Group Project 1

Huanyu Chen, Shaolei Ma, Ruiqi Xue

Abstract

This study compares the performance of three hypothesis tests for time-to-event data: the conventional log-rank test and two variants of the weighted log-rank test. We assess their effectiveness under scenarios involving both proportional and non-proportional hazard functions, using Monte Carlo simulation techniques to evaluate power across a range of coefficient values. Our analysis highlights nuanced differences in their ability to detect treatment effects, providing insights into selecting appropriate statistical methodologies for analyzing time-to-event data in clinical trials.

1 Introduction

In clinical trials, one common type health outcome used to assess the treatment effect is the time-to-event outcome, where the measurement is usually the hazard ratio under the proportional hazards assumption. The log-rank test is used to compare observed and expected event counts. However, in real-world scenarios, the proportional hazards assumption may not always hold, requiring statistical adjustments. To address this issue, researchers have proposed weighted log-rank tests that incorporate parameters for different emphases on early or late events. In this project, the scientific question we aim to address is that given assumptions of proportional hazards and non-proportional hazard, which type of log-rank test is the most suitable to use, based on the simulation of the survival distributions. And we give the answer by comparing the performance of different log-rank tests.

2 Methods

2.1 Log Rank Test

The log-rank test statistic calculates the difference in observed versus expected failures over time.

$$\chi^2 = \frac{[\sum_{t=1}^{D}(o_t - e_t)]^2}{\sum_{t=1}^{D} v_t}$$

where o_t, observed number of deaths in treatment group at time t; e_t, expected number of deaths in treatment group at time t; v_t, variance of expected number of deaths in treatment group at time t.

A weighted log-rank test incorporates a weight function w_t that may change over time, allowing for the testing of differences between the survival curves under alternatives that differ from proportional hazards.

$$\chi^2 = \frac{[\sum_{t=1}^{D} w_t(o_t - e_t)]^2}{\sum_{t=1}^{D} w_t^2 v_t}$$

2.2 Proportional-Hazard Assumption

Under proportional-hazards assumption, the hazard function (Cox model) can be written as:

$$h(t|x) = h_0(t)exp(\beta' x)$$

where t is the time, x the vector of covariates, β the vector of regression coefficients, $h_0(t)$ is the baseline hazard function. Then, the survival function is

$$S(t|x) = exp[-H_0(t)exp(\beta' x)]$$

where

$$H_0(t) = \int_0^t h_0(u)du$$

Thus, the distribution function is

$$F(t|x) = 1 - exp[-H_0(t)exp(\beta' x)]$$

Let Y be a random variable with distribution function F, then $U = F(Y) \sim U(0,1)$, $(1-U) \sim U(0,1)$, i.e.

$$U = exp[-H_0(t)exp(\beta' x)] \sim U(0,1)$$

if $h_0(t) > 0$ for all t, then H_0 can be inverted and the survival time T of the model can be written as

$$T = H_0^{-1}[-log(U)exp(-\beta' x)]$$

where $U \sim U(0,1)$.

To simply the problem, here we only consider one covariate x, which indicates whether the sample belongs to the control arm ($x = 0$) or the treatment arm ($x = 1$), and set a negative β under the assumption that the treatment has a consistent positive effect.

Now, we only need to know H_0^{-1} to simulate the survival time. To do so, we consider two commonly used survival time distributions: **Exponential distribution** and **Weibull distribution**.

2.2.1 Exponential Distribution

For exponential distribution with scale parameter $\lambda > 0$, the possibility density function is $f_0 = \lambda exp(-\lambda t)$.

Thus, $T = -\lambda^{-1} log(U)exp(-\beta' x)$ where $U \sim U(0,1)$.

2.2.2 Weibull Distribution

For Weibull distribution with the scale parameter λ, and is the shape parameter γ, the possibility density function is $f_0 = \lambda\gamma t^{\gamma-1} exp(-\lambda t^{\gamma})$. Thus, $T = (-\lambda^{-1} log(U)exp(-\beta' x))^{1/\gamma}$ where $U \sim U(0,1)$.

2.3 Non-Proportional-Hazard Assumption

Under Non-Proportional-Hazard Assumption, we still consider the exponential model and Weibull model.

2.3.1 Piecewise Exponential Model

We suppose the hazard function for the treatment arm is: $h(t|x=1) = \begin{cases} \lambda_0 & t < 1 \\ \lambda_1 & t \geq 1 \end{cases}$

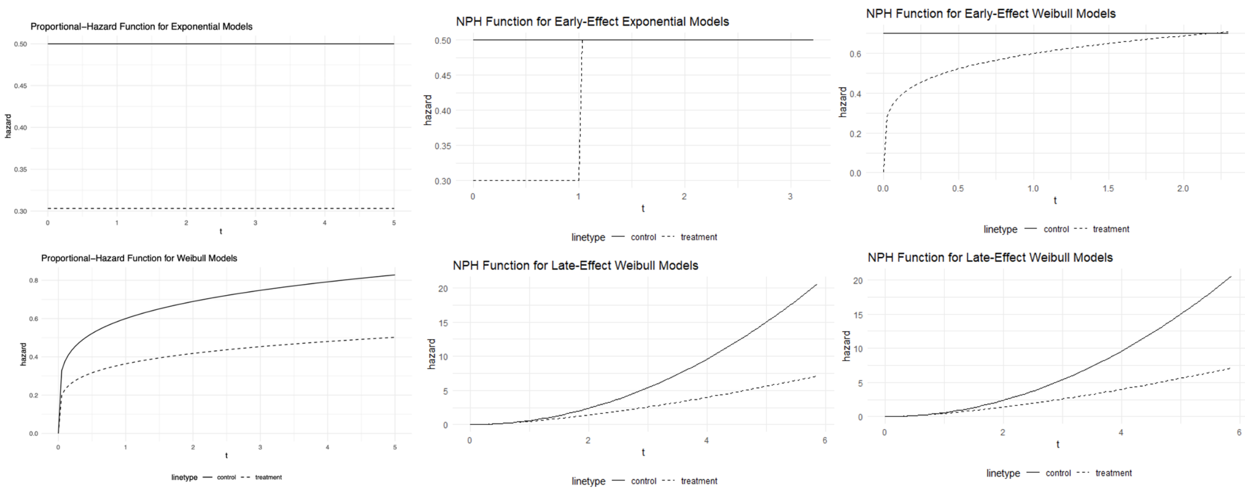
Thus, $T = \begin{cases} -\lambda_0^{-1} log(U) & U > exp(-\lambda_0) \\ \dfrac{\lambda_1 - log(U)}{\lambda_0 + \lambda_1} & U \leq exp(-\lambda_0) \end{cases}$ where $U \sim U(0,1)$

2.3.2 Weibull Model

To simplify the problem, we assume the control and treatment arm share the same scale parameter λ. For the control arm, suppose the hazard function is: $h(t|x=0) = \lambda\gamma_0 t^{(\gamma_0 - 1)}$. Thus, $T = (-\lambda^{-1} log(U))^{1/\gamma_0}$.

Similarly, we can write the hazard function for the treatment arm as: $h(t|x=1) = \lambda\gamma_1 t^{(\gamma_1 - 1)}$. We can derive that $T = (-\lambda^{-1} log(U))^{1/\gamma_1}$.

2.4 Visualization



2.5 Simulation Setup

Based on the above formulas, we can simulate the value of T by setting U for random number generation, where $U \sim U(0,1)$. To ensure its generalizability, we plug in different parameters and evaluate the test performance respectively for the data generated under different value settings.

3 Results

3.1 Proportional-Hazard Assumption

In this scenario, two tables have been presented, illustrating the baseline function following either an exponential **(Table 1)** or Weibull distribution **(Table 2)**. The results reveals that for assessing the Proportional-Hazard Assumption, the overall log-rank test (test 1_specificity) exhibits superior specificity, indicating more robust results. Furthermore, an increase in sample size tends towards higher specificity rates.

Table 1: Specificity of 3 Log-Rank Tests based on PH Assumption

n	baseline	lambda	beta	test1_specificity	test2_specificity	test3_specificity
200	Exponential	1.0	-5.0	1.00	1.00	1.00
100	Exponential	1.0	-5.0	1.00	1.00	1.00
200	Exponential	0.8	-5.0	1.00	1.00	1.00
100	Exponential	0.8	-5.0	1.00	1.00	1.00
200	Exponential	0.5	-5.0	1.00	1.00	1.00
100	Exponential	0.5	-5.0	1.00	1.00	1.00
200	Exponential	1.0	-1.0	1.00	1.00	1.00
100	Exponential	1.0	-1.0	0.94	0.96	0.94
200	Exponential	0.8	-1.0	1.00	1.00	1.00
100	Exponential	0.8	-1.0	0.98	0.92	0.98
200	Exponential	0.5	-1.0	1.00	1.00	1.00
100	Exponential	0.5	-1.0	0.98	0.94	0.96
200	Exponential	1.0	-0.5	0.90	0.68	0.84
100	Exponential	1.0	-0.5	0.60	0.52	0.54
200	Exponential	0.8	-0.5	0.80	0.74	0.74
100	Exponential	0.8	-0.5	0.64	0.48	0.58
200	Exponential	0.5	-0.5	0.84	0.74	0.74
100	Exponential	0.5	-0.5	0.62	0.46	0.60

Table 2: Specificity of 3 Log-Rank Tests based on PH Assumption

n	baseline	lambda	gamma	beta	test1_specificity	test2_specificity	test3_specificity
200	Weibull	1.0	1.5	-5	1.00	1.00	1.00
100	Weibull	1.0	1.5	-5	1.00	1.00	1.00
200	Weibull	0.5	1.5	-5	1.00	1.00	1.00
100	Weibull	0.5	1.5	-5	1.00	1.00	1.00
200	Weibull	1.0	1.2	-5	1.00	1.00	1.00
100	Weibull	1.0	1.2	-5	1.00	1.00	1.00
200	Weibull	0.5	1.2	-5	1.00	1.00	1.00
100	Weibull	0.5	1.2	-5	1.00	1.00	1.00
200	Weibull	1.0	1.5	-1	1.00	1.00	1.00
100	Weibull	1.0	1.5	-1	0.98	0.96	0.94
200	Weibull	0.5	1.5	-1	1.00	1.00	1.00
100	Weibull	0.5	1.5	-1	1.00	0.94	1.00
200	Weibull	1.0	1.2	-1	1.00	1.00	1.00
100	Weibull	1.0	1.2	-1	0.98	0.96	0.98
200	Weibull	0.5	1.2	-1	1.00	0.98	1.00
100	Weibull	0.5	1.2	-1	1.00	0.98	1.00

3.2 Non-Proportional-Hazard Assumption

In the first scenario of addressing the Non-Proportional-Hazard Assumption, we assume a stepwise exponential distribution. The subsequent two tables demonstrate that employing the corresponding weighted log-rank test for late and early effects does indeed yield higher specificity rates. The weighted log-rank test for late effects (test 2_specificity) in **Table 3** demonstrates superior performance with increasing hazard ratios between the control and treatment groups. Likewise, the weighted log-rank test for early effects (test 3_specificity) in **Table 4** exhibits better efficacy with decreasing hazard ratios between the control and treatment groups. In addition, similar to the previous table, there is a tendency toward higher specificity rates as the sample size increases.

Table 3: Specificity of 3 Log-Rank Tests based on NPH Assumption (Late)

n	lambda0	lambda1	test1_specificity	test2_specificity	test3_specificity
200	0.8	0.4	0.10	0.28	0.04
100	0.8	0.4	0.14	0.28	0.06
200	0.5	0.4	0.86	1.00	0.34
100	0.5	0.4	0.58	0.80	0.18
200	0.8	0.3	0.14	0.18	0.06
100	0.8	0.3	0.10	0.08	0.08
200	0.5	0.3	0.58	0.82	0.20
100	0.5	0.3	0.38	0.60	0.18

Table 4: Specificity of 3 Log-Rank Tests based on NPH Assumption (Early)

n	lambda0	lambda1	test1_specificity	test2_specificity	test3_specificity
200	0.9	0.7	0.32	0.24	0.90
100	0.9	0.7	0.16	0.14	0.70
200	0.8	0.7	0.22	0.26	0.88
100	0.8	0.7	0.12	0.10	0.66
200	0.9	0.6	0.30	0.34	0.80
100	0.9	0.6	0.10	0.16	0.58
200	0.8	0.6	0.10	0.40	0.62
100	0.8	0.6	0.08	0.16	0.32

In the second scenario of dealing with the Non-Proportional Hazard Assumption, we involve the Weibull distribution and consider two scenarios to generate late effect and early effect models respectively. For the late effect model, we assume that the survival time for the treatment group and the control group both follow Weibull distribution with the same scale parameter λ and different shape parameters γ. When γ is greater than 2 and the control arm has a higher γ, the hazard for the control arm increases faster, resulting in a late effect for the treatment. Thus, the weighted log-rank test for late effects (test 2_specificity) shows the best performance in **Table 5**, consistent with our intuitive findings. For the early effect model, we use exponential distribution for control arm and Weibull distribution for treatment arm, resulting in an early effect as the hazard for the treatment arm is lower at the beginning. The weighted log-rank test for early effects (test 3_specificity) shows the best performance, as shown in **Table 6**. Furthermore, increasing the sample size further improves the specificity.

Table 5: Specificity of 3 Log-Rank Tests based on NPH Assumption

n	lambda	gamma0	gamma1	test1_specificity	test2_specificity	test3_specificity
200	0.4	5	3.0	1.00	1.00	0.84
100	0.4	5	3.0	0.94	1.00	0.44
200	0.4	4	3.0	0.86	1.00	0.32
100	0.4	4	3.0	0.68	0.96	0.18
200	0.4	5	2.5	1.00	1.00	0.86
100	0.4	5	2.5	1.00	1.00	0.64
200	0.4	4	2.5	1.00	1.00	0.78
100	0.4	4	2.5	0.96	1.00	0.50

Table 6: Specificity of 3 Log-Rank Tests based on NPH Assumption

n	lambda0	lambda1	gamma	test1_specificity	test2_specificity	test3_specificity
200	1.0	0.5	1.2	1.00	0.98	1.00
100	1.0	0.5	1.2	0.98	0.82	1.00
200	0.8	0.5	1.2	0.86	0.54	0.94
100	0.8	0.5	1.2	0.66	0.28	0.76
200	1.0	0.5	1.1	1.00	1.00	1.00
100	1.0	0.5	1.1	0.98	0.86	0.94
200	0.8	0.5	1.1	0.96	0.86	0.96
100	0.8	0.5	1.1	0.74	0.66	0.70

4 Conclusion and Discussion

In conclusion, our study explored the performance of three hypothesis tests for time-to-event data analysis: the regular log-rank test and two variations of the weighted log-rank test addressing late and early effects, under both proportional and non-proportional hazard assumptions.

Under the Proportional Hazard Assumption, we found that the overall log-rank test consistently demonstrated superior specificity compared to the weighted log-rank tests.

And under the Non-Proportional Hazard Assumption, our findings revealed consistent results. In the first scenario with a piece-wise exponential distribution, employing corresponding weighted log-rank tests for late and early effects resulted in higher specificity rates. In the second scenario, where we assume Weibull distribution for the treatment arm, the different tests yield results correspondingly for late effect and early effect models, which is aligned with the previous result. This highlights the importance of selecting appropriate statistical tests in problem solving.

In addition, the trend of increasing specificity with larger sample sizes persisted, emphasizing the critical role of adequate sample sizes in ensuring reliable statistical inference.

Besides, when considering the measure of log-rank test performance, there could be other instrument, for example, the test power. However, after practice in software, we found that it is hard to be obtained by the test result object directly, when applying the log-rank test by calling the test function. The calculation of power through other ways and other measure of test performance are expected to be explored.

Overall, this study offers valuable insights into choosing suitable statistical methodologies for time-to-event data analysis in clinical trials, with careful choice of models using Monte Carlo simulation. These insights contribute to the continuous improvement of statistical practices in clinical research, ultimately enhancing the reliability and interpretability of trial results.

5 References

Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. Statistics in Medicine, 31(29), 3946–3958. https://doi.org/10.1002/sim.5452

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine, 24(11), 1713–1723. https://doi.org/10.1002/sim.2059

Bardo, M. F., Huber, C., Benda, N., Brugger, J., Fellinger, T., Vaidotas Galaune, Heinz, J., Heinzl, H., Hooker, A. C., Florian Klinglmüller, Franz König, Mathes, T., Mittlböck, M., Posch, M., Ristl, R., & Friede, T. (2023). Methods for non-proportional hazards in clinical trials: A systematic review. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2306.16858