confidence intervals from the same data set used in Jeong et al. (2008). At each given time point, the median residual lifetime would be the median of the mortality distribution of the surviving population beyond t_0 in node-negative and node-positive patients, respectively. The median residual life times in node-positive patients are significantly shorter than ones in node-negative patients up to year 8. The limitation of this summary measure, however, would be that it can be estimated only at time points where the median of the residual life distribution exists. More importantly, this summary measure would produce unstable estimates when there is heavy censoring at the tail of the distribution. This explains, in Table 3.6.3, why the ratios of the median residual lifetimes were able to be compared only up to the first 12 years.

Table 3.6.3 also displays the hazard ratio estimates and their 95% confidence intervals from the Cox's proportional hazards model among survivors beyond t_0. The event times were truncated at t_0 so new event times were defined as $T - t_0$. The hazard ratio in this case can be interpreted conditionally as the ratio, assumed constant over time, of the two instantaneous failure rates given that a patient survived up to the current time point *in the distribution of survivors beyond* t_0, so that its interpretation is not always straightforward to the laymen. The significance of the hazard ratios also holds up to year 8, which is consistent with ones from the median residual life analysis in this case.

Table 3.6.3: *Estimated median residual lifetimes in node-negative and node-positive groups, the ratios, hazard ratios, and 95% confidence intervals (NSABP B-04 data)*

	Median Residual Lifetime			Hazard Ratio (95% CI)
t_0	Node-Negative	Node-Positive	Ratio (95% CI)	
0	12.46	6.87	0.55 (0.49, 0.63)	1.53 (1.36, 1.71)
2	12.44	6.93	0.56 (0.47, 0.70)	1.41 (1.24, 1.60)
4	13.05	8.24	0.63 (0.49, 0.81)	1.36 (1.18, 1.57)
6	13.40	8.75	0.65 (0.54, 0.81)	1.40 (1.19, 1.63)
8	12.91	10.19	0.79 (0.66, 0.93)	1.22 (1.01, 1.46)
10	12.48	9.66	0.77 (0.62, 1.00)	1.19 (0.97, 1.46)
12	11.85	9.66	0.82 (0.63, 1.08)	1.14 (0.91, 1.44)

The hazard ratios were also estimated at the time points corresponding to the lost lifespan analysis in Table 3.6.2. The events that occurred after t_0 were administratively censored at that time, so that the hazard function summarizes the information prior to each fixed time point $t_0 = 13$, 15, 20, 24, and 26 in terms of the conditional instantaneous failure rate. As shown in Table 3.6.4, the hazard ratios range from 1.54 (95% CI; 1.38-1.73) at $t_0 = 26$ to 1.67 (95% CI; 1.47-1.90) at $t_0 = 13$, all indicating that there were significantly higher hazard rates on average in the node-positive population. The confidence intervals generally become narrower as t_0 increases as in the case of the lost lifespan analysis, because more information is gained as time progresses. Again, grasping the concept of the hazard function as the conditional instantaneous failure rate could be challenging to non-statisticians.

Table 3.6.4: *Hazard ratio estimates from the proportional hazards model corresponding to life loss time points*

t_0	Hazard Ratio	95% CI
13	1.67	(1.47, 1.90)
15	1.65	(1.46, 1.87)
20	1.62	(1.44, 1.82)
24	1.57	(1.40, 1.76)
26	1.54	(1.38, 1.73)

3.7 DISCUSSION ON NONPARAMETRIC INFERENCE

In this paper, a new summary measure for time-to-event outcome, lost lifespan, was introduced and an inference procedure was proposed to estimate and compare the quantile lost lifespans. In contrast to the traditional residual life analysis, the proposed method is less affected by heavy censoring, gaining higher efficiency, toward the end of the study period. The proposed method is nonparametric in nature yet does not require estimation of the density function of the underlying event distribution to evaluate the variance of the quantile estimator. The clinical interpretation of the lost lifespan as time lost due to occurrence of an event of interest is straightforward.

A practical question may arise regarding how to select the fixed time points where the statistical analyses are performed. Because the proposed method is based on the quantiles of the distribution of the lost lifespan, the minimum time point for analysis should be selected, so that the quantiles of interest exist. Here the quantile of interest might be determined based on an investigator's study goal. For example, if the investigator is interested in patients with poor prognosis, the higher quantile of the lost lifespan distribution would be appropriate and vice versa.

A statistical aspect of the proposed method can be directly compared with an existing method closely related to the hazard function such as the log-rank test. The log-rank test may be viewed as a method to assess a group effect on the hazard function, because it is well-known that the score test statistic from the partial likelihood under the proportional hazards model is equivalent to the log-rank test when there are no ties. Therefore the log-rank test is known to be optimal when the proportional hazards assumption holds, but the proposed method does not need such assumption. Rather, for the proposed test to be optimal, an assumption of a linear covariate effect on the quantile lost lifespan on a log-scale in a semiparametric setting might be needed, which might also merit further investigation.

The main advantage of using the lost lifespan would be its straightforward interpretation compared to an approach using the hazard function, as described in the last two paragraphs at the end of Section 3.6. In this era of patient-centered outcomes research, this might be very helpful in practice for the stakeholders such as physicians and patients to understand

potential benefits and harms of an experimental drug in clinical studies. In addition to this advantage of easy interpretation, the proposed method based on the lost lifespan would allow for the routine analysis of time-to-event data by using the cumulative information prior to a given time point, unlike the residual lifetime analysis.

4.0 REGRESSION ON QUANTILE LOST LIFESPAN

4.1 INTRODUCTION

In the following chapter, we propose a novel quantile regression model on the quantiles of the distribution of the lost lifespan under right censoring. The consistency and asymptotic normality of the regression parameters are established. To avoid estimation of the probability density function of the lost lifespan distribution under censoring, the estimating equation for the quantile lost lifespan is directly used to construct the test statistics for the regression parameters. To test a subset of the regression parameters, the minimum dispersion statistic is adopted to eliminate the nuisance parameters not being tested. Simulation results are presented to validate the finite sample properties of the proposed estimators and test statistics. The proposed method is illustrated with a real dataset from a clinical trial on cancer.

4.2 QUANTILE LOST LIFESPAN FUNCTION

The lost lifespan, as defined in the previous chapter, considers the time lost due to an event occurring prior to a specified timepoint, t_0. Recall the definition of the λ-percentile of the lost lifespan distribution as

$$\theta_{\lambda|t_0} = \lambda\text{-percentile}\{t_0 - T_i | T_i \leq t_0\}.$$

Then $\theta_{\lambda|t_0}$ satisfies $P(t_0 - T_i \leq \theta_{\lambda|t_0} | T_i \leq t_0) = \lambda$, or

$$\frac{P(T_i \geq t_0 - \theta_{\lambda|t_0}) - P(T_i \geq t_0)}{1 - P(T_i \geq t_0)} = \lambda,$$

which can be rewritten in terms of the survival function as

$$\frac{S(t_0 - \theta_{\lambda|t_0}) - S(t_0)}{1 - S(t_0)} = \lambda.$$

Here given observed data and λ, $\theta_{\lambda|t_0}$ can be nonparametrically estimated after replacing $S(t)$ with $\hat{S}(t)$. Throughout the remainder, Y_i will represent the observed time as the minimum of T_i and censoring time C_i, and Δ_i will be an event indicator ($\Delta_i = 1$ if $Y_i = T_i$). The censoring distribution will be denoted by $G(t) = P(C \geq t)$ and will be estimated by the Kaplan-Meier estimator denoted $\hat{G}(t)$ (Kaplan and Meier, 1958). We will assume independence between event times and censoring times. In the following section, the concept of percentile lost lifespan ($\theta_{\lambda|t_0}$) will be extended to a regression setting.

4.3 REGRESSION MODEL

4.3.1 Formulation of Estimating Equation

We propose the following log-linear regression model on the λ-percentile of a distribution of lost lifespans at t_0:

$$\lambda\text{-percentile}\{\ln(t_0 - T_i)|T_i \leq t_0, Z_i\} = \beta'_{\lambda|t_0} Z_i, \tag{4.3.1}$$

where $\beta'_{\lambda|t_0}$ is a vector of the regression coefficients, $(\beta_{\lambda|t_0,0}, \beta_{\lambda|t_0,1}, ..., \beta_{\lambda|t_0,p})'$, and Z_i is a vector of covariates for the ith individual, $(1, X_{1i}, ..., X_{pi})$. When censoring is not present, the regression parameter can be estimated from minimizing the following absolute deviation function:

$$\sum_{i=1}^{n} |\ln(t_0 - T_i) - \beta'_{\lambda|t_0} Z_i)| =$$

$$\sum_{i=1}^{n} (\{\ln(t_0 - T_i) - \beta'_{\lambda|t_0} Z_i\} \times I(\ln(t_0 - T_i) - \beta'_{\lambda|t_0} Z_i > 0) - \{\ln(t_0 - T_i) - \beta'_{\lambda|t_0} Z_i\}$$

$$\times I(\ln(t_0 - T_i) - \beta'_{\lambda|t_0} Z_i < 0)).$$

The right side of the equation can be rewritten as

$$= \sum_{i=1}^{n} (2\{\ln(t_0 - T_i) - \beta' Z_i\} \times I\{\ln(t_0 - T_i) - \beta' Z_i \geq 0\} - \{\ln(t_0 - T_i) - \beta' Z_i\})$$

and in terms of λ,

$$= \left(\frac{1}{1-\lambda}\right) \sum_{i=1}^{n} \{\ln(t_0 - T_i) - \beta' Z_i\} \times (I\{\ln(t_0 - T_i) - \beta' Z_i \geq 0\} - (1-\lambda)).$$

Differentiating with respect to β results in a general form of the estimating function for the noncensored case as

$$-(\frac{1}{1-\lambda}) \sum_{i=1}^{n} Z_i(\tau - I(T_i > t_0 - \exp(\beta'_{\lambda|t_0} Z_i)) \times I(T_i < t_0),$$

where the indicator $I(T_i < t_0)$ has been added to ensure existence of the natural logarithm. Assuming conditional independence of T_i and C_i given Z_i, and independence of C_i from covariates Z_i, the following equation holds conditionally given Z_i under right censoring,

$$\begin{aligned} E(I\{\ln(t_0 - Y_i) - \beta'_{\lambda|t_0} Z_i < 0\}|Z_i) &= P(\ln(t_0 - Y_i) - \beta'_{\lambda|t_0} Z_i < 0) \\ &= P(T_i > t_0 - \exp(\beta'_{\lambda|t_0} Z_i)) \times P(C_i > t_0 - \exp(\beta'_{\lambda|t_0} Z_i)). \end{aligned}$$

Since the λ-percentile lost lifespan function is defined under model (4.3.1) as

$$\lambda = P(t_0 - T_i \leq \exp(\beta'_{\lambda|t_0} Z_i)|T_i \leq t_0)$$

$$= \frac{P(T_i \geq t_0 - \exp(\beta'_{\lambda|t_0} Z_i)) - P(T_i \geq t_0)}{1 - P(T_i \geq t_0)},$$

we have

$$\begin{aligned} &P(T_i > t_0 - \exp(\beta'_{\lambda|t_0} Z_i)) \times P(C_i > t_0 - \exp(\beta'_{\lambda|t_0} Z_i)) \\ &= \left(\frac{P(T_i \geq t_0 - \exp(\beta'_{\lambda|t_0} Z_i)) - P(T_i \geq t_0)}{1 - P(T_i \geq t_0)}\right) \times P(C_i > t_0 - \exp(\beta'_{\lambda|t_0} Z_i)) \\ &= \lambda \times P(C_i > t_0 - \exp(\beta'_{\lambda|t_0} Z_i)) \\ &= \lambda \times G(t_0 - \exp(\beta'_{\lambda|t_0} Z_i)), \end{aligned}$$

leading to

$$E[I(\ln(t_0 - Y_i) - \beta'_{\lambda|t_0} Z_i < 0)] = \lambda \times G(t_0 - \exp(\beta'_{\lambda|t_0} Z_i))$$

assuming $Y_i < t_0$ to satisfy the natural logarithm. Therefore the regression parameter $\beta_{\lambda|t_0}$ can be estimated from the following equation under right censoring:

$$\begin{aligned} S_{\lambda|t_0}(\beta_{\lambda|t_0}) &= \sum_{i=1}^{n} Z_i \left[\frac{I(\ln(t_0 - Y_i) - \beta'_{\lambda|t_0} Z_i < 0)}{\hat{G}(t_0 - \exp(\beta'_{\lambda|t_0} Z_i))} - \lambda \right] \times I(Y_i < t_0) \\ &= \sum_{i=1}^{n} Z_i \left[\frac{I(Y_i > t_0 - \exp(\beta'_{\lambda|t_0} Z_i))}{\hat{G}(t_0 - \exp(\beta'_{\lambda|t_0} Z_i))} - \lambda \right] \times I(Y_i < t_0) \approx 0. \end{aligned} \quad (4.3.2)$$

A solution of the estimating equation (4.3.2) can be found by minimizing $||S_{\lambda|t_0}(\beta_{\lambda|t_0})||$, where $||\cdot||$ denotes the square root of the sum of squares. In the next section, we show that the estimator from equation (4.3.2), $\hat{\beta}_{\lambda|t_0}$, is a consistent estimator for the true parameters, $\beta^0_{\lambda|t_0}$, under certain regularity conditions.

4.3.2 Consistency of Regression Parameter Estimates

We start by defining

$$\tilde{S}_{\lambda|t_0,n}(\beta_{\lambda|t_0}) = \sum_{i=1}^{n} Z_i \times [P(T_i \geq t_0 - \exp(\beta'_{\lambda|t_0} Z_i)|Z_i) - \lambda].$$

When $\beta_{\lambda|t_0}$ is replaced with $\beta^0_{\lambda|t_0}$, the true value in the interior of a bounded convex region D, the above equation reduces to 0. We will denote $F_i(\cdot|Z)$ to be the cumulative distribution function of $\log(t_0 - T_i) + \beta^{0\prime}_{\lambda|t_0} Z_i$ given covariate Z_i, and its corresponding derivative will be denoted by $f_i(\cdot|Z_i)$. Additionally, $G(\cdot|Z_i)$ will denote the survival function of the censoring distribution, such that the derivative of $-G(\cdot|Z_i)$ will be $g(\cdot|Z_i)$. Note, that both the derivatives and vector Z_i are uniformly bounded. Following Csorgo and Horvath (1983), we know that for all $\epsilon > 0$,

$$\sup_{s \leq \tilde{t}} |\hat{G}(s) - G(s)| = o(n^{-1/2+\epsilon}), \ a.s.$$

where $\tilde{t}$ is a constant satisfying $P\{\log(t_0 - Y_i) \leq \tilde{t}\} > 0$ and $\beta'_{\lambda|t_0} Z \leq \tilde{t}$, with probability 1. This can be used to show that for $\beta_{\lambda|t_0} \in D$, a bounded convex region,

$$S_{\lambda|t_0,n}(\beta_{\lambda|t_0}) - \tilde{S}_{\lambda|t_0,n}(\beta_{\lambda|t_0}) = \sum_{i=1}^{n} Z_i \times \left[\frac{I\{Y_i \geq t_0 - \exp(\beta'_{t_0} Z_i\} }{G\{t_0 - \exp(\beta'_{t_0} Z_i)\}} - \frac{P\{Y_i \geq t_0 - \exp(\beta'_{t_0} Z_i)\}}{G\{t_0 - \exp(\beta'_{t_0} Z_i)\}} \right]$$

$$= \sum_{i=1}^{n} Z_i \times \left[\frac{1}{G\{t_0 - \exp(\beta'_{t_0} Z_i)\}} \right] \times [I\{Y_i \geq t_0 - \exp(\beta'_{t0} Z_i)\} - P\{Y_i \geq t_0 - \exp(\beta'_{t_0} Z_i)\}]$$

Because

$$\sup_{\beta_{\lambda|t_0} \in D} \left| \sum_{i=1}^{n} G^{-1}\{t_0 - \exp(\beta'_{t_0} Z_i)\} \times [I\{Y_i \geq t_0 - \exp(\beta'_{t_0} Z_i)\} - P\{Y_i \geq t_0 - \exp(\beta'_{t_0} Z_i)\}] \right|$$

$$= o(n^{1/2+\epsilon})$$

it follows that

$$\sup_{\beta_{\lambda|t_0} \in D} \left|\left| n^{-1} S_{\lambda|t_0,n}(\beta_{\lambda|t_0}) - n^{-1} \tilde{S}_{\lambda|t_0,n}(\beta_{\lambda|t_0}) \right|\right| = o(n^{-1/2+\epsilon}), \ a.s. \tag{4.3.3}$$

Now let us define

$$A_n(\beta_\lambda|t_0) = -\frac{1}{n} \sum_{i=1}^{n} f_i\{(\beta_{\lambda|t_0} - \beta^0_{\lambda|t_0})' Z_i | Z_i\} Z_i Z_i',$$

where $f(\cdot|Z)$ is the conditional density of T given $Z = z$. Then, supposing $E\{ZZ'f(0|Z)\}$ is positive definite implies $A_n(\beta_{\lambda|t_0})$ is nonpositive definite. Then, $A_n(\beta^0_{\lambda|t_0})$ converges to $-E(ZZ'f(0|Z))$ with probability equal to 1, which is negative definite. Using Taylor series expansion around $\beta^0_{\lambda|t_0}$ and letting $\beta^*_{t_0}$ be some point between $\hat{\beta}_{\lambda|t_0}$ and $\beta^0_{\lambda|t_0}$, we have

$$n^{-1}\{\tilde{S}_{\lambda|t_0}(\hat{\beta}_{\lambda|t_0}) - \tilde{S}_{\lambda|t_0}(\beta^0_{\lambda|t_0})\} \approx (\hat{\beta}_{\lambda|t_0} - \beta^0_{\lambda|t_0})' A_n(\beta^*_{t_0}). \tag{4.3.4}$$

From the definition of $\hat{\beta}_{\lambda|t_0}$, we know $n^{-1} S_{\lambda|t_0,n}(\hat{\beta}_{\lambda|t_0}) = 0$, and so by (4.3.3) $n^{-1}\tilde{S}_{\lambda|t_0,n}(\hat{\beta}_{\lambda|t_0})$ will converge to 0, almost surely, as n $\to \infty$. Together, with (4.3.4), this shows $\hat{\beta}_{\lambda|t_0} \to \beta^0_{\lambda|t_0}$, a.s. as $n \to \infty$.

4.3.3 Test Statistic for Significance of Regression Parameters

Now, we construct global and local test statistics to test the regression coefficients under specified null values. To avoid estimation of the probability density function of $(t_0 - T_i)I(T_i < t_0)|Z_i$ under censoring, an inference procedure directly based on the asymptotic distribution of the estimating equation (4.3.2) is proposed. First for the global test, consider the null hypothesis of $H_0 : \beta_{\lambda|t_0} = \beta_{\lambda|t_0,0}$.

Here, we establish asymptotic normality of $n^{-\frac{1}{2}} S_{\lambda|t_0,n}(\beta^0_{\lambda|t_0})$, used to develop the test statistic. We can begin by approximating the equation with a sum of independent zero-mean random variables. Recall that

$$n^{-1/2} S_{\lambda|t_0}(\beta^0_{\lambda|t_0}) = n^{-1/2} \sum_{i=1}^{n} Z_i \times \left[\frac{I\{Y_i \geq t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}}{\hat{G}\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}} - \lambda \right],$$

where the right-hand side can be re-written as

$$= n^{-1/2} \sum_{i=1}^{n} Z_i \times \left\{ \left[\frac{I\{Y_i \geq t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}}{\hat{G}\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}} \right] + \left[\frac{I\{Y_i \geq t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}}{G\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}} \right] \right.$$
$$\left. - \left[\frac{I\{Y_i \geq t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}}{G\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}} \right] - \lambda \right\}$$

$$= n^{-1/2} \sum_{i=1}^{n} Z_i \times \left\{ \left[\frac{I\{Y_i \geq t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}}{G\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}} \right] - I\{Y_i \geq t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i))\} \right.$$
$$\left. \times \left[\frac{\hat{G}\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\} - G\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}}{G\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\} \times \hat{G}\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}} \right] - \lambda \right\}. \tag{4.3.5}$$

Let us define

$$Q_1(s) = n^{-1} \sum_{i=1}^{n} Z_i I\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i) \leq \min(s, Y_i)\}$$

such that the second term of (4.3.5) is asymptotically equivalent to,

$$- \int_{-\infty}^{\infty} \left[\frac{n^{-1/2}\{\hat{G}(s) - G(s)\}}{G(s)^2} \right] dq_1(s),$$

where $q_1(s) = \lim_{n\to\infty} Q_1(\cdot)$. The random process $-n^{-1/2}\{\hat{G}(s) - G(s)\}/G(s)$ can be represented as a martingale integral, following arguments of Fleming and Harrington (Corollary 3.2.1, 1991),

$$\int_{-\infty}^{s} \frac{n^{-1/2}\sum_{i=1}^{n}\{dI(Y_i \le v, \Delta_i = 0) - I(Y_i \ge v)d\Lambda_G(v)\}}{n^{-1}\sum_{i=1}^{n} I(Y_i \ge v)}, \tag{4.3.6}$$

where $\Lambda_G(\cdot)$ represents the cumulative hazard function of the censoring distribution, and $h(v)$ is the limit of $\sum_{i=1}^{n} I(Y_i \ge v)/n$ as $n \to \infty$. Then, (4.3.6) is asymptotically equivalent to

$$\int_{-\infty}^{s} h^{-1}(v)n^{-1/2}\sum_{i=1}^{n}[dI(Y_i \le v, \Delta_i = 0) - I(Y_i \ge v)d\Lambda_G(v)],$$

and so, the second term of (4.3.5) is asymptotically equivalent to

$$\int_{-\infty}^{\infty} G^{-1}(s)\int_{-\infty}^{s} h^{-1}(v)n^{-1/2}\sum_{i=1}^{n}[dI(Y_i \le v, \Delta_i = 0) - I(Y_i \ge v)d\Lambda_G(v)]dq_1(s).$$

Finally, (4.3.5) is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^{n}\tau_{\lambda|t_0,i}$, where

$$\tau_{\lambda|t_0,i} = Z_i \times \left[\frac{I\{Y_i \ge t_0 - \exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\}}{G\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\}} - \lambda\right] + \int_{-\infty}^{\infty} G^{-1}(s)\int_{-\infty}^{s} h^{-1}(v)$$

$$\times[dI\{Y_i \le v, \Delta_i = 0\} - I(Y_i \ge v)d\Lambda_G(v)]dq_1(s).$$

Following the Multivariate Central Limit Theorem and because $\tau_{\lambda|t_0,i}$ for $i = 1, ..., n$ are independent random vectors with mean 0, the distribution of $n^{-1/2}S_{\lambda|t_0}(\beta^0_{\lambda|t_0})$ is asymptotically normal with mean 0 and variance-covariance matrix $\Gamma_{\lambda|t_0} = n^{-1}\sum_{i=1}^{n}\tau_{\lambda|t_0,i}\tau'_{\lambda|t_0,i}$. Then a consistent estimator of $\Gamma_{\lambda|t_0}$ can be obtained by replacing $\beta^0_{\lambda|t_0}$ with $\hat{\beta}^0_{\lambda|t_0}$, G with $\hat{G}$, $h(s)$ with $\sum_{i=1}^{n} I(Y_i \ge s)/n$, $q_1(s)$ with $Q_1(s)$, and $d\Lambda_G(S)$ with $[\sum_{i=1}^{n} I(Y_i \ge s)]^{-1}d[\sum_{i=1}^{n} I(Y_i \le s, \Delta_i = 0)]$, which leads to our estimator $\hat{\Gamma}_{\lambda|t_0} = n^{-1}\sum_{i=1}^{n}\hat{\tau}_{\lambda|t_0,i}\hat{\tau}'_{\lambda|t_0,i}$, where

$$\hat{\tau}_{\lambda|t_0,i} = Z_i \times \left[\frac{I\{Y_i \ge t_0 - \exp(\hat{\beta}'Z_i)\}}{\hat{G}\{t_0 - \exp(\hat{\beta}'Z_i)\}} - \lambda\right] + \sum_{l=1}^{n} Z_l\left[\frac{I\{t_0 - \exp(\hat{\beta}'Z_i) \le Y_l\}}{\hat{G}\{t_0 - \exp(\hat{\beta}'Z_i)\}}\right] \times$$

$$\left\{(1-\Delta_i) \times \frac{I\{Y_i \le t_0 - \exp(\hat{\beta}'Z_i)\}}{\sum_{m=1}^{n} I(Y_m \ge Y_i)} - \sum_{j=1}^{n}\frac{(1-\Delta_j)I(Y_j \le \min\{t_0 - \exp(\hat{\beta}'Z_i), Y_i\})}{\{\sum_{m=1}^{n} I(Y_m \ge Y_j)\}^2}\right\}.$$

Now that we have established asymptotic normality of $n^{-\frac{1}{2}} S_{\lambda|t_0,n}(\beta^0_{\lambda|t_0})$, we can derive a test statistic for testing the null hypothesis as

$$n^{-1} S'_{\lambda|t_0,n}(\beta_{\lambda|t_0,0}) \hat{\Gamma}^{-1}_{\lambda|t_0} S_{\lambda|t_0,n}(\beta_{\lambda|t_0,0}),$$

which approximately follows a χ^2-distribution with $p+1$ degrees of freedom. Recall that p corresponds to the number of covariates in the model.

Next, we establish asymptotic normality of regression parameter estimates based on local linearity for $S_{\lambda|t_0,n}(\beta_{\lambda|t_0})$. From Section 4.3.2, $A_n(\beta_{\lambda|t_0}) = \frac{1}{n}\frac{\partial}{\partial \beta_{\lambda|t_0}} \tilde{S}_{\lambda|t_0}(\beta_{\lambda|t_0})$ and $A_n(\beta^0_{\lambda|t_0})$ converges to $A = -E[ZZ' f(0|Z)]$ as $n \to \infty$, a nonsingular matrix. We will show that

$$S_{\lambda|t_0,n}(\beta_{\lambda|t_0}) = S_{\lambda|t_0,n}(\beta^0_{\lambda|t_0}) + nA(\beta_{\lambda|t_0} - \beta^0_{\lambda|t_0}) + o_p(\max(n^{1/2}, n||\beta_{\lambda|t_0} - \beta^0_{\lambda|t_0}||)) \quad (4.3.7)$$

for all β in $||\beta_{\lambda|t_0} - \beta^0_{\lambda|t_0}|| < cn^{-1/3}$, where c is any fixed constant. It will follow from (4.3.7) that the distribution of $(\hat{\beta}_{\lambda|t_0} - \beta^0_{\lambda|t_0})$ is approximately normal with mean 0 and covariance matrix $\Lambda = n^{-1}A^{-1}\Gamma(A^{-1})'$. Two important lemmas will be used to show the previous (Lai and Ying, 1988).

Lemma 1. Let μ be a continuously differentiable function. Then

$$\sup_{|s-t|\le cn^{-1/3}, s,t\le \tilde{t}} |\mu\{\hat{G}(t)\} - \mu\{G(t)\} - \mu\{\hat{G}(s)\} + \mu\{G(s)\}| = o_p(n^{-1/2})$$

Lemma 2. Let ν_i be a sequence of constants. Then, for a fixed t_0

$$\sup_{||\beta_{\lambda|t_0} - \beta^0_{\lambda|t_0}|| < cn^{-1/3}} \Bigg| \sum_{i=1}^n \nu_i I\{Y_i \ge t_0 - \exp(\beta'_{\lambda|t_0} Z_i)\} - \sum_{i=1}^n \nu_i I\{Y_i \ge t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\} -$$

$$\sum_{i=1}^n \nu_i [1 - F_i\{\exp(\beta'_{\lambda|t_0} Z_i) - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}] G(t_0 - \exp(\beta'_{\lambda|t_0} Z_i)) +$$

$$\sum_{i=1}^n \nu_i \lambda G\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\} \Bigg| = o_p(n^{1/2})$$

In particular, we have

$$\sup_{||\beta_{\lambda|t_0} - \beta^0_{\lambda|t_0}|| < cn^{-1/3}} \sum_{i=1}^n |I\{Y_i \ge t_0 - \exp(\beta'_{\lambda|t_0} Z_i)\} - I\{Y_i \ge t_0 - \exp(\beta^{0\prime}_{\lambda|t_0} Z_i)\}| = O_p(n^{2/3}). \quad (4.3.8)$$

Since $\beta_{\lambda|t_0}$ is in the $n^{-1/3}$ -neighborhood of $\beta^0_{\lambda|t_0}$, *Lemma 1* gives

$$S_{\lambda|t_0,n}(\beta_{\lambda|t_0}) = S_{\lambda|t_0,n}(\beta^0_{\lambda|t_0}) + \sum_{i=1}^{n}[G\{t_0 - \exp(\beta'_{\lambda|t_0}Z_i)\}^{-1}I\{Y_i \geq t_0 - \exp(\beta'_{\lambda|t_0}Z_i)\} - \lambda]Z_i$$

$$-\sum_{i=1}^{n}[G\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\}^{-1}I\{Y_i \geq t_0 - \exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\} - \lambda]Z_i + \sum_{i=1}^{n}[\hat{G}\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\}^{-1}$$

$$-G\{t_0-\exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\}^{-1}]\times[I\{Y_i \geq t_0-\exp(\beta'_{\lambda|t_0}Z_i)\}-I\{Y_i \geq t_0-\exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\}]Z_i+o_p(n^{1/2}). \tag{4.3.9}$$

By using (4.3.8) and $|\hat{G}\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\}^{-1} - G\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\}^{-1}| = o_p(n^{-1/2+\epsilon})$ for all $\epsilon > 0$, the fourth term of the right-hand side of (4.3.9) is $o_p(n^{1/2})$. Furthermore, by *Lemma 2*, it can be shown that

$$\sum_{i=1}^{n}[G\{t_0 - \exp(\beta'_{\lambda|t_0}Z_i)\}^{-1}I\{Y_i \geq t_0 - \exp(\beta'_{\lambda|t_0}Z_i)\} - \lambda]Z_i$$

$$-\sum_{i=1}^{n}[G\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\}^{-1}I\{Y_i \geq t_0 - \exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\} - \lambda]Z_i$$

$$=\sum_{i=1}^{n}[G\{t_0 - \exp(\beta'_{\lambda|t_0}Z_i)\}^{-1}P\{Y_i \geq t_0 - \exp(\beta'_{\lambda|t_0Z_i})|Z_i\} - \lambda]Z_i + o_p(n^{1/2}),$$

because

$$G\{t_0 - \exp(\beta^{0\prime}_{\lambda|t_0}Z_i)\}^{-1}P\{Y_i \geq t_0 - \exp(\beta^{0\prime}_{\lambda|t_0Z_i})|Z_i\} - \lambda = 0.$$

This and (4.3.9) imply that

$$S_{\lambda|t_0,n}(\beta_{\lambda|t_0}) = S_{\lambda|t_0,n}(\beta^0_{\lambda|t_0}) + \tilde{S}_{\lambda|t_0,n}(\beta_{\lambda|t_0}) + o_p(n^{1/2}) \tag{4.3.10}$$

Thus local linearity in (4.3.7) follows by taking Taylor's expansion of $\tilde{S}_{\lambda|t_0,n}(\beta_{\lambda|t_0})$ in (4.3.10) at $\beta^0_{\lambda|t_0}$, i.e.

$$n^{-1}\{\tilde{S}_{\lambda|t_0,n}(\beta_{\lambda|t_0}) - \tilde{S}_{\lambda|t_0,n}(\beta^0_{\lambda|t_0})\} \approx (\beta_{\lambda|t_0} - \beta^0_{\lambda|t_0})'A.$$

4.3.4 Partitioning Regression Coefficients

For the local test, suppose we are now interested in testing a subset of the regression parameters explicitly, say $H_0 : \beta^{(1)}_{\lambda|t_0} = \beta^{(1)}_{\lambda|t_0,0}$, from a partition of the regression coefficients $\beta'_{\lambda|t_0} = (\beta^{(1)}_{\lambda|t_0}, \beta^{(2)}_{\lambda|t_0})$, where $\beta^{(1)}_{\lambda|t_0}$ is a vector with $r \times 1$ elements. The regression parameters $\beta^{(2)}_{\lambda|t_0}$ would be still involved in the test statistic, but not specified under the null hypothesis, so that they need to be treated as nuisance parameters. One way of eliminating those nuisance parameters would be to minimize the test statistic over them (Basawa and Koul, 1988). Therefore, we have the minimum test statistic for the local test as

$$V(\beta^{(1)}_{\lambda|t_0}) = \min_{\beta^{(2)}_{\lambda|t_0})} \{n^{-1} S'_{t_0,n}((\beta^{(1)\prime}_{\lambda|t_0}, \beta^{(2)\prime}_{\lambda|t_0}))\hat{\Gamma}^{-1}_{\lambda|t_0} S_{t_0,n}((\beta^{(1)\prime}_{\lambda|t_0}, \beta^{(2)\prime}_{\lambda|t_0}))\} \tag{4.3.11}$$

Following arguments in Wei et al. (1990) and Ying et al. (1995), this statistic would follow an asymptotic χ^2-distribution with r degrees of freedom. We can further obtain a $100 \times (1-\alpha)\%$ confidence interval for $\beta^{(1)}_{\lambda|t_0}$ by inverting the minimum dispersion statistic $V(\beta^{(1)}_{\lambda|t_0})$ such that

$$\{\beta^{(1)}_{\lambda|t_0} : V(\beta^{(1)}_{\lambda|t_0}) < \chi^2_{r,1-\alpha}\}. \tag{4.3.12}$$

4.4 SIMULATION STUDIES

4.4.1 Empirical Estimates

Several simulation studies were performed to assess the proposed estimators and test statistics with the finite samples. To generate data, we assumed a parametric proportional hazards model (Cox, 1972) with a Weibull distribution as the baseline distribution and one group indicator as a covariate. Thus, the survival function is specified as

$$S(t) = \exp(-(\rho t)^{\eta} \exp(\beta Z_i)), \tag{4.4.1}$$

where ρ and η are the the Weibull parameters set to 0.2 and 2, respectfully, and β is the regression parameter associated with the group indicator Z_i ($Z_i = 0$ for control and $Z_i = 1$ for

an intervention, or treatment). By using the probability integral transformation, potential failure times were generated from

$$T_i = (1/\rho)(-\exp(-\beta Z_i)(\log(1-u_i))^{1/\eta},$$

where u_i is a uniform random variable on $[0,1]$. Potential censoring times C_i were generated from a uniform distribution on $[a,b]$, where a and b determine the desired censoring proportions. Observed survival times Y_i were then determined as the minimum of potential failure times and potential censoring times, i.e. $\min(T_i, C_i)$. Under the parametric Cox model (4.4.1), the true median lost lifespan (θ_{t_0}), as shown in Chapter 3, can be derived as

$$\theta_{t_0}(z) = t_0 - \frac{1}{\rho}[\exp(-\beta z)\{\log(2) - \log(1+\exp(-(\rho t_0)^{\eta}\exp(\beta z))\}]^{\eta}. \tag{4.4.2}$$

First, we evaluate performance of our proposed method of estimation. The true values of θ_{t_0} in (4.4.2) when $\beta = 0$ would be the same for both control and treatment groups as 10.8, 9.8, 8.8, and 7.8 at $t_0 = 15$, 14, 13, and 12, respectively. Let us consider a simple log-linear regression model on the median lost lifespan,

$$\text{med}(\ln(t_0 - T_i)|T_i \le t_0) = \beta_{t_0}^{(0)} + \beta_{t_0}^{(1)} Z_{1i}, \tag{4.4.3}$$

where Z_{1i} is a binary covariate indicating treatment group ($Z_{1i} = 1$) or control group ($Z_{1i} = 0$), and $\beta_{t_0}^{(0)}$ and $\beta_{t_0}^{(1)}$ are the intercept and a regression coefficient associated with Z_{1i}, respectively. Following the invariance property of the log-transformation, the model is equivalent to

$$\text{med}(t_0 - T_i|T_i \le t_0) = \exp(\beta_{t_0}^{(0)} + \beta_{t_0}^{(1)} Z_{1i}),$$

implying that $\exp(\beta_{t_0}^{(0)})$ and $\exp(\beta_{t_0}^{(0)} + \beta_{t_0}^{(1)})$ can be interpreted as the median lost lifespan in the control group and in the treatment group, respectively. Thus, the difference in median lost lifespans between groups is given by $\exp(\beta_{t_0}^{(0)})(\exp(\beta_{t_0}^{(1)}) - 1)$, and the ratio of two lost lifespans by $\exp(\beta_{t_0}^{(1)})$, so that testing a null hypothesis of $\beta_{t_0}^{(1)} = 0$ will be equivalent to testing whether the ratio of two median lifespans equals 1.

In order to evaluate our parameter estimates, we compare $\hat{\beta}_{t_0}^{(1)}$ to 0 and $\hat{\beta}_{t_0}^{(0)}$ to the logarithm of the true median lost lifespan from (4.4.2) under H_0. At time point 15, for example, the true median lost lifespan of 10.8 corresponds to $\beta_{t_0}^{(0)} = 2.38$ and $\beta_{t_0}^{(1)} = 0$ under

the proposed log-linear regression model. We used the grid search method to minimize the proposed estimating equation in (4.3.2). In situations where $\hat{G}(\cdot)$ is not defined or is equal to 0, the minimum Kaplan-Meier survival estimate of the censoring distribution was used. The mean and standard deviation of the parameter estimates were used to evaluate the empirical distribution of $\beta_{t_0}^{(0)}$ and $\beta_{t_0}^{(1)}$ under various time points (15, 14, 13, and 12) and censoring proportions (0%, 10%, 20%, and 30%).

The results are displayed in Table 4.4.1 based on 1000 simulations with 100 observations per group. As the censoring proportion increases, the differences between parameter estimates and true values slightly increase. The empirical standard deviations also inflate as the censoring proportion increases and as t_0 decreases. In general, the proposed method provides estimates very close to the true values.

Table 4.4.1: *Mean and standard deviation of the empirical estimates of true regression parameters* $\beta_{t_0}^{(0)}$ *= 2.38, 2.29, 2.18, and 2.06 and* $\beta_{t_0}^{(1)}$*=0 at* t_0 *= 15, 14, 13, and 12; estimated median lost lifespan in control group (*$\hat{\theta}^{(0)}$*); and estimated median lost lifespan in treatment group (*$\hat{\theta}^{(1)}$*)*

t_0	$c\%$	$\beta_{t_0}^{(0)}$	SD($\beta_{t_0}^{(0)}$)	$\beta_{t_0}^{(1)}$	SD($\beta_{t_0}^{(1)}$)	$\theta^{(0)}$	$\theta^{(1)}$
15	0	2.376	0.028	0.005	0.038	10.764	10.814
	10	2.376	0.028	0.005	0.037	10.765	10.814
	20	2.378	0.027	0.003	0.038	10.781	10.812
	30	2.377	0.038	0.004	0.049	10.770	10.812
14	0	2.279	0.032	0.004	0.043	9.772	9.809
	10	2.280	0.032	0.004	0.042	9.773	9.808
	20	2.281	0.030	0.002	0.041	9.790	9.805
	30	2.281	0.040	0.002	0.055	9.783	9.802
13	0	2.172	0.035	0.004	0.047	8.775	8.810
	10	2.172	0.035	0.004	0.047	8.774	8.806
	20	2.173	0.036	0.002	0.048	8.787	8.802
	30	2.172	0.045	0.004	0.060	8.774	8.806
12	0	2.052	0.039	0.005	0.054	7.780	7.815
	10	2.051	0.039	0.003	0.054	7.776	7.803
	20	2.053	0.039	0.002	0.054	7.788	7.800
	30	2.052	0.055	0.002	0.076	7.780	7.792

4.4.2 Type I Errors

We then assessed the proposed test statistic (4.3.11) in terms of type I error probabilities to test locally the null hypothesis of $H_0 : \beta_{t_0}^{(1)} = 0$ with a significance level of 0.05 under various time points (15, 14, 13, and 12), censoring proportions (0%, 10%, 20%, and 30%), and sample sizes (50, 100, and 200). The results are displayed in Table 4.4.2. Empirical type I error probabilities are generally conservative regardless of the censoring proportion or sample size. These results are similar to those presented in previous papers using the minimum dispersion test statistic (Su and Wei, 1993; Wei et al., 1990; Ying et al., 1995; Jeong et al., 2008; Jung et al., 2009).

Table 4.4.2: *Type 1 Errors for testing the null hypothesis* $H_0 : \beta_{t_0}^{(1)} = 0$

t_0	$c\%$	n=50	n=100	n=200
15	0	0.023	0.022	0.022
	10	0.024	0.017	0.022
	20	0.014	0.015	0.021
	30	0.005	0.008	0.015
14	0	0.021	0.021	0.022
	10	0.022	0.017	0.022
	20	0.013	0.016	0.022
	30	0.010	0.010	0.015
13	0	0.026	0.024	0.019
	10	0.025	0.017	0.022
	20	0.015	0.016	0.021
	30	0.006	0.008	0.016
12	0	0.026	0.022	0.019
	10	0.021	0.017	0.022
	20	0.014	0.015	0.020
	30	0.007	0.009	0.014

4.4.3 Power Analysis

For power analysis, we generated data under the parametric proportional hazards model in (4.4.1) by increasing the values of β_{t_0} to induce differences between control and treatment groups. We assumed that $\beta_{t_0} = -0.44, -0.82, -1.18$, and -1.60 in (4.4.2), which is equivalent to increasing the differences in median lost lifespan between control and treatment by 1, 2, 3, and 4. To further illustrate the effect of beta, Figures 4.4.3 and 4.4.3 show event times generated under different scenarios. Figure 4.4.3 assumes no difference between lost lifespans in the control and treatment group, while Figure 4.4.3 assumes a difference of 4 years between lost lifespans in the control and treatment group. The event times for the treatment group in Figure 4.4.3 are noticeably longer than those for the control group, while event times in Figure 4.4.3 are similar between groups. Introducing a smaller beta increases event times in the treatment group, thus simulating larger differences in lost lifespans.

Table 4.4.3 summarizes the probabilities of rejecting the null hypothesis of $H_0 : \beta_{t_0}^{(1)} = 0$ from the simple model (4.4.3) under various scenarios. Power decreases as t_0 decreases since less observations are included in the analysis, and it increases as β_{t_0} decreases, indicating greater power to detect a larger difference between groups. Power varied slightly among different censoring proportions, occasionally increasing as the censoring proportion increases. This is possibly due to an increased number of observations included in the analysis as more observed times occur before t_0 when the censoring proportion increases. Power also increases as sample size increases, as expected.

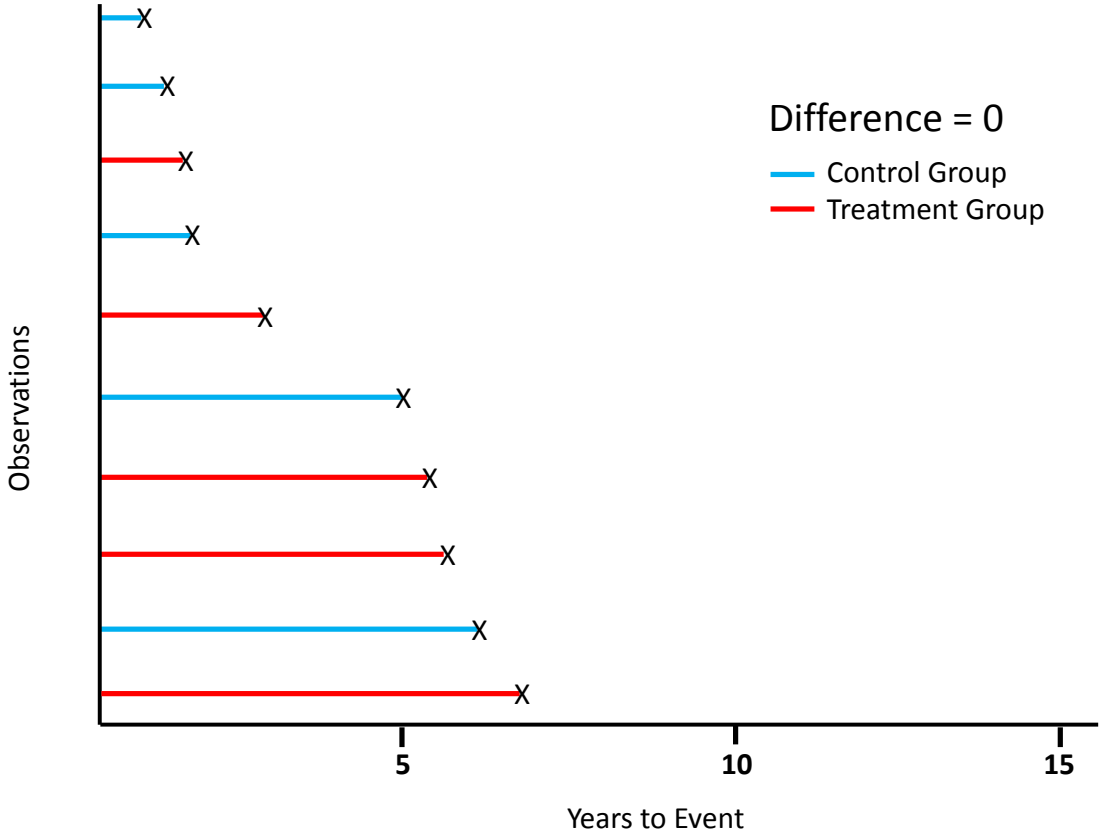

Figure 4.4.1: *Event times simulated assuming no difference between groups*

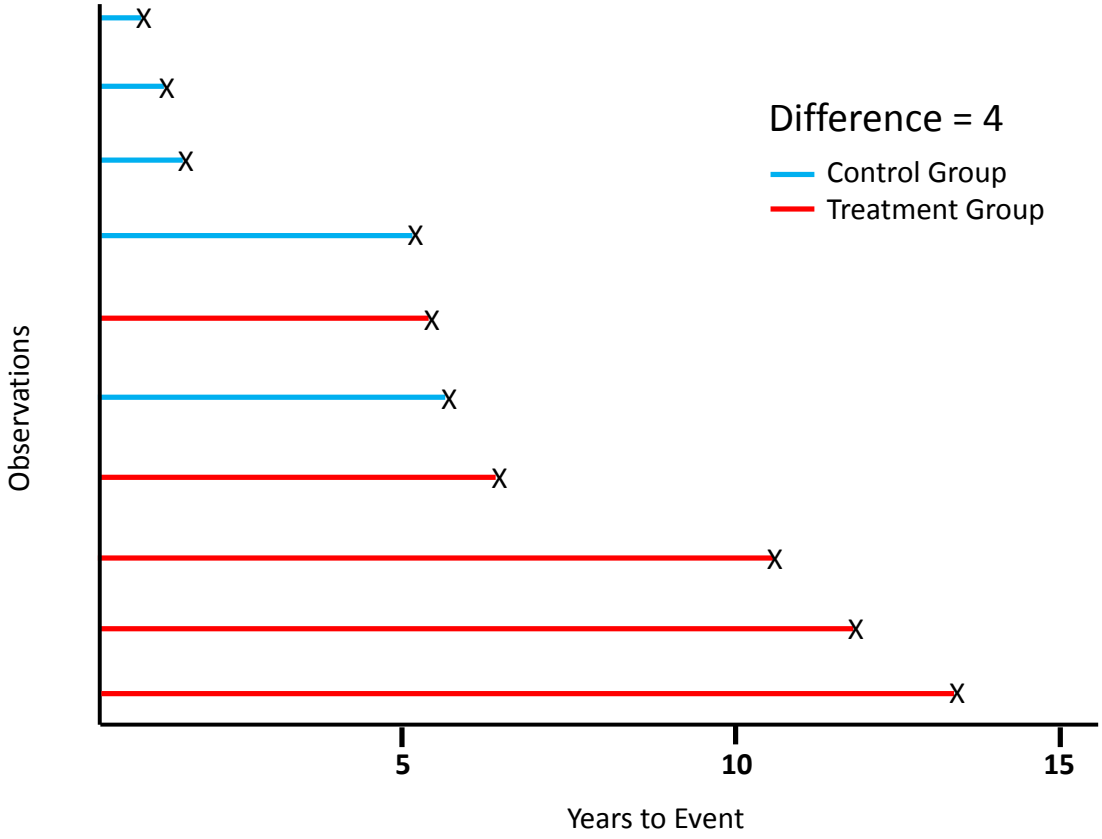

Figure 4.4.2: *Event times simulated assuming a difference of 4 years between groups*

Table 4.4.3: *Powers for testing the null hypothesis* $H_0 : \beta_{t_0}^{(1)} = 0$

		n=100				n=200			
		-0.44	-0.82	-1.18	-1.60	-0.44	-0.82	-1.18	-1.60
15	0	0.158	0.401	0.648	0.803	0.304	0.768	0.948	0.991
	10	0.150	0.409	0.624	0.786	0.314	0.795	0.969	0.993
	20	0.139	0.434	0.715	0.881	0.380	0.840	0.979	0.998
	30	0.175	0.525	0.791	0.938	0.540	0.959	0.995	1.000
14	0	0.151	0.371	0.593	0.756	0.293	0.740	0.936	0.983
	10	0.146	0.395	0.570	0.706	0.309	0.785	0.959	0.976
	20	0.139	0.429	0.722	0.828	0.367	0.839	0.983	0.995
	30	0.185	0.517	0.801	0.928	0.527	0.937	0.996	0.999
13	0	0.127	0.340	0.536	0.664	0.278	0.680	0.905	0.958
	10	0.133	0.390	0.545	0.613	0.307	0.780	0.928	0.951
	20	0.126	0.418	0.684	0.729	0.341	0.836	0.973	0.982
	30	0.167	0.523	0.793	0.898	0.520	0.934	0.995	0.999
12	0	0.125	0.290	0.448	0.555	0.262	0.618	0.849	0.913
	10	0.139	0.336	0.400	0.518	0.298	0.726	0.835	0.894
	20	0.123	0.408	0.606	0.601	0.328	0.815	0.959	0.947
	30	0.156	0.492	0.752	0.820	0.503	0.934	0.995	0.998

4.5 APPLICATION TO NSABP B-04 DATA

In this section, we apply the proposed estimation procedure and test-statistic to a real dataset from a clinical trial on breast cancer, i.e. NSABP (National Surgical Adjuvant Breast and Bowel Project) B-04 dataset (Fisher et al. 2002), introduced in the previous chapter. In addition to follow-up information, surgery type, and nodal status, the dataset also contains other covariates including age at diagnosis and pathological tumor size. Both simple regression models and multiple regression models will be considered. Three covariates will be included in the analyses to be performed in this section; nodal status as a binary covariate with 0 for node-negative and 1 for node-postive, and both age at diagnosis and pathological tumor size as continuous covariates. There were 1,079 node-negative women and 586 node-positive women. Age at diagnosis ranged from 20 to 87 years with the mean of 55.4, and pathological tumor size ranged from 0 to 250mm with the mean of 34.1mm. Additionally the median follow-up was 26 years with the overall censoring proportion of 23%. In the multivariable models, the continuous covariates were multiplied by 0.01, to satisfy the regularity conditions. The main outcome from the analysis utilizing the proposed method will be how many more years the node-positive patients are expected to lose compared to the node-negative patients at various time points post treatment, adjusted for age and tumor size.

First, we used the proposed minimum dispersion statistic in (4.3.11) to evaluate the significance of nodal status in a simple log-linear regression model (4.4.3). The test statistic was calculated at 5 time points (t_0 = 13, 15, 20, 24, and 26 years post-surgery). Table 4.5.1 summarizes the results, including the parameter estimates ($\hat{\beta}^{intercept}$ and $\hat{\beta}^{node}$), 95% confidence intervals for β^{node} calculated from equation (4.3.12), and corresponding median lost lifespans for both groups. Here, θ^{node-} represents the median lost lifespan for the node-negative group equal to $exp(\beta^{intercept})$, and θ^{node+} represents the median lost lifespan for the node-postive group equal to $exp(\beta^{intercept} + \beta^{node})$.

Note that regardless of any time point specified, the median lost lifespans were significantly different between the two nodal groups. The node positive group had consistently longer median lost lifespans across all time points indicating worse prognosis in survival. The

difference between nodal status groups also increased as time point increased, as evident by the increasing values of β^{node}.

Table 4.5.1: *Parameter estimates, 95% CIs, and corresponding median lost lifespans from simple regression models*

t_0	$\hat{\beta}^{intercept}$	$\hat{\beta}^{node}$	95% CI	$\hat{\theta}^{node-}$	$\hat{\theta}^{node+}$
13	2.054	0.125	(0.035, 0.215)	7.799	8.837
15	2.166	0.172	(0.070, 0.255)	8.723	10.360
20	2.475	0.196	(0.100, 0.300)	11.882	14.454
24	2.679	0.204	(0.135, 0.300)	14.571	17.868
26	2.762	0.217	(0.160, 0.305)	15.831	19.668

While the estimates and inference procedure from the proposed model cannot be compared directly to ones from other models, we provide the p-values from the proposed method (4.3.11) compared to the p-values from testing the significance of the nodal status parameter using the proportional hazards model and the bootstrap method in Table 4.5.2. The proportional hazards model was evaluated at the same time points, with events occurring after t_0 being administratively censored at t_0. Thus, the hazard function summarizes the cumulative information up to each specified time point in terms of the conditional instantaneous failure rate. Table 4.5.2 additionally shows the p-values from testing significance of the nodal status parameter based on the bootstrap method of variance estimation (Efron, 1979). We resampled 500 times from the original dataset with replacement and estimated the regression parameters with the proposed method. The p-values were calculated from the Wald tests by using the variance of those parameter estimates, which provided similar results at all time points. The results from the simple regression model presented here are also consistent with those from the two-sample test statistic proposed in Chapter 3.

Table 4.5.2: *P-values from the proposed minimum dispersion statistic (p-value$_{new}$), Cox model (p-value$_{cox}$), and the bootstrap method (p-value$_{bs}$)*

t_0	p-value$_{new}$	p-value$_{cox}$	p-value$_{bs}$
13	0.011	<0.0001	0.01
15	0.001	<0.0001	0.001
20	<0.001	<0.0001	<0.0001
24	<0.001	<0.0001	<0.0001
26	<0.001	<0.0001	<0.0001

For further comparison, the same data have also been previously analyzed under the quantile regression model for the median residual life (Jung et al., 2009). This method summarizes the remaining life conditioned on survival beyond a specified time point. The results of the original analysis are displayed in Table 4.5.3. The nodal status parameter was only significant through 7 years post-treatment, indicating longer remaining life in the node-negative group. Analyses were limited through 10 years post mastectomy due to heavy censoring in the tail of the distribution.

Table 4.5.3: *Regression parameter estimates from median residual life simple regression model, 95% confidence intervals, and p-values for testing* $H_0 : \beta_{t_0}^{(node)} = 0$

t_0	$\hat{\beta}_{t_0}^{(intercept)}$	$\hat{\beta}_{t_0}^{(node)}$	95% CI	p-value
0	2.54	-0.62	(-0.74, -0.47)	<0.0001
2	2.53	-0.59	(-0.77, -0.37)	<0.0001
4	2.56	-0.50	(-0.72, -0.21)	0.0003
6	2.59	-0.44	(-0.71, -0.17)	0.001
7	2.57	-0.26	(-0.57, -0.09)	0.008
8	2.54	-0.22	(-0.42, 0.05)	0.116
10	2.46	-0.09	(-0.48, 0.11)	0.343

Now we extend our analysis to a multiple log-linear regression model containing nodal status, age at diagnosis, and pathological tumor size as covariates. Each covariate was tested separately for its significance using the proposed test statistic (4.3.11). The parameter estimates and corresponding 95% confidence intervals are shown in Table 4.5.4. Except at time point 13, the nodal status covariate remained statistically significant in all other multivariable models. Additionally, the difference between node-negative and node-positive groups generally increased as time point increased, similar to the results from the simple log-linear regression models (Table 4.5.1). The covariate of age at diagnosis was significant through 15 years post-surgery, while the covariate of pathological tumor size was consistently significant in all models. The proposed regression model allows for predicting a patient's median lost lifespan for a given time point based on significantly important factors, i.e. nodal status and age at diagnosis. For example, a 30 year old woman with positive lymph nodes and tumor size of 50mm is expected to have a median lost lifespan of 11.8 years $(= \exp\{2.274 + 0.127 \times 1 - 0.390 \times (0.01 \times 30) + 0.364 \times (0.01 \times 50)\})$ at 15 years after diagnosis. In comparison, a 30 year old patient with negative lymph nodes and tumor size of 50mm is expected to have a median lost lifespan of 10.4 years at 15 years after diagnosis.

Table 4.5.4: *Parameter estimates and corresponding confidence intervals (95% CI) from multiple regression models using the proposed minimum dispersion statistic*

t_0	$\hat{\beta}^{intercept}$	$\hat{\beta}^{node}$	$\hat{\beta}^{age}$	$\hat{\beta}^{size}$
13	2.365 (2.14, 2.57)	0.072 (-0.01, 0.15)	-0.685 (-0.94, -0.34)	0.245 (0.08, 0.43)
15	2.274 (2.12, 2.65)	0.127 (0.02, 0.21)	-0.390 (-0.93, -0.22)	0.364 (0.10, 0.52)
20	2.567 (2.25, 2.72)	0.182 (0.12, 0.27)	-0.395 (-0.61, 0.06)	0.371 (0.24, 0.55)
24	2.647 (2.41, 2.86)	0.188 (0.11, 0.26)	-0.134 (-0.49, 0.17)	0.364 (0.23, 0.52)
26	2.715 (2.39, 2.88)	0.189 (0.12, 0.27)	-0.114 (-0.37, 0.37)	0.376 (0.25, 0.54)

4.6 DISCUSSION ON REGRESSION

We have proposed a new method for time-to-event analysis in a regression setting that allows for analyzing covariate effects on the quantiles of the distribution of lost lifespan. Asymptotic properties were derived for the regression parameter estimators and test statistics. Simulation studies validated the estimation and inference procedure under various scenarios, and the proposed method was illustrated with an application to a breast cancer dataset. While prognosis is well known to be worse for breast cancer patients with positive lymph nodes, the information gained from this particular application could allow a physician to explain the difference in terms of years lost at various time points after surgery.

The proposed log-linear regression model provides benefits over traditional survival models such as the proportional hazards model and accelerated failure time model. It can be easily seen that the lost lifespan-based methods are less sensitive to heavy censoring towards the end of study than the residual life-based methods. Also, the proposed summary measure provides a straightforward clinical interpretation that is invaluable to clinicians, patients, and other stakeholders. The quantile-based approach would be more robust than the mean-based one for the analysis of time-to-event data.

While the proposed regression model avoids many assumptions of the traditional survival models, it does assume a log-linear relationship between covariates and the quantile lost-lifespan. Next steps should include methods for testing this model assumption. Ying et al. (1995) proposed a graphical check, which was also implemented in the median residual life analysis of the NSABP B-04 data (Jung et al., 2009). A zero-mean Gaussian process of cumulative sums of median residuals was defined as

$$W(s) = n^{-1/2} \sum_{i=1}^{n} e_i I(\hat{\beta}'_{\epsilon|t_0} Z_i \leq z), \tag{4.6.1}$$

where e_i takes the following form

$$e_i = \frac{I\{Y_i \geq t_0 + \exp(\hat{\beta}'_{\epsilon|t_0} Z_i)\}}{\hat{G}\{t_0 + \exp(\hat{\beta}'_{\epsilon|t_0} Z_i)\}} - (1-\epsilon)\frac{I(Y_i \geq t_0)}{\hat{G}(t_0)}$$

Plotting $W(s)$ against predicted residual lifetimes could evaluate the model assumption. This method could be extended to the lost lifespan by replacing e_i with the proposed estimating equation 4.3.2 to evaluate $W(s)$ in 4.6.1.

5.0 DISCUSSION AND FUTURE RESEARCH

The primary goal of this dissertation was to develop a new method of analyzing time-to-event data that allows for more effective communication of results to clinicians and patients. Here, we consider the life lost due to an event of interest occurring before some specified time point. Thus, the time lost can be captured at various time points after diagnosis or treatment. We've extended the nonparametric based methods for quantile residual life to estimating the life lost in the presence of censoring, comparing the lost lifespans between groups, and examining covariate effects on the lost lifespan.

The proposed method provides potential for numerous extensions. Future research can involve developing methods to account for competing risks. In situations where a particular outcome is of interest, say death from breast cancer, deaths from other causes would prevent observance of the event of interest. Methods for quantile residual life analysis under competing risks have been proposed for both parametric and non-parametric settings (Jeong, 2014). The non-parametric methods proposed here could potentially be extended to the quantile lost lifespan for analysis under competing risks.

Future research could also consider regression models allowing random effects, for situations where covariates are considered to be representative of a general population. The traditional cox proportional hazards model has been extended to allow for clustered time-to-event data (Sargent, 1998; Vaida and Xu, 2000); however, these methods still rely on the proportional hazards assumption. The extension to random effects in lost lifespan modeling could be particularly useful for scenarios involving genetically or environmentally related patients or trials conducted at multiple centers.

Additionally, one limitation of the proposed methods is the computationally intensive grid search method of estimation, particularly for models with multiple covariates. Development of a technique of smoothing the estimating equation for more efficient estimation of the regression parameters might merit future research.

BIBLIOGRAPHY

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726.

Andersen, P. (2013). Decomposition of number of life years lost according to causes of death. *Statistics in Medicine*, 32(30):5278–5285.

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Andersen, P. K., Hansen, M. G., and Klein, J. P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis*, 10(4):335–50.

Balmert, L. and Jeong, J.-H. (2017). Nonparametric inference on quantile lost lifespan. *Biometrics*, 73(1):252–259.

Basawa, I. V. and Koul, H. L. (1988). Large-sample statistics based on quadratic dispersion. *International Statistical Review / Revue Internationale de Statistique*, 56(3):199–219.

Berger, R. L., Boos, D. D., and Guess, F. M. (1988). Tests and confidence sets for comparing two mean residual life functions. *Biometrics*, 44(1):103–115.

Block, H. W., Savits, T. H., and Singh, H. (2009). The reversed hazard rate function. *Probability in the Engineering and Informational Sciences*, 12(1):69–90.

Chandra, N. and Roy, D. (2001). Some results on reversed hazard ratio. *Probability in the Engineering and Informational Sciences*, 15:95–102.

Chen, Y. Q., Jewell, N. P., Lei, X., and Cheng, S. C. (2005). Semiparametric estimation of proportional mean residual life model in presence of censoring. *Biometrics*, 61(1):170–178.

Chen, Y.-Y., Hollander, M., and Langberg, N. A. (1983). Tests for monotone mean residual life, using randomly censored data. *Biometrics*, 39(1):119–127.

Chiang, C. L. (1960). A stochastic study of the life table and its applications: I. probability distributions of the biometric functions. *Biometrics*, 16(4):618–635.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Csorgo, S. and Horvath, L. (1983). The rate of strong uniform consistency for the product-limit estimator. *Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 62(3):411–426.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26.

Fisher, B., Jeong, J. H., Anderson, S., Bryant, J., Fisher, E. R., and Wolmark, N. (2002). Twenty-five-year follow-up of a randomized trial comparing radical mastectomy, total mastectomy, and total mastectomy followed by irradiation. *New England Journal of Medicine*, 347(8):567–75.

Fisher, B., Wolmark, N., Redmond, C., Deutsch, M., and Fisher, E. (1981). Findings from nsabp protocol no. b-04: Comparison of radical mastectomy with alternative treatments. ii. the clinical and biologic significance of medial-central breast cancers. *Cancer*, 48(8):1863–1872.

Fleming, T. and Harringtion, D. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Fligner, M. A. and Rust, S. W. (1982). A modification of mood's median test for the generalized behrens–fisher problem. *Biometrika*, 69(1):221–226.

Gelfand, A. E. and Kottas, A. (2003). Bayesian semiparametric regression for median residual life. *Scandinavian Journal of Statistics*, 30:651–665.

Irwin, J. O. (1949). The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *The Journal of hygiene*, 47(2):188–9.

Jeong, J.-H. (2014). *Statistical Inference on Residual Life*. Springer, New York.

Jeong, J.-H., Jung, S.-H., and Costantino, J. P. (2008). Nonparametric inference on median residual life function. *Biometrics*, 64(1):157–63.

Jung, S.-H. (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association*, 91(433):251–257.

Jung, S.-H., Jeong, J.-H., and Bandos, H. (2009). Regression on quantile residual life. *Biometrics*, 65(4):1203–12.

Kalbfleisch, J. D. and Lawless, J. F. (1989). Inference based on retrospective ascertainment: An analysis of the data on transfusion-related aids. *Journal of the American Statistical Association*, 84(406):360–372.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Karrison, T. (1987). Restricted mean life with adjustment for covariates. *Journal of the American Statistical Association*, 82(400):1169–1176.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis Techniques for Censored and Truncated Data, Second Edition*. Springer, New York.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrics*, 46:33–50.

Lai, T. L. and Ying, Z. (1988). Stochastic integrals of empirical-type processes with applications to censored regression. *Journal of Multivariate Analysis*, 27(2):334–358.

Li, R., Huang, X., and Cortes, J. (2016). Quantile residual life regression with longitudinal biomarker measurements for dynamic prediction. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(5):755–773.

Li, X. and Lu, J. (2003). Stochastic comparisons on residual life and inactivity time of series and parallel systems. *Probability in Engineering and Informational Sciences*, 17(2):267–275.

Lim, J. Y. and Jeong, J. H. (2015). Cause-specific quantile residual life regression. *Statistical Methods in Medical Research*.

Lindgren, A. (1997). Quantile regression with censored data using generalized l1 minimization. *Computational Statistics and Data Analysis*, 23:509–524.

Maguluri, G. and Zhang, C.-H. (1994). Estimation in the mean residual life regression model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):477–489.

McKeague, I., Subramanian, S., and Sun, Y. (2001). Median regression and the missing information principle. *Journal of Nonparametric Statistics*, 13:709–727.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966.

Oakes, D. and Dasu, T. (1990). A note on residual life. *Biometrika*, 77(2):409–410.

Peng, L. and Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association*, 103:637–649.

Portnoy, S. and Koenker, R. (1997). The gaussian har and the laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statistical Science*, 12:279–196.

Raja rao, B., Damaraju, C. V., and Alhumoud, J. M. (1992). Covariate effect on the life expectancy and percentile residual life functions under the proportional hazards and the accelerated life models. *Communications in Statistics - Theory and Methods*, 22(1):257–281.

Royston, P. and 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):2409–21.

Ruiz, J. M. and Navarro, J. (1996). Characterizations based on conditional expectations of the doubled truncated distribution. *Annals of the Institute of Statistical Mathematics*, 48(3):563–572.

Sargent, D. (1998). A general framework for random effects survival analysis in the cox proportional hazards setting. *Biometrics*, 54:1486–1497.

Schmittlein, D. C. and Morrison, D. G. (1981). The median residual lifetime: A characterization theorem and an application. *Operations Research*, 29(2):392–399.

Su, J. Q. and Wei, L. J. (1993). Nonparametric estimation for the difference or ratio of median failure times. *Biometrics*, 49(2):603–7.

Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, 19(24):3309–3324.

Wang, J.-L. and Hettmansperger, T. P. (1990). Two-sample inference for median survival times based on one-sample procedures for censored survival data. *Journal of the American Statistical Association*, 85(410):529–536.

Wei, L. J., Ying, Z., and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika*, 77(4):845–851.

Yin, G. and Cai, J. (2005). Quantile regression models with multivariate failure time data. *Biometrics*, 61(1):151–61.

Ying, Z., Jung, S. H., and Wei, L. J. (1995). Survival analysis with median regression models. *Journal of the American Statistical Association*, 90(429):178–184.