University of South Carolina
Scholar Commons

Theses and Dissertations

1-1-2013

A New Method For the Comparison Of Survival Distributions

Jaymie Shanahan
University of South Carolina

Follow this and additional works at: https://scholarcommons.sc.edu/etd

 Part of the Biostatistics Commons

Recommended Citation

Shanahan, J.(2013). *A New Method For the Comparison Of Survival Distributions.* (Master's thesis). Retrieved from https://scholarcommons.sc.edu/etd/555

This Open Access Thesis is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.

A New Method For The Comparison Of Survival Distributions

by

Jaymie Shanahan

Bachelor of Arts
Elon University, 2010

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in Public Health in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2013

Accepted by:

Jiajia Zhang, PhD, Director of Thesis

James Hussey, PhD, Reader

Andrew Ortaglia, PhD, Reader

Lacy Ford, Vice Provost and Dean of Graduate Studies

© Copyright by Jaymie Shanahan, 2013
All Rights Reserved.

DEDICATION

To my Mom and Dad, thank you for your endless love, encouragement, and unconditional support with my studies.

ACKNOWLEDGEMENTS

I am beyond grateful for my committee chair, Dr. Jiajia Zhang, for this thesis would not have been possible without her excellent guidance, support and patience throughout this process, not to mention her advice and unsurpassed knowledge of Survival Analysis. The good advice, support and friendship of my second committee member and advisor, Dr. James Hussey, has been invaluable on both an academic and a personal level, for which I will always be extremely grateful. Also, to my third committee member, Dr. Andrew Ortaglia, his guidance through this process and my academic career will always be greatly appreciated.

To my friends in the program who traveled every step of this journey with me, your friendship, support, love and memories will be something I will treasure forever. You all played a significant role in my life these past two years and I consider myself blessed and forever grateful for your friendships.

Finally, my family, whom I am most grateful for, this thesis is simply impossible without them. I am indebted to my parents, Bill and Stacey Shanahan, for their unconditional love and for having my absolute best interest at heart in every aspect of my life. I am honored to have you as parents, I love you.

ABSTRACT

The assessment of overall homogeneity of time-to-event curves is a key element in survival analysis in biomedical research. The currently commonly used testing methods, e.g. log-rank test, Wilcoxon test, and Kolmogorov–Smirnov test, may have a significant loss of statistical testing power under certain circumstances. In this thesis we replicate a testing method (Lin & Xu, 2009) that is robust for the comparison of the overall homogeneity of survival curves based on the absolute difference of the area under the survival curves using normal approximation by Greenwood's formula, and propose a new weight component to their test statistic. The weight component is added to Lin and Xu's test statistic to better fit the data at hand (i.e. emphasizing more weight on earlier data). Monte Carlo simulations are conducted to investigate the performance of the new testing method compared against the log-rank, Wilcoxon, Kolmogorov–Smirnov, and Lin & Xu's tests under a variety of circumstances. The proposed new weighted method has robust performance compared to the common test statistics, with greater power to detect the overall differences than the log-rank, Wilcoxon, Kolmogorov–Smirnov, and Lin & Xu's (2009) tests in many scenarios resulting from the simulations. Furthermore, the applicability of the new testing approach is illustrated in a real data example from a Leukemia analysis trial.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. INTRODUCTION

Survival analysis techniques are typically used in medicine, biology, public health, clinical trials, and epidemiology. The unique characteristic of survival data is the existence of censoring. That is, some subjects may experience the event of interest, while other subjects may not. Not experiencing the event may be due to the study period ending, loss to follow up, or a subject's withdrawal from the study. This is also known as right-censored data. An event of interest can be defined as death, disease occurrence, disease recurrence, or recovery; any of which is referred to as a "failure". Traditional statistical models, such as multiple linear regression, cannot handle censoring directly, which is why survival models are specifically designed to manage data with censored observations. If censoring is ignored in a study, the results will tend to lead to underestimation of the survival probability.

The basic objectives in survival analysis include estimating and interpreting survival probabilities, comparing survival probabilities between groups, and assessing the possible risk factors that are related to survival probability. In this thesis, we focus on methods comparing survival curves between two groups. These methods are evaluated based on the power of the method. Section 1.2 introduces specific survival notation and definitions that will be used throughout this thesis. Section 1.3 introduces data of

Leukemia patients, and the Kaplan Meier estimation method. Section 1.4 outlines the organization of the thesis.

1.2 Definition and Notation

The survival function gives the probability that a person survives longer than some specified time, t. The survival probability is a decreasing function from 1 to 0, which means that at the beginning of the study, all patients are alive and have not yet experienced the event (Klein 1997). The survival function, S(t), can be written as:

$$S(t) = P(T > t)$$

where T is defined as a random variable for an individual's survival time, and *t* represents any specific value of interest for the random variable T (Klein 1997).

The hazard function gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time *t* **(**Klein 1997**)**. The Hazard function can be written as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

where h(t) equals the limit, as Δt approaches zero of a probability statement about survival, divided by Δt, where Δt denotes a small interval of time.

The Kaplan Meier estimator is a nonparametric method that incorporates censoring for estimating survival probabilities (Klein 1997). This is the estimate of the unconditional probability of surviving beyond time t. The Kaplan Meier estimator is used to compare the survival information between different groups. For a sample of size *n*, let t_i represent distinct event times in the sample, where $t_1 \leq t_2 \leq \ldots \leq t_k$

The Kaplan Meier estimator, given below, is the estimated survival probability for any particular one of the *t* time periods. The Kaplan–Meier estimate does not change between events, nor at times when only censorings occur. It drops only at times when a failure has been observed.

The Kaplan-Meier estimator of the survival function is given by:

$$\hat{S}(t)_{KM} = \prod_{t_i \le t}(1 - \frac{m_i}{n_i})$$

Where n_i is the number of subjects at risk at the beginning of time period t_i, and m_i is the number of subjects who had the event occur during time period t_i. The ratio m_i/n_i is the overall proportion in the population failing at time t_i, (i.e., the hazard). One minus this ratio gives the probability of the hazard at time t_i.

From Greenwood's formula, the variance of the Kaplan Meier estimate of the survival function is written as:

$$\widehat{Var}\left(\widehat{S_{km}}(t)\right) = \hat{S}_{km}(t)^2 \sum_{i|t_i \le t} \frac{m_i}{n_i(n_i - m_i)}$$

The variance allows us to obtain confidence intervals for all ordered failure times. This gives us a point wise confidence interval for the survival probability, $S_{km}(t)$. Brookmeyer and Crowly (1982) constructed the confidence interval for the median survival time based on the confidence interval for $S_{km}(t)$. The methodology is generalized to construct the confidence interval for the 100*p* percentile based on a *g*-transformed confidence interval for S(t). The transformation that was used in this situation was the Loglog

transformation, where g(x) = log(-log(x)). This is also referred to as the log cumulative hazard transformation since it applies to the logarithmic function to the cumulative hazard function (Brookmeyer and Crowly 1982). This is interpreted as the true $S_{km}(t)$ at time *t* will be within the interval for 95% of experiments conducted.

1.3. EXAMPLE OF LEUKEMIA DATA

The dataset we will use as our example of real data analysis was a study designed to test the ability of a therapy to prolong the duration of remission in acute leukemia patients (Freireich et al. 1963). Ninety-two patients under age 20 entered the study and were accepted for analysis. Of those 92 patients, 62 had complete or partial remissions (Freireich et al. 1963). Patients in remission were randomly assigned to maintenance therapy with either 6-Mercaptopurine treatment applied or a placebo. A sequential experimental design was used to analyze remission times while the study was in progress, which resulted in the study being stopped after analysis of the remission times of 21 pairs of patients. Those 21 pairs of patients (42 individual patients) were used to construct the dataset for our study. The dataset described in Table 1.1 consists of 42 patients ($\leq$ 20 years) with leukemia, divided evenly into two groups: the first group had a 6-Mercaptopurine treatment applied (X_1=1), and group 2 was designated as a placebo ($X_{1=}$0). The values given for each group consist of time in weeks a patient is in remission, up to the point of the patient's either going out of remission (event of interest) or being censored (t_i). An individual was censored (δ_i=0) if they remained in remission until the end of the study, was lost to follow-up, or withdrew before the end of the study, else δ_i=1. We assume that censoring is noninformative, meaning that each subject's censoring time

is statistically independent of their failure time. Gender is indicated by X_2, where 0=males and 1=females. According to this dataset, the proportion of patients who received treatment (6-Mercaptopurine) that were censored is 12 out of 21 (9 out of 21 patients had the event occur), while 0 out of 21 patients receiving the placebo were censored (21 out of 21 had the event occur). The proportion of females who were censored is 6 out of 22 (16 out of 22 had the event occur), while the proportion of males who were censored is 6 out of 20 (14 out of 20 had the event occur).

TABLE 1.1. Description of Leukemia Data remission time in weeks (Freireich et al., *Blood*, 1963)

t_i	δ_i	X_1	X_2
1	1	1	1
1	1	1	1
2	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
4	1	1	1
5	1	1	1
5	1	1	0
6	0	2	0
6	1	2	0
6	1	2	1
6	1	2	0
7	1	2	0
8	1	1	0
8	1	1	0
8	1	1	0
8	1	1	1
9	0	2	0
10	0	2	0
10	1	2	0
11	0	2	0
11	1	1	0
11	1	1	0
12	1	1	0
12	1	1	0
13	1	2	0
15	1	1	0
16	1	2	1
17	0	2	0
17	1	1	0
19	0	2	0
20	0	2	1
22	1	2	1
22	1	1	0
23	1	1	1
23	1	1	1
25	0	1	1
32	0	1	1
32	0	1	1
34	0	1	1
35	0	1	1

For our first scenario, we consider the comparison of the treatment (Group 1) and placebo (Group 2) subjects. We test the null hypothesis that the patient groups have the same survival distribution against the alternative that the survival distributions are different. The null hypotheses are given by:

$$\text{Ho: } S_1(t) = S_2(t) \quad \text{-OR-} \quad \text{Ho: } h_1(t) = h_2(t)$$

$$\text{Ha: } S_1(t) \neq S_2(t) \qquad \text{Ha: } h_1(t) \neq h_2(t).$$

The survival and the logarithm of cumulative hazard function plots for the group comparison are shown below in Figures 1.1 and 1.2.

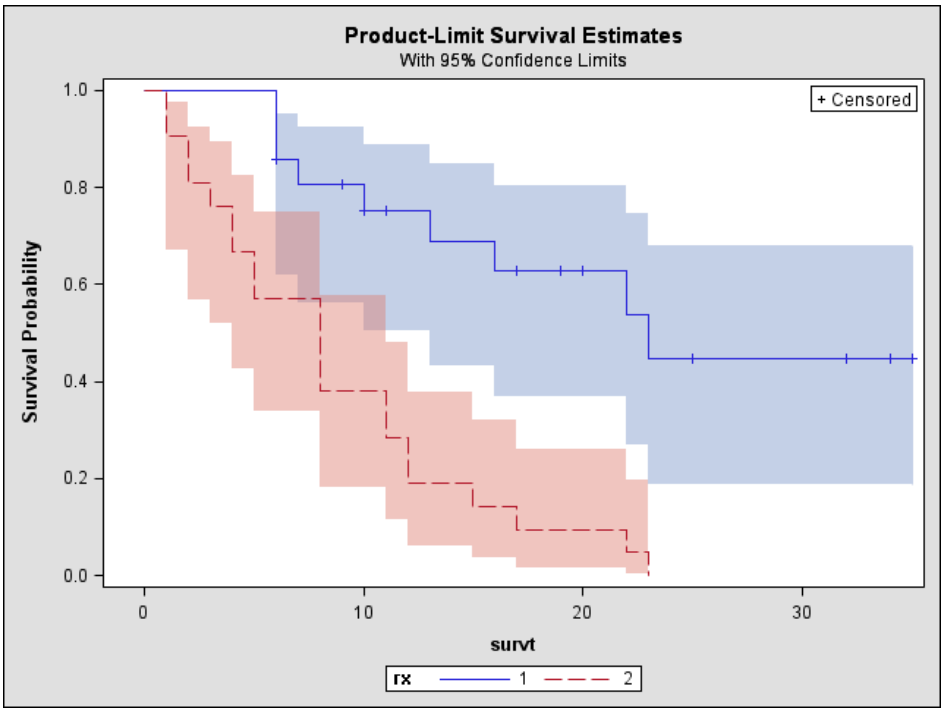

FIGURE 1.1. Kaplan Meier survival curves for Group 1 (treatment) and Group 2 (placebo)

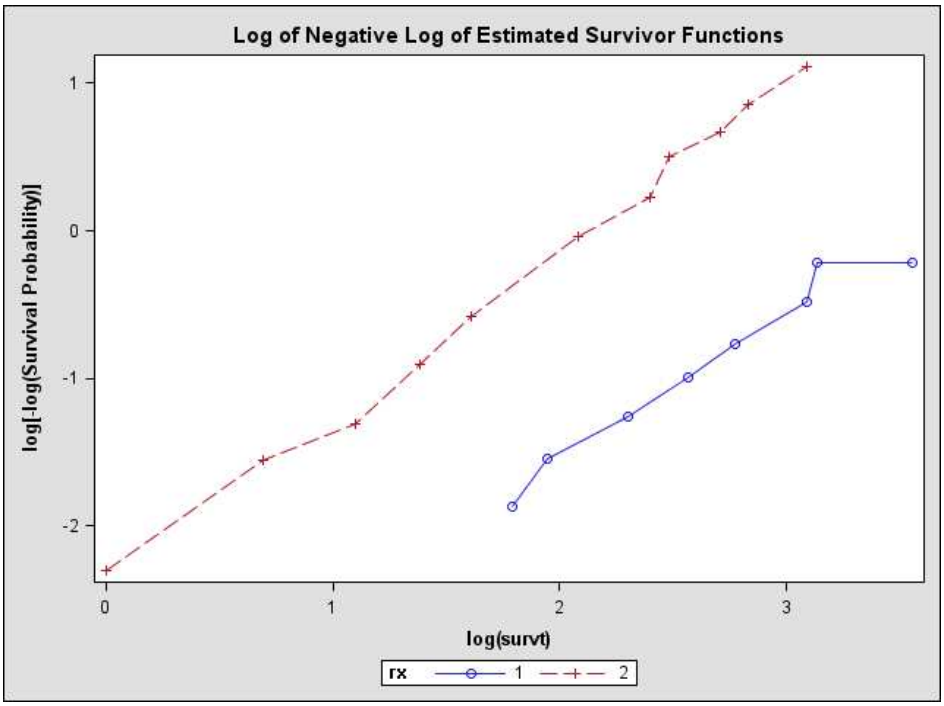

FIGURE 1.2 Logarithm of Cumulative Hazard Function curves for Group 1 (treatment) and Group 2 (placebo)

Figure 1.1 and 1.2 plots the two treatment groups by survival probability versus the survival time and display proportional hazards. From these figures, we can conclude that patients who receive the 6-Mercaptopurine treatment have a longer survival rate than the patients in the placebo group. Also, based on the fact that there is little overlap between the confidence bands in Figure 1.1, we can conclude that there may be a difference in survival probability between the two treatment groups. Various tests have been proposed for testing statistical differences in survival between categorical covariates, such as the long rank test and Wilcoxon test. These tests will be explained in further detail in chapter 2.

For our second example, we use the same dataset but stratify by gender rather than by group. Figure 1.3 and 1.4 represent the Kaplan Meier survival function and the logarithm of cumulative hazard function curves for males and females.

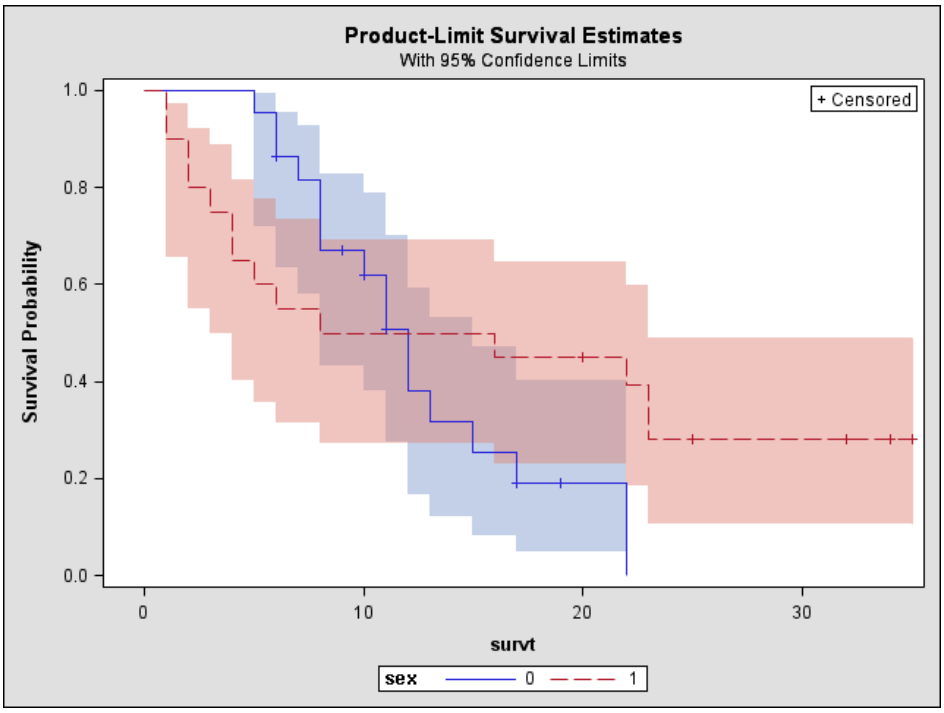

FIGURE 1.3 Kaplan Meier survival curves for males and females for males (0) and females (1)

.

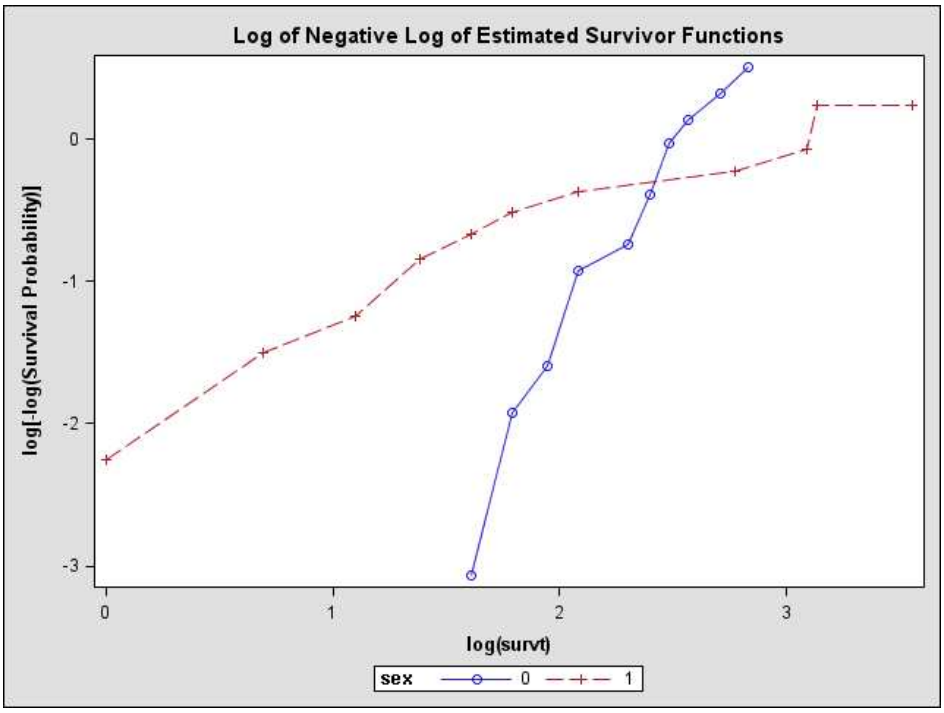

FIGURE 1.4 Cumulative Hazard Function curves for males (0) and females (1)

The survival probability and cumulative hazard function plots suggests that we cannot adjust for sex in the PH model. From Figure 1.4, we can see that while adjusting for gender, plotting the survival probability versus the survival time result in intersecting hazards. From Figure 1.3, it appears that male patients begin with a larger survival probability, but over time (after about 12 weeks) women appear to have a higher survival probability. Also, the confidence interval bands overlap each other substantially, which further supports our claim that it isn't clear there is a difference in survival between males and females.

1.4. OUTLINE OF THESIS

In Chapter 2, we describe at length the log rank test, the weighted log rank test (i.e. Wilcoxon), the Kolmogorov-Smirnov test, Lin and Xu's proposed test, and our proposed weighted version of Lin and Xu's test. For each statistical procedure introduced, we will be testing the same hypotheses; as stated in the example prior, we test the null hypothesis that the patient groups have the same survival distribution against the alternative that the survival distributions are different. The null hypotheses are given by:

$$\text{Ho: } S_1(t) = S_2(t) \quad \text{-OR-} \quad \text{Ho: } h_1(t) = h_2(t)$$
$$\text{Ha: } S_1(t) \neq S_2(t) \qquad\qquad \text{Ha: } h_1(t) \neq h_2(t).$$

Where $S_i(t)$ denotes the survival function where $i = 1$ or 2 representing the treatment and placebo group, respectively. $h_i(t)$ denotes the hazard function where $i = 1$ or 2 representing the treatment and placebo group, respectively. Chapter 3 presents a

comprehensive simulation study and describes the simulation design of the weighted Lin and Xu's test statistic under different situations at length. Chapter 4 evaluates the performances of the weighted Lin and Xu's test statistic and draws conclusions based on the results. The weighted Lin and Xu's test statistic is applied to the Leukemia dataset in Chapter 5. Finally, conclusions and potential future work are summarized in Chapter 6.

CHAPTER 2

TEST STATISTICS

Survival analysis includes a wide variety of methods for analyzing the timing of events and comparing survival functions. Throughout this chapter we will discuss the following test procedures: the log rank, the Wilcoxon test, the Kolmogorov-Smirnov test, Lin and Xu's proposed test statistic, and our extended version of Lin and Xu's test statistic.

2.1. LOG RANK TEST STATISTIC

The log rank test (Peto R. and Peto J. 1972) is a procedure used to compare the survival distributions of two samples. It is a nonparametric test and appropriate to use with right censored data (Peto 1972). The test is sometimes called the Mantel–Cox test, named after Nathan Mantel and David Cox (Mantel 1967). The log rank test statistic has a distribution that is approximately chi square in large samples (Klein 1997). According to Peto (1972), the log-rank test statistic is calculated by obtaining each distinct event time, and comparing the hazard rates between the two groups, conditional on the number at risk in the groups (Klein 1997).

The log-rank test statistic is given by:

$$x^2_{log} = \frac{(O_j - E_j)^2}{Var(O_j - E_j)} \sim x^2_{(1)},$$

where $(O_j - E_j)$ is sum of the observed minus expected counts over all failure times for one of the two groups (Klein 1997). Let j= 0 or 1 indicate the control and treatment group, respectively. When j=0, let t_{i0} be the event time of the i^{th} patient in the control group, and δ_{i0} be the censoring indicator ($\delta_{i0} = 0$) of the i^{th} patient in the control group, where $i = 1, \ldots, m_0$, where m_0 indicates the last distinct event time in the control group. When j=1 and m=1, let t_{i1} be the event time of the i^{th} patient in the treatment group, and δ_{i1} be the censoring indicator ($\delta_{i1} = 0$) of the i^{th} patient in the treatment group, where $i = 1, \ldots, m_1$, where m_1 indicates the last distinct event time in the treatment group. $(O_j - E_j)$ is given by the formula:

$$(O_j - E_j) = \sum_{i=1}^{\#\, of failure times} (m_{ij} - e_{ij}) = \sum_{i=1}^{\#\, of failure times} (m_{ij}) - \sum_{i=1}^{\#\, of failure times} (e_{ij}),$$

Where m_{ij} denotes the total number of events in group j at failure time t_i, and e_{ij} indicates the expected total number of events in group j at failure time t_i.

If H_0 is true, e_{ij} can be obtained by the following formula:

$$e_{ij} = \frac{n_{ij}}{ni1+ni2} * (m_{i1} + m_{i2}),$$

where $n_{ij}$is defined by the number of subjects at risk at time t_i, separated by group *j*. The expected cell counts must be computed for both groups.

The variance of the difference between observed and expected values can be obtained by the following formula (Klein 1997):

$$Var(O_j - E_j) = \sum_j \frac{n_{i1}n_{i2}(m_{i1}+m_{i2})(n_{i1}+n_{i2}-m_{i1}-m_{i2})}{(n_{i1}+n_{i2})^2(n_{i1}+n_{i2}-1)}$$

An approximation to the log rank statistic can be calculated using observed and expected values for each group without having to compute the variance formula given above. The approximate formula sums over each group being compared the square of the observed minus expected value divided by the expected value. The approximate formula is given by:

$$\chi^2_{approx} = \frac{(O_j-E_j)^2}{E_j} \sim \chi^2_{(1)}$$

Where O_j is the number of observed events for group *j* and *Ej* is the number of expected events for group *j*. For a two-sided test with significance level α, the null hypothesis is rejected if $\chi^2_{log} > \chi^2_{(1,1-\alpha)}$. The log rank test is will be optimal test when the

underlying hazard rates are proportional to each other. (X. Lin, Q. Xu, 2010). This test statistic has been implemented in most statistical softwares, such as, SAS and R, for easy use in practice. For example, in our study, we implemented PROC LIFETEST in SAS to obtain the log rank test statistics and used the statistical software program, R, by using the R package Survival, and the command "logrank(Surv(time, status))".

2.2. GENERALIZED WILCOXON TEST STATISITC

An additional test statistic for detecting statistical differences between survival curves is the generalized Wilcoxon test. This is a variation of the log rank test that applies a different weight at the j^{th} failure time (R. Peto and J. Peto 1972). While the log rank test gives equal weight to early and late failures, the Wilcoxon test statistic gives more weight to earlier failure times. In other words, the Wilcoxon test places more emphasis on the information at the beginning of the survival curve where the number at risk is large allowing early failures to receive more weight than later failures. An example of when one might choose to use the Wilcoxon test would be when we are interested in investigating whether a treatment is most effective in the earlier phases of administration and tends to be less effective over time. Therefore, this test statistic is considered more powerful than the log rank test in detecting early survival differences (Klein 1997). The Wilcoxon test statistic has high power when the failure times are lognormally distributed, with equal variance in both groups but a different mean (Klein 1997).

The Wilcoxon Test Statistic is given by:

$$\chi^2_{wilcoxon} = \frac{(\sum_{i=1}^{k} w(t_i)(m_{ij} - e_{ij}))^2}{Var(\sum_{i=1}^{k} w(t_i)(m_{ij} - e_{ij})} \quad , \quad \text{where, } w(t_i) = n_{ij}$$

Note: if $w(t_i) = 1$ then this formula reduces to the log rank test.

At a significance level α, the null hypothesis is rejected if $\chi^2_{wilcoxon} > \chi^2_{(1,1-\alpha)}$.

While the log rank and Wilcoxon are the two tests that are used most frequently, there are several other test statistics with a variety of weight settings that will be introduced later in this chapter. These tests will lack power if the survival curves are not proportional (Schoenfeld 1981). However, that does not necessarily mean that the tests are invalid. This test statistic has been implemented in most statistical software, such as SAS and R, for easy use in practice. For example, in our study, we implemented PROC LIFETEST in SAS to obtain the Wilcoxon test statistics and used the statistical software program, R, using the Survival function, and the command "Wilcoxon.test()".

2.3. KOLMOGOROV-SMIRNOV TEST STATISTIC

A method that can be used when the underlying PH assumption is not met is the Kolmogorov-Smirnov (KS) test statistic (Massey 1951). The KS-test is suitable for this "purpose so that it robustly works in the condition where the hazard functions $h_1(t)$ and $h_2(t)$ cross over through time *t*" (Massey 1951). This test statistic has been known to have greater power than the log-rank and Wilcoxon tests when survival curves are not

proportional. However, it is not guaranteed to have more power when the survival curves cross.

The Kolmogorov-Smirnov test is based on the following equation.

$$D[0,\tau] = sup_{0 \le t \le \tau} |\hat{s}_1(t) - \hat{s}_2(t)|$$

Where *sup* represents a supremum of a set that gives the smallest real number that is greater than or equal to every number in the set, and *D* represents the largest absolute vertical deviation (Massey 1951). τ denotes the last time point by which the areas under the survival curves can be calculated for both groups based on the data available. $|\hat{s}_1(t) - \hat{s}_2(t)|$ represents the absolute difference between survival estimates for group 1 and group 2. The Kolmogorov-Smirnov test has the advantage of making no assumption about the distribution of the data. It is a non-parametric and distribution free test statistics. The hypothesis regarding the distributional form is rejected if the test statistic, D, is greater than the critical value obtained from a table. There are several variations of these tables in the literature that use somewhat different scalings for the K-S test statistic and critical regions (Massey 1951). These alternative formulations should be equivalent, but it is necessary to ensure that the test statistic is calculated in a way that is consistent with how the critical values were tabulated. For our study, this test statistic has been implemented using R package Surv2Sample, and the command "surv2.ks".

2.4. LIN &XU'S TEST STATISTIC

Xun Lin and Qiang Xu (2009) proposed a new testing method that is robust for the comparison of the overall homogeneity of survival curves. They based this test statistic on the absolute difference of the area under the survival curves using a normal approximation based on the Kaplan Meier Estimator. Their objective with this test statistic was to be able to use it for all-purpose situations whether the hazards are proportional to each other or not. Their new testing method "not only considers the difference in Kaplan Meier estimates between the groups at the actual event time points, it also takes into consideration the length of time intervals to better capture the differences" (2009). Some notation that needs to be introduced for this test statistic are as follows. The estimated survival probabilities used for this test statistic are from three sources: the control group, the treatment group, and the pooled data set of both groups. The notation used in chapter 1 for the Kaplan Meier estimator is consistent with the Kaplan Meier estimator for Lin and Xu's test statistic, given below. The survival distribution for each group at time *t*, $S_j(t)$, can be estimated by the Kaplan-Meier estimator $\hat{S}_j(t_i)$, where

$$\hat{S}_j(t) = \prod_{i|t_i \le t}\left(1 - \frac{m_{ij}}{n_{ij}}\right)$$

The observed absolute difference of the areas under the two survival curves is defined below

$$\Delta = \int_0^{\tau} \left|\hat{S}_1(t) - \hat{S}_2(t)\right| dt = \sum_{i|t_i \le t} \left|\hat{S}_1(t_i) - \hat{S}_2(t_i)\right| (t_{i+1} - t_i)$$

where τ is the last time point by which the areas under the survival curves can be calculated for both groups based on the data available (Lin and Xu 2009). In detail, $\tau = \min_j(X_{j(Nj)})$ if the last time points in the two groups are both censored, where $X=\min(t_i,\delta_i)$ assuming δ_i is independent of t_i.and where N_j denotes the sample size in group j. For example, two Kaplan–Meier curves are both open at the right tails; $\tau = \max_j(X_{j(Nj)}(1-\delta_{j(Nj)}))$ if the last time point in one group is an actual event, and the one in the other group is censored, i.e. one survival curve is open, and the other one is closed; and $\tau = \max_j(X_{j(Nj)})$ if the last time points in both groups are actual events, i.e. both survival curves are closed at the right tails. Note, in the right hand side of equation, the t_{j+1} for the last element in the summation is defined as τ instead. (Lin and Xu 2009)

The expected value of delta (E(Δ)), can be estimated by:

$$\hat{E}(\Delta) = \sum_{i|t_i \le t} \{\frac{2}{\pi}[\hat{\sigma}^2_{s_1}(t_i) + \hat{\sigma}^2_{s_2}(t_i)]\}^{1/2}(t_{i+1} - t_i),$$

where $\hat{\sigma}^2_{s_1}$ = var($\widehat{S_1(t)}$), and $\hat{\sigma}^2_{s_2}$ = var($\widehat{S_2(t)}$).

The variance of delta, Var(Δ) can be estimated as follows

$$\widehat{Var}(\Delta)=\sum_{i|t_i\le\tau}(t_{i+1}-t_i)^2(1-\frac{2}{\pi})[\hat{\sigma}^2_{s_1}(t_i)+\hat{\sigma}^2_{s_2}(t_i)]+\sum_{i<i'|t_i,t_i'\le t}(t_{i+1}-t_i)\quad(t_{i'+1}-$$

$$t_{i'})\quad(1-\frac{2}{\pi})*\{[\,\hat{\sigma}^2_{s_1}(t_i)+\hat{\sigma}^2_{s_2}(t_i)]\,\hat{\sigma}^2_{s_1}(t_i')+\hat{\sigma}^2_{s_2}(t_i')]\}^{1/2}$$

where i≠i'

Given what's above, the following test statistic proposed by Lin and Xu for the comparison of the survival curves between the two groups is

$$\Delta^* = \frac{\Delta-\widehat{E(\Delta)}}{\sqrt{\widehat{Var}(\Delta)}}$$

Δ^* is asymptotically standard normal (Lin and Xu 2009). The null hypothesis is rejected at a significance level α if $\Delta^* > z_{1-\alpha}$, where $z_{1-\alpha}$ is the point with cumulative probability (1-α) in a standard normal distribution.

2.5. Proposed Weighted Lin & Xu's Test Statistic

The proposed weighted test statistic is a variation of Lin and Xu's test statistic that applies a different weight at the i^{th} failure time. The purpose of adding this weight component to the new test statistic is to compare the power of the test to Lin and Xu's original propose test statistic to see if a specific weight function is considered more powerful for all-purpose situations. The various tests that will be applied to our simulation are shown below along with their corresponding weight components.

TABLE 2.1 Test statistics whose weights were applied to Lin and Xu's test statistic

Test Statistic	**W(t_i)**
Log rank	1
Gehen	n_i
Tarone-Ware	$\sqrt{n_i}$
Peto	$\tilde{s}(t_i)$
MPeto	$\frac{n_i}{n_i+1} * \tilde{s}(t_i)$

Similar to the Gehen (i.e. Wilcoxon) test, the Tarone-Ware test statistic applies more weight to the early failure times by weighting the observed minus expected score at time t_i by the square root of the number at risk, $\sqrt{n_i}$. The Peto test weights the i[th] failure time by the survival estimate, $\tilde{s}(t_i)$, calculated overall groups combined. The survival estimate $\tilde{s}(t_i)$ is similar but not exactly equal to the Kaplna-Meier survival estimate. The MPeto weight takes the ratio of the number at risk at time t_i divided by the number at risk at time t_i *plus 1* multiplied by the survival estimate, $\tilde{s}(t_i)$, calculated overall groups combined (Klein 1997)**.**

The observed absolute difference of the areas under the two survival curves with the applied weight is defined below.

$$\Delta_w = \int_0^\tau W(t_i) * \left|\hat{S}_1(t) - \hat{S}_2(t)\right| dt = \sum_{i|t_i \le t} W(t_i) * \left|\hat{S}_1(t_i) - \hat{S}_2(t_i)\right| (t_{i+1} - t_i)$$

The expected value of the weighted delta can be estimated by:

$$\hat{E}(\Delta_w) = \sum_{i|t_i \le t} W(t_i) * \{\frac{2}{\pi}[\hat{\sigma}^2_{s_1}(t_i) + \hat{\sigma}^2_{s_2}(t_i)]\}^{1/2}(t_{i+1} - t_i)$$

Furthermore, the estimated variance of the weighted delta can be estimated by

$$\widehat{Var}(\Delta_w) = \sum_{i|t_i \le t} W(t_i)^2 * (t_{i+1} - t_i)^2(1 - \frac{2}{n})[\hat{\sigma}^2_{s_1}(t_i) + \hat{\sigma}^2_{s_2}(t_i)] + \sum_{i<i'|t_i, t_i' \le t}(t_{i+1} -$$

$$t_i)\ \ (t_{i'+1} - t_{i'})\ \ (1-\frac{2}{n})*\{[\ \hat{\sigma}^2_{s_1}(t_i) + \hat{\sigma}^2_{s_2}(t_i)]\ \hat{\sigma}^2_{s_1}(t_i') + \hat{\sigma}^2_{s_2}(t_i')]\}^{1/2} * W(t_i)^2$$

Where i≠i'

Therefore, we obtain a weighted test statistic given by

$$\Delta_w^* = \frac{\Delta_w - E\widehat{(\Delta_w)}}{\sqrt{\widehat{Var}(\Delta_w)}}$$

Δ_w^* is asymptotically standard normal (Lin and Xu, 2009). The null hypothesis is rejected at a significance level α if $\Delta_w^* > z_{1-\alpha}$, where $z_{1-\alpha}$ is the point with cumulative probability (1-α) in a standard normal distribution.

CHAPTER 3

SIMULATION STUDY

Lin and Xu (2009) conducted a simulation study to examine the statistical power of their proposed test statistic, Δ^*, under a variety of situations. They claimed that Δ^* was robust in the comparison of overall homogeneity of survival curves whether or not the underlying PH assumption was met. However, they did not investigate the weighted version of this test statistic. Therefore, in our study, we performed a similar Monte Carlo simulation to compare the statistical power of Lin and Xu's testing method, the new weighted version of their testing method, the log rank test, the Wilcoxon test, and the Kolmogorov-Smirnov test.

3.1 SIMULATION DESIGN

As stated earlier, we consider the same three situations utilized by Lin and Xu (2009). In Situation 1, we have two survival curves that intersect one another. In Situation 2, the two survival curves are identical in the beginning, then separate as time goes on. In Situation 3, the two survival curves have proportional hazard rates. Figure 3.1 displays each situation by plotting the survival time versus the survival probability. Similar to Lin and Xu's simulation study (2009), we conducted 1000 iterations in each simulation study and exhibited the statistical power of the log rank test,

Wilcoxon test, Kolmogorov-Smirnov test, Lin and Xu's test (2009), and the proposed weighted test statistic. Comparable to Lin and Xu (2009), the estimated statistical power was calculated as the proportion of 1000 repeated random samples where we reject the null hypothesis at the 0.05 significance level.

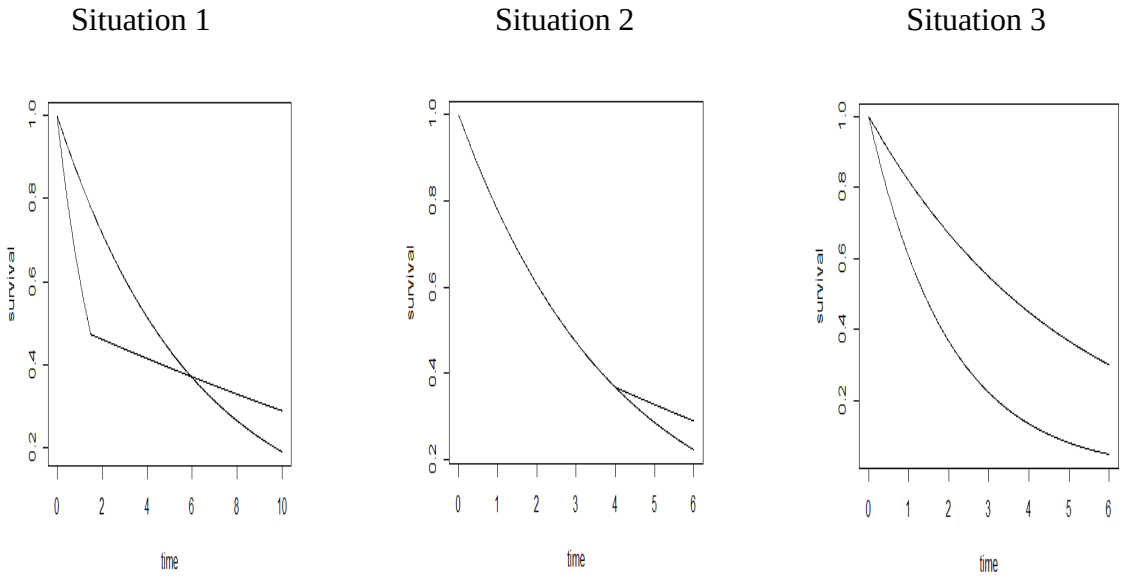

FIGURE 3.1.Three situations that are considered in the simulation study.

For Situation 1, we considered the same situation as Lin and Xu (2009) in where the survival times in Group I follow an exponential distribution with mean of 6, and in Group II the survival times follow an exponential distribution with mean of 2. Yet, if the survival time in Group II is greater than or equal to 1.5, then the survival time is simulated to follow an exponential distribution with mean of 40 (see Figure 3.1, Situation 1). Also akin to Lin and Xu's (2009) study, we considered censoring to better evaluate the performance of the tests in Situation 1. In the first situation we considered no censoring, while the second scenario we based censoring on a fixed period of follow-up.

If the survival time is greater than 12, then the observation was censored. The censoring survival time for this situation plus each of the following situations were chosen based on Lin and Xu's (2009) simulation study to keep the studies as comparable as possible with a censoring rate for group 1 and group 2 of about 14 percent and 19 percent, respectively.

For Situation 2 (see Figure 3.1, Situation 2), both Groups I and II have survival times that follow an exponential distribution with mean of 4. However, if the survival time in Group II is greater than or equal to 4, then the survival time is simulated to follow an exponential distribution with a mean of 20. Again, for better assessment of performance of the tests, the following censoring scenarios were considered. For scenario 1, no censoring was considered, while for scenario 2 censoring was based on a fixed period of follow-up. If the survival time is greater than 12, then the observation was censored. The censoring rates for group 1 and group 2 are about 5 percent and 14 percent, respectively.

For Situation 3 (see Figure 3.1, Situation 3), Group I follows an exponential distribution with mean of 2, while Group II follows an exponential distribution with a mean of 5. For better judgment of performance for each test, the following censoring scenarios were considered. For scenario 1 no censoring was considered, while for scenario 2 censoring was based on a fixed period of follow-up. If the survival time is greater than 10, then the observation was censored. The censoring rate for group 1 and group 2 are about 1 percent and 14 percent, respectively.

Not only are we interested in comparing the power of the test, we also want to investigate the estimation of the Type I error for each test statistic. The Type I error is the probability of rejecting the null hypothesis given the null is true. In terms of this

study, it is the probability that the survival curves are different when in actuality, they are the same. We want this probability to be near the nominal value of 0.05. Not only is it important to have a high power for the newly proposed test statistic, but we also want a small probability of falsely rejecting the null hypothesis. The test can have high power, but with a high probability of false rejections, which would make for a poor testing procedure.

To estimate the Type I error we created Situation 4 where we considered the same scenario as Lin and Xu (2009) where two random samples were generated independently from an exponential distribution with mean of 4. Again, for better judgment of performance, the following censoring scenarios were considered. For scenario 1, no censoring was considered, while for scenario 2 censoring was based on a fixed period of follow-up. If the survival time is greater than 10, then the observation was censored. According to Lin and Xu (2009), the censoring rate is approximately 8% in each of the two groups.

Situation 5 is another simulation created to test the Type I error. We generated two random samples independently that follow an exponential distribution with mean of 2. If the data point was greater than or equal to 1.5 then the data point was re-generated to follow an exponential distribution with mean of 6. The same censoring scenarios were considered as the previous Type I error simulation. According to Lin and Xu (2009), the censoring rate is approximately 9% in each of the two groups.

The same sample sizes that were created in Lin and Xu's study were replicated for our study. These vary from 20 to 100 in each group, representing equal and unequal sample sizes. The following sample size combinations were used for each situation and

both censoring scenarios, (group1, group2): (20, 20), (40, 40), (50, 50), (60, 60), (80, 80), (100, 100), (20, 50), (50, 20), (50, 100), (100, 50).

For our study we are expecting to see several different outcomes. While we believe that the new weights added to Lin and Xu's test statistic will perform better, we expect different weights to work better for different situations. For Situation 1 and 2, we believe that the Peto weight will perform the best because it is the most flexible among the weights we are testing. We are expecting the Wilcoxon and Tarone-Ware weights will perform the worst for Situation 2 since we would want the weight to depend more on later failure times, where the Wilcoxon and Tarone-Ware weigh more on early failure times. In Chapter 4 we will present the results of the power for each of the three situations, as well as, the Type I error estimations

CHAPTER 4

SIMULATION RESULTS

In this chapter we present the results based on the simulation design that was presented in Chapter 3. Tables 4.1 – 4.6 show the results of the power of the four existing tests plus the four newly proposed weighted tests that were applied to each of the three situations.

4.1. SIMULATION RESULTS

Tables 4.1 and 4.2 below show the power results for Situation 1 (see description, p. 27). Based on previous research, the log rank test statistic is the least powerful in a situation where the survival curves overlap (Schoenfeld 1981). This is apparent in Tables 4.1 and 4.2 where the log rank test does not exceed a power 0.715, and an average power of 0.282, averaging across both censoring and non-censoring scenarios. The Wilcoxon test also performed poorly for situation 1 for both censoring and non-censoring scenarios with a power not exceeding 0.303. As noted earlier, it is expected that the Kolmogorov-Smirnov test to perform reasonably well in this situation. However, in Tables 4.1 and 4.2, we see that results for Lin & Xu's test statistic for the non-censoring scenario performs considerably better than the Kolmogorov-Smirnov test. These results are consistent with Lin and Xu's simulation results (2009).

TABLE 4.1 Power of the four test plus weighted tests in Situation 1 (without censoring)

Sample size	**Log-Rank Test**	**Wilcoxon Test**	**Kolmogorov-Smirnov Test**	**Lin &Xu Log-Rank**	**Weight Gehen**	**Weight Tarone**	**Weight Peto**	**Weight MPeto**
(20,20)	0.186	0.100	0.207	0.679	0.691	0.699	0.691	0.681
(40, 40)	0.344	0.120	0.545	0.971	0.981	0.981	0.981	0.979
(50, 50)	0.429	0.121	0.712	0.990	0.995	0.995	0.995	0.995
(60, 60)	0.482	0.134	0.877	0.999	0.998	0.999	0.998	0.998
(80, 80)	0.597	0.164	0.986	1	1	1	1	1
(100, 100)	0.715	0.167	1	1	1	1	1	1
(20,50)	0.110	0.049	0.318	0.973	0.948	0.967	0.948	0.945
(50, 20)	0.396	0.180	0.491	0.747	0.781	0.800	0.781	0.772
(50, 100)	0.376	0.091	0.875	1	1	1	1	1
(100, 50)	0.662	0.211	0.978	0.999	1	1	1	1

TABLE 4.2 Power of the four test plus weighted tests in Situation 1 (with censoring)

Sample size	Log-Rank Test	Wilcoxon Test	Kolmogorov-Smirnov Test	Lin &Xu Log-Rank	Weight Gehen	Weight Tarone	Weight Peto	Weight MPeto
(20,20)	0.090	0.122	0.224	0.223	0.282	0.247	0.282	0.284
(40, 40)	0.104	0.157	0.537	0.567	0.637	0.597	0.637	0.639
(50, 50)	0.128	0.158	0.683	0.722	0.790	0.772	0.790	0.790
(60, 60)	0.140	0.186	0.805	0.858	0.906	0.890	0.906	0.904
(80, 80)	0.165	0.227	0.942	0.973	0.985	0.979	0.985	0.985
(100, 100)	0.182	0.303	0.987	1	1	1	1	1
(20,50)	0.046	0.074	0.370	0.430	0.465	0.447	0.465	0.467
(50, 20)	0.153	0.191	0.385	0.291	0.358	0.327	0.358	0.358
(50, 100)	0.098	0.142	0.852	0.921	0.955	0.939	0.955	0.955
(100, 50)	0.231	0.266	0.884	0.924	0.958	0.951	0.958	0.958

For this study, we are concerned with how the weighted tests performed for Situation 1. Table 4.1 shows each of the four weighted tests perform just as well as Lin and Xu's test statistic in terms of power. However, there is no evidence of one weighted test performing better than the other. When considering censoring, Table 4.2 suggests that the tests using Gehen weight, Peto weight, and MPeto weight perform better than Lin and Xu's test statistic, as well as, the log rank, Wilcoxon, and Kilmogorov-Smirnov test statistics. The Tarone weight test suggests that the power is similar to the power of Lin and Xu's test, however, it not as powerful as the other three proposed weighted tests. The Tarone weight emphasizes weight on earlier failure times which does not apply for this situation. Based on the results of the tables above, there does appear to be an influence on sample size

Tables 4.3 and 4.4 show the power results for Situation 2 (see description, p. 28). In this situation, the two survival curves are close in the beginning then separate as time goes on. Based on previous research, it is expected that the log rank and Wilcoxon tests are less powerful than all the other tests. In Table 4.3 and 4.4, the log rank test does not exceed a power of 0.1. In this particular situation, a logical weight emphasis should be considered on later failure times, thus, the Wilcoxon test performed poorly in this situation with a power that doesn't surpass 0.32 and averages a power of 0.171 for both censoring and non-censoring scenarios. Table 4.3 and 4.4 suggests that the Kolmogorov-Smirnov test performs better than log rank and Wilcoxon tests, however, it does not perform as well as Lin & Xu's test statistic, which, like Situation 1, is consistent with Lin & Xu's results (2009).

TABLE 4.3 Power of the four test plus weighted tests in Situation 2 (without censoring)

Sample size	Log-Rank Test	Wilcoxon Test	Kolmogorov-Smirnov Test	Lin &Xu Log-Rank	Weight Gehen	Weight Tarone	Weight Peto	Weight MPeto
(20,20)	0.052	0.094	0.075	0.116	0.113	0.115	0.113	0.118
(40, 40)	0.050	0.148	0.181	0.331	0.257	0.278	0.257	0.245
(50, 50)	0.041	0.183	0.220	0.502	0.337	0.425	0.337	0.329
(60, 60)	0.052	0.241	0.305	0.651	0.457	0.596	0.457	0.443
(80, 80)	0.061	0.294	0.409	0.864	0.569	0.794	0.569	0.561
(100, 100)	0.055	0.334	0.481	0.936	0.725	0.883	0.725	0.716
(20,50)	0.028	0.104	0.159	0.403	0.245	0.312	0.245	0.244
(50, 20)	0.097	0.158	0.123	0.142	0.158	0.145	0.158	0.159
(50, 100)	0.025	0.206	0.336	0.848	0.526	0.747	0.526	0.514
(100, 50)	0.097	0.266	0.324	0.555	0.412	0.540	0.412	0.408

TABLE 4.4 Power of the four test plus weighted tests in Situation 2 (with censoring)

Sample size	Log-Rank Test	Wilcoxon Test	Kolmogorov-Smirnov Test	Lin &Xu Log-Rank	Weight Gehen	Weight Tarone	Weight Peto	Weight MPeto
(20,20)	0.058	0.111	0.109	0.097	0.129	0.115	0.129	0.132
(40, 40)	0.059	0.170	0.210	0.172	0.208	0.186	0.208	0.211
(50, 50)	0.049	0.204	0.261	0.197	0.267	0.231	0.267	0.267
(60, 60)	0.058	0.217	0.317	0.249	0.299	0.276	0.299	0.300
(80, 80)	0.049	0.270	0.398	0.386	0.389	0.386	0.389	0.390
(100, 100)	0.054	0.312	0.490	0.505	0.466	0.476	0.466	0.465
(20,50)	0.034	0.107	0.174	0.134	0.169	0.153	0.169	0.170
(50, 20)	0.092	0.143	0.121	0.120	0.153	0.133	0.153	0.155
(50, 100)	0.046	0.233	0.384	0.339	0.357	0.345	0.357	0.356
(100, 50)	0.084	0.252	0.310	0.278	0.315	0.398	0.315	0.317

When considering non-censoring, Table 4.3 shows that the power of Lin and Xu's test statistic performs better than the four proposed weighted tests. The weighted Tarone test shows similar power results, however, for this situation, it would not make sense to add weight to earlier failure times when the survival curves are analogous. When considering censoring, Table 4.4 suggests that the power using Kolmogorov-Smirnov, Lin &Xu, Gehen weight, Peto weight, Tarone weight, and MPeto weight perform similarly. However, the four weighted tests performing slightly better than Lin & Xu's test statistic. Based on the results of the tables above, there does appear to be an influence on sample size.

Tables 4.5 and 4.6 show the power results for Situation 3 (see description, p. 28). In this situation, the two survival curves are proportional to one another. Based on previous research, it is expected that the Log Rank Test would perform best in this type of scenario, while Wilcoxon and Kolmogorov-Smirnov would not perform nearly as well. We would also expect Lin and Xu's test to perform as well as Log Rank in this situation when the survival curves are proportional. In Table 4.5 and 4.6, the Log Rank test is consistently the most powerful with a power reaching 1 nearly every time. In this particular situation, a logical weight emphasis should be distributed evenly over the time (hence, why Log Rank is most powerful). Therefore, the Wilcoxon test was less powerful with an average power of 0.900 for both censoring and non-censoring scenarios. Table 4.5 and 4.6 suggests that the Kolmogorov-Smirnov test performs about the same as the Wilcoxon, with the log rank test surpassing both these tests. However, Lin & Xu's test statistic performs just as well as the Log Rank Test with powers consistently reaching 1, which is consistent with Lin and Xu's study (2009).

Now that we have confirmed that our results are similar to Lin and Xu's study (2009), we are concerned with the following weighted tests. When considering Situation 3, it is evident in Tables 4.5 and 4.6, that there isn't a particular weighted test that is consistently most powerful, nor are they better than Lin & Xu's or Log Rank test. These results provide evidence that the four newly proposed weighted tests can be considered just as powerful as the Log Rank and Lin & Xu's test when comparing survival curves that are proportional to one another. Based on the results of the tables above, there does appear to be an influence of sample size.

TABLE 4.5 Power of the four test plus weighted tests in Situation 3 (without censoring)

Sample size	Log-Rank Test	Wilcoxon Test	Kolmogorov-Smirnov Test	Lin &Xu Log-Rank	Weight Gehen	Weight Tarone	Weight Peto	Weight MPeto
(20,20)	0.776	0.680	0.697	0.744	0.725	0.749	0.725	0.719
(40, 40)	0.976	0.931	0.962	0.965	0.945	0.963	0.945	0.945
(50, 50)	0.994	0.973	0.989	0.989	0.980	0.986	0.980	.979
(60, 60)	0.998	0.988	0.994	0.995	0.990	0.993	0.990	0.990
(80, 80)	0.999	0.998	0.999	0.999	1	1	1	1
(100, 100)	1	1	1	0.999	1	1	1	1
(20,50)	0.939	0.876	0.812	0.933	0.903	0.922	0.903	0.900
(50, 20)	0.906	0.787	0.908	0.843	0.825	0.847	0.825	0.821
(50, 100)	0.999	0.997	0.999	0.999	0.998	0.998	0.998	0.998
(100, 50)	0.999	0.994	0.999	0.997	0.997	0.997	0.997	0.997

TABLE 4.6 Power of the four test plus weighted tests in Situation 3 (with censoring)

Sample size	Log-Rank Test	Wilcoxon Test	Kolmogorov-Smirnov Test	Lin &Xu Log-Rank	Weight Gehen	Weight Tarone	Weight Peto	Weight MPeto
(20,20)	0.781	0.677	0.685	0.775	0.718	0.751	0.718	0.714
(40, 40)	0.978	0.933	0.965	0.971	0.941	0.967	0.941	0.938
(50, 50)	0.992	0.981	0.984	0.990	0.983	0.986	0.983	0.983
(60, 60)	1	0.988	0.999	1	0.994	0.999	0.994	0.994
(80, 80)	1	0.998	1	1	0.999	1	0.999	0.999
(100, 100)	1	1	1	1	1	1	1	1
(20,50)	0.923	0.866	0.818	0.932	0.891	0.918	0.891	0.890
(50, 20)	0.904	0.803	0.913	0.871	0.851	0.874	0.851	0.852
(50, 100)	1	0.998	1	1	0.997	1	0.997	0.997
(100, 50)	1	0.992	1	0.999	0.993	0.998	0.993	0.993

When evaluating a test, in addition to the power of the test, appropriate control of the probability of a Type I error should also be considered. Tables 4.7-4.10 present the results for Situations 4 and 5, where we look at the Type I Error estimation for each of the test statistics.

Tables 4.7 and 4.8 display the Type I Error results for situation 4 (see description, p. 29). These results are a bit inflated compared to the results given in Lin and Xu's article (2009). Therefore, we constructed single proportions 95% confidence interval to check if the nominal value of 0.05 is within the interval. Based on these results, we are 95% confident that the nominal value of the Type I Error lies between the values 0.036 and 0.0635.

When considering Situation 4 without censoring, Table 4.7 shows that Lin & Xu's test has a small amount of error (<0.05) along with log rank, Wilcoxon, and Kolmogorov-Smirnov tests. This is consistent with Lin & Xu's results (2009). Each of the weighted test statistics have an observed alpha level that lies within the interval of 0.036 and 0.0635. The results show that the eight tests have similar Type I error in this situation and the Type I error of the newly weighted testing method is controlled at the specified confidence interval (0.036, 0.0635)

When considering Situation 4 with censoring, Table 4.8 shows that each of the existing test statistics have a relatively small amount of error which all lie within the interval of 0.036 and 0.0635. Each of the weighted test statistics also show similar results of smaller error amounts that lie within this interval. The simulation results are a bit inflated compared to Lin & Xu's results (2009). The results show that the eight tests have

similar Type I error in this situation and the Type I error of the newly weighted testing method is controlled based on the 95% confident interval constructed for this situation.

TABLE 4.7 Type I Error estimation of the four test plus weighted tests in Situation 4 (without censoring)

Sample size	Log-Rank Test	Wilcoxon Test	Kolmogorov-Smirnov Test	Lin &Xu Log-Rank	Weight Gehen	Weight Tarone	Weight Peto	Weight MPeto
(20,20)	0.058	0.052	0.037	0.051	0.062	0.053	0.062	0.060
(40, 40)	0.051	0.055	0.049	0.037	0.056	0.047	0.056	0.057
(50, 50)	0.050	0.045	0.040	0.035	0.047	0.041	0.047	0.047
(60, 60)	0.061	0.044	0.044	0.038	0.047	0.045	0.047	0.047
(80, 80)	0.040	0.040	0.032	0.023	0.041	0.031	0.041	0.043
(100, 100)	0.048	0.046	0.041	0.030	0.046	0.039	0.046	0.046
(20,50)	0.058	0.054	0.046	0.045	0.057	0.052	0.057	0.057
(50, 20)	0.059	0.054	0.052	0.052	0.066	0.057	0.066	0.067
(50, 100)	0.060	0.057	0.055	0.047	0.055	0.047	0.055	0.055
(100, 50)	0.064	0.052	0.055	0.050	0.048	0.049	0.048	0.048

TABLE 4.8 Type I Error estimation of the four test plus weighted tests in Situation 4 (with censoring)

Sample size	Log-Rank Test	Wilcoxon Test	Kolmogorov-Smirnov Test	Lin &Xu Log-Rank	Weight Gehen	Weight Tarone	Weight Peto	Weight MPeto
(20,20)	0.054	0.057	0.038	0.060	0.062	0.061	0.062	0.063
(40, 40)	0.045	0.050	0.042	0.057	0.056	0.059	0.056	0.056
(50, 50)	0.056	0.050	0.040	0.054	0.051	0.049	0.051	0.051
(60, 60)	0.061	0.055	0.056	**0.069**	0.059	0.061	0.059	0.059
(80, 80)	0.063	0.053	0.060	**0.066**	0.055	0.056	0.055	0.055
(100, 100)	0.051	0.057	0.054	0.062	0.056	0.062	0.056	0.056
(20,50)	0.051	0.052	0.044	0.053	0.057	0.054	0.057	0.057
(50, 20)	**0.067**	0.059	0.059	**0.072**	**0.070**	**0.069**	**0.070**	**0.070**
(50, 100)	0.054	0.041	0.043	0.059	0.046	0.053	0.046	0.045
(100, 50)	0.055	0.057	0.056	**0.067**	0.062	0.058	0.062	0.062

*BOLD INDICATES OUTSIDE OF CONFIDENCE INTERVAL

Tables 4.9 and 4.10 display the Type I error estimates for Situation 5 (see description, p. 29). When considering Situation 5 without censoring, Table 4.9 shows that Lin & Xu's test has a small amount of error along with log rank, Wilcoxon, and Kolmogorov-Smirnov test statistics, based on the 95% confidence interval (0.036, 0.0635). This is consistent with Lin & Xu's results (2009). The four weighted tests also show that the Type I error results mainly fall within the interval of 0.036 and 0.0635. Therefore, the weighted tests claim to have small amount of error, as well as, the four pre-existing tests.

When considering Situation 5 with censoring, Table 4.10 shows that the three pre-existing test statistics have relatively small Type I error based on our 95% confidence interval. Lin & Xu's and the four newly proposed weighted test statistics have a few more results that are above our liking for Type I error probability. The Type I error results seem to be inflated compared to Lin & Xu's results (2009). The results show that the log rank, Wilcoxon, and Kolmogorov-Smirnov tests demonstrate similar Type I error in this situation. Also based on the 95% confidence interval, we conclude that the four weighted tests have a better type I error and Lin and Xu's test by consistently having smaller amounts of error. There are a few examples where the Type I error falls outside of our interval, but the majority of the results fall within the confident interval of 0.036 and 0.0635.

TABLE 4.9 Type I Error estimation of the four test plus weighted tests in Situation 5 (without censoring)

Sample size	Log-Rank Test	Wilcoxon Test	Kolmogorov-Smirnov Test	Lin &Xu Log-Rank	Weight Gehen	Weight Tarone	Weight Peto	Weight MPeto
(20,20)	0.060	0.047	0.037	**0.068**	0.060	0.065	0.060	0.061
(40, 40)	0.056	0.047	0.043	0.051	0.055	0.055	0.055	0.056
(50, 50)	0.056	0.052	0.046	0.047	0.055	0.058	0.055	0.057
(60, 60)	0.055	0.047	0.042	0.047	0.049	0.046	0.049	0.049
(80, 80)	0.044	0.043	0.034	0.038	0.043	0.042	0.043	0.043
(100, 100)	0.052	0.053	0.057	0.048	0.041	0.039	0.041	0.041
(20,50)	0.052	0.060	0.00	**0.068**	**0.067**	0.064	**0.067**	**0.069**
(50, 20)	0.058	0.062	0.044	**0.069**	**0.067**	**0.066**	**0.067**	**0.70**
(50, 100)	0.052	0.048	0.042	0.045	0.051	0.042	0.051	0.051
(100, 50)	0.054	0.049	0.044	0.048	0.047	0.054	0.047	0.048

*BOLD INDICATES OUTSIDE OF CONFIDENCE INTERVAL

TABLE 4.10 Type I Error estimation of the four test plus weighted tests in Situation 5 (with censoring)

Sample size	Log-Rank Test	Wilcoxon Test	Kolmogorov-Smirnov Test	Lin &Xu Log-Rank	Weight Gehen	Weight Tarone	Weight Peto	Weight MPeto
(20,20)	0.054	0.052	0.037	0.064	0.059	**0.067**	0.059	0.060
(40, 40)	0.052	0.049	0.047	**0.070**	0.060	**0.067**	0.060	0.057
(50, 50)	0.052	0.048	0.036	**0.067**	0.057	0.056	0.057	0.057
(60, 60)	0.048	0.038	0.043	0.063	0.050	0.058	0.050	0.051
(80, 80)	0.050	0.057	0.051	**0.078**	0.057	**0.068**	0.057	0.057
(100, 100)	0.047	0.050	0.047	**0.073**	0.054	0.063	0.054	0.053
(20,50)	0.055	0.061	0.046	**0.093**	**0.073**	**0.090**	**0.073**	**0.075**
(50, 20)	0.059	0.050	0.048	**0.085**	**0.068**	**0.083**	**0.068**	**0.068**
(50, 100)	0.047	0.048	0.046	**0.075**	0.059	**0.069**	0.059	0.060
(100, 50)	0.044	0.039	0.034	0.060	0.049	0.053	0.049	0.049

*BOLD INDICATES OUTSIDE OF CONFIDENCE INTERVAL

4.2. CONCLUSION AND DISCUSSION

In Conclusion, based on the results of the Type I error estimates in Tables 4.7-4.10, we suggest that the four weighted Lin and Xu's test statistics perform just as well, if not slightly better, than Lin and Xu's original test statistic and the three other pre-existing tests. Not only were they slightly more powerful than Lin and Xu's test, but with our 95% confidence interval, it was also evident that these weighted tests had proper Type I error values as well. Also, in tables 4.1-4.10, the Gehen and Peto weight had the same power and Type I error results for each situation. This may be due to the censoring rate of the simulation data. If there is an increase in censoring, that may affect the results of the power and Type I error. We further conclude that the larger the sample size, the more powerful the test, which is to be expected.

In the end we believe that the weighted test statistics is at least comparable to the Log rank, Wilcoxon, Kolmogorov-Smirnov, and Lin & Xu's test statistics, sometime is better. However, we would only suggest using these weighted tests if it was necessary with the data at hand. If it doesn't make sense to apply these weight components, then we suggest to simply use Lin and Xu's original test statistic, however, there adding the weight component wouldn't harm the analysis.

CHAPTER 5

REAL DATA RESULTS

5.1. REAL DATA EXAMPLE

For illustration, we apply our new method to the Leukemia data introduced in Chapter 1, Section 1.3. First, we consider the comparison of the treatment (Group 1) and placebo (Group 2) subjects. We consider the null hypothesis that the leukemia patient groups have the same survival distribution against the alternative that the survival distributions are different. Referring back to Figure 1.2, we see that the two treatment groups appear to be proportional to each other because the plot shows that their survival curves are not intersecting. Similarly, Figure 1.2 plots the logarithm cumulative hazard function versus log of survival time graph and result in proportional curves as well.

Based on the simulation results in Chapter 3 for proportional curves, we expect Lin & Xu's and the newly weighted test statistics to perform just as well as the log rank test. Table 5.1 shows the p-value of each test statistic when stratifying by group.

TABLE 5.1.P-value of each test statistic when comparing survival curves stratified by group

Test Statistic	P-Value
Log Rank	0.00004
Wilcoxon	0.00014
Kolmogorov-Smirnov	0.00050
Lin & Xu	0.00000
Gehen	0.00000
Tarone	0.00000
Peto	0.00000
Mpeto	0.00000

From these p-values for each test statistic, we conclude, at the 0.05 significance level, that patients who receive the 6-Mercaptopurine treatment have a longer survival rate than the patients in the placebo group.

For the second example, instead of stratifying by group, we stratify by gender. We consider the null hypothesis that the leukemia patients adjusted by gender have the same survival distribution against the alternative that the survival distributions are different. Referring back to Figure 1.3, this plot displays that these two hazards are not proportional to each other. Similarly, Figure 1.4 displays the cumulative hazard function versus log of survival time graph and result in intersecting hazards which as a result claim that the PH assumption not being satisfied. As a reminder, we stated earlier that when stratifying by gender, it appears that male patients begin with a larger survival probability, but over time women appear to have a higher survival probability. Therefore, it is more difficult to determine if there is in fact a difference in survival probability. We apply our new testing method by using Lin and Xu's method along with

the newly proposed weighted test statistics to analyze this hypothesis. Based on the simulation results in Chapter 4 for non-proportional curves, we obtain less accurate estimates while adjusting for gender for the log rank test and Wilcoxon test statistic. We expect the weighted component test statistics to yield the most accurate results based on this situation. However, while these weight components will still have accurate results, these weights do not necessarily apply to this specific dataset, therefore, it would make most sense to apply Lin and Xu's original test statistic. Table 5.2 shows the p-value of each test statistic when stratifying by gender.

TABLE 5.2. P-value of each test statistic when comparing survival curves stratified by gender

Test Statistic	P-Value
Log Rank	0.3872
Wilcoxon	0.8200
Kolmogorov-Smirnov	0.2515
Lin & Xu	0.0158
Gehen	0.0122
Tarone	0.0146
Peto	0.0129
Mpeto	0.0128

Based on the p-values presented in Table 5.2, the log rank, Wilcoxon, and Kolmogorov-Smirnov test statistics conclude that there is no difference detected in survival curves when stratifying by gender. However, we know that these tests lack power, therefore, may not represent accurate results. Lin and Xu's test statistic, as well as, the four weighted test statistics, lead us to conclude that there is a difference in survival curves when stratifying by gender. Based on our simulation study, we know these test statistics

have high power and significant Type I error values. Therefore, based on Figures 1.3 and 1.4, as well as, the p-value presented in Table 5.2 for Lin &Xu's test statistic, we conclude that there is a significant difference in survival between male and female leukemia patients, with a p-values for Lin and Xu's test and the four weight components tests being less and 0.05.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In this thesis, we investigated the efficiency of adding a weight component to a test statistic based on the method proposed by Lin and Xu (2009) that would be effective for all-purpose situations when comparing survival curves that may or may not be proportional. The difference between this new method compared to existing test statistics, is that this test takes into consideration the length of time intervals to better capture the differences. Our proposed method was to apply a different weight at the i^{th}failure time. In this study, we compared the power of the weighted test statistics to Lin and Xu (2009) original testing method. We also tested the accuracy of these weighted test statistics by analyzing the Type I error estimations.

After we investigated the power and Type I error of each weighted test statistic, we conclude that each weighted approach had slightly better results when compared to the power of Lin and Xu's test statistic in most of cases. There was not a specific weight component that was consistently better than the other for all three of our situations presented throughout the thesis. Therefore, we conclude that the best test statistic to use when comparing survival distributions for all-purposes is Lin and Xu's original testing method, or use the applied weight components if necessary. Adding the weight components to Lin and Xu's test statistic is not only very powerful in all circumstances, the Type I error estimation proves to be efficient as well.

For future work, we suggest investigating the Flemington-Harrington weight component. The Flemington-Harrington test uses the Kaplan-Meier survival estimate over all groups to calculate its weights for the i^{th} failure time. This weight component is appealing because it allows the most flexibility in terms of the choice of weights because the user can provide specific values to the weight component. For example, this weight component can apply more weight on later failure times rather than early failure times.

There were also a few limitations to our simulation study. For example, we only generated data from exponential distributions. For future work, we can generate data from a variety of distributions, which would capture different shapes of the hazard risk functions. We would also want to consider different censoring rates to see how that would directly affect the results of the power and Type I error values. For example, if there was an increase in censoring would that increase or decrease the power and Type I error values of the weight components.

REFERENCES

Brookmeyer, R. and Crowley, J. 1982. A Confidence Interval for the Median Survival Time. *Biometrics*, 38: 29-41.

Freireich, E. et al. 1963.The Effect of 6-Mercaptopurine on the Duration of Steroid-induced Remissions in Acute Leukemia: A Model for Evaluation of Other Potentially Useful Therapy,*Blood*.1: 699-716.

Klein, J. P. and Moeschberger, M. L. (1997), *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer-Verlag.

Lin, X. and Xu, Q. 2009. A new method for the comparison of survival distributions. *Pharmaceutical Statistics*. 9: 67-76.

Mantel, N. 1967. Ranking procedures for arbitrarily restricted observation. *Biometrics*. 23: 65-78,

Massey, F. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*. 46: 68-78.

Peto, R. and Peto, J. 1972.Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society. Series A*. 135:185-207.

Schoenfeld, D. 1981. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*. 68: 316-319.