(Weighted) Log-rank Tests with Application Scenarios

Youlan Shen, Zijian Xu, Zhengwei Song

Contents

Objective

Survival analysis, which is a branch of statistics that studies the duration of time before a particular event occurs, such as death. Survival function usually means the probability of living longer than t for a group or individual sometimes. In general, people use some built proportional hazards models to describe survival function, such as Exponential, Weibull, Cox. In the meantime, non-proportional hazards models are in general use.

Treatment effect, is a concept that describes the survival probability difference between control and treatment group, related to hazard ratio in proportional hazards model. Logrank test, weighted Logrank test, and Logrank maximum test are four hypothesis tests that measure the existence of treatment effect in survival analysis.

To examine the accuracy and efficiency of the four hypothesis tests, we conducted a simulation project. We simulated survival datas under different scenarios, both proportional hazards model and non-proportional hazards model, to compare the performance of the four hypothesis tests, in order to give a suggestion of the usage of the tests.

Background

Survivial function

T = time to death, with density function f(t)

S(t) = probability of living longer than t

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(s)ds = 1 - F(t) \tag{1.1}$$

Hazard Function

h(t) = instantaneous risk of failure at time t (giving that a patient has survived until time t):

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(T \in (t, t + \Delta t) | T > t)}{\Delta t} = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} \tag{1.2}$$

Hazard Ratio

Hazard Ratio for the i-th patient at time t:

$$h_i(t) = h_0(t) e^{\beta_1 x_i + \beta_2 x_i f(t)} \tag{1.3}$$

where

- $h_0(t)$: the baseline hazard function; There are different ways to formulate the baseline hazard function $h_0(t)$, which lead to different models
- x_i: coded 0 for control; 1 for the treatment
- β_1: the log hazard ratio for the treatment effect. β_1 measures the relative hazard reduction due to treatment in comparison to the control
- β_2: coded 0 for Proportional Hazards; otherwise for Non-proportional Hazards

- f(t): a function of t, we choose to apply piecewise constant for data simulation

The treatment effect is summarized by the hazard ratio (HR) between the control and treatment arms:

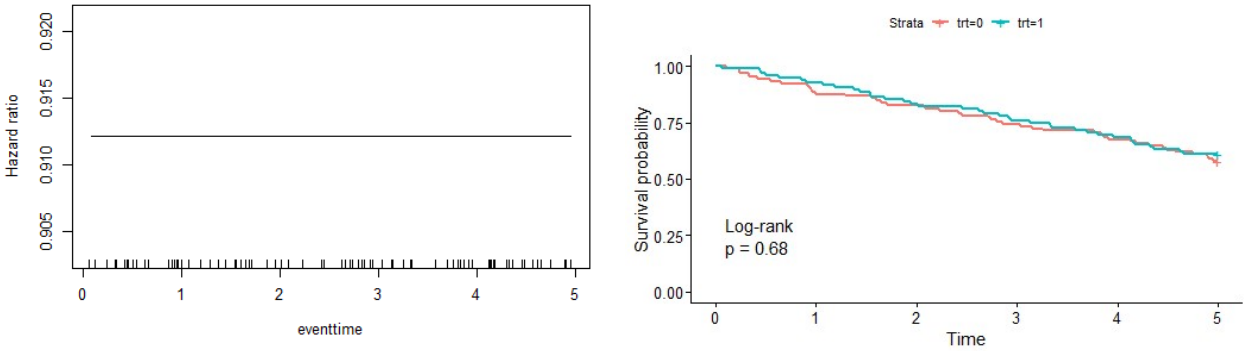
- HR < 1: Treatment arms have higher S(t)
- HR $= 1$: Two arms share similar S(t)
- HR > 1: Control arms have higher S(t)

Proportional Hazards

In proportional hazards models, the hazard ratio (HR) between two groups is constant over time, meaning that the relative risk of an event occurring in one group compared to another remains the same throughout the entire follow-up period. This can be expressed as HR = exp(b), where b is the coefficient associated with the group variable in the proportional hazards model. It's worth noting that while proportional hazards models are commonly used, they may not always be appropriate for the data at hand. In some cases, non-proportional hazards models may be a better fit, particularly if there is evidence of changing relative risks over time.

$$\frac{h(t|x_1)}{h(t|x_2)} = e^{\beta_1^T(x_1 - x_2)} \tag{1.4}$$

does not depend on t



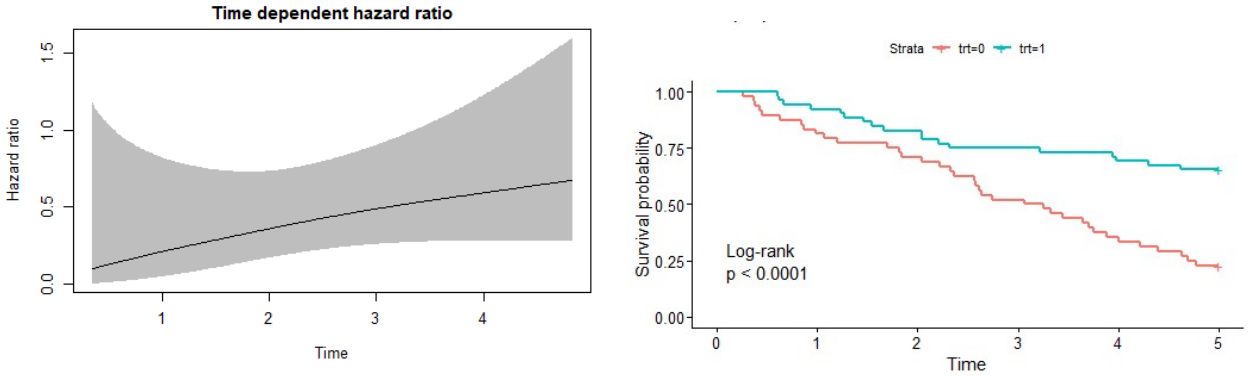
We can see the two survival probability plots are almost parallel by time under constant hazard ratio.

Non-proportional Hazards

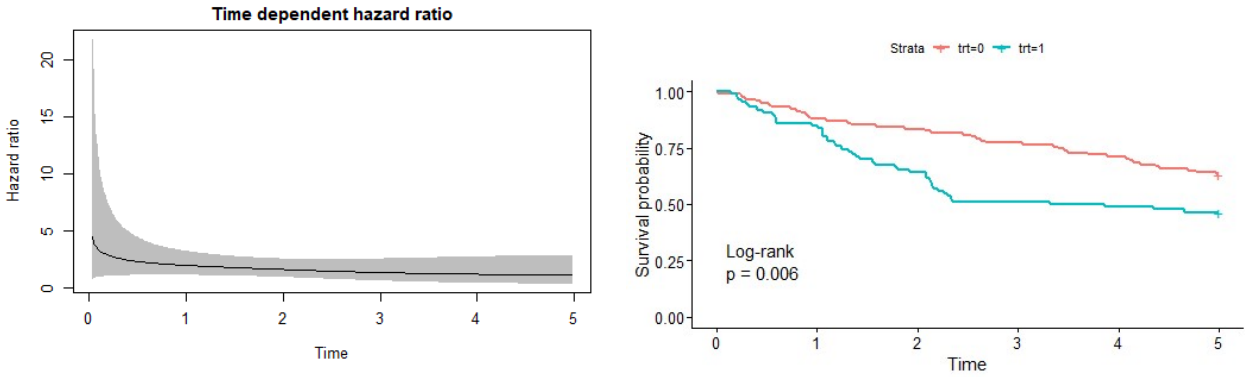
In contrast, non-proportional hazards models do not assume a constant hazard ratio over time. This means that the relative risk of an event occurring in one group compared to another changes over time, and the hazard ratio may depend on the specific follow-up time. Non-proportional hazards models often involve interactions between time and one or more covariates, indicating that the effect of the covariate(s) on the hazard of the event changes over time.

$$\frac{h(t|x_1)}{h(t|x_2)} = e^{\beta_1^T(x_1 - x_2) + \beta_2^T f(t)(x_1 - x_2)} \tag{1.5}$$

does depend on t



We can see the two survival probability plots are not parallel by time, with a sharp decline of the treatment curve in the late period, under increasing hazard by time.



We can see the two survival probability plots are not parallel by time, with a sharp decline of the treatment curve in the early period, under decreasing hazard by time.

Methods

Survival Data Simulation Overview

This study involves the simulation of right-censored survival data, with the inclusion of a binary treatment indicator x. Our response variable is dichotomous, taking the value of 1 when an event occurred and 0 when an individual censored within a given maximum observation period. The follow-up time is measured from time zero until the event of interest transpires, the study concludes, or the participant becomes lost, whichever comes first.

Proportional Models: Weibull and Exponential

An exponential proportional-hazards model assumes the baseline hazard function is a constant

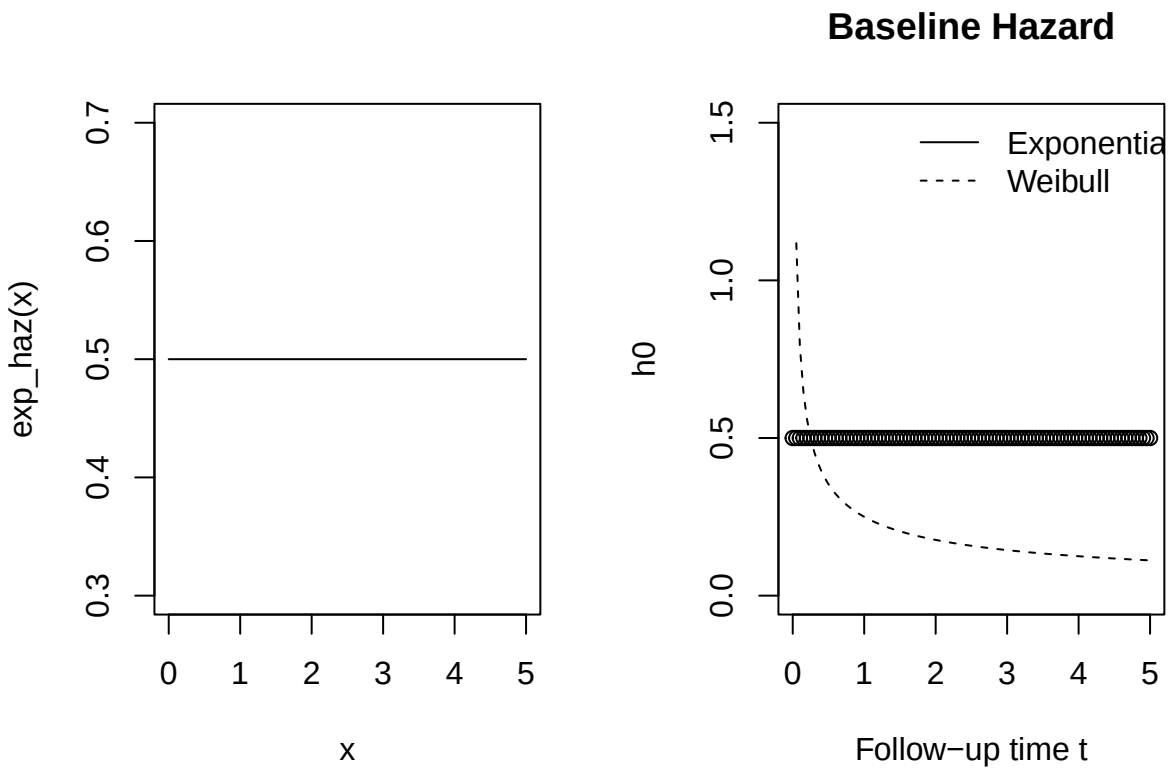
$$h_0(t) = \lambda \tag{2.1}$$

A Weibull proportional-hazards model assumes the baseline hazard function follows Weibull distribution, where

$$h_0(t) = \lambda\gamma t^{\gamma-1} \tag{2.2}$$

for $\gamma > 0$

We then plot the baseline hazard function h_0's by time t of the two proportional hazard models. The first subplot shows the baseline hazard function for an exponential distribution($\gamma = 1$), which has a constant hazard rate over time. The second subplot shows the baseline hazard function for a Weibull distribution, which has a hazard rate that increases or decreases over time depending on the value of the shape parameter γ.



Distribution Formulas

By (1.2) we can obtain the cumulative distribution function of t,

$$F(t) = 1 - \exp\left(-\int_0^t h(s)ds\right) = 1 - \exp\left[-(\int_0^t h_0(s)ds) \cdot \exp(\beta^T X)\right] \tag{2.3}$$

Generating Data

In this study, we applied Weibull and the Exponential baseline hazard functions, by utilizing the inverse transformation method

$$\begin{cases} F(t) = 1 - exp[-\lambda t^{\gamma} exp(\beta^T X)] \\ U = S(t) = 1 - F(t) \end{cases} \tag{2.4}$$

The dataset used in this study comprises information on treatment assignment, status indicator, and observed time. The treatment assignment variable X_i is obtained by generating random samples from a Bernoulli distribution with a probability of success (p) equal to 0.5. The event time t is then derived by applying the

inverse transformation method to the generated data. We obtain the event time t

$$t = S^{-1}(u) = (-\frac{log(u)}{\lambda exp(\beta^T X)})^{\frac{1}{\gamma}} \tag{2.5}$$

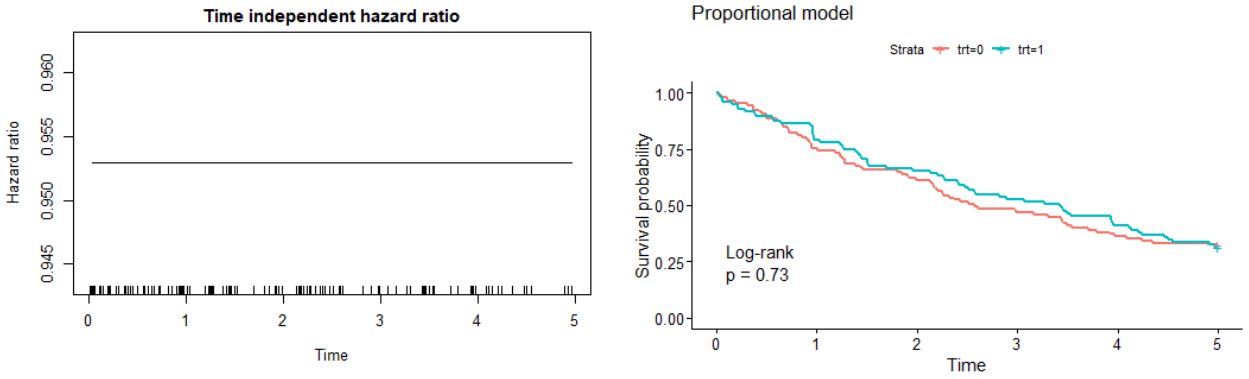
for Weibull version

Let $\gamma = 1$, we have

$$t = -\frac{log(u)}{\lambda exp(\beta^T X)} \tag{2.6}$$

for exponential version

We repeated each simulation process given times to generate survival datasets, with given sample sizes. Here is one example under exponential hazard function $\gamma = 1$, where the sample size n=200, repeat 100 times to get 100 dataframes in total, the whole observation time 5 with $\lambda = 0.2$, $\beta_1 = 0$:



Non-proportional Models

There are many versions of non-proportional hazard functions, we choose to generate data from a non-proportional hazard model with the hazard function defined as

$$h_i(t) = h_0(t)e^{\beta_1 x_i + \beta_2 x_i f(t)} \tag{2.7}$$

in this study we also choose to use piecewise constant fucntion to define the f(t), and the Weibull baseline hazard function to define $h_0(t)$, and $\beta_2 = -\beta_1$ to better show early and late effect mathematically.

Distribution Formulas

We set a cut-off time point t_0, the f(t) is defined as follows

$$f(t) = I(t) = \begin{cases} 1 & t > t_0 \\ 0 & otherwise \end{cases} \tag{2.8}$$

for early effect

$$f(t) = I(t) = \begin{cases} 0 & t < t_0 \\ 1 & otherwise \end{cases} \tag{2.9}$$

for late effect

Thus (2.3) becomes

$$h_i(t) = \lambda\gamma t^{\gamma-1} exp[\beta_1(1 - I(t))x_i] \tag{2.10}$$

Generating Data

Simiarly as (2.2), we applied Weibull and the Exponential baseline hazard functions, by utilizing the inverse transformation method

$$\begin{cases} F(t) = 1 - exp[-\lambda t^{\gamma} exp(\beta_1(1 - I(t))x_i)] \\ U = S(t) = 1 - F(t) \end{cases} \tag{2.11}$$

The dataset used in this study comprises information on treatment assignment, status indicator, and observed time. The treatment assignment variable X_i is obtained by generating random samples from a Bernoulli distribution with a probability of success (p) equal to 0.5. The event time t is then derived by applying the inverse transformation method to the generated data. We obtain the event time t

$$t_i = S^{-1}(u) = (-\frac{log(u)}{\lambda exp(\beta_1(1 - I(t))x_i)})^{\frac{1}{\gamma}} \tag{2.12}$$

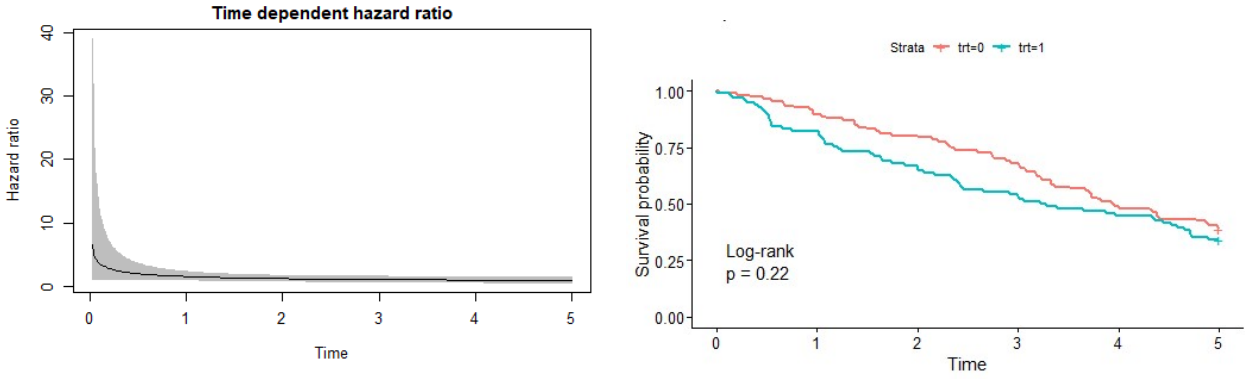
for Weibull version

Let $\gamma = 1$, we have

$$t_i = -\frac{log(u)}{\lambda exp(\beta_1(1 - I(t))x_i)} \tag{2.13}$$

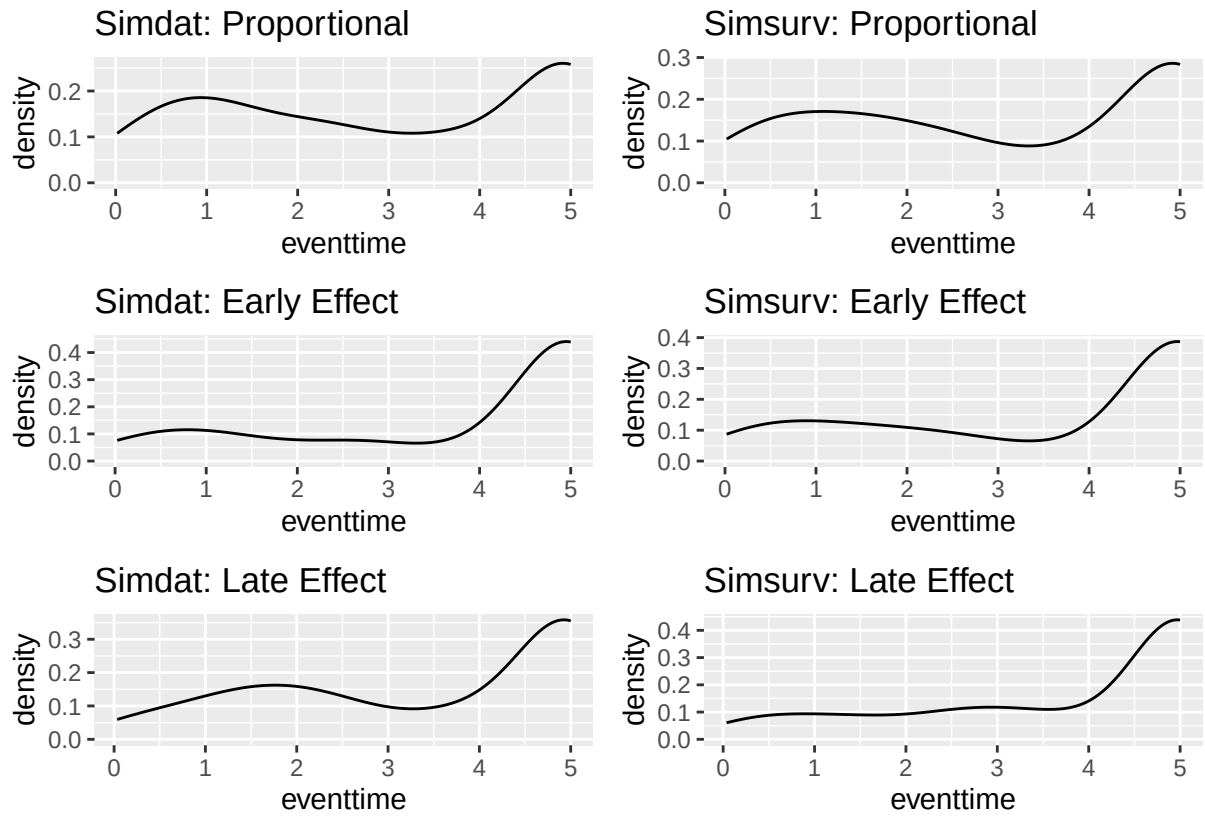
for exponential version

We repeated each simulation process given times to generate survival datasets, with given sample sizes. Here is one example under early effect $t_0 = 2.5$, where the sample size n=200, repeat 20 times to get 20 dataframes in total, the whole observation time 5 with $\lambda = 0.1$, $\gamma = 1$, $\beta_1 = 0.75$, $\beta_2 = -\beta_1 = -0.75$:



Comparison with the mature R function *simsurv*

simsurv::simsurv is a useful tool for generating data under the exponential, Weibull models that allow for time-dependent covariates and/or non-proportional hazards. Now we can compare the performance for the generated data in our R function ***simdat*** to that in ***simsurv***.



There could be many factors at play why the output from our function ***simdat*** might differ from the mature R function ***simsurv***. However, there are several differences in the way that the two functions generate survival data:

- ***simsurv*** uses a piecewise-constant hazard function with random knot locations, while simdat uses a Weibull hazard function with a fixed shape parameter and covariate effects.
- ***simsurv*** uses a more complex method to generate censored data, including stratified sampling to ensure a certain proportion of censored observations.
- ***simdat*** has the default values for some parameters (such as the standard deviation of the error term used to jitter values near 1/365) may differ between the two functions.
- All of these differences could contribute to differences in the output of the two functions. It's worth noting that the two functions are not necessarily intended to produce identical results, but rather to provide flexible ways to simulate survival data that can be tailored to different research questions and settings.

Log Rank-test

In our study we want to compare the following four hypothesis tests under two versions for the various proportional and non-proportional hazard models.

Ordinary version

Let $t_1 < t_2 < \cdots < t_J$ are J distinct failure times; The log-rank test statistics calculate the difference between "observed" and "expected" number of faitures (under H_0) at each observed failure time, and aggregate them overtime.

$$S_{Logrank} = \frac{\sum_{j=1}^{J}(O_j - E_j)}{\sqrt{\sum_{j=1}^{J} V_j}} \sim N(0,1) \text{ under } H_0 \tag{1.6}$$

where O_j and E_j are "observed" and "expected" numbers of failures at the jth failure time, and V_j is the variance of the observed number of failures.

Weighted version

To adjust for potential non-proportional hazard functions, Fleming and Harrington considered "weighted log rank tests"

$$S^w_{Logrank} = \frac{\sum_{j=1}^{J} W_j(O_j - E_j)}{\sqrt{\sum_{j=1}^{J} W_j V_j}} \tag{1.7}$$

where the weight function $w(t) = \widehat{S}(t)^\rho(1 - \widehat{S}(t))^\gamma$ include two parameters ρ and γ.

The only change compared to the ordinary log-rank test is to multiply a weight function W, which is built on survival function S(t). If we let

- $\rho = 0$ and $\gamma = 1$, the weight function focuses on the late period.
- $\rho = 1$ and $\gamma = 0$, the weight function focuses on the early period.
- $\rho = 0$ and $\gamma = 0$, the weight function does not play a role, the test becomes Ordinary Log-rank test.

The weighted log-rank test is much fair for non-proportional hazard model because it can gives more weight to the test statistic $S^w_{Logrank}$ on the specific effect period (early or late effect). Furthermore, we also have the Maximum Log-Rank Test to take the maximum test statistic $S^w_{Logrank}$ over the previous three log-rank tests with multiple comparison control.

Measure of Test Peformance

We set two criteria to compare the four hypothesis tests.

First is the type I error α, the percentage of wrong results from the four tests (p-value <= 0.05) when the simulated data show no treatment effect. If the Type I error rate is too high, we may conclude that the test is too liberal and may be incorrectly identifying differences in survival that are due to chance.

Second is the power β, the percentage of correct results from the four tests (p-value < 0.05) when the simulated data show treatment effect. Power reflects tests's ability to detect a true difference in survival between two groups. If the power is low, we may conclude that the test is too conservative and may be failing to identify differences in survival that are clinically important.

Type I error and α:

Type I error is the rejection of the true null hypothesis, while the null hypothesis in four Logrank test is that there is no treatment effect. In order to obtain this result, we simulated survival data of zero treatment

effect repeatedly, and continuously applied four tests. The percentage of four tests rejecting the true null hypothesis is our α.

$$\alpha_k = \frac{\sum_{i=1}^{n} I(p - value < 0.05 \; in \; test_{ik})}{n},$$

where n is the number of repeated simulation, i is the ith test in $test_k$,

k is 1 to 4, e.g. k = 1, α_1 is the type I error for ordinary Logrank test

Power β:

Power is the probability that the test correctly rejects the null hypothesis, while the alternative hypothesis is true. In four Logrank test, the alternative hypothesis is that there exists treatment effect. In order to obtain this result, we simulated survival data of none zero treatment effect repeatedly from small to large, and continuously applied four tests. The percentage of four tests correctly rejecting the null hypothesis is our β.

$$\beta_k = \frac{\sum_{i=1}^{n} I(p - value < 0.05 \; in \; test_{ik})}{n},$$

where n is the number of repeated simulation, i is the ith test in $test_k$,

k is 1 to 4, e.g. k = 2, β_2 is the power for Weighted Logrank test-Late

Consistent low Type I error α and high power β are what we look for, therefore, we compared the four hypothesis tests performance in different scenarios.

Results

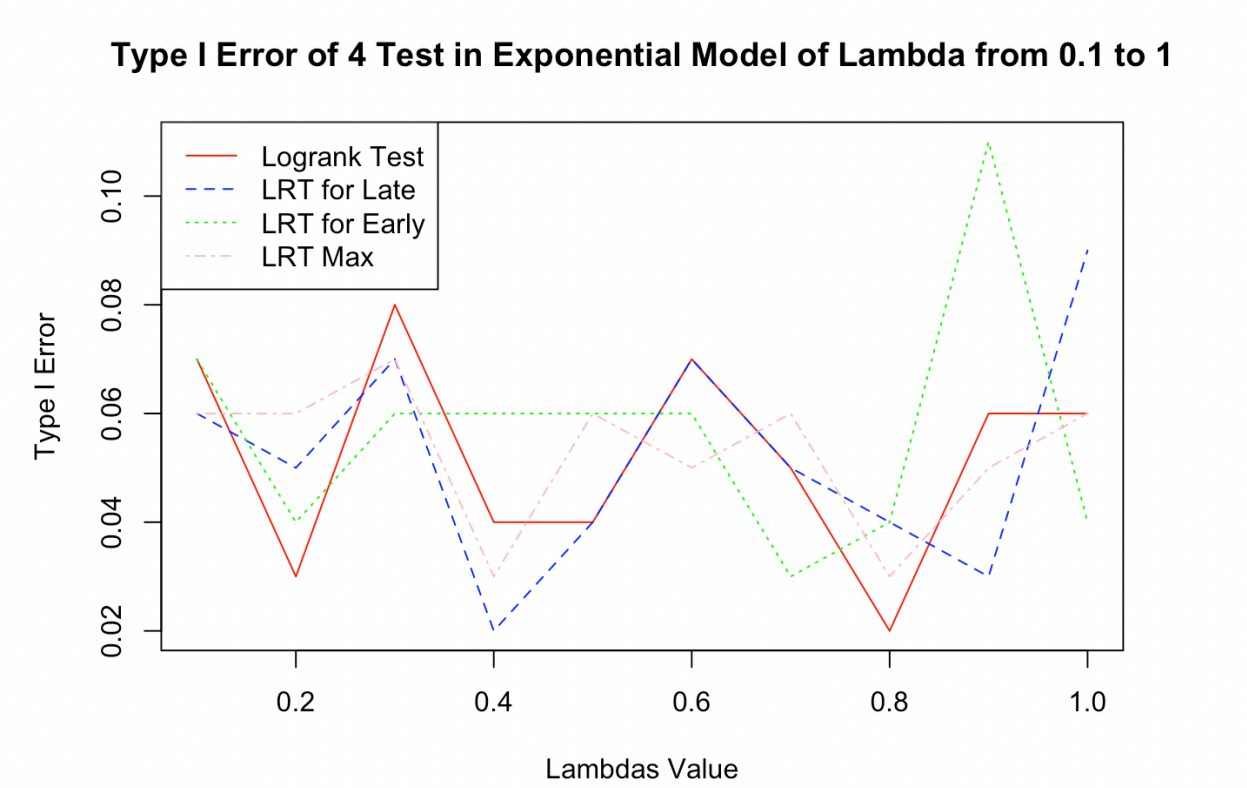
Tests for Proportional Hazards model

Exponential proportional-hazards model:

Exponential proportion-hazards model's hazard function is listed below:

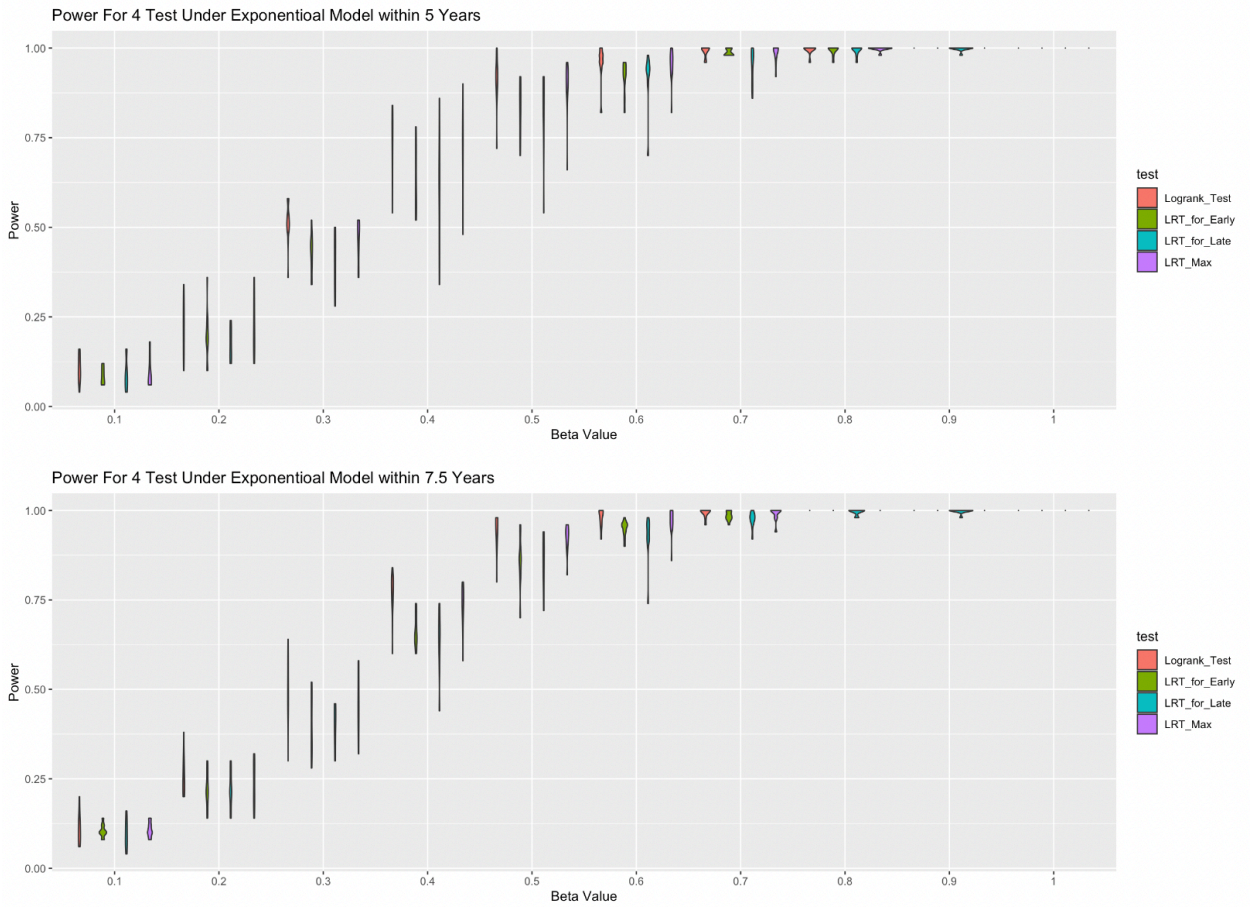
$$h_i(t) = \lambda e^{\beta X_i}$$

Set $\beta = 0$, varying λ from 0.1 to 1 by 0.1, sample size of 200, maximum recording time to 7.5 years, simulated 100 data sets in each simulation. We repeated the hypothesis test for each data set.



From the figure above, the Type I error of four hypothesis tests are around 0.05, it fluctuates a bit, but not very much. We drive the conclusion that the Type I error is consistent and small for four hypothesis tests with small variance. Therefore, on this comparison standard, all four tests performed well.

Set $\beta \neq 0$ and varying from 0.1 to 1 by 0.1, indicating the increase of the absolute value of the treatment effect. Set λ from 0.1 to 1 by 0.1, sample size of 200, maximum recording time to 5 and 7.5 years, simulated 50 data sets in each simulation. We repeated the hypothesis test for each data set.



From the figure above, the power of four hypothesis tests are all increasing as the absolute value of the treatment effect increases. When the treatment effect becomes more obvious, the power of four tests are all near 100%. When maximum recording survival time changes from 5 to 7.5 years, we can see from the plot that the length of each small violin figures shrinks, which means holding everything else costant, the variance of the Type I error will decrease as the maximum recording survival time increases.

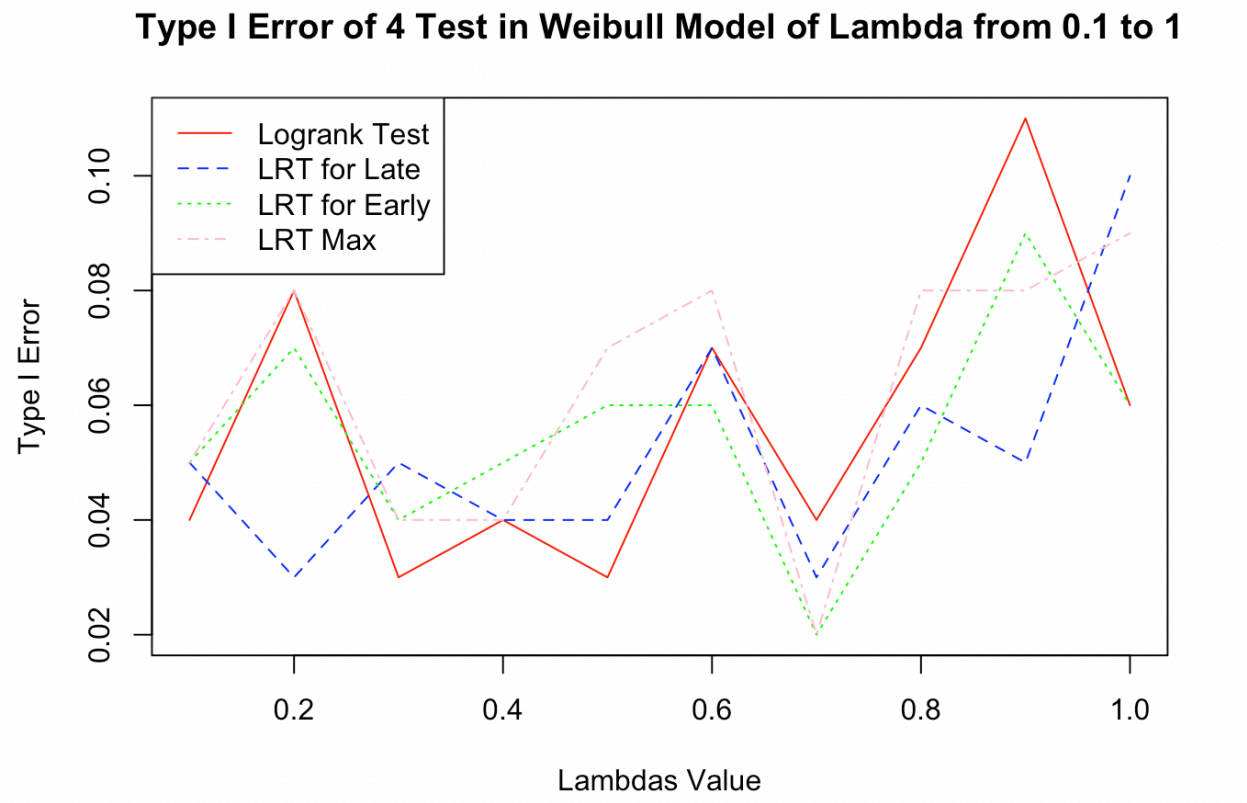
Moreover, the red one which represents ordinary Logrank test, is always having the highest mean with a relatively small variance, compared to other three tests. We drive the conclusion that the four tests performs bad in power measure when the treatment effect is not obvious However, as the treatment effect becomes more obvious, ordinary Logrank test' power is always the highest one. Therefore, on this comparison standard, ordinary Logrank Test performed best.

Weibull proportional-hazards model:

Weibull proportion-hazards model's hazard function is listed below:

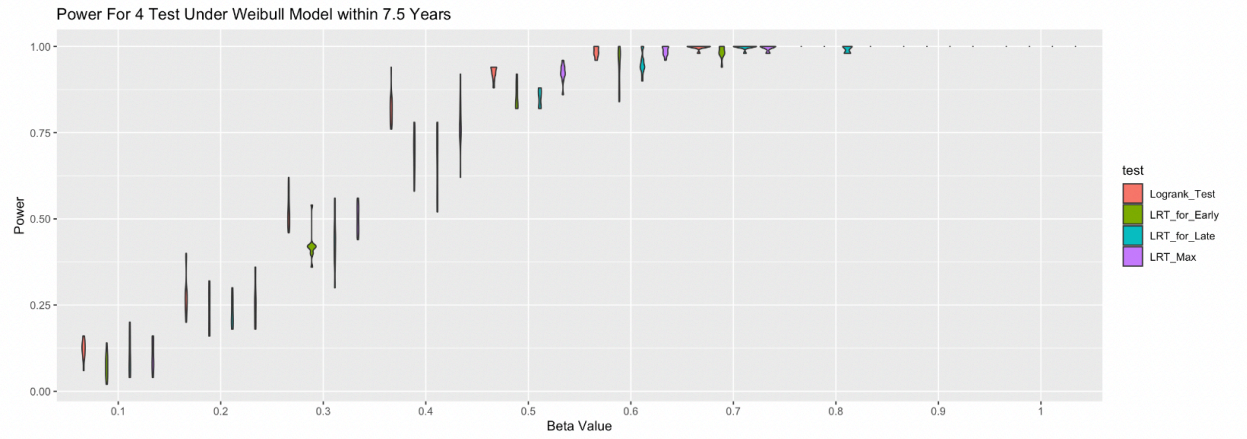
$$h_i(t) = \lambda\gamma t^{\gamma-1} e^{\beta X_i}$$

Set $\beta = 0$, $\gamma = 1.5$, varying λ from 0.1 to 1 by 0.1, sample size of 200, maximum recording time to 7.5 years, simulated 100 data sets in each simulation. We repeated the hypothesis test for each data set.



From the figure above, the Type I error of four hypothesis tests are around 0.05, it fluctuates a bit, but not very much. At around $\lambda = 0.9$, we can see that while ordinary Logrank test and two weighted Logrank tests have a large Type I error, Logrank Maximum test has a relative small Type I error. Moreover, Logrank maximum test has the lowest Type I error almost on every point. We drive the conclusion that the Type I error is consistent and small for four hypothesis tests with small variance. On this comparison standard, all four tests performed well, but Logrank Maximum test performs slightly better than other three tests.

Set $\beta \neq 0$ and varying from 0.1 to 1 by 0.1, indicating the increase of the absolute value of the treatment effect. Set λ from 0.1 to 1 by 0.1, sample size of 200, maximum recording time to 7.5 years, simulated 50 data sets in each simulation. We repeated the hypothesis test for each data set.

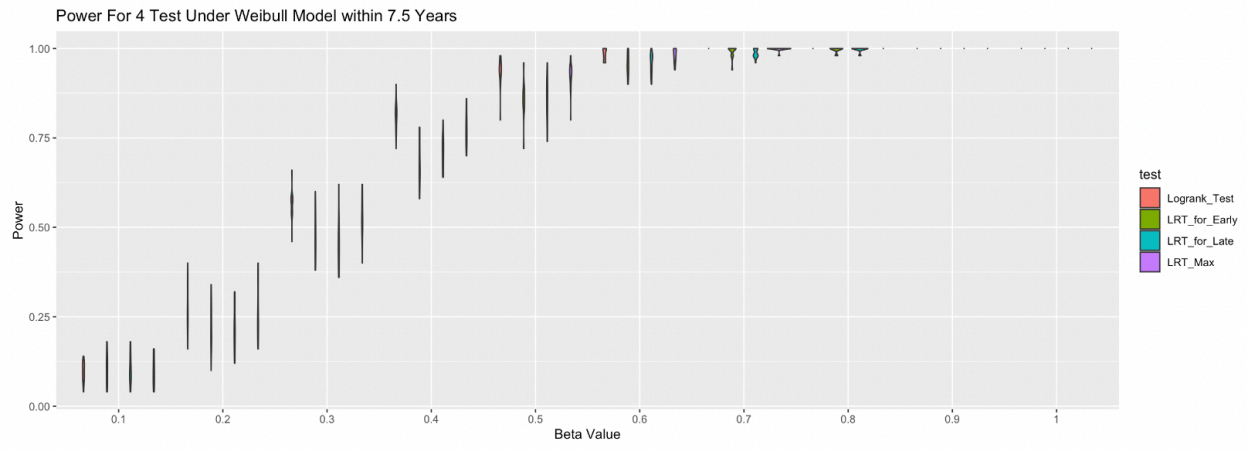


From the figure above, the power of four hypothesis tests are all increasing as the absolute value of the treatment effect increases. When the treatment effect becomes more obvious, the power of four tests are all near 100%.

Moreover, the red one which represents ordinary Logrank test, is always having the highest mean with a relatively small variance, compared to other three tests. We drive the conclusion that the four tests performs bad in power measure when the treatment effect is not obvious However, as the treatment effect becomes more obvious, ordinary Logrank test' power is always the highest one. Therefore, on this comparison standard, ordinary Logrank Test performed best.

To check if this conclusion is consistent when the other parameter, γ also changes while λ changes, we repeated the power measure in a new parameter setting.

Set $\beta \neq 0$ and varying from 0.1 to 1 by 0.1, indicating the increase of the absolute value of the treatment effect. Set λ from 0.1 to 1 by 0.1, γ from 1.5 to 2 by 0.1, sample size of 200, maximum recording time to 7.5 years, simulated 50 data sets in each simulation. We repeated the hypothesis test for each data set.



From the figure above, we can conclude that the result is consistent. The red one which represents ordinary Logrank test, is always having the highest mean with a relatively small variance, compared to other three tests. Therefore, on this comparison standard, ordinary Logrank Test performed best.

Tests for Non-Proportional Hazards model

In our study, we needed to compare the performance of the weighted log-rank test and the log-rank test under various scenarios of non-proportional hazards using simulation. If there is truly no treatment effect, then the non-proportional hazards model is identical to proportional hazards model. There is no need for us to measure Type I error in non-proportional hazards model. Thus we only used power to totally represent the performance of four version of tests in the next discussion.

Test performance design

Data illustration, Test Simulated Data Validity

The first step is to inspect and verify the simulated data for its plausibility and validity.(eg. whether or not we have simulated non-proportional data with early effect and late effect correctly, and what is the corresponding hazard ratio.)

We have two data generation function, one is our design, the other is simsuv, which derive from r package. First of all, we compare the performance of two generation function, and choose the one which can fulfill our experimental requirement: generate time dependent survival data, and make early or late effect as obviously as possible.

The below flow chart shows the design of the test performance procedure in non-proportional model.

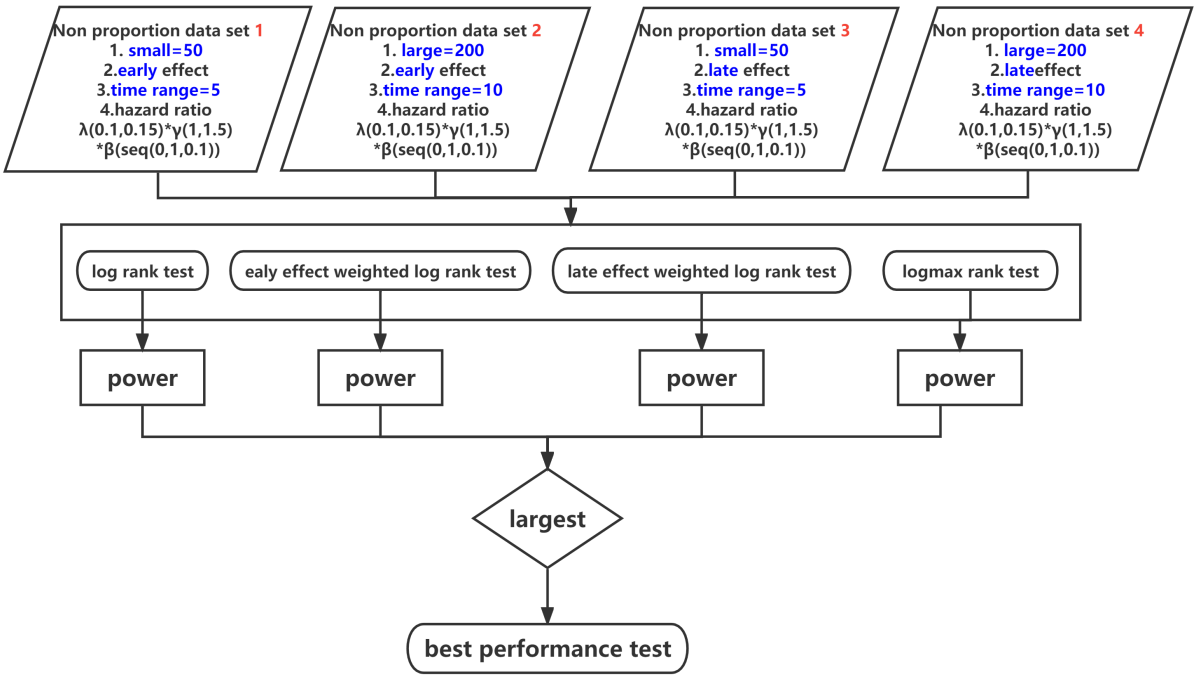
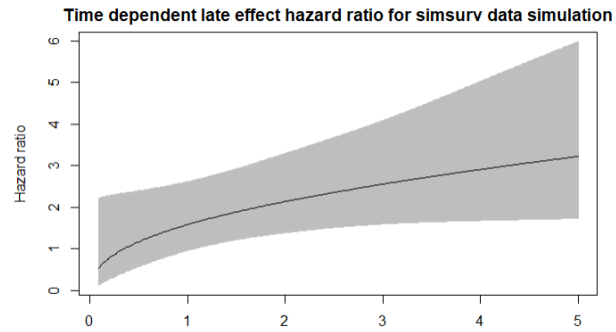
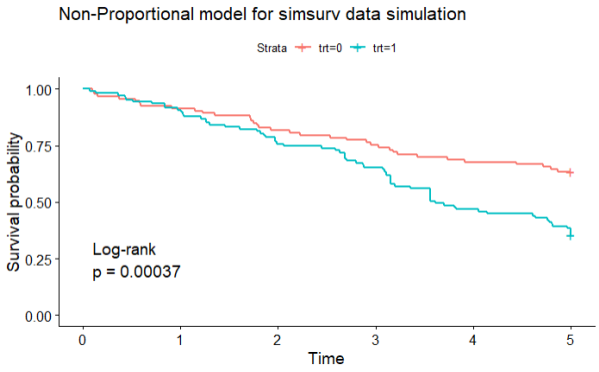


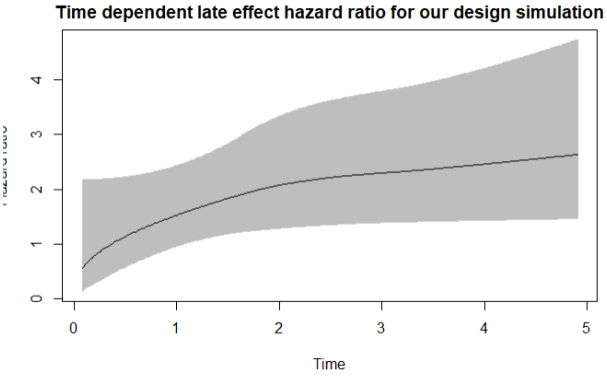
Figure 1: The logic process of design our datasets and step of manipulation of test

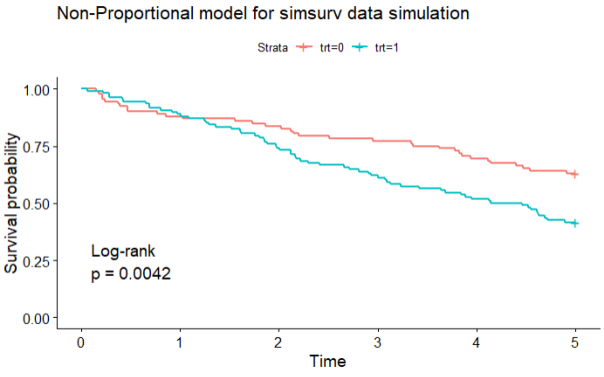
The late effect for simsurv simulation function data is



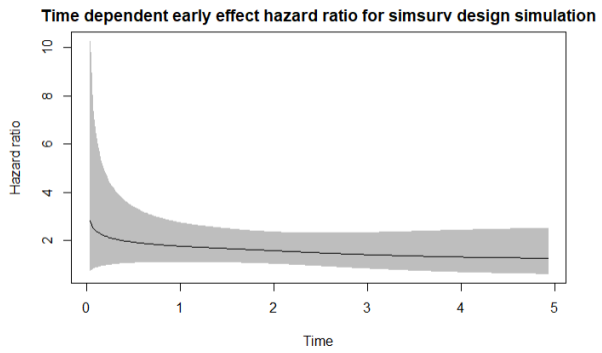


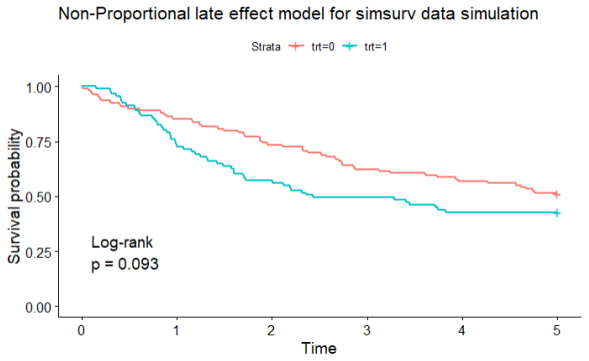
The late effect for our design simulation function data is



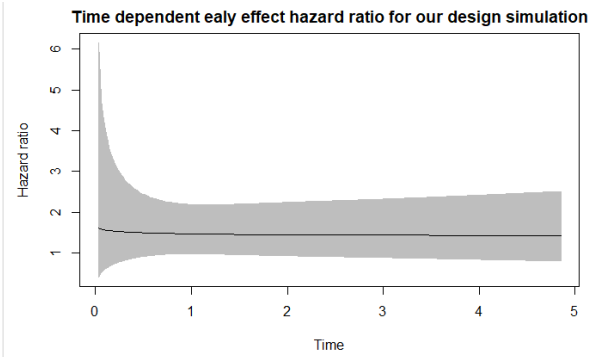


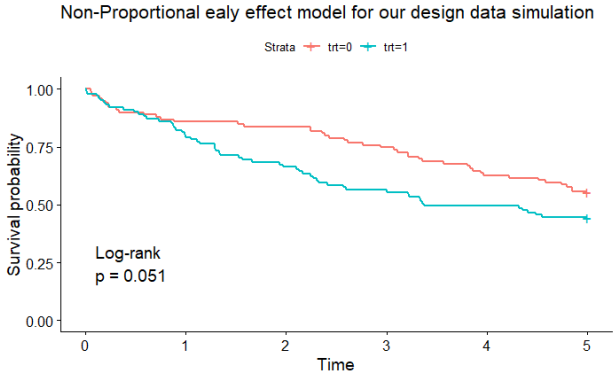
The early effect for simsurv simulation function data is





The early effect for our design simulation function data is



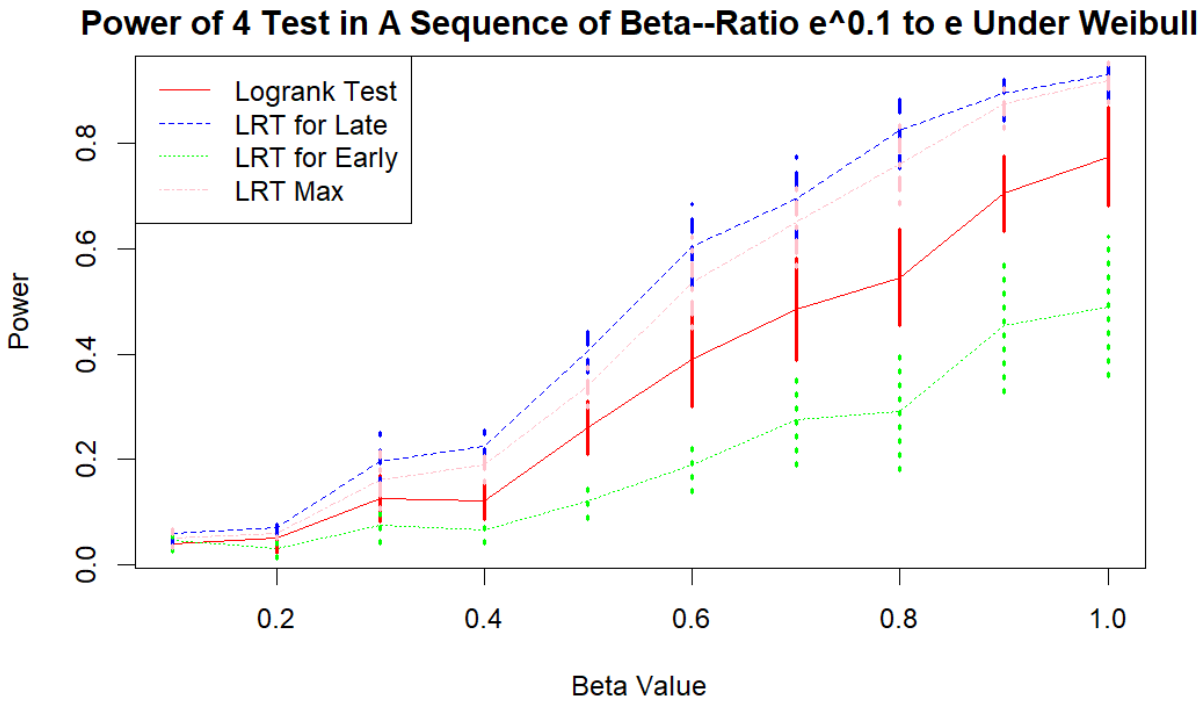


By comparison, it can be seen that our designed data simulation algorithm performs fairly well for non-proportional data simulation, especially in terms of early and late effects, as compared to the Simsurv algorithm. However, Simsurv outperforms our algorithm. Specifically, for data with early effects, the hazard function curve simulated by our algorithm does not show a clear decrease in slope as that simulated by Simsurv, while in some circumstances, the hazard ratio showed rapid decrease or increase by our algorithm, which means the data simulated from simsurv is more robust from the stratifying sampling.

Thus, in order to make the experimental results more significant, we have decided to use the Simsurv method for the non-proportional test performance experiment, and our functions are also attached in the code comment.

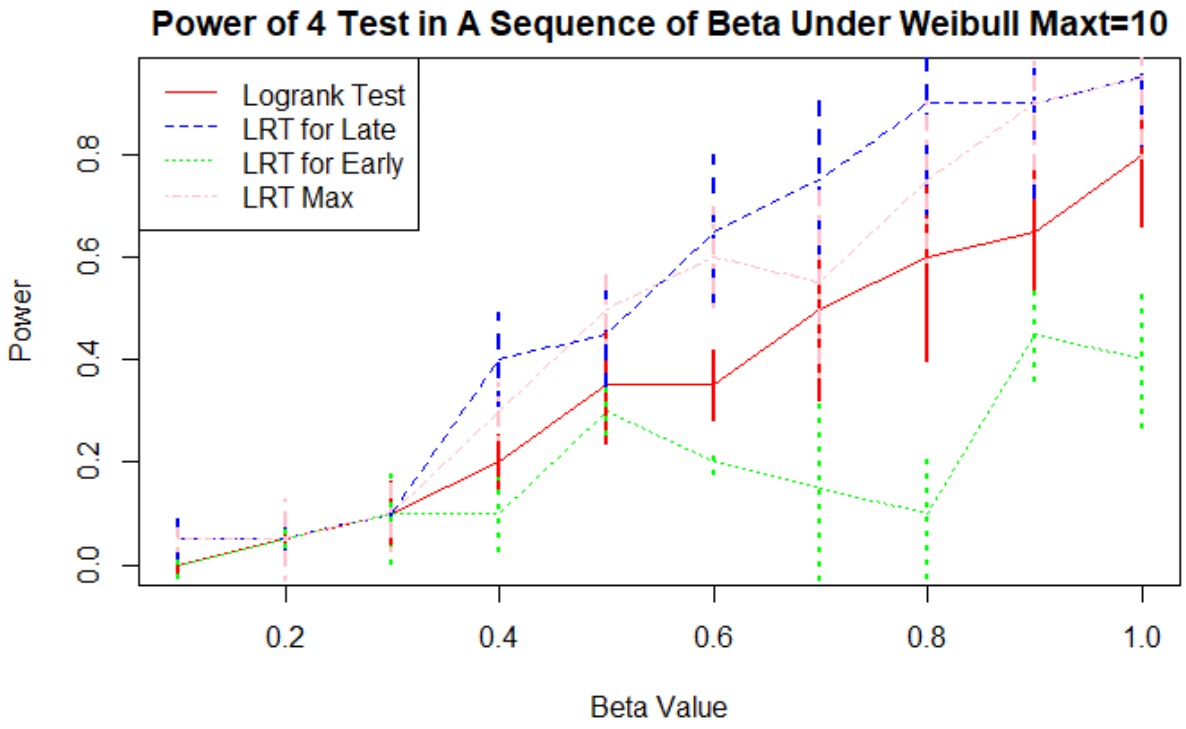
Tests' Performance Comparasion

Late obvious data with small datasets = 50, time series from 0 to 5 (maxt) = 5 In this scenario, we have a small datasets and limited time series data, we changed lambdas and gamma for different combination and changed beta from 0 to 1, by 0.1. By calculating the variance of test power change, We created a confidence interval with the beta change under different parameters sets.



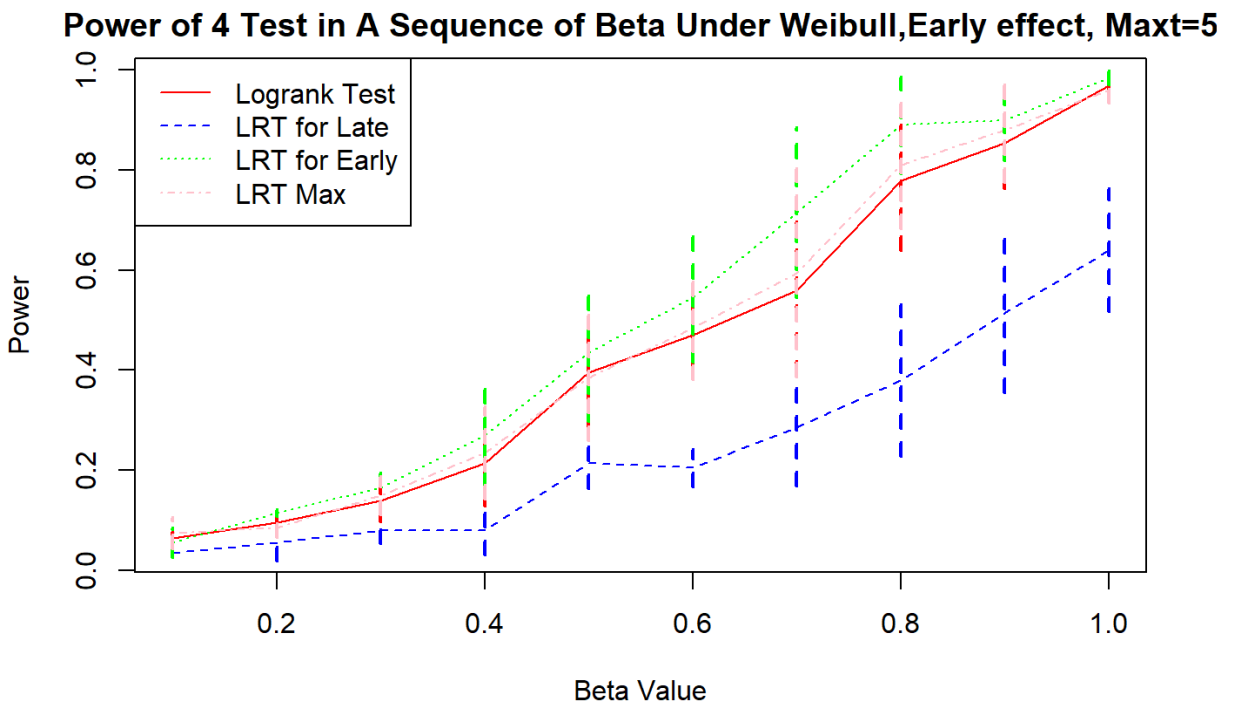
Observing the graph, the vertical lines in the graph represent the confidence intervals and the blue curve is consistently higher and significantly higher than the other curves, indicating that the late weight log-rank test outperforms other versions of the test at any beta level. We can also see that the curve increases quickly at the beginning and slowly at the end. For the blue curve, a beta of at least 0.8 is required to achieve a power level greater than 0.8. For the pink curve, a beta of at least 0.9 is required to achieve a power level greater than 0.8. The red and green curves fail to reach a power level of 0.8. This suggests that the late weight log-rank test can still be effective even when the beta level is not very significant, demonstrating its robust detection power.

Late obvious data with small datasets = 200, time series from 0 to 10 (maxt) = 10 In this scenario, we have large data sets and wide time series data, we changed lambdas and gamma for different combination and changed beta from 0 to 1, by = 0.1. By calculating the variance of test power change, We created a CI with the beta change under different parameters sets.



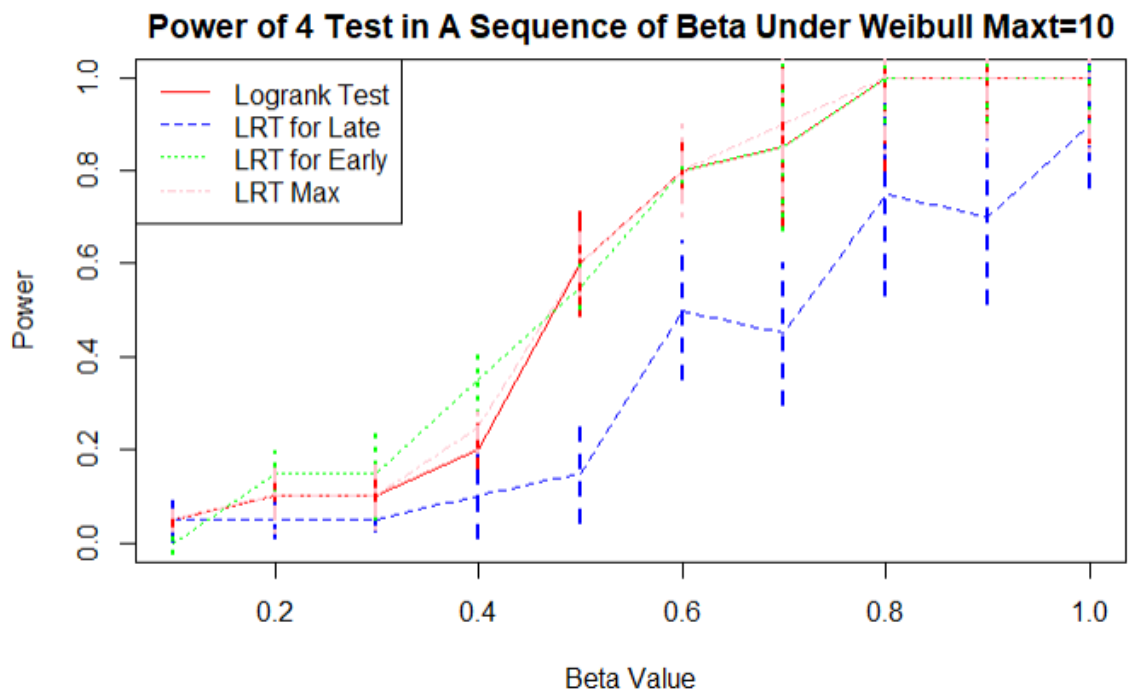
The vertical lines in the graph represent the confidence intervals. Based on the observation of the graph, the blue line performs similarly to the other lines at some points, and outperforms them at most points. As the dataset expands and the timeline lengthens, the advantage of the late effect log weight rank test will weaken, but it still performs better than other versions of the test. Furthermore, by examining the points at which the power of the late effect logrank test reaches its peak as the beta value increase, it is noted that its peak occurs earlier than previous graph. This indicates that as the dataset expands, the detection power becomes more pronounced. In other words, even when the hazard ratio is not particularly significant, as long as the dataset is large enough and the time is long enough, late effect log weight rant test can still detect it.

Eealy effect with small datasets = 50, time series from 0 to 5 (maxt) = 5 In this scenario, we have small data sets and limited time series data, we changed lambdas and gamma for different combination and changed beta from 0 to 1. By calculating the variance of test power change, we created a CI with the beta change under different other parameters sets.



Observing the graph, the green curve is consistently higher and significantly higher than the other curves, indicating that the early weight log-rank test outperforms other versions of the test at any beta level. We can also see that the curve increases quickly at the beginning and slowly at the end. For the green curve, a beta of at least 0.7 is required to achieve a power level greater than 0.8. For the pink and red curve, a beta of at least 0.9 is required to achieve a power level greater than 0.8. The blue curves fail to reach a power level of 0.8. This suggests that the early weight log-rank test can still be effective even when the beta level is not very significant, demonstrating its robust detection power.

Eealy effect with large datasets = 200, time series from 0 to 10 (maxt) = 10 In this scenario, we have large data sets and wide time series data, we changed lambdas and gamma for different combination and changed beta from 0 to 1. By calculating the variance of test power change, we created a CI with the beta change under different other parameters sets.



Apart from the blue line, all other lines exhibit similar detection capabilities. This indicates that, as the absolute value of hazard ratio (beta value) increases, all tests, except for the late weight log-rank test, have almost same power at the same level. Moreover, as beta approaches 1, especially when beta is greater than 0.8, the power of the other three types of tests tends towards 1. This suggests that as the dataset and time series increase in size, the advantage of the early weight log-rank test is no longer significant (But at some point, early effect weighted log rank test still performs better). In a nutshell, early weight log-rank test still has the strongest detection power, compared to the other three tests, but the difference of power among them decreases as the dataset size and time series length increase.

Conclusions & Suggestions

For real survival analysis, if we believe the survival data is under proportional assumption, and we can use built hazards models to fit the survival data:

If we believe two groups have similar survival function, which means there is no treatment effect, all four logrank test are recommended. Generally, if we want to minimize the type I error or want to have higher accuracy, we would recommend maximum Logrank test, since under all scenarios, its type I error is almost the smallest one.

If we believe two groups have different survival function, which means there exists treatment effect, ordinary Logrank test is the best choice, since it has the highest power to have the correct result.

If we believe two groups have different survival function, and the absolute hazard ratio is much different from 1, for example, higher than 1.6 or lower than 0.6, all four tests performs similar. So we would recommended all four of them.

If we believe the survival data is not under proportional assumption:

First, use the corresponding logrank test based on the shape of the curve:

In a real test situation, we can use the shape of the survival probability curve to determine whether to use the early weight logrank test or the late weight logrank test. If the survival curve has an “convex” shape, we

should use the early weight logrank test. If the survival curve has a trumpet open-ending shape, we should use the late weight logrank test. Alternatively, if the hazard ratio decreases over time, we use an early effect weighted test, and if the hazard ratio increases over time, we use a late effect weighted test.

Second, increase the volume of data and the length of time series:

The power of the weighted logrank test is related to the hazard ratio and data sets size and length of time (end time). Therefore, if we have a longer time series of survival data and a larger hazard ratio, the test result will be better. Thus, more data volume and longer time series are always welcome for increasing the accuracy of the logrank test.

To conclude, the recommendation is to use the appropriate logrank test based on the shape of the survival curve, and the hazard ratio, and if possible, we recommend increasing the volume of data, and the length of the time series to improve the accuracy of the hypothesis test results.

Contribution

All members contributed and actively participated in the group project. All of us explored the background and formulas for simulating survival data. Zhengwei Song took the major part of the methodology and data simulation. Youlan Shen contributed to the design of test performance and proportional model test results. Zijian Xu contributed to the non-proportional model design and test results, and conclusion.

Reference

Tasnime Hamdeni & Soufiane Gasmi (2022) A proportional-hazards model for survival analysis and long-term survivors modeling: application to amyotrophic lateral sclerosis data, Journal of Applied Statistics, 49:3, 694-708, DOI: 10.1080/02664763.2020.1830954

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. Biometrika 69, 553-566.

Anthony Y.C. Kuk, Estimating Monotonic Hazard Ratio Functions of Time, International Statistical Review, 10.1111/insr.12483, 90, 2, (285-305), (2021).