Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yuqing Chen Date

Comparison of Methods for Two Crossing Survival Curves

By

Yuqing Chen
Master of Science in Public Health

Biostatistics and Bioinformatics

_____

Jose Binongo, PhD
(Thesis Advisor)

_____

Yi-An Ko, PhD
(Reader)

Comparison of Methods for Two Crossing Survival Curves

By

Yuqing Chen

Bachelor of Science
China Agricultural University
2016

Thesis Advisor: Jose Binongo, PhD
Reader: Yi-An Ko, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2018

Abstract

Comparison of Methods for Two Crossing Survival Curves

By Yuqing Chen

In clinical research, it is not uncommon to see that the proportional hazard assumption is violated, particularly when survival curves cross. The log-rank test which is widely used in comparing survival curves is known to have less power under such circumstance. In this thesis, we introduce three methods to test the difference between two survival curves which are (1) Gehan's weighted log-rank test, (2) Renyi type test, and (3) Lin and Xu's test. We give a brief background of the statistical theory underpinning these methods. Then, we conduct a simulation study to compare the three tests with log-rank test under different situations and censor rates. We also apply all three methods to a real data example from a kidney dialysis trial. In our comparison of the three methods, we found that Gehan's test does not perform well in the setting situations. The weight function of the weighted log-rank test must be specified prior to the analysis, and an inappropriate weight function may result in a misleading conclusion. Renyi test is suitable to use when two survival curves separate largely and cross, because its power stays above 90% even under a high censor rate. Lin and Xu's test is appropriate when two survival curves do not separate largely, and its power is influenced by the censor rate.

Comparison of Methods for Two Crossing Survival Curves

By

Yuqing Chen

Bachelor of Science
China Agricultural University
2016

Thesis Advisor: Jose Binongo, PhD
Reader: Yi-An Ko, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2018

Table of Contents

1. Introduction

In time-to-event data, it is often important to compare the overall equality of two survival distributions under censoring. To achieve such a goal, the log-rank test has been widely used under the assumption of proportional hazard rates. However, this assumption may be violated, especially when two survival curves cross. Because early differences in favor of one group are canceled out by late differences in favor of the other treatment[1], the log-rank test may lose power when survival curves cross.

The phenomenon of crossing survival curves is observed when a treatment may offer a short-term benefit but does not provide long-term advantages[2]. It is also seen in survival after surgery because the effect of surgery can last for a long time, but the risk of death may be high right after surgery and decreases as the patient recovers.

Kristiansen[3] conducted a survey including all publications in five notable journals with 175 studies that had a time-to-event endpoint. Among 175 included studies, 47% had survival curve crossings. Of those studies where crossing survival curves were presented, the log-rank test was performed in 70% of the tests, and only 31% of them reported testing for proportional hazards.

Using the log-rank test under conditions of non-proportional hazards may lead to misleading results. Therefore, it is necessary to use other alternative tests with greater power when survival curves cross.

Statisticians have developed several tests since the 1960. Those existing methods can be divided into three types[4]. The first type of tests assigns different weights to the log-rank

test. These tests include Gehan's generalized Wilcoxon test[5], Tarone-Ware test, Peto-Prentice test, and Fleming-Harrington test, etc. The second type of tests are supremum versions of the log-rank tests, including modified Kolmogorov-Smirnov[6], and Renyi type tests[7], etc. These tests are based on the maximum difference between estimates of two survivor functions. The third type of tests is the modified log-rank test, such as Lin and Wang's test[8] using squared differences at each time point, Lin and Xu's test[9] using absolute differences at each time point and Levene type test focusing on variance differences. Methods based on splitting the analysis at the crossing point and reporting separate p-values have also been proposed.

In this article, we select one specific test in each of the three types for comparison, which are Gehan's test, Renyi test, and Lin and Xu's test. We perform Monte Carlo simulation to assess the statistical power and type I error rate of each tests. We also apply all three methods to a real data example. Our goal is to evaluate the strengths and weaknesses of the tests in a variety of situations and censoring rates in order to better understand the strength and weaknesses of each tests.

This thesis is structured as follows: In chapter 2-4, we introduce the statistical theory behind Gehan's test, Renyi test, and Lin and Xu's test. In chapter 5, we perform Monte Carlo simulation to compare the statistical power and type I error rate of the three methods. We apply all three methods to a real data in chapter 6. Finally, we summarize our findings and provide recommendations for the use of these tests in chapter 7.

2. Weighted Log-rank Test

Let $t_1<t_2<\ldots<t_D$ be the distinct death time in the pooled sample. At time t_j we observe d_{ij} events in the ith sample out of n_{ij} individuals at risk, $i = 1, \ldots, K, j = 1, \ldots, D$. Let $d_j = \sum_{i=1}^{K} d_{ij}$ and $n_j = \sum_{i=1}^{K} n_{ij}$ be the number of deaths and number at risk in combined sample at time t_j.

Table 1 Number of failures in two groups at observed failure time t_j

Group	# of failure	# of non-failure	# at risk
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

For a two sample test, we have the null hypothesis and alternative hypothesis as follows:

H_0: $h_1 = h_2$ for all $t_j \leq \tau$

H_1: $h_1 \neq h_2$ for some $t_j \leq \tau$

Here h_1 and h_2 is the hazard rate in group I and group II respectively, and τ is the longest time at which both of the groups have at least one subject at risk.

The test of H_0 is based on weighted comparisons of the estimated hazard rate of the ith population under the null and alternative hypotheses. Under the null hypothesis, an estimator of the expected hazard rate in the ith group at time t_j is the pooled sample estimator of the hazard rate d_j/n_j. Using only data from the ith group, the estimator of the hazard rate is d_{ij}/n_{ij}. To make comparisons of the two estimators, let $W_i(t)$ be a positive weight function with the property that $W_i(t_j)$ is zero whenever n_{ij} is zero. The test of H_0 is based on

$$Z_i(\tau) = \sum_{j=1}^{D} W_i(t_j)\left(\frac{d_{ij}}{n_{ij}} - \frac{d_j}{n_j}\right), i = 1,2$$

The weight function that is commonly used in practice is $W_i(t_j) = n_{ij}W(t_j)$, where $W(t_j)$ is a common weight shared by both groups. Thus, we have

$$Z_i(\tau) = \sum_{j=1}^{D} W(t_j)(d_{ij} - n_{ij}\frac{d_j}{n_j}) \, , i = 1,2$$

The variance of $Z_i(\tau)$ is

$$\hat{\sigma}_{ii} = \sum_{j=1}^{D} W(t_j)^2 \frac{n_{ij}}{n_j}(1 - \frac{n_{ij}}{n_j})(\frac{n_j - d_j}{n_j - 1}) \, d_j, i = 1,2$$

For the two sample test, the test statistics can be written as

$$Z = \frac{Z_i(\tau)}{\hat{\sigma}_{ii}} \, ,$$

where Z follows a standard normal distribution. We reject H_0 when $|Z| > Z_{\alpha/2}$. Note that when $W(t_j) = 1$ we have the log-rank test.

Many different weight functions have been applied to the log-rank test. The following is the weight functions that is commonly used[10]:

1. Log-Rank: $W(t) = 1$. This weight is the most general and applies equal weight to all parts of the survival curve. The log-rank test has optimum power to detect alternatives where hazard rates of survival curves are proportional to each other.

2. Gehan (Wilcoxon): $W(t) = n_j$. This weight emphasizes early differences in the survival curve as each statistic is weighted by the at-risk set at a given time point.

3. Tarone-Ware: $W(t) = n_j^{1/2}$. This weights gives more weight to differences between the observed and expected number of deaths in sample i at time points where there is the most data.

4. Peto-Peto: $W(t) = \tilde{S}(t_j)$. This is a more robust weight when the censoring between the two groups follow different patterns.

5. Modified Peto-Peto: $W(t) = \frac{\tilde{S}(t_j)n_j}{(n_j+1)}$. This weight is based on slight modification to the Peto-Peto.

6. Fleming-Harrington: $W(t) = \hat{S}(t_{j-1})^p\left[1 - \tilde{S}(t_{j-1})\right]^q$. Here the survival function at the previous death time is used as a weight to ensure that these weights are known just prior to the time at which the comparison is to be made. This is a family of weights determined by the powers p and q. Setting p and q to 0 gives the log-rank test. When p = 1 and q = 0 we have a weight very similar to the Peto-Peto weight. When p=0 and q>0, these tests give more weights to the later departures; while p>0 and q<0 gives more weight to early departures. By choosing p and q appropriately, one can construct tests that are most suitable against alternatives.

In this thesis, we choose Gehan's generalized Wilcoxon test as a representative of weighted log-rank tests, which emphasizes the early stage difference when there is a larger at-risk set.

3. Renyi Type Test

The Renyi type test is a combination of non-parametric survival analysis and Standard Brownian motion, which was designed specifically for situations where crossing hazard function appears. In this test, we are testing the same hypothesis as in weighted log-rank test.

H_0: $h_1 = h_2$ for all $t_j \leq \tau$

H_1: $h_1 \neq h_2$ for some $t_j \leq \tau$

Here h_1 and h_2 is the hazard rate in group I and group II respectively. We need to find a test statistic for some weighted function at each time of failure. The weighted functions we mentioned in the weighted log-rank test can also be applied in Renyi type test. Under H_0, although the survival curves may cross, there should be a maximum absolute value of the test statistic, which represents the difference between the two groups, at some time point before τ. When the value is too large, we reject H_0 in favor of H_1 for some *t*.

When we calculate $Z(t_j)$, which is the numerator of the statistic in the log-rank test, we only include the data up to t_j (compared to include all data up to τ in the log-rank test). Thus, we have

$$Z(t_j) = \sum_{t_k \leq t_j} W(t_k)(d_{1k} - n_{1k}\frac{d_k}{n_k}) \, , j = 1, \dots, D$$

$$\sigma^2(\tau) = \sum_{t_k \leq \tau} W(t_k)^2 \left(\frac{n_{1k}}{n_k}\right)\left(\frac{n_{2k}}{n_k}\right)\left(\frac{n_k - d_k}{n_k - 1}\right)(d_k)$$

where τ is the largest t_k with $n_{1k}, n_{2k} > 0$.

The test statistic is

$$Q = \sup\{|Z(t)|, t \leq \tau\} / \sigma(\tau)$$

In order to find the critical value of Q, standard Brownian Motion is introduced. Brownian Motion is a special type of Markov Chain. In addition to Markov Chain, Brownian Motion also has the following criteria:

1. For all time h>0, the difference $X(t_j + h) - X(t_j)$ has a normal distribution
2. The difference $X(t_j + h) - X(t_j)$, $0<t_1<t_2<\ldots<t_n$ are mutually independent.
3. The mean difference is 0.

The standard Brownian Motion is intended to describe random noise that follows a standard normal distribution. Thus the distribution of Q can be approximated by the standard Brownian motion process. The p-value is

$$Pr[\sup|Z(t)| > y] = 1 - \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{2k\ +1}\exp\left[-\frac{\pi^2(2k+1)^2}{8y^2}\right]$$

4. Lin & Xu's Test

Lin and Wang[8] developed a new statistical testing approach to compare the overall homogeneity of survival curves by measuring the squared differences between the number of observed failures and the number of expected failures over time. Later, Lin and Xu[9] suggested a new method for comparing survival distributions based on the absolute difference between the area under two survival curves. Both of the two methods suggest better performance in statistical power than log-rank test when there are two survival curves that cross. Here, we only introduce Lin and Xu's test in this article.

Our null hypothesis and alternative hypothesis are:

H_0: $h_1 = h_2$ for all $t_j \leq \tau$

H_1: $h_1 \neq h_2$ for some $t_j \leq \tau$

Here h_1 and h_2 is the hazard rate in group I and group II respectively. Let $X = \min(T, C)$ and $\delta = I(T \leq C)$, where T denotes survival time, and C denotes censoring time. Let $t_1 < t_2 < \ldots < t_D$ be the distinct death time in the pooled sample. At time t_j we observe d_{ij} events in the ith sample out of n_{ij} individuals at risk, $i = 1, \ldots, K, j = 1, \ldots, D$. Let $d_j = \sum_{i=1}^{K} d_{ij}$ and $n_j = \sum_{i=1}^{K} n_{ij}$ be the number of deaths and number at risk in combined sample at time t_j.

The survival distribution for each group at time t, $S_i(t)$, is estimated by Kaplan-Meier estimator $\hat{S}_i(t)$.

$$\hat{S}_i(t) = \prod_{j|t_j \leq t} \left(1 - \frac{d_{ij}}{n_{ij}}\right)$$

To measure the difference between survival curves, we calculate the absolute difference of the area between the two curves, which is defined as

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

Here τ is the last time point by which the areas under the survival curves can be calculated for both groups. We can further divide τ into three different situations. If the two groups are both censored at the last time point, $\tau = \min(X_{iD})$. If the last time point in one group is a failure, and censored in another, $\tau = \max(X_{iD}(1 - \delta_{iD}))$. If there is failure for both groups at the last time point, $\tau = \max(X_{iD})$. Note, t_{j+1} for the last element in the summation is defined as τ.

Using Greenwood's formula, the estimate of the variance of $\hat{S}_i(t)$ is

$$\hat{\sigma}^2_{\hat{S}_i}(t) = \hat{S}_i^2(t) \sum_{j|t_j \le t} \frac{d_{ij}}{n_{ij}(n_{ij} - d_{ij})}$$

For standard normal random variable Z, we have $E(|Z|) = \sqrt{(2/\pi)}$, and $Var(|Z|) = 1 - 2/\pi$. Under the null hypothesis, using normal approximation, $\hat{S}_1(t) - \hat{S}_2(t)$ follows a normal distribution with mean 0 and variance $\hat{\sigma}^2_{\hat{S}_1}(t) + \hat{\sigma}^2_{\hat{S}_2}(t)$. We can then find the estimate of the expectation and variance of $\hat{S}_1(t) - \hat{S}_2(t)$,

$$\hat{E}(|\hat{S}_1(t) - \hat{S}_2(t)|) \doteq \left\{2/\pi\left[\hat{\sigma}^2_{\hat{S}_1}(t) + \hat{\sigma}^2_{\hat{S}_2}(t)\right]\right\}^{1/2}$$

$$\widehat{Var}(|\hat{S}_1(t) - \hat{S}_2(t)|) \doteq \left(1 - \frac{2}{\pi}\right)\left[\hat{\sigma}^2_{\hat{S}_1}(t) + \hat{\sigma}^2_{\hat{S}_2}(t)\right]$$

Based on the normal approximation, $\hat{E}(\Delta)$ can be estimated as the following:

$$\hat{E}(\Delta) \doteq \sum_{j|t_j<\tau} \{2/\pi[\hat{\sigma}^2_{S_1}(t_j) + \hat{\sigma}^2_{S_2}(t_j)]\}^{1/2}(t_{j+1} - t_j)$$

The variance of Δ is

$$Var(\Delta) \doteq \sum_{j|t_j<\tau} (t_{j+1} - t_j)^2 \left(1 - \frac{2}{\pi}\right)[\sigma^2_{S_1}(t_j) + \sigma^2_{S_2}(t_j)]$$

$$+ \sum_{j<j'|t_j,t_{j'}<\tau} 2\rho_{j,j'}(t_{j+1} - t_j)(t_{j'+1} - t_{j'})\left(1 - \frac{2}{\pi}\right)$$

$$\times \{[\sigma^2_{S_1}(t_j) + \sigma^2_{S_2}(t_j)][\sigma^2_{S_1}(t_{j'}) + \sigma^2_{S_2}(t_{j'})]\}^{1/2}$$

where $\rho_{j,j'}$ is the correlation coefficient between $|\hat{S}_1(t_j) - \hat{S}_2(t_j)|$ and $|\hat{S}_1(t_j) - \hat{S}_2(t_j)|$, $j \neq j'$.

The estimation of $Var(\Delta)$ depends on the value of $\rho_{j,j'}$. Lin and Xu suggested to set $\rho_{j,j'} = 0.5$ for all j and j', because the test statistic has a high power and does not inflate type I error. Then the estimate of $Var(\Delta)$ is

$$\widehat{Var}(\Delta) \doteq \sum_{j|t_j<\tau} (t_{j+1} - t_j)^2 \left(1 - \frac{2}{\pi}\right)[\hat{\sigma}^2_{S_1}(t_j) + \hat{\sigma}^2_{S_2}(t_j)]$$

$$+ \sum_{j<j'|t_j,t_{j'}<\tau} (t_{j+1} - t_j)(t_{j'+1} - t_{j'})\left(1 - \frac{2}{\pi}\right)$$

$$\times \{[\hat{\sigma}^2_{S_1}(t_j) + \hat{\sigma}^2_{S_2}(t_j)][\hat{\sigma}^2_{S_1}(t_{j'}) + \hat{\sigma}^2_{S_2}(t_{j'})]\}^{1/2}$$

By standardizing Δ, we now have the test statistic:

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

We reject H_0 in favor of H_1 when $|\Delta^*| > Z_{\alpha/2}$.

5. Simulation Study

To evaluate the performance of Gehan's test, Renyi test and Lin's method, Monte Carlo simulations are carried out to study the statistical power and the type I error under a variety of situations. We compare these three methods with log-rank test in the simulations.

5.1 Estimation of Type I Error

The number of iteration is 3000. Two random samples are generated independently from exponential distribution with mean 5, with 4 different censor rate 0, 20%, 40% and 60%. The type I error rate is calculated as the proportion of 3000 repeated samples in which we reject the null hypothesis at 0.05 significant level.

Table 2 shows that all four tests have similar type I error rates which are close to 0.05. Gehan's test and Renyi test are more conservative compared to log-rank test and Lin and Xu's test. As censor rate increase, Lin and Xu's type I error rate exhibit slight increase while other tests remain at 0.05 level.

Table 2 Type I error rate of four tests at different censor rate

censor rate	*log-rank test*	*Gehan's test*	*Renyi test*	*Lin and Xu's test*
0	0.048	0.048	0.043	0.031
20%	0.053	0.048	0.042	0.044
40%	0.052	0.046	0.045	0.05
60%	0.047	0.047	0.044	0.056

5.2 Estimation of Power

We consider the three different situations. In Situation 1, we create two survival curves with proportional hazards. In Situation 2, we design two crossing survival curves. In Situation 3, we have two survival curves are very close at the beginning, and diverge then.

The simulations will demonstrate the statistical power of the log-rank test, Gehan's test, Renyi test, and Lin and Xu's method. The number of iterations in each simulation study is 3000. The estimated statistical power is calculated as the proportion of 3000 repeated random samples in which we reject the null hypothesis at 0.05 significance level.

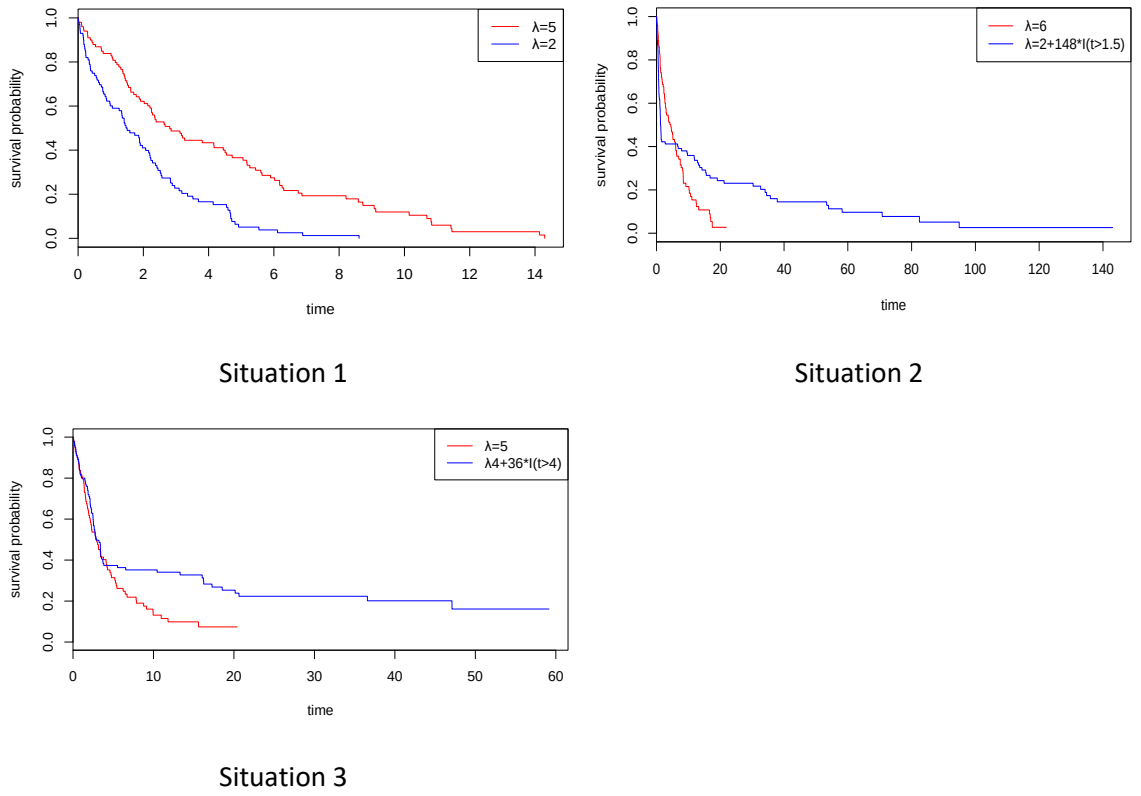

Figure 1 Survival curves of samples in simulation study under 3 situations

5.2.1 Situation 1

We first consider a situation where two survival curves with proportional hazard. We generate samples from an exponential distribution with mean 5 and 2 independently for group I and group II.

Furthermore, we also include different scenarios of censoring to better evaluate the performance of the tests in this situation. We first include the situation without censoring. Then we obtain censor rates of 20%, 40%, and 60% approximately for both of the groups by generating censoring time from uniform distribution Unif(0, 28), Unif(0, 13), Unif(0, 7) separately for group I, and Unif(0, 150), Unif(0, 55), Unif(0, 30) for group II respectively.

Table 3 Power of four tests at different censor rate for when survival curves have proportional hazard

censor rate	*log-rank test*	*Gehan's test*	*Renyi test*	*Lin and Xu's test*
0	1	0.999	0.989	1
20%	1	0.998	0.998	1
40%	0.998	0.993	0.994	0.998
60%	0.964	0.942	0.951	0.960

Table 3 shows that all of the four tests perform well in situation I. As censor rate increases, there is a slight decrease in power for all four tests, but they still have great power which are above 90%. It also shows that among all tests, log-rank test has the optimal power when survival curves show proportional hazards.

5.2.2 Situation 2

We next consider a situation where two survival curves cross. In Group I the survival times follow an exponential distribution with mean of 6. In Group II the survival times follow an exponential distribution with mean of 2. However, if the survival time in Group II is greater than 1.5, then the survival time is re-generated from an exponential distribution with mean of 40.

Furthermore, we also include different scenarios of censoring to better evaluate the performance of the tests in this situation. We generate censoring time from uniform distribution Unif(0, 28), Unif(0, 13), Unif(0, 7) separately for group I, and Unif(0, 150), Unif(0, 55), Unif(0, 30) for group II in order to achieve censor rates of 20%, 40%, and 60% approximately for both of the groups. We also consider the situation without censoring.

Table 4 Power of four tests at different censor rate for when survival curves cross

censor rate	*log-rank test*	*Gehan's test*	*Renyi test*	*Lin and Xu's test*
0	0.714	0.168	0.999	1
20%	0.21	0.407	0.996	0.999
40%	0.082	0.705	0.987	0.945
60%	0.539	0.924	0.99	0.682

Table 4 confirms that the log-rank test has limited power when survival curves cross. A high censor rate seems to have a considerable influence on the log-rank test and Lin and Xu's test. Renyi test and Lin and Xu's test both perform better than other tests at low and moderate censor rates. Renyi test performs the best among all four tests with a high power at all censor rate, while Lin and Xu's test lose power when censor rate in high. Gehan's test has an increasing power as censor rate increases because the Gehan's test assigns a greater weight to earlier failure times, making the test insensitive to differences at later times.

5.2.3 Situation 3

Finally, we consider a situation where two survival curves are close at the early stage, and diverge later. We generate samples for group I from an exponential distribution with mean 5. For group II, we generate samples from an exponential distribution with mean 4.

However, if survival time is greater than 4, we regenerate the sample from an exponential distribution with mean 40.

We consider different scenarios of censoring to better evaluate the performance of the tests in this situation as well. In addition to the scenario without censoring, we consider other situations where we have censor rate of 20%, 40%, and 60% approximately for both of the groups. We generate censor time from uniform distribution Unif(0, 25), Unif(0, 11.5), Unif(0, 4.5) for group I, and Unif(0, 60), Unif(0, 12), Unif(0, 4.3) for group II respectively.

Table 5 Power of four tests at different censor rate for when survival curves are close at early stage and diverge later

censor rate	*log-rank test*	*Gehan's test*	*Renyi test*	*Lin and Xu's test*
0	0.602	0.059	0.707	1
20%	0.178	0.097	0.358	0.983
40%	0.057	0.187	0.229	0.291
60%	0.246	0.218	0.217	0.269

In situation 3, we create two survival curves that are close at the beginning, and separate later. As far as the log-rank test is concerned, the simulation results here are similar to those in situation 2 in that the log-rank test has low power. Gehan's test also shows a similar trend as censor rate increases. It has an increasing power as censor rate increases because the Gehan's test assigns a greater weight to earlier failure times, causing the test to be insensitive to differences at later times. A high censor rate affects the log-rank test, Renyi test, and Lin and Xu's test. Renyi test exhibits moderate performance when there is no censoring with a 70.7% power. However, its power decrease apparently under the influence of increasing censor rate. Lin and Xu's test performs better than other tests at all censor rate. However, it loses power to below 30% when the censor rate is above 40%.

6. Real Data Example from a Kidney Dialysis Trial

We apply the log-rank test, Gehan's test, Renyi test and Lin and Xu's test to a real dataset from a kidney dialysis trail. The data can be found in Klein and Moeschberger[1]. Scientists designed the trial to study the time to first exit-site infection (in months) in patients with renal insufficiency. Patients are divided into two groups based on how their catheter is replaced. 43 patients accepted surgically replacement, and 76 patients had a percutaneous replacement. Data in table 6 include time to first exit-site infection (or censoring time) in both groups.

We apply the log-rank test, Gehan's test, Renyi test and Lin and Xu's test to the data, and get the corresponding p-values. Log-rank test, Gehan's test, and Renyi test all give a non-significant p-value which is 0.112, 0.964, and 0.225. We fail to reject H_0. Lin and Xu's test give a significant p-value of $0.010<0.05$, so we reject H_0 and claim there is significant difference between the two survival curves.

In this example, the log-rank test fails because the two survival curves cross. The earlier differences are cancelled out by the later opposite direction differences. Gehan's test fails because it puts more weight to the earlier time points. The two curves actually have smaller differences at earlier time points and bigger differences at later points, which the Wilcoxon test is not able to detect. Renyi test is based on detecting the largest difference between two survival curves, while Lin and Xu's test, however, sum up the absolute difference at all time point. It is reasonable that Lin and Xu's test suggests to reject H_0 while Renyi test does not.

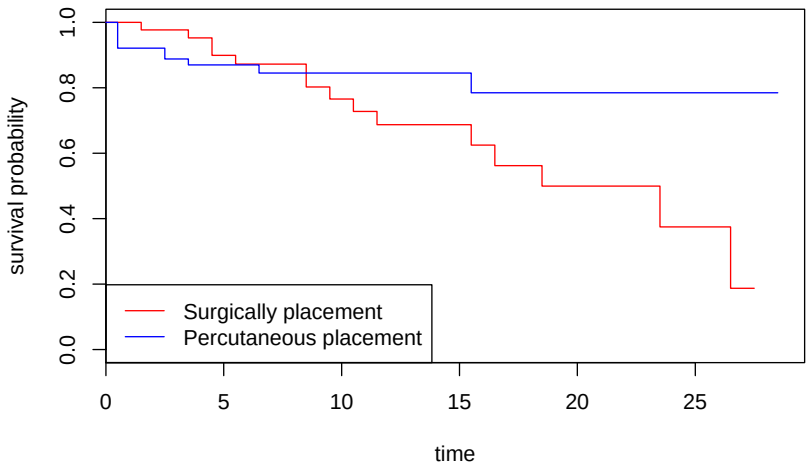

Figure 2 Survival curves of the kidney dialysis data

Table 6 Time to first exit-site infection from kidney dialysis data

Surgically placed catheter

Infection times: 1.5, 3.5, 4.5, 4.5, 5.5, 8.5, 8.5, 9.5, 10.5, 11.5, 15.5, 16.5, 18.5, 23.5, 26.5
Censored observations: 2.5, 2.5, 3.5, 3.5, 3.5, 4.5, 5.5, 6.5, 6.5, 7.5, 7.5, 7.5, 7.5, 8.5, 9.5, 10.5, 11.5, 12.5, 12.5, 13.5, 14.5, 14.5, 21.5, 21.5, 22.5, 22.5, 25.5, 27.5

Percutaneous placed catheter

Infection times: 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 2.5, 2.5, 3.5, 6.5, 15.5
Censored observations: 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1.5, 1.5, 1.5, 1.5, 2.5, 2.5, 2.5, 2.5, 2.5, 3.5, 3.5, 3.5, 3.5, 3.5, 4.5, 4.5, 4.5, 5.5, 5.5, 5.5, 5.5, 5.5, 6.5, 7.5, 7.5, 7.5, 8.5, 8.5, 8.5, 9.5, 9.5, 10.5, 10.5, 10.5, 11.5, 11.5, 12.5, 12.5, 12.5, 12.5, 14.5, 14.5, 16.5, 16.5, 18.5, 19.5, 19.5, 19.5, 20.5, 22.5, 24.5, 25.5, 26.5, 26.5, 28.5

7. Discussion and Conclusion

In clinical research, it is not uncommon that the hazards rates are unequal, particularly when two survival curves cross each other. In this thesis, we considered three different tests – Gehan's test, Renyi test, and Lin and Xu's test - for the comparison of two survival distributions under various situations. The objective of this study was to suggest hypothesis tests that is appropriate for use when survival curves cross. Because survival functions are more common and intuitive than hazard functions when investigating the survival differences between two groups, simulations were performed for situations in which the survival distributions 1) show proportional hazards; 2) cross; 3) are close at early stage and separate. The simulation results demonstrated that Renyi test is suitable when two survival curves have very different patterns and cross each other. Renyi test perform well under a high range of censor rate. Lin and Xu's test becomes optimal when two survival curves do not separate well, and its power is largely influenced by censor rate.

Weighted log-rank test is not particularly designed for crossing survival curves. It is important to choose the correct weight before using the test because inappropriate weight may lead to a misleading conclusion. Furthermore, it is also important to decide the weight based on prior knowledge or other information without looking into the data. It is inappropriate to choose weight after seeing the data.

In this thesis, we only generate censoring time from uniform distribution, and survival time from exponential distribution. More situations need to be considered in further study. We only compare three methods, while there are more tests particularly designed for the situation with survival curves crossing. Fleming et al.[6] have developed the modified

Kolmogorov-Smirnov test for the correct comparison of crossing survival curves. Koziol[11] has generalized the Cramér-von Mises test to censored data. Qiu and Sheng[12] have suggested a two-stage procedure in which the log-rank test serves as the first stage and a proposed procedure for addressing the crossing hazard rates is applied in the second stage. Kraus[11] has constructed a class of Neyman's smooth tests based on the concept of Neyman's embedding and a data-driven strategy. Li[2] et al. have compared the methods using Monte Carlo simulations under different sample size, censor rate, and crossing time. In practice, researchers should decide a proper test to use based on the characteristics of the data.

8. Reference

1. Klein, J.P. and M.L. Moeschberger, *Survival analysis: techniques for censored and truncated data*. 2005: Springer Science & Business Media.
2. Li, H., et al., *Statistical inference methods for two crossing survival curves: a comparison of methods.* PLoS One, 2015. **10**(1): p. e0116774.
3. Kristiansen, I., *PRM39 Survival curve convergences and crossing: a threat to validity of meta-analysis?* Value in health, 2012. **15**(7): p. A652.
4. Bouliotis, G. and L. Billingham, *Crossing survival curves: alternatives to the log-rank test.* Trials, 2011. **12**(Suppl 1).
5. Gehan, E.A., *A generalized Wilcoxon test for comparing arbitrarily singly-censored samples.* Biometrika, 1965. **52**(1-2): p. 203-224.
6. Fleming, T.R., et al., *Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data.* Biometrics, 1980: p. 607-625.
7. Gill, R., *Censoring and stochastic integrals, Mathematical Centre Tract, 124, Mathematisch Centrum, Amsterdam.* MATH Google Scholar, 1980.
8. Lin, X. and H. Wang, *A New Testing Approach for Comparing the Overall Homogeneity of Survival Curves.* Biometrical Journal, 2004. **46**(5): p. 489-496.
9. Lin, X. and Q. Xu, *A new method for the comparison of survival distributions.* Pharm Stat, 2010. **9**(1): p. 67-76.
10. Davis, M. and S.X. Xie, *Caution: hazards crossing! Using the Renyi test statistic in survival analysis.* Pharma SUG, 2011: p. 7-8.
11. Schumacher, M., *Two-Sample Tests of Cramér--von Mises-and Kolmogorov--Smirnov-Type for Randomly Censored Data.* International Statistical Review/Revue Internationale de Statistique, 1984: p. 263-281.
12. Qiu, P. and J. Sheng, *A two‐stage procedure for comparing hazard rate functions.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008. **70**(1): p. 191-208.

9. Appendix

R code for Lin and Xu's test:

```
lin2<-function(time, event,group){
  raw<-data.frame(time,event,group)
  raw<-raw[with(raw, order(group, time)), ]
  raw1<-raw[which(raw$group==1),]
  raw2<-raw[which(raw$group==2),]

  d<-data.frame(summary(survfit(Surv(time,event)~group, conf.int = 0.95))$time,
summary(survfit(Surv(time,event)~group, conf.int =
0.95))$surv,summary(survfit(Surv(time,event)~group, conf.int =
0.95))$strata,summary(survfit(Surv(time,event)~group, conf.int = 0.95))$std.err)
  colnames(d)<-c("time","surv","strata","std.err")

  l<-length(d$strata[which(d$strata==levels(d$strata)[1])])
  ll<-nrow(d)
  strata<-c(rep(1,l),rep(0,(ll-l)))
  group1.dat<-head(d,l)
  group2.dat<-tail(d,(ll-l))

  mergedata<-merge(group1.dat, group2.dat, by.x=colnames(group1.dat)[1],
by.y=colnames(group2.dat)[1],all = T)
```

```
n=nrow(mergedata)

index2<-min(which(!is.na(mergedata[,2])))
index5<-min(which(!is.na(mergedata[,5])))

for (i in index5:n){
  if (is.na(mergedata[i,5])){
    mergedata[i,5]=mergedata[i-1,5]
    mergedata[i,7]=mergedata[i-1,7]}
}
for (i in index2:n){
  if (is.na(mergedata[i,2])){
    mergedata[i,2]=mergedata[i-1,2]
    mergedata[i,4]=mergedata[i-1,4]}
}
if(index5!=1){
  mergedata[1:(index5-1),5]<-1
  mergedata[1:(index5-1),7]<-0}
if(index2!=1){
  mergedata[1:(index2-1),2]<-1
  mergedata[1:(index2-1),4]<-0}

mergedata<-mergedata[,-6]
```

```
mergedata<-mergedata[,-3]
colnames(mergedata)<-c("time","surv1","se1","surv2","se2")


if (tail(raw1$event,1)==0 && tail(raw2$event,1)==0)
{tao=min(tail(raw1$time,1),tail(raw2$time,1))
mergedata[n+1,]<-c(tao,mergedata[n,2:5])}

if (tail(raw1$event,1)==1 && tail(raw2$event,1)==0)
{tao=tail(raw2$time,1)}

if (tail(raw1$event,1)==0 && tail(raw2$event,1)==1)
{tao=tail(raw1$time,1)}

if ((tail(raw1$event,1)==1 && tail(raw2$event,1)==0) ||(tail(raw1$event,1)==0 &&
tail(raw2$event,1)==1))
{mergedata$se1[which(is.na(mergedata$se1))]=0
mergedata$se2[which(is.na(mergedata$se2))]=0
mergedata[n+1,]<-c(tao,mergedata[n,2:5])}

if (tail(raw1$event,1)==1 && tail(raw2$event,1)==1)
{tao=max(tail(raw1$time,1),tail(raw2$time,1))
mergedata$se1[which(is.na(mergedata$se1))]=0
```

```
 mergedata$se2[which(is.na(mergedata$se2))]=0}


 delta=0
 edelta=0
 var2=0
 var1=0
 for (i in 1:(nrow(mergedata))){
  # print(i)
  # print(var1)
  if(mergedata$time[i]<tao){
   delta<-delta+abs(mergedata$surv1[i]-mergedata$surv2[i])*(mergedata$time[i+1]-
mergedata$time[i])
   edelta<-
edelta+sqrt(2/pi*(mergedata$se1[i]^2+mergedata$se2[i]^2))*(mergedata$time[i+1]-
mergedata$time[i])
   for (j in 1:(nrow(mergedata)-1)){
    if (i<j && mergedata$time[j]<tao){
     var2<-var2+(mergedata$time[i+1]-mergedata$time[i])*(mergedata$time[j+1]-
mergedata$time[j])*(1-
2/pi)*sqrt((mergedata$se1[i]^2+mergedata$se2[i]^2)*(mergedata$se1[j]^2+mergedat
a$se2[j]^2))}
   }
```

```
    var1=var1+(mergedata$time[i+1]-mergedata$time[i])^2*(1-
2/pi)*(mergedata$se1[i]^2+mergedata$se2[i]^2)}
  }

  var=var1+var2
  delt<-(delta-edelta)/sqrt(var)
  1-pnorm(delt)

}
```