MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Mini Review

Open Access

CrossMark

A review of statistical methods on testing time-to-event data

Abstract

The proportional hazard (PH) is commonly assumed for claiming efficacy and planning sample size in randomized clinical trials with time-to-event (TTE) type of endpoints. It is well known that the log-rank test is the most powerful testing method when the PH assumption holds. In recent years, with the advancement of immuno-oncology therapies, the non-PH scenarios, such as the delayed treatment effect and the diminished treatment effect, are frequently observed. A variety of alternative methods have been proposed for testing the time-to-event data while there is no uniformly most powerful method under the non-PH setting. In this paper, six popularly used methods for testing the TTE data are reviewed followed by a numerical comparison.

Keywords: time-to-event, proportional hazard, log-rank test, restricted mean survival time

Volume 7 Issue 6 - 2018

Tu Xu,[1] Danting Zhu[2]
[1]Agios Pharmaceuticals, Cambridge, USA
[2]Department of Biostatistics, University of Michigan, Ann Arbor, USA

Correspondence: Tu Xu, Agios Pharmaceuticals, Cambridge, MA, 02139, USA, Email xutu1116@gmail.com

Received: November 30, 2018 | **Published:** December 19, 2018

Introduction

In oncology randomized clinical trials, the time-to-event(TTE) type of endpoints such as progression-free survival (PFS) and overall survival(OS), are commonly used as the primary or key secondary endpoints for comparing the experimental treatment and active control/ placebo. In practice, the proportional hazard (PH) is usually assumed to characterize the treatment benefit over time of TTE endpoints and calculate the required sample size. With the PH assumption, the hazard ratio (HR) between treatment arms is a constant over time, and the corresponding testing hypothesis is expressed as

$$H_0 : \mathrm{HR}(t) = 1, H_1 : \mathrm{HR}(t) = c,$$

where $c < 1$. Recently, it is observed that the PH assumption does not hold in clinical trials investigating the cancer immunotherapies. Instead, a variety of non-PH pattens, such as the delayed treatment effect and the diminished treatment effect (e.g. Nivolumab)[1,2] were observed.

It is well known that the log-rank test is most powerful when the PH assumption holds. However, when the PH assumption is violated, the log-rank test may yield suboptimal power when compared with other testing procedures.[3,4] Therefore, extensive research has been conducted to explore alternative methods for handling the non-PH scenarios. In literature, multiple alternative methods have been proposed. In this paper, we provide a review of six popularly used testing methods including the log-rank test,[5] the weighted log-rank test,[6] the optimal weighted log-rank test,[7] the restricted mean survival time,[8] the weighted Kaplan-Meier,[9] and minimum P-value.[10,11]

Statistical methods for testing time-to-event data

Log-rank test (LR)

The log-rank test[5] is the most popularly used methods for comparing two distributions of the TTE data. Its popularity is largely due to the fact that log-rank test is the most powerful test with PH assumption as the test statistic is essentially equivalent to the score test.[5,12] Also the sample size planning based on the asymptotic property of log-rank test is convenient. However, researchers have pointed out that the log-rank test may yield suboptimal testing power in many non-PH cases.[4]

Weighted log-rank test (WLR)

The weighted log-rank test is a natural extension of the log-rank test in order to provide a more flexible test statistic for handling the non-PH cases. The log-rank test becomes a special case of weighted log-rank test if a constant weight function over time is adopted. The Fleming-Harrington family $L^{p,q}$ [6] is the most popular approach to specify weight functions as it consists a variety of weight functions to tackle different types of non-PH cases. The weighted log-rank test can yield robust power if a proper weight function is applied. Otherwise, the test statistics may cause a severe power loss, such as applying $L^{1,0}$ for the scenarios with delayed treatment effect.

Optimal weighted log-rank test (OptLR)

Recently, Liu et al.,[7] further explore the optimal weight for the weighted log-rank test. The optimal weight refers to the weight function that maximizes the testing power within the weighted log-rank test family. It is showed that with mild asymptotic assumptions, the optimal weight is expressed as

$$w_{opt} = \frac{\lambda_1(t) - \lambda_0(t)}{\lambda(t)},$$

where $\lambda_0(t), \lambda_1(t), \lambda(t)$ are the respective hazard functions of control arm, treatment arm, and overall population. When PH assumption holds, it is straightforward that the optimal weight becomes a constant and thus log-rank test is most powerful. The key challenge for implementing the OptLR is regarding the estimation of hazard functions over time. The simulation results in Lin et al.,[10] show that the optimal weighted log-rank test yields robust performance when delayed treatment effect is present. It is not surprising as the close form of optimal weight function can be derived in this setting. However, for other non-PH cases, further research needs to be conducted.

Restricted mean survival time (RMST)

In contrast to LR, WLR, and OptLR, which essentially estimate the difference of hazard functions, the RMST[9] directly estimates the difference D of two survival functions over a follow-up period $(0, L)$. That is,

$$D = \int_0^L \{S_1(t) - S_0(t)\}dt,$$

where $S_0(t)$ and $S_1(t)$ are survival functions of treatment arm and control arm, respectively. The use of RMST is favored by physicians as its estimate D has a clear clinical interpretation. For example, if $D = 3$ month with $L = 2$ years, then its clinical interpretation is that subjects receiving treatment is expected to survive 3 months longer during the first 2 year of treatment. In the setting of non-PH, Huang & Kuan[4] show by the numerical simulation that the RMST yields more robust power performance than the log-rank test if parameter L is properly chosen. Furthermore, Tian et al.,[13] show the power advantage of RMST over LR in certain non-PH scenarios theoretically.

Weighted Kaplan-Meier (WKM)

The WKM (Pepe & Fleming[9]) is a extension of the RMST by imposing a weight function on the RMST test statistic. The WKM is motivated by the fact that the estimation of survival function $S(t)$ becomes unstable when t gets close to L due to the high censoring and event rate. The WKM test statistic is expressed as

$$D_{WKM} = \int_0^L w(t)\{S_1(t) - S_0(t)\}du,$$

where the weight function is recommended to be a function of non-censoring rate $C_1(t)$ and $C_0(t)$,[9] such as

$$w(t) = \frac{C_1(t)C_0(t)}{C_1(t) + C_0(t)}.$$

That is to say, more weight is given to the survival estimates where censoring rate is low. It is shown that the WKM works better than LR and RMST when the overall censoring is heavy based on the simulation.[9]

Minimum P-value (MinP)

As there are no uniformly most powerful test for testing the non-PH data, a versatile testing method MinP is proposed by Karrison.[11] The test statistic of MinP is expressed as

$$Z_{MinP} = \max(|Z_1|,\ldots,|Z_m|),$$

where $Z_i, i = 1,\ldots,m$ could be any type of standardized test statistics. Clearly, the Z_{MinP} aims to include different types test statistics, and therefore becomes sensitive to a variety of non-PH scenarios. For example,

$$Z_{MinP} = \max(|Z_{G_{1,0}}|,|Z_{G_{0,0}}|,|Z_{G_{0,1}}|),$$

where $Z_{G_{\cdot,\cdot}}$ are WLR test statistic from Fleming-Harrington family. Intuitively, the power of Z_{MinP} should be robust in the scenarios of diminished treatment effect, PH, and delayed treatment effect scenarios, which has been demonstrated in Karrison.[11] However, the challenge for implementation MinP is regarding the estimation of covarance matrix of $(Z_1,\ldots,Z_m)$ if the test statistics are not WLR within Fleming-Harrington family. In literature, Lin et al.,[10] proposed a permutation method while it is quite computational intensive.

Simulations

To evaluate the numerical performance of the aforementioned methods, we perform numerical evaluation by simulation in different settings of PH and non-PH. Due to the nature of this review paper, we only include the following examples for illustration purpose. There are five scenarios considered as shown in Figure 1: 1. the PH; 2. the delayed treatment effect with a high cure rate; 3. the delayed treatment; 4. the diminished treatment effect 5. the diminished treatment effect with crossed survival curves. For each scenario, the total sample size is 300 with analysis performed when 200 events are observed. For simplicity, no censoring is considered in the simulation examples and therefore WKM performs the same as RMST. Each simulation is replicated 5,000 times. The details of simulation scenarios are described as follows, and the simulation performance of LR, WLR ($L^{(0,1)}, L^{(1,0)}$), OptLR, RMST, MinP are summarized in Table 1.

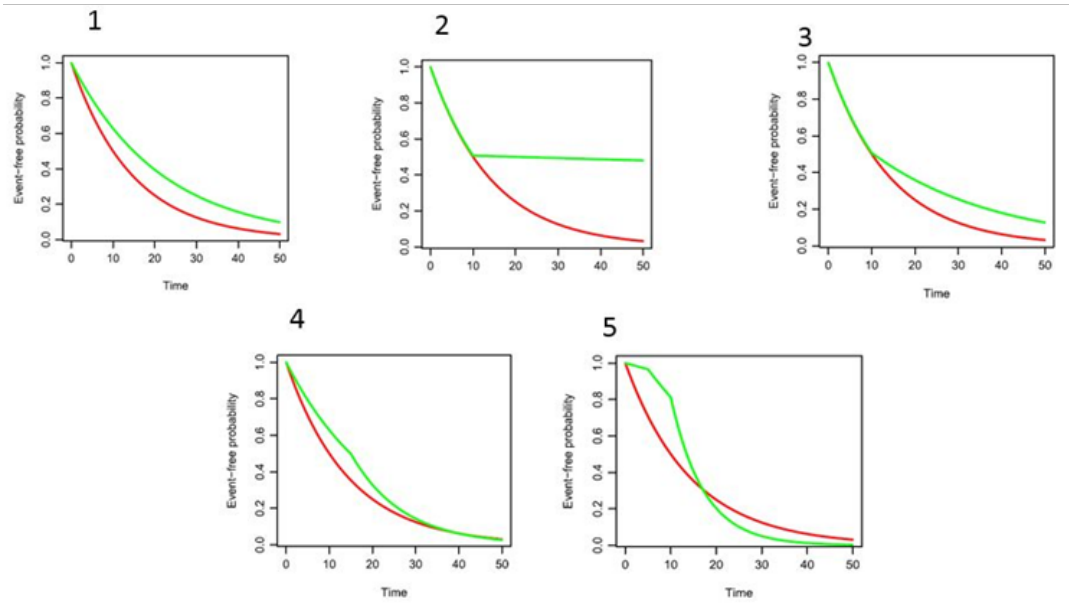

Figure 1 Simulation examples.

Table 1 The Simulation results (Power/Type 1 error)

Method	Scenario 0	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
LR	0.053	0.791	0.94	0.262	0.486	0.492
$L^{(0,1)}$	0.049	0.662	1	0.494	0.148	0.414
$L^{(1,0)}$	0.051	0.741	0.62	0.132	0.585	0.973
OptLR	0.051	0.79	1	0.674	0.745	1
RMST	0.053	0.795	0.968	0.305	0.458	0.589
MinP	0.051	0.765	1	0.42	0.516	0.998

The control arm of each scenario follows exponential distribution with median survival 10 months. The Scenario 0 evaluates the performance of Type 1 error by setting the survival distribution of treatment arm in the same way.

a) For the Scenario 1, the treatment arm follows exponential distribution with hazard ratio (HR) 0.67.

b) For the Scenario 2, the treatment arm follows piecewise exponential distribution with $HR_1 = 1, t < 10; HR_2 = 0.02$, otherwise.

c) For the Scenario 3, the treatment arm follows piecewise exponential distribution with $HR_1 = 1, t < 10; HR_2 = 0.5$, otherwise.

d) For the Scenario 4, the treatment arm follows piecewise exponential distribution with $HR_1 = 0.67, t < 15; HR_2 = 1.2$, otherwise.

e) For the Scenario 5, the treatment arm follows piecewise exponential distribution with $HR_1 = 0.1, t < 5; HR_2 = 0.5, 5 \le t < 10; HR_3 = 2$ otherwise.

From Table 1, it is evident that all the methods preserve the Type 1 error strictly. The LR yields robust performance when the PH holds while a significant power loss is observed in multiple non-PH settings. The performance of WLR depends on the selection of weight functions, which is difficult to be properly pre-specified. The RMST performs similarly as LR, and its estimate has straightforward clinical interpretation. The OptLR and MinP yield strong performance in all scenarios. It worths noting that the good performance of OptLR could be due to the fact that the optimal weights in the simulation examples are easy to estimate.

Discussion

As illustrated in Section 2, a rich set of methods has been available for testing TTE data. However, there is no uniformly most powerful method in the non-PH setting. Therefore, it is important to conduct intensive numerical evaluation for understanding the operating characterics of different methods before decision making. Otherwise, a significant power loss could be caused due to the selection of testing method. Furthermore, the numerical evaluation needs to carefully designed via reviewing the historical data and the mechanism of action of investigational drug. For instance, the numerical evaluation for immuno-oncology compounds may focus on the delayed effect scenarios. For the further work, the theoretical development on comparing the efficacy of existing methods as Tian at al.,[13] and the software development to facilitate the use of existing methods, such as R Shiny, could be interesting.

Acknowledgments

None.

Conflicts of interest

Authors declare that there is no conflicts of interest.

References

1. Ferris R, George B, Blumenschein, et al. Nivolumab for recurrent squamous–cell carcinoma of the head and neck. *The New England Journal of Medicine*. 2016;375:1856–1867.
2. Carbone D, Martin R, Luis P, et al. First–line nivolumab in stage IV or recurrent non–small–cell lung cancer. *The New England Journal of Medicine*. 2017;376:2415–2426.
3. Xu Z, Zhen B, Park Y, et al. Designing therapeutic cancer vaccine trials with delayed treatment effect. *Statistics in Medicine*. 2017;36:592–605.
4. Huang B, Kuan P. Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time–to–event endpoint. *Pharmaceutical Statistics*. 2018;17(3):202–213.
5. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistics Society, Series A*. 1972;135:185–207.
6. Harrington D, Fleming T. A class of rank test procedures for censored survival data. *Biometrika*. 1982;69:553–566.
7. Liu S, Chu C, Rong A. Weighted log–rank test for time–to–event data in immunotherapy trials with random delayed treatment effect and cure rate. *Pharmaceutical Statistics*. 2018;17(5):541–554.
8. Karrison T. Restricted mean life with adjustment for covariate. *Journal of the American Statistical Association*. 1987;82:1169–1175.
9. Pepe M, Fleming T. Weighted Kaplan–Meier statistics:a class of distance tests for censored survival data. *Biometrics*. 1989;45:497–507.
10. Lin Y, Zhou K, Ganju J. A single test for rejecting the null hypothesis in subgroups and in the overall sample. *Journal of Biopharmaceutical Statistics*. 2017;27(1):101–110.
11. Karrison T. Versatile tests for comparing survival curves based on weighted log–rank statistics. *Stata Journal*. 2016;16:678–690.
12. Strawderman R. An asymptotic analysis of the logrank test. *Lifetime Data Analysis*. 1997;3(3):225–249.
13. Tian L, Fu H, Ruberg S, et al. Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics*. 2017;74:694–702.