Received 26 April 2016; | Revised 6 June 2016; | Accepted 6 June 2016

DOI: xxx/xxxx

ARTICLE TYPE

$Z_{\max}$ test for delayed effect in immuno-oncology clinical trials

Miao Yang[1] | Zhaowei Hua[2] | Lan Xue*[1] | Mingxiu Hu[3]

[1]Department of Statistics, Oregon State University, Corvallis, OR 97330, US
[2]Takeda Pharmaceutical Company Limited, Cambridge, MA 02139, US
[3]Nektar Therapeutics, San Francisco, CA 94158, US

Correspondence
*Lan Xue, Department of Statistics, Oregon State University, Corvallis, OR 97330, US
Email: xuel@science.oregonstate.edu

Summary

Delayed separation in survival curves has been observed in immuno-oncology clinical trials. Under this situation, the classic log-rank test may confront high power loss. In this paper, we consider a $Z_{\max}$ test, which is the maximum of the log-rank test and a Fleming-Harrington test. Simulation study indicates that the $Z_{\max}$ test controls the Type I error rate and maintains good power under all kinds of delayed effect models. The properties of the $Z_{\max}$ test are also proven in theory, which further supports its robustness. We apply the $Z_{\max}$ test to two data sets reported in recent immuno-oncology clinical trials, in which $Z_{\max}$ has exhibited remarkable improvement over the conventional log-rank test.

KEYWORDS:
$Z_{\max}$ test; log-rank test; weighted log-rank test; Fleming-Harrington; delayed effect; immuno-oncology; clinical trials

1 | INTRODUCTION

Cancer immunotherapy is currently boosting and dominating drug development in the oncology field. It has achieved unprecedented clinical benefits in treating life-threatening cancers such as melanoma and non-small cell lung cancer. These innovative therapies work by stimulating the immune system thereby imparting substantial benefits in tumor response and long-term survival (Hoos, 2012). However, the special mechanism results in a lag in the translation of immune and anti-tumor response into a survival benefit (Hoos, 2012). Consequently, the randomized clinical trials show delayed separation of the Kaplan-Meier survival curves (Chen, 2013). For example, the overall survival curves from CheckMate 141 trial targeting recurrent squamous-cell carcinoma of the head and neck demonstrate delayed separation around 4 months (Ferris et al., 2016).

The issue of delayed separation of the Kaplan-Meier survival curves presents unique challenges in using the standard log-rank test statistic for trial analysis. The conventional way to use log-rank test statistic to analyze time- to-event endpoints in randomized oncology clinical trials assumes proportional hazards between the two arms (Lachin and Foulkes, 1986). The log-rank test statistic is the most powerful test under proportional hazards model (Peto and Peto, 1972). However, proportional hazards assumption often does not hold and thus the log-rank test may not be as powerful when there is delayed effect. The use of log-rank test under delayed separation can cause power loss and increase the risk of trial failures.

There is a rich development in the literature to address design and analysis issues related to delayed treatment benefit in immune-oncology clinical trials. A common approach is to use weighted log-rank test statistic with appropriately pre-specified weights or weight function to allocate more weight to late events to maximize power under the alternative of delayed separation. For example, Self et al. (1988) pre-specified linear weight using weighted log-rank test statistics to incorporate increasing risk of breast cancer for a health trial. The more general G family with weights of the form $G^{r_1,r_2} = \left\{\hat{S}(t-)\right\}^{r_1}\left\{1-\hat{S}(t-)\right\}^{r_2}$ specifies the parameters, i.e. $r_1 = 0$ and $r_2 > 0$, to have the test more sensitive to delayed separation (Fleming and Harrington, 1991; Hasegawa, 2014). Unfortunately, mis-specified weights or weight function can lead to decreased sensitivity to the actual observed

[0]**Abbreviations:** LR: the log-rank test; FH: the Fleming-Harrington test with weight $1-\hat{S}(t)$; WLR: the weighted log-rank test; PH: proportional hazard

treatment effect. Xu et al. (2017) proposed a piecewise weighted log-rank test with weights proportional to log hazard ratio of treatment versus control to optimize power if time point for separation can be pre-specified correctly. Hence mis-specifying time point for change can result in less-than-optimal power. Another approach is to consider combinations of weighted log-rank tests. For example, Zucker and Lakatos (1990) proposed to use a linear combination of weighted log-rank tests or a combination of maximun efficiency to account for a broad range of lags time functions. A similar idea of taking the maximum of a collection of weighted log-rank tests was considered by Fleming and Harrington (1991), Lee (1996), Lee (2007), and Karrison (2016) for a selection of alternatives of interest. The most recent FDA workshop on non-proportional hazards (2018) considers a maximum test of $G^{0,0}$, $G^{0,1}$, $G^{1,0}$, and $G^{1,1}$ and finds around $3-4\%$ power loss compared with the optimal test under proportional hazards model and survival models with diminishing effects. A recent idea was proposed by Sit et al. (2016) to use an intersection-union test to handle delayed effect with a non-inferiority log-rank test for the period prior to the pre-specified lag time τ and a superiority log-rank test for the period after the lag time τ. This method is sensitive to the choices of the non-inferiority margin and the time change point.

Even though delayed separation in survival curves commonly exists in the immuno-oncology clinical trials, some survival curves do not show delayed separation. For example, the overall survival curves from Checkmate 025 trial targeting pre-treated renal-cell carcinoma did not show delayed separation (Motzer et al., 2015). Using weighted log-rank test with mis-specified weight function to allocate more weight on late events under the alternative of no delayed separation could lead to power loss as well. A robust test statistic is needed to account for both alternatives of no delayed separation and delayed separation.

This article considers a combination test $Z_{\max}$ to handle both alternatives of no delayed separation and delayed separation. This combination test $Z_{\max}$ takes the maximum of the standard log-rank test and the weighted log-rank test of weight function $1-\hat{S}(t-)$. It favors the standard log-rank test under the alternative of no delayed separation and favors the weighted log-rank test under the alternative of delayed separation. Therefore, this $Z_{\max}$ test is robust to provide satisfying power under the alternative of proportional hazards and the alternative of delayed separation. Theoretical work proves power gain for the $Z_{\max}$ test within the framework of local asymptotics, assuming logarithm of hazard ratio decreases with sample size at the rate of $n^{-1/2}$: (1) $Z_{\max}$ test is more powerful than the log-rank test under delayed separation; (2) $Z_{\max}$ test is more powerful than the weighted log-rank test under proportional hazards; (3) Power gaining of the $Z_{\max}$ test versus the log-rank test under delayed separation decreases when sample size increases. Simulation studies were performed to show that power loss for $Z_{\max}$ is small compared with the log-rank test under proportional hazards or compared with the weighted log-rank test under delayed separation. The asymptotic distribution of the $Z_{\max}$ test was derived based on Theorem 7.5.1 of Fleming and Harrington (1991). Hence, we can conveniently use $Z_{\max}$ for clinical trial design and analysis. A computational R package was developed as well to determine the sample size and power for clinical trial design.

This paper is constructed as follows. Section 2 describes the $Z_{\max}$ test and derives its theoretical properties. Performance of $Z_{\max}$ is illustrated in Section 3 via simulation studies in terms of type I error, power, sample size, and follow-up time. Section 4 shows the results of applying the $Z_{\max}$ test to two real examples. Section 5 describes estimation in a delayed effect model. Section 6 concludes the paper with discussions.

2 | METHODS

2.1 | Preliminaries

Let the data be generated from the standard two-sample random censoring model with a total of n individuals randomly allocated to either the control or the treatment group. Denote the survival functions for the control and treatment groups as $S_0(t)$ and $S_1(t)$ respectively. In this paper, we are interested in comparing the survival curves between the two groups and testing the hypotheses that

$$H_0 : S_0(t) = S_1(t) \text{ for all } t \text{ versus } H_a : S_0(t) \neq S_1(t) \text{ for some } t.$$

Let $\{T_i, \delta_i, X_i\}_{i=1}^n$ be an independent sample of right-censored survival data from two groups, where T_i is a possibly right censored event time; δ_i is the censoring indicator with $\delta_i = 1$ if T_i is an event time, and $\delta = 0$ if T_i is censored; and X_i is the group indicator that takes value 1 if the individual belongs to the treatment group, and 0 otherwise. The numbers of individuals in the control and treatment groups are denoted as n_0 and n_1 respectively with $\sum_{i=1}^n X_i = n_1$. In addition, let D_n be the event set which contains the indices of individuals in the pooled sample who have had an event, and t_j denotes the observed event time of individual j in D_n. For a given time t, let $n_k(t)$ be the number of individuals in the risk set of group k ($k = 0, 1$) and $p(t) = n_1(t)/\{n_1(t) + n_0(t)\}$ be the fraction of individuals from the treatment group.

For the above hypothesis testing problem, a family of weighted log-rank tests (WLR) has been proposed

$$U_{w,n} = \frac{\sum_{j\in D_n} w_{n,j}\left\{X_j - p\left(t_j\right)\right\}}{\sqrt{\sum_{j\in D_n} w_{n,j}^2 p\left(t_j\right)\left\{1-p\left(t_j\right)\right\}}} \tag{1}$$

where $w_{n,j}$ is a predefined weight at time t_j. Different choices of weights correspond to different test statistics. The well-known test statistics, such as, log-rank (Mantel 1966; Cox 1972), Gehan-Breslow (Gehan 1965), Tarone-Ware (Tarone and Ware 1977), Peto-Peto (Peto and Peto 1972), and Fleming-Harrington tests (Fleming and Harrington 1982) all belong to this family. By Proposition A.1 in the appendix, the test statistic $U_{w,n}$ asymptotically follows a standard normal distribution under the null hypothesis. Therefore one rejects the null hypothesis when $\left|U_{w,n}\right| \geq z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the standard normal.

In this paper, we focus on two tests from this family. The first is the log-rank (LR) test with $w_{n,j} = 1$ for all j, and the other one is a Fleming-Harrington (FH) test with $w_{n,j} = 1 - \hat{S}\left(t_j-\right)$, where $\hat{S}$ is the Kalplan-Meier estimator using the pooled sample. We denote the LR test and the FH test as $U_{1,n}$ and $U_{F,n}$ respectively. The LR test assigns the same weight to all time points, while the FH test assigns more weight to later time points. Therefore, the FH test is more sensitive or powerful for late differences in survival curves.

In the traditional design of a randomized two-arm clinical trial, proportional hazard (PH) is often assumed. However, delayed separation of survival curves has been observed in many clinical trials of cancer immunotherapy. Therefore, for these trials, it is important to investigate the properties of these test statistics under both the PH and delayed treatment effect alternatives. Let $\lambda_k(t)$ be the hazard function for group k, $k = 0, 1$. Then the PH alternative is of the form $H_a^{\mathrm{PH}} : \lambda_1(t)/\lambda_0(t) = e^{\theta}$ for a constant $\theta \neq 0$. For the delayed alternative, $H_a^{\mathrm{Delay}} : \lambda_1(t)/\lambda_0(t) = 1 - \left(1-e^{\theta}\right) I_{(t\geq t_0)}$, where I is the indicator function and t_0 is a pre-determined separation time. Examples of survival curves under the PH and delayed alternatives are given in Figure 1 . Under both alternatives, the control arm (solid lines in the figures) follows an Exponential distribution with rate 0.05. Under the PH alternative with hazard ratio of 0.8, the survival curve from treatment arm is always higher than that from the control arm; but for the delayed treatment model, the two survival curves are the same before the separation time $t_0 = 4$ months and the treatment group has a higher survival probability than the control group after t_0.

Insert Figure 1 here.

The theoretical properties of the weighted log-rank test given in Equation (1) have been well studied in the literature (Fleming and Harrington, 1991). In particular, Proposition A.2 in the appendix shows that WLR tests given in Equation (1) are consistent. That is, for any fixed alternative, the power of the WLR test go to one as $n \to \infty$. Therefore, the limiting value of the power function cannot serve as a criterion to compare tests. In this paper, we compare the behavior of the LR and FH tests under the local asymptotics instead. Schoenfeld (1981) used the first-order approximation to study the power function of the WLR tests derived under the local proportional hazards alternatives. Peto and Peto (1972) claimed that log-rank test has greater local power than any other rank-invariant test procedure for detecting Lehmann-type differences between groups of independent observations subject to possible right-censoring. Peto (1972) showed that the log-rank test is the locally most powerful rank-invariant test in the absence of ties. Based on the score function statistics, Kalbfleisch (1978) showed the log-rank scores are the locally optimum under proportional hazards. Gill (1980) discussed the Pitman efficiency of the WLR test statistics.

Section 7a.7 of Rao (2001) introduced four different criteria to measure asymptotic efficiency of a test, including the first and second derivatives of the limiting power function. Different measures characterize different local behaviors of the power curve near the null hypothesis in large samples. Let $\Psi_{U_{w,n}}(\theta)$ be the power function of a test $U_{w,n}$ evaluated at parameter θ. In this paper, we consider $e(U_{w,n}) = \lim_{n\to\infty} \Psi_{U_{w,n}}\left(\theta_0 + \delta/\sqrt{n}\right)$ to describe the local asymptotics for different tests, where θ_0 gives the null hypothesized value of θ, and $\delta \neq 0$ is a given constant. If $e(U_{w_1,n}) > e(U_{w_2,n})$, then the power curve of $U_{w_1,n}$ is higher than that of $U_{w_2,n}$ in a local neighborhood of θ_0, and we call $U_{w_1,n}$ locally more efficient than $U_{w_2,n}$. In the following, we compare the local asymptotic properties of the LR and FH tests under both proportional hazard and delayed treatment effect alternatives.

We note that we are not the first ones to develop the properties for the LR and FH tests. The (locally) most optimal property of the log-rank test has been discussed before in Peto and Peto (1972), Peto (1972) and Kalbfleisch (1978). Related results regarding the power of FH test can also be found in Zucker and Lakatos (1990), Fine (2007) and Zhang and Quan (2009). We are summarizing the results and presenting them in a different way in Theorems 1 and 2.

Theorem 1. Suppose that the Assumption (A1) in the appendix holds. For testing $H_0 : \lambda_1(t) = \lambda_0(t)$ versus $H_{a,n}^{\mathrm{PH}} : \lambda_1(t)/\lambda_0(t) = e^{\theta_n}$ with $\theta_n = \delta/\sqrt{n}$ and $\delta < 0$, the log-rank test is locally more efficient than the Fleming-Harrington test in the neighborhood of H_0.

Under the PH alternatives, the treatment effect is assumed to be a constant over time. Therefore, all events should be treated evenly. The fact that LR test assigns the same weights to all time points makes it a more appropriate test to use under the PH alternatives. Theorem 1 shows that under the local proportional hazards alternatives, the LR test is asymptotically more efficient than the FH test in the neighbor of $\theta_0 = 0$.

Theorem 2. Suppose that the Assumption (A1) in the appendix holds. Consider testing $H_0 : \lambda_1(t) = \lambda_0(t)$ versus $H_{a,n}^{\text{Delay}} : \lambda_1(t)/\lambda_0(t) = 1 - \left(1 - e^{\theta_n}\right) I_{(t \geq t_0)}$,where $\theta_n = \delta/\sqrt{n}$ and $\delta < 0$. When the separation time t_0 is large enough, the Fleming-Harrington test is locally more efficient than the log-rank test in the neighborhood of H_0.

Theorem 2 compares the efficiencies of the LR and FH tests under the local delayed treatment effect alternatives, in which the treatment effect only exists after the separation time t_0. Under local delayed alternatives, the magnitudes of $e\left(U_{1,n}\right)$ and $e\left(U_{F,n}\right)$ are uniquely determined by the asymptotic means of the LR and FH test statistics. In the proof of Theorem 2, we show that the ratio of the asymptotic means of the LR and FH test statistics is a continuous and decreasing function of separation time t_0, under assumption (A1). In addition, when t_0 goes to zero and the delayed alternative becomes the PH, the ratio of asymptotic means of the LR and FH tests is greater than one as proven in Theorem 1. Consequently, the LR test is locally more efficient than FH test when t_0 is small. When t_0 approaches the upper bound of the support of the survival time, the ratio gets smaller than one. This means that the FH enjoys higher local efficiency than the LR test when t_0 is large.

An example of how the powers of the LR and FH change with separation time t_0 is given in Figure 2 . We select a sequence of separation time t_0 ranging from 0 to 4 months by 0.2 months. For a fixed t_0, we use the same model in Section 3 to generate 10, 000 data sets to approximate the powers of the LR and FH tests. In particular, We set the number of events to be 100. In the figure, the LR possesses higher power than the FH when t_0 is small. The powers of both LR and FH decrease as t_0 increases, but the LR decreases faster than the FH. After a certain separation time point (around 2 months in the figure), the power curve of the FH is higher than that of the LR, which is consistent with the results in Theorem 2.

Insert Figure 2 here.

This property makes it challenging to pick a better test between the LR and FH under delayed alternatives since the separation time is generally unknown in advance in practice. Therefore it is necessary to develop a test which is more robust to the specification of the separation time of the delayed treatment effect, and has better overall performance under a wide range of alternatives compared with the individual LR or FH test.

2.2 | $Z_{\max}$ Test

In Subsection 2.1, we compared the local efficiencies of the LR and FH tests under both the PH and delayed treatment effect alternatives. The LR test was found to be locally more efficient under the PH alternative or when the separation time is close to the time origin. On the other hand, the FH test is locally more efficient when the separation time is large enough. Therefore, to combine the strength of both LR and FH tests, we consider a new test statistic $Z_{\max} = \max\left\{\left|U_{1,n}\right|, \left|U_{F,n}\right|\right\}$. A larger value of $\left|U_{1,n}\right|$ or $\left|U_{F,n}\right|$ indicates stronger evidence against the null hypothesis. The test statistic $Z_{\max}$ combines the evidence in both the LR and PH test statistics and maintains the power of the better individual test of the two under both the PH and delayed treatment effect alternatives. We shall reject the null hypothesis when $Z_{\max}$ is large. The critical value of $Z_{\max}$ can be obtained using the following discussion.

By Theorem 7.5.1 in Fleming and Harrington (1991), under $H_0 : \lambda_0(t) = \lambda_1(t)$, the LR and FH jointly follow a bivariate normal distribution asymptotically with

$$\begin{pmatrix} U_{1,n} \\ U_{F,n} \end{pmatrix} \xrightarrow{d} N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right\}, \tag{2}$$

where F_{Z_1,Z_2} is the joint distribution function of $\left(Z_1, Z_2\right)$. Therefore for a given α, we can use the quantile function for bivariate normal distribution to find the critical value $c_{\alpha,\rho}$. On the other hand, for a given $Z_{\max}$ statistic, we can also use the distribution

Therefore, the test statistic $Z_{\max}$ asymptotically follows the same distribution as $\max\left(\left|Z_1\right|, \left|Z_2\right|\right)$, where $\left(Z_1, Z_2\right)^\top$ is bivariate Normal with zero mean, unit variance and $\text{corr}\left(Z_1, Z_2\right) = \rho$. Let $c_{\alpha,\rho}$ be a critical value such that $P\left\{\max\left(\left|Z_1\right|, \left|Z_2\right|\right) \geq c_{\alpha,\rho}\right\} = \alpha$. To obtain an asymptotically level α test, we reject the null hypothesis when $Z_{\max} \geq c_{\alpha,\rho}$. According to the definition of $c_{\alpha,\rho}$, one has

$$\begin{aligned} 1 - \alpha &= P\left\{\max\left(\left|Z_1\right|, \left|Z_2\right|\right) < c_{\alpha,\rho}\right\} = P\left(\left|Z_1\right| < c_{\alpha,\rho}, \left|Z_2\right| < c_{\alpha,\rho}\right) \\ &= F_{Z_1,Z_2}\left(c_{\alpha,\rho}, c_{\alpha,\rho}\right) - F_{Z_1,Z_2}\left(c_{\alpha,\rho}, -c_{\alpha,\rho}\right) - F_{Z_1,Z_2}\left(-c_{\alpha,\rho}, c_{\alpha,\rho}\right) + F_{Z_1,Z_2}\left(-c_{\alpha,\rho}, -c_{\alpha,\rho}\right), \end{aligned}$$

where F_{Z_1,Z_2} is the joint distribution function for (Z_1, Z_2). Therefore we can use the quantile function for bivariate normal distribution to calculate $c_{\alpha,\rho}$. On the other hand, for a given $Z_{\max}$ statistic, we can also use the distribution function of the bivariate normal to compute the p-value.

An illustration of $Z_{\max}$ is given in Figure 3 . It displays the plot of the calculated values of LR test statistic against FH test statistic using 500 data sets simulated from the same model as described in the simulation study in Section 3. In particular, the number of events is set to be 200 and we evaluate the two test statistics under both null hypothesis and two different alternatives: PH and delayed treatment effect with separation time $t_0 = 4$. For each panel, the highlighted square denotes the acceptance region of $Z_{\max}$ at level $\alpha = 0.05$. In the first panel, most points are within the square, confirming low rejection rate of $Z_{\max}$ under the null hypothesis. On the other hand, in the second and third panels, most points move out of the square, indicating high rejection rate under either the PH or delayed alternatives. In addition, under the PH alternative, more points are above the dotted line in panel 2, indicating that the LR test tends to have larger absolute value and is thus more powerful than the FH test. In this figure, we also highlight the cases with the null hypothesis erroneously accepted by $Z_{\max}$ (points in the square) and the FH tests (points shaped with "A"). It clearly shows that the FH has a much higher acceptance rate and is thus less powerful than $Z_{\max}$ under the PH alternative. Similarly, under delayed treatment effect, the acceptance rate of the LR test is larger and thus it is less powerful than $Z_{\max}$.

Insert Figure 3 here.

The following theorems compare the local efficiency of the proposed $Z_{\max}$ test under both PH and delayed alternatives, The detailed proofs are given in the appendix.

Theorem 3. Suppose the Assumptions (A1) and (A3) in the appendix hold. For testing $H_0 : \lambda_1(t) = \lambda_0(t)$ versus $H_{a,n}^{\mathrm{PH}} : \lambda_1(t)/\lambda_0(t) = e^{\theta_n}$, where $\theta_n = \delta/\sqrt{n}$ and $\delta < 0$, the $Z_{\max}$ test is locally more efficient than the Fleming-Harrington test in the neighborhood of H_0.

Theorem 3 shows under PH, the $Z_{\max}$ test is more efficient than the FH. In the second panel of Figure 3 , there are more points in the rejection region of $Z_{\max}$ (points outside the square) than that of the FH (points labeled "R"), meaning that $Z_{\max}$ has higher rejection rates than the FH, which confirms Theorem 3.

Theorem 4. Suppose the Assumptions (A1) and (A4) in the appendix hold. For testing $H_0 : \lambda_1(t) = \lambda_0(t)$ versus $H_{a,n}^{\mathrm{Delay}} : \lambda_1(t)/\lambda_0(t) = 1 - \left(1 - e^{\theta_n}\right) I_{(t \geq t_0)}$ where $\theta_n = \delta/\sqrt{n}$ and $\delta < 0$, when t_0 is large enough, $Z_{\max}$ test is locally more efficient than the log-rank test in the neighborhood of H_0.

Theorem 4 states $Z_{\max}$ is locally more efficient than that of LR under delayed treatment effect. The third panel of Figure 3 shows that $Z_{\max}$ rejects more often than the LR test under this scenario and further validates Theorem 4.

3 | SIMULATION

In this section, different sets of simulation studies are performed to compare the performances of the $Z_{\max}$ test with the LR and FH tests. We first check the Type I error for the three tests of interest, and then compare their powers under different alternatives. We also compare the sample size or follow-up time required to achieve the same level of power for the three tests, as they are also key factors for clinical practitioners.

Unless specified, data are generated according to the same mechanism for all simulations in this section. In particular, patients are enrolled into the study with a constant rate of 30 subjects per month . In our simulation, the patients are subject to right censoring due to either dropping out or the termination of the study. For each individual, the dropping out time follows an Exponential distribution with rate 0.01 and is independent of the patient survival time. Follow-up time is either a pre-determined value or chosen randomly to achieve a fixed censoring rate. In addition, we assume equal randomization and fix Type I error α at 5%.

In Lee (2007) and Karrison (2016), the calculation of critical values for their maximum tests, such as $Z_{\max}$, involves the integration of multivariate normal distributions, which can be computationally slow. Yang et al. (2005) generated a table with the calculated critical values for $\alpha = 0.05$ with a sequence of ρ. Lee (2007) and Karrison (2016) applied the tables in Yang et al. (2005) to approximate the critical values. New tables also need to be generated when one changes the level α. In our codes, we write a function using the distribution function of bivariate normal distributions as described in Subsection 2.2. The function can directly and quickly compute $c_{\alpha,\rho}$ for every α and ρ.

3.1 | Type I Error

Three different distributions have been considered: Exponential with rate 0.05, piecewise Exponential with rate $\lambda(t) = 0.05I_{\{t<4\}} + 0.1I_{\{t\geq 4\}}$, and Weibull with rate 0.05 and shape 0.5. We consider different sample sizes with total number of patients $n = 84, 168, 334, 500, 668$ or 834. For each run, the follow-up time is determined to have approximately 40% censoring, or equivalently the number of events to be $50, 100, 200, 300, 400, 500$ respectively. For each scenario, we simulate $10,000$ data sets, and performed the three tests on each data set. Figure 4 summarized the empirical rejection rates for the three tests for different survival distribution and sample size combinations.

Insert Figure 4 here.

Figure 4 shows that for all the scenarios we have considered, the three tests always have empirical rejection rate close to $\alpha = 5\%$, with values between 4.5% and 5.5%. It shows that there is no Type I error inflation for any of the three tests under consideration.

3.2 | Power Comparison

Four types of alternatives are considered for power comparison: proportional hazards (PH) and three delayed treatment effect with separation time $t_0 = 3, 4$ or 5 months. Note that the PH alternative can also be viewed as a delayed treatment effect with $t_0 = 0$. For each separation time t_0, we consider the hazard functions for the control arm to be $\lambda_0^{t_0}(t) = 0.05I_{(t<t_0)} + 0.1I_{\{t\geq t_0\}}$, and for the treatment arm to be $\lambda_1^{t_0}(t) = 0.05I_{(t<t_0)} + 0.1e^{\theta}I_{\{t\geq t_0\}}$ respectively. The four alternatives are denoted as $H_a^0, H_a^3, H_a^4, H_a^5$ respectively. For all four scenarios, the hazard functions between the two arms are the same before the separation time t_0, and proportional with ratio e^{θ} after t_0.

We consider a wide range of number of events D from 100 to $1,000$ by 10 events. For a given D, we set the total sample size n to be $2\lceil D/1.2\rceil$, where $\lceil x\rceil$ is the smallest integer greater than x. For each alternative $H_a^{t_0}$ and event size D combination, a hazard ratio e^{θ} is chosen such that the LR test has approximately 80% power. Then the empirical powers of the FH and $Z_{\max}$ tests are compared against this benchmark. In this simulation study, the follow-up time for each replication is individually determined so that the censoring proportion is close to 40% for each replication. The LR, FH and $Z_{\max}$ tests are applied to each data set, and their rejection rates out of $10,000$ replications (empirical powers) have been recorded.

We also add another test, named "Delay" in our simulation. The test belongs to the family given by Equation (1) and we denote the test as $U_{t_0,n}$. The weights of $U_{t_0,n}$ are set to be zero before time t_0 and one afterwards. That is, in Equation (1), $w_{n,j} = I(t_j \geq t_0)$. According to Proposition A.3 in the appendix, $U_{t_0,n}$ is the locally most efficient test among the family given by Equation (1) under delayed alternatives.

Figure 5 plots the empirical power against the number of events for four testing methods under the four different alternatives. The first panel compares their performances under the PH alternative. It shows that the LR and Delay are identical and are the most powerful one. But the performances of $Z_{\max}$ and LR are close, and both are more powerful than FH. For example, under the PH alternative and the number of events $D = 100$, the empirical powers of $Z_{\max}$ and LR are 76.81% and 79.01% respectively, which are much higher than the power of the FH (66.23%). The remaining three panels of Figure 5 show that under alternatives of delayed effect, the LR is always the least powerful one, and the performance of $Z_{\max}$ is close to that of FH test. In addition, larger power gain of $Z_{\max}$ and FH tests is observed for later separation time. Compared to the Delay, which is the optimal test, the power loss of $Z_{\max}$ is around 5%. For example, when $t_0 = 5$ months and $D = 1,000$, The Delay test has 91.65% power, and the FH test has the second highest empirical power of 87.30%, followed by $Z_{\max}$ with 86.07%, but the LR has the lowest power of 79.87%. Although $Z_{\max}$ can not out-perform the FH test in each scenario, Figure 5 shows that the power loss of $Z_{\max}$ to the FH test is quite small regardless of the number of events. These findings are consistent with our theoretical results in Section 2.

Insert Figure 5 here.

3.3 | Sample Size

We perform the third simulation study to calculate the required sample size to achieve a power of 80% for different tests under various alternatives. In this subsection, we fix the follow-up time to be 10 months. For each of the four alternatives, we consider different hazard ratios e^{θ} in $\{0.55, 0.6, 0.65, 0.7, 0.75\}$ The binary search algorithm in Yang et al. (2018+) is used to find the required sample size.

Table 1 presents the sample size needed to achieve 80% power for different tests under various scenarios. The corresponding empirical powers are also displayed in the parenthesis in Table 1 , which are all very close to the target power 80%.

Table 1 here.

The first column of Table 1 shows under the PH alternative, the LR test always requires the smallest sample size, and a slightly larger number is required for $Z_{\max}$, but for the FH, a much greater number is needed to achieve the same power. For the delayed treatment effect alternative, FH test always requires the smallest sample size , followed by $Z_{\max}$, which is almost the same as the FH when separation time $t_0 = 3, 4$; the LR test requires much larger sample sizes than the other two.

3.4 | Follow-up Time

In this subsection, we still consider four types of alternatives, but with the hazard ratios e^{θ} only being 0.65 or 0.7. For each alternative and hazard ratio combination, we use the maximum sample size among the three tests presented in Table 1 to ensure each test can achieve 80% power. The algorithm is similar to the sample size determination algorithm described in Yang at al. (2018+).

Insert Table 2 here.

Table 2 show that with fixed sample size, the follow-up time for the $Z_{\max}$ test always is between that for the LR and FH, and is longer than the LR under the PH alternative, shorter than the LR under delayed treatment effect. Therefore with the same number of sample size, $Z_{\max}$ test always requires a reasonably efficient follow-up time among the three tests.

4 | APPLICATION

In this section, we illustrate the application of $Z_{\max}$ test via two examples: Nivolumab CheckMate 025 trial and Pembrolizumab Keynote 040 trial. The patient level data for the two trials are not original but reproduced based on the published survival curves using the method of Guyot et al. (2012). As shown in Figure 6, the original survival curve and the reproduced survival curve from digitized data are very similar.

4.1 | Digitized Data Based on Nivolumab CheckMate 025 Trial

Nivolumab is a prominent cancer immunotherapy recently approved for multiple un-curable cancer indications including renal-cell carcinoma, non-small cell lung cell, melanoma, etc. It is a programmed death 1 (PD-1) checkpoint inhibitor antibody which selectively blocks the interaction between PD-1 on activated T cells and PD-1 ligand 1 (PD-L1) and 2 (PD-L2) on immune cells and tumor cells to allow activated T cells to fight against the tumor cells (Motzer et al., 2015). The special mechanism of nivolumab to stimulate the immune system to attack cancer results in delayed clinical benefits (Hoos et al., 2010). The progression-free survival (PFS) curves in clinical trials exhibit severely delayed separation, which can lead to statistically insignificant results with the conventional log-rank test (Motzer et al., 2015, Ferris et al., 2016) although the treatment may provide significant clinical benefits. Therefore, PFS with the conventional log-rank test may not be an appropriate primary endpoint for Nivolumab trials (Kaufman et al., 2017).

CheckMate 025 trial was a randomized phase 3 trial to enable registration of nivolumab in pre-treated renal-cell carcinoma (Motzer et al., 2015). It compared nivolumab with everolimus in patients with renal-cell carcinoma who had received previous treatments. Due to the mechanism of nivolumab, it was expected that PFS is not an optimal primary endpoint and therefore overall survival (OS) was designated as the primary endpoint. Objective response rate (ORR) and PFS were two key secondary endpoints. A sequential testing procedure was applied. The original order was to test OS first, then PFS if OS was statistically significant, and finally ORR if PFS was statistically significant. Later the testing order was amended to test OS first, followed by ORR, and then by PFS. This amendment of the testing order was likely due to the concern of the delayed PFS separation, which may lead to a missed opportunity for ORR.

CheckMate 025 trial observed statistically significant OS results (p-value of 0.002), statistically significant ORR results (p-value < 0.001), and statistically non-significant PFS results (hazard ratio of 0.88, a p-value of 0.11) (Motzer et al., 2015). The PFS curves showed delayed separation around 7 months. We reproduced the PFS patient-level data by digitizing the published PFS curve using the method of Guyot et al. (2012). Figure 6 is the PFS curves reconstructed from the digitized dataset, which is almost identical to the original PFS curves. Applying the original log-rank test to the digitized dataset, we obtained hazard ratio of 0.89 and p-value of 0.12, which were very close to the originally published results of 0.88 and 0.11, respectively. Next, we applied $Z_{\max}$ test to the digitized dataset. The observed correlation between LR and FH was 0.85, and the p-value of the

$Z_{\max}$ test was 0.001, which was highly statistically significant. Similarly we apply $Z_{\max}$ test to the reconstructed OS dataset and obtain a similar p-value of 0.002 as the p-value of 0.002 based on log-rank test.

Insert Figure 6 here.

Therefore, if $Z_{\max}$ had been used to test PFS, PFS would have been able to show significant treatment difference and thus could successfully serve as the primary endpoint in CheckMate 025, which would have accelerated the approval considerably to bring this critical treatment to patients much earlier.

4.2 | Digitized Data Based on Pembrolizumab Keynote 040 Trial

Pembrolizumab is another famous PD-1 inhibitor which was recently approved in multiple cancer indications including second-line head and neck squamous cell carcinoma (HNSCC), refractory Hodgkin's lymphoma, melanoma, etc. Due to a similar mechanism to nivolumab, delayed clinical benefits were observed in pembrolizumab clinical trials as well. For example, PFS curves and OS curves in pembrolizumab clinical trials showed delayed separation (Cohen et al., 2017, Herbst et al., 2016).

Keynote 040 trial is a randomized phase 3 confirmatory trial to compare pembrolizumab with a standard of care in patients with recurrent or metastatic HNSCC with disease progression on or after platinum-containing chemotherapy in the US. The primary endpoint of Keynote 040 trial was OS in the intent-to-treat (ITT) population. The pre-specified efficacy boundary on OS in the ITT population was one-sided p-value of 0.0175 with log-rank test (Cohen et al., 2017).

The OS results in the ITT population from Keynote 040 trial has a hazard ratio of 0.81 and a one-sided p-value of 0.0204 with log-rank test (Cohen et al., 2017), which is statistically not significant at the pre-specified level of 0.0175. Hence Keynote 040 trial failed to reach the primary endpoint of OS in the ITT population. The OS curves presented delayed separation around 5 months. Similarly, we reconstructed the OS patient-level data by digitizing the published OS curves from a conference presentation (Cohen et al., 2017). The OS curves generated from the digitized OS dataset (Figure 7) were almost identical to the original OS curves. The OS hazard ratio from the digitized OS dataset was 0.81, and the one-sided p-value with log-rank test was 0.0208, which were quite close to the original results with a hazard ratio of 0.81 and a p-value of 0.0204. When we apply the $Z_{\max}$ test to the digitized data, the observed correlation between LR and FH was 0.86, and the p-value associated with the $Z_{\max}$ test was 0.0055, which was highly statistically significant. Similarly we applied $Z_{\max}$ test to the reconstructed PFS dataset and obtain a statistically non-significant p-value of 0.0946, which is similar to the statistically non-significant p-value of 0.3037 with the log-rank test.

Insert Figure 7 here.

Keynote 040 trial did not reach the primary endpoint of OS in the ITT population because the primary analysis was specified to apply the standard log-rank test to analyze OS. If $Z_{\max}$ test had been specified to analyze OS as the primary analysis, the trial would have been able to meet the primary endpoint.

5 | ESTIMATION

In the previous sections, we have discussed different testing procedures (LR, FH and $Z_{\max}$ tests) and their power comparisons. However, it is also of interest to estimate the treatment effect over the control. Under the proportional hazards assumption, the constant hazard ratio over time is a genetic evaluation of the treatment effect. For example, a hazard ratio of 0.5 in a clinical trial means the conditional hazard rate of the treatment arm is half that of the control arm. Under non-proportional hazard assumptions, the hazard ratio is no longer a constant over time. Kalbfleisch and Prentice (1981) considered estimating the effect by integrating the hazard ratio over time and defined it as the average hazard ratio. Schemper, Wakounig and Heinze (2009) discussed the properties of several average hazard ratios under different situations. However, the meaning of average HR is misleading. For example, two trials may end up having the same average HR but with different hazard ratio functions. Moreover, the interpretation of average HR is challenging.

Under delayed treatment effect, it is more meaningful first to estimate the separation time t_0 and then calculate the HR after t_0. We consider two different approaches to estimate the separation time. The first method is based on the difference between the survival curves. We define $\hat{t}_0 = \max_t \left\{\hat{S}_1(t) \leq \hat{S}_0(t)\right\}$, where $\hat{S}_1(t)$ and $\hat{S}_0(t)$ are the Kaplan-Meier survival curves for the treatment and control arms respectively. That is, the separation time is estimated to be the maximum time where treatment arm has a survival rate no greater than the control. In the other method we apply the Bayesian information criterion (BIC). Given

the separation time t, the BIC is defined as

$$\mathrm{BIC}(t) = -2\left\{l\left(\hat{\theta}_t\right) - l\left(\hat{\theta}_0\right)\right\} + \left(p_t - p_0\right)\log(n), \tag{3}$$

where $l_1\left(\hat{\theta}_t\right), l_0\left(\hat{\theta}_0\right)$ are the maximized log-likelihood under the delayed effect model with separation time t and a PH model, respectively. Their corresponding numbers of parameters are p_t, p_0, and n is the sample size. The separation time is estimated to be the time that minimizes the BIC in Equation (3). That is, $\tilde{t}_0 = \min_t\{\mathrm{BIC}(t)\}$. After the estimation of separation time t_0 using either methods, we then fit a Cox model with the data after separation to estimate the corresponding HR.

We apply the methods as mentioned above to the two examples in Section 4. For the Nivolumab CheckMate 025 trial, if we use the Kaplan-Meier approach, the separation time is 6.1 months, and the hazard ratio after 6.1 months is 0.587; with the BIC method, the separation time is estimated to be 5.9 months, after which the HR is calculated to be 0.575. These two results are not too different from each other. For the Pembrolizumab Keynote 040 trial, both methods estimate the separation time to be 4.7 months, and the HR after separation is 0.636. From the two real trials, the approaches using Kaplan-Meier curves and BIC yield similar estimates.

6 | DISCUSSION

Due to the mechanism of action of immuno-therapies, PFS or OS curves have demonstrated delayed separation in many clinical trials. The conventional log-rank test assumes proportional hazard over time and often lacks power in these types of trials. Weighted log-rank tests can be powerful if the extent of delay and corresponding weights can be reasonably accurately pre-specified. However, based on what have been observed so far, the extent of delay in immune-oncology trials varies from endpoint to endpoint, from indication to indication, and even from trial to trial with the same endpoints in the same indications. This makes it difficult and even impractical to apply these weighted statistical tests that require pre-specification of time delays. It may also pose major challenges in reaching agreement with regulatory agencies on what pre-specified weights should be used in the statistical tests.

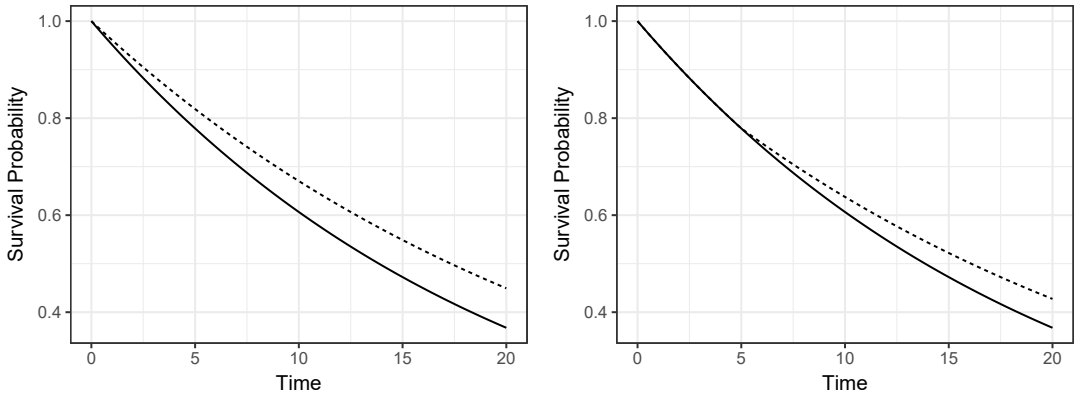
This paper demonstrated that the $Z_{\max}$ test, the maximum of log-rank test and FH test, possesses important properties and it does not require accurate pre-specification of delay time. Under proportional hazard model, the $Z_{\max}$ test is almost as powerful as the most powerful log-rank test and can be substantially more powerful than the FH test. On the other hand, under delayed effect model, the $Z_{\max}$ test is almost as powerful as the FH test while can be much more powerful than the conventional log-rank test. In addition, $Z_{\max}$ is robust in terms of power across different delayed or non-delayed survival models. These properties can make $Z_{\max}$ the most practical candidate for immune-oncology trials in which delayed effects may or may not exist and the extent of delay is unknown in advance. Software will be made available for clinical trial design and analysis using $Z_{\max}$.

When the $Z_{\max}$ test is applied to digitized PFS data from Checkmate 025 trial, PFS becomes highly significant with p-value of 0.0014, which is close to the p-value of 0.0008 for the FH test, while the conventional log-rank test has a non-significant p-value of 0.12. This means, if $Z_{\max}$ had been used to test PFS in Checkmate 025, PFS would have been able to show significant treatment difference and thus could successfully serve as the primary endpoint, which would have accelerated the approval considerably to bring this critical treatment to patients months to years earlier. In the Keynote-040 trial, OS curves displayed delayed separation. When the conventional log-rank test is used, the one-sided p-value was 0.0204, which is not statistically significant based on pre-specified rule; while the p-value corresponding to the $Z_{\max}$ test was 0.0055, which would be statistically significant.

Under proportional hazard model, it is straightforward to estimate the constant hazard ratio (HR) over time and its meaning is clear. The current convention in clinical trials is to provide one average treatment effect or one HR estimate. However, under delayed effect model, one average HR estimate can be misleading. To better understand the magnitude of treatment effect under delayed treatment effect model, it is more meaningful to provide an estimate for delay time, and an HR after delayed effect time point. The delayed effect time point $\hat{t}_0$ can be estimated via either the KM curve difference or the BIC approach. The HR after separation is estimated via a Cox model with the data after $\hat{t}_0$.

$Z_{\max}$ test has some weaknesses as well. It will lose power comparing to the weighted log-rank test if we can reasonably accurately pre-specified the delay time and thus the corresponding weight. In addition, unlike log-rank test or weighted log-rank test, there is no corresponding weight associated with $Z_{\max}$ test and thus there is no treatment effect estimate directly connected to the $Z_{\max}$ test. There is a gap between hypothesis testing and estimation, which may pose some regulatory and practical challenges in drug development based on current convention.

FIGURE 1 Examples of survival curves under the PH and delayed effect models. The solid and dashed lines correspond to the survival curves of the control and treatment arms respectively.



(a) Survival curves under a PH alternative

(b) Survival curves under a delayed alternative

FIGURE 2 Example of power and separation time relationship. The solid and dashed lines display the empirical power curves for the LR and FH tests with different separation time t_0.

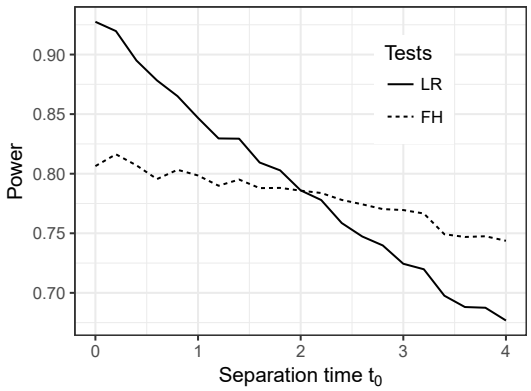
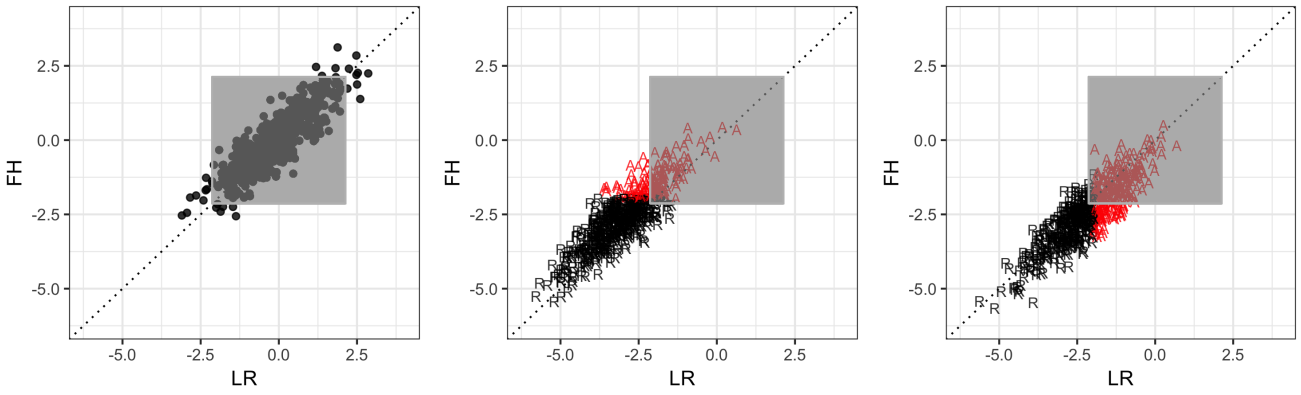


FIGURE 3 Illustration of the $Z_{\max}$ test under the null and two alternatives: PH and delayed treatment effect with separation time $t_0 = 4$ months. The highlighted squares are the acceptance regions for $Z_{\max}$ test with $\rho = 0.855$. The shape of the points specifies the testing results for the FH and LR in the second and third panels respectively. The symbols "A", "R" correspond to "fail to reject" , "reject", respectively.



(a) Under a null

(b) Under a PH alternative

(c) Under a delayed treatment alternative

FIGURE 4 Type I error for the LR, FH and $Z_{\max}$ tests. The x-axis labels different tests and y-axis shows the empirical rejection rates based on 10,000 simulations. Each row represents a different distribution of survival time , and each column lists the number of events.

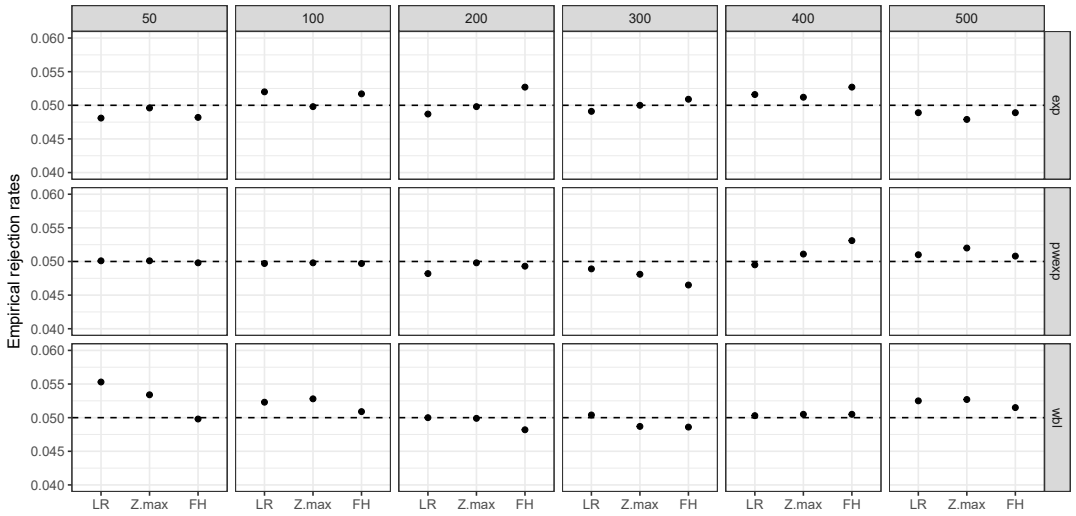


FIGURE 5 Power comparison of the LR, FH, $Z_{\max}$ and Delay tests under the PH and delayed treatment effect alternatives. The x-axis is the number of events, and the y-axis shows the empirical power of different tests based on 10,000 simulations. Each panel corresponds the separation time (0,3,4,5 months), where 0 is the PH case.

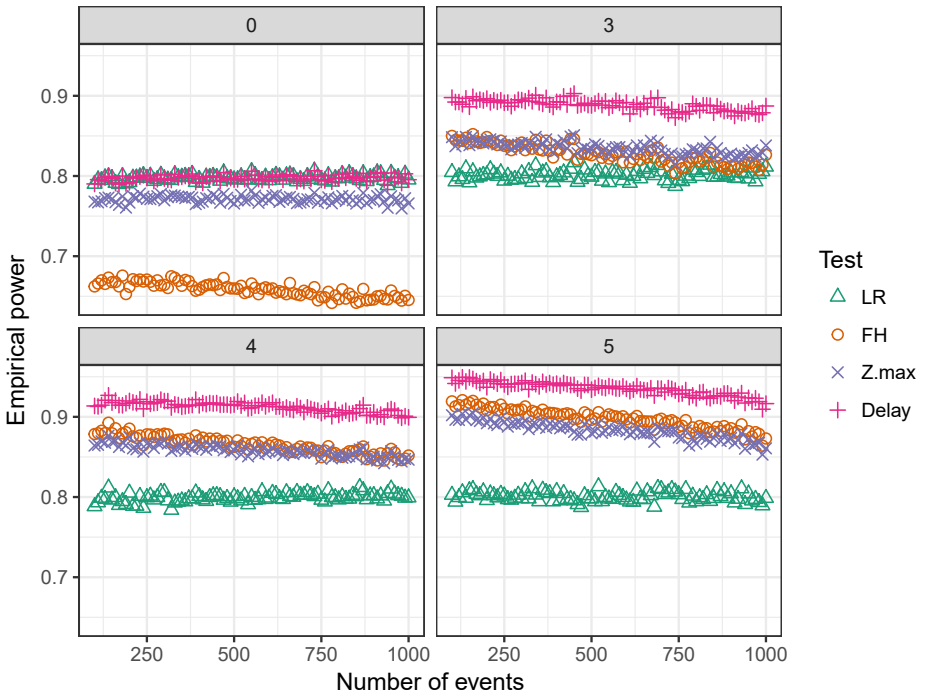


TABLE 1 Sample size required for different tests to achieve 80% power when the follow-up time is fixed at 10 months. Each column corresponds to a separation time t_0, and each row represents different hazard ratios e^θ after t_0. The integer values represent the sample sizes required and empirical powers are in parentheses.

e^θ	Tests	Separation time t_0			
		0	3	4	5
0.55	LR	152 (0.7955)	256 (0.7963)	308 (0.7976)	366 (0.7974)
	FH	204 (0.8084)	236 (0.8078)	254 (0.8005)	282 (0.8024)
	$Z_{\max}$	164 (0.7986)	236 (0.7982)	268 (0.8084)	298 (0.8012)
0.6	LR	200 (0.8000)	328 (0.7998)	382 (0.7977)	440 (0.7912)
	FH	256 (0.7973)	302 (0.8095)	324 (0.7967)	352 (0.7976)
	$Z_{\max}$	210 (0.7921)	300 (0.8057)	330 (0.7957)	370 (0.7944)
0.65	LR	264 (0.8017)	418 (0.8033)	478 (0.7949)	556 (0.8014)
	FH	344 (0.7919)	392 (0.8045)	424 (0.8081)	452 (0.7995)
	$Z_{\max}$	284 (0.8045)	388 (0.8014)	434 (0.8067)	468 (0.7964)
0.7	LR	366 (0.8064)	560 (0.8079)	626 (0.7963)	720 (0.7985)
	FH	470 (0.7924)	530 (0.8032)	564 (0.7999)	594 (0.8024)
	$Z_{\max}$	382 (0.8030)	524 (0.7954)	568 (0.7976)	628 (0.8035)
0.75	LR	520 (0.7982)	766 (0.7963)	878 (0.7972)	984 (0.7900)
	FH	690 (0.8064)	752 (0.7956)	784 (0.7973)	832 (0.7965)
	$Z_{\max}$	540 (0.7923)	734 (0.7984)	792 (0.7961)	864 (0.7972)

TABLE 2 Follow-up time required to achieve 80% power for different tests with the fixed sample size. Empirical power is provided in parenthesis.

e^θ	Tests	Separation time t_0			
		0	3	4	5
0.65	LR	3.5776 (0.7990)	10.2015 (0.8028)	10.2816 (0.7980)	9.8128 (0.7998)
	FH	9.2769 (0.7934)	8.2589 (0.8006)	6.8579 (0.8060)	5.3797 (0.7962)
	$Z_{\max}$	4.5574 (0.7923)	8.2065 (0.8039)	7.4172 (0.8009)	6.4756 (0.8059)
0.7	LR	2.7208 (0.8004)	9.6226 (0.8064)	10.3569 (0.7983)	9.6786 (0.7907)
	FH	9.8497 (0.7964)	7.8271 (0.7960)	6.6402 (0.7973)	5.1048 (0.7924)
	$Z_{\max}$	3.9541 (0.8023)	7.3868 (0.7974)	7.5533 (0.8095)	5.7801 (0.8009)

FIGURE 6 The progression-free survival curves for Nivolumab CheckMate 025 trial. The solid and dashed lines correspond to control and treatment arms, respectively.

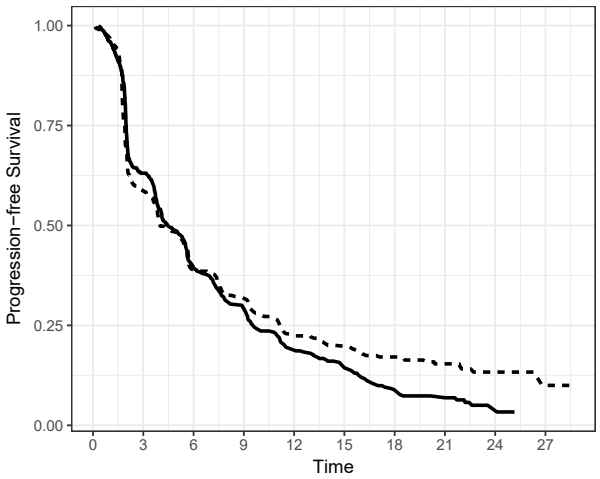
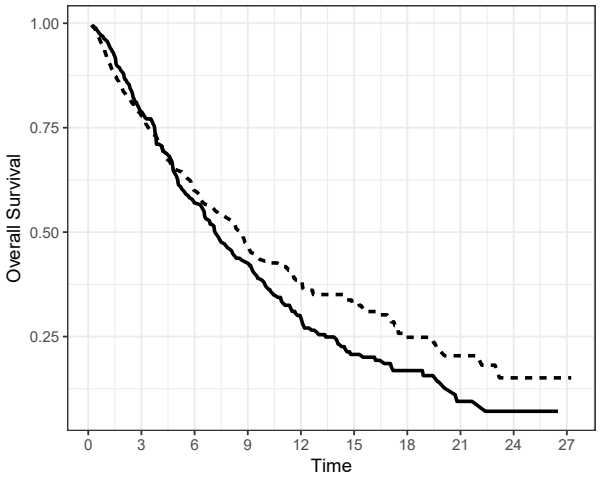


FIGURE 7 The overall survival curves for Pembrolizumab Keynote 040 trial. The solid and dashed lines correspond to control and treatment arms, respectively.



Appendix

A.1 | PROPERTIES OF TEST STATISTIC $U_{W,N}$

In this section, we study the properties of the tests of the family given in Equation (1). For the j^{th} event in D_n, we denote $\mu_j = \dfrac{n_1\left(t_j\right)\lambda_1\left(t_j\right)}{n_1\left(t_j\right)\lambda_1\left(t_j\right)+n_0\left(t_j\right)\lambda_0\left(t_j\right)}$. By Schoenfeld (1981),

$$U_{w,n} = \frac{\sum_{j\in D_n} w_{n,j}\left(X_j-\mu_j\right)}{\sqrt{\sum_{j\in D_n} w_{n,j}^2\mu_j\left(1-\mu_j\right)}} + \frac{\sum_{j\in D_n} w_{n,j}\log\left\{\lambda_1\left(t_j\right)/\lambda_0\left(t_j\right)\right\}p\left(t_j\right)\left\{1-p\left(t_j\right)\right\}}{\sqrt{\sum_{j\in D_n} w_{n,j}^2 p\left(t_j\right)\left\{1-p\left(t_j\right)\right\}}}. \tag{A.1}$$

The first term has a limiting standard normal distribution. Let

$$\tau_w = \sqrt{P_0P_1}\int w(t)\log\left\{\frac{\lambda_1(t)}{\lambda_0(t)}\right\}V(t)\,dt\left\{\int w^2(t)V(t)\,dt\right\}^{-1/2},$$

where $\lim_{n\to\infty} w_{n,j} = w\left(t_j\right)$, $V(t) = P_1S_1(t)G_1(t) + P_0S_0(t)G_0(t)$ with $G_k(t)$ being the survival of censoring in group k, $F_k(t) = 1-S_k(t)$, and P_k is the percentage of individuals allocated to Group k, $k=0,1$. We define the limit of the second term as $\sqrt{n}\tau_w$, which depends on $w(t)$, P_0, $V(t)$ and hazard ratios. Note that for the LR test, $w_{n,j}=1$, and $w(t)=1$. For the FH test, $\lim_{n\to\infty} w_{n,j} = 1-S\left(t_j\right) = F\left(t_j\right)$. Therefore in what follows, we use $\sqrt{n}\tau_1$ and $\sqrt{n}\tau_F$ to denote the asymptotic means of the LR and FH test statistics, respectively.

To develop the theoretical properties of the tests, we need the following assumptions:

(A1). The survival and censoring distributions are independent from each other, and both have finite supports.

(A2). The sequence of weights $\left\{w_{n,j}\right\}_{j\in D_n}$ is the realization of an adapted bounded nonnegative predictable process at event times $\left\{t_j\right\}_{j\in D_n}$.

Assumption (A1) are standard assumptions on the failure time and censoring distributions. Assumption (A2) adds constraints on the weights in Equation (1) such that the test statistics are well defined. Same assumption can be found in Gill (1980) and Fleming and Harrington (1991).

Under the null $H_0 : \lambda_1(t)/\lambda_0(t) = 1$, $\sqrt{n}\tau_w = 0$. Therefore under H_0, $U_{w,n} \xrightarrow{d} N(0,1)$ .The asymptotical normality of the test statistics in Equation (1) can trace back to Cox (1972). Similar conclusions can also be found in Gill (1980), Fleming and Harrington (1991).

Proposition A.1**.** Suppose that the Assumptions (A1), (A2) hold. Under the null hypothesis $H_0 : \lambda_1(t)/\lambda_0(t) = 1$, the test statistic in Equation (1) has an asymptotic normal distribution with mean zero and variance one.

Proposition A.1 establishes the asymptotic normality for test statistic $U_{w,n}$ under H_0. Therefore, we can apply it to find the asymptotic rejection region for the tests in Equation (1).

On the other hand, for any fixed alternative H_a, for example, $H_a^{\mathrm{PH}} : \lambda_1(t)/\lambda_0(t) = e^{\theta}$ or $H_a^{\mathrm{Delay}} : \lambda_1(t)/\lambda_0(t) = 1-\left(1-e^{\theta}\right)I_{(t\geq t_0)}$, one has $\lim_{n\to\infty}\sqrt{n}\tau_w = \infty$. Then the power of $U_{w,n}$ goes to one for any test. That is, any test $U_{w,n}$ in Equation (1) is consistent.

Proposition A.2**.** Suppose that the Assumptions (A1), (A2) hold. Any test in Equation (1) is consistent. That is, under any fixed alternative H_a, $\lim_{n\to\infty}\Psi_{U_{w,n}}(\theta) = 1$.

Proposition A.2 shows that each test by Equation (1) is consistent. Therefore it's not helpful to compare different tests using their limiting powers, which states the necessities of using local asymptotics.

A.2 | PROOF OF THEOREM 1

Under proportional hazards (PH) alternative $H_{a,n}^{\mathrm{PH}}: \lambda_1(t)/\lambda_0(t) = e^{\theta_n}$. For $\theta_n = \delta/\sqrt{n}$ with $\delta < 0$,

$$\tau_w = \sqrt{P_0 P_1}\frac{\delta}{\sqrt{n}}\frac{\int w(t)V(t)\,dt}{\sqrt{\int w^2(t)V(t)\,dt}} < 0.$$

In addition, for any w, we have $\sqrt{n}\tau_w$ converges to a constant. Let $\xi_w = \lim_{n\to\infty}\sqrt{n}\tau_w$. Furthermore by Cauchy-Schwartz inequality,

$$\frac{\xi_1}{\xi_F} = \frac{\int V(t)\,dt}{\sqrt{\int V(t)\,dt}}\frac{\sqrt{\int F^2(t)V(t)\,dt}}{\int F(t)V(t)\,dt} = \frac{\sqrt{\int V(t)\,dt\int F^2(t)V(t)\,dt}}{\int F(t)V(t)\,dt} \geq 1.$$

The equality holds if and only if $V(t) = F^2(t)V(t)$. That is $F(t) = 1$. Therefore $|\xi_1| > |\xi_F|$. Thus $\xi_1 < \xi_F < 0$.

By applying Equation (7a.7.4) in Rao (2001) to measure the asymptotical efficiency of $U_{1,n}$ and $U_{F,n}$, we have

$$\begin{aligned} e\left(U_{1,n}\right) &= \lim_{n\to\infty}\psi_{U_{1,n}}\left(\theta_n\right) = 1 - \Phi\left(-\xi_1 + z_{\alpha/2}\right) + \Phi\left(-\xi_1 - z_{\alpha/2}\right), \\ e\left(U_{F,n}\right) &= \lim_{n\to\infty}\psi_{U_{F,n}}\left(\theta_n\right) = 1 - \Phi\left(-\xi_F + z_{\alpha/2}\right) + \Phi\left(-\xi_F - z_{\alpha/2}\right). \end{aligned}$$

Note that $1 - \Phi\left(-\mu + z_{\alpha/2}\right) + \Phi\left(-\mu - z_{\alpha/2}\right)$ in an increasing function of μ on $(-\infty, 0)$. Therefore $e\left(U_{1,n}\right) > e\left(U_{F,n}\right)$. Thus under the PH alternative, the LR test is asymptotically more efficient than the FH test in the neighborhood of H_0. ■

A.3 | PROOF OF THEOREM 2

Under delayed treatment alternative, consider a simple case: $H_{a,n}^{\mathrm{Delay}}: \lambda_1(t)/\lambda_0(t) = 1 - \left(1 - e^{\theta_n}\right) I_{(t\geq t_0)}$ with $\theta_n = \delta/\sqrt{n}$ and $\delta < 0$. Then under $H_{a,n}^{\mathrm{Delay}}$,

$$\tau_w = \tau_w\left(t_0\right) = \sqrt{P_0 P_1}\frac{\delta}{\sqrt{n}}\frac{\int I_{\{t\geq t_0\}}w(t)V(t)\,dt}{\sqrt{\int w^2(t)V(t)\,dt}} < 0,$$

and let $\xi_w\left(t_0\right) = \lim_{n\to\infty}\sqrt{n}\tau_w\left(t_0\right)$. We would like to compare the asymptotic means of the LR and FH tests

$$\frac{\xi_1^2\left(t_0\right)}{\xi_F^2\left(t_0\right)} = \frac{\left\{\int I_{\{t\geq t_0\}}V(t)\,dt\right\}^2}{\int V(t)\,dt}\frac{\int F^2(t)V(t)\,dt}{\left\{\int I_{\{t\geq t_0\}}F(t)V(t)\,dt\right\}^2} = \frac{\int F^2(t)V(t)\,dt}{\int V(t)\,dt}\Big/\left\{\frac{\int I_{\{t\geq t_0\}}F(t)V(t)\,dt}{\int I_{\{t\geq t_0\}}V(t)\,dt}\right\}^2.$$

Notice that $\xi_1^2\left(t_0\right)/\xi_F^2\left(t_0\right)$ is a continuous and differentiable function of t_0. We denote $t_{\mathrm{sup}} = \sup_t\{F(t) < 1\}$. Then one observes that as $t_0 \to 0$, $H_{a,n}^{\mathrm{Delay}}$ degenerates to $H_{a,n}^{\mathrm{PH}}$, $\xi_1(0)/\xi_F(0) > 1$. In addition, as $t_0 \to t_{\mathrm{sup}}$, the limit of $\xi_1^2\left(t_0\right)/\xi_F^2\left(t_0\right)$ only depends on the denominator,

$$\lim_{t_0\to t_{\mathrm{sup}}}\frac{\int I_{\{t\geq t_0\}}F(t)V(t)\,dt}{\int I_{\{t\geq t_0\}}V(t)\,dt} = 1 - \lim_{t_0\to t_{\mathrm{sup}}}\frac{\int I_{\{t\geq t_0\}}S(t)V(t)\,dt}{\int I_{\{t\geq t_0\}}V(t)\,dt} = 1 - \lim_{t_0\to t_{\mathrm{sup}}}\frac{S\left(t_0\right)V\left(t_0\right)}{V\left(t_0\right)} = 1.$$

Therefore $\lim_{t_0\to t_{\mathrm{sup}}}\frac{\xi_1^2\left(t_0\right)}{\xi_F^2\left(t_0\right)} = \frac{\int F^2(t)V(t)dt}{\int V(t)dt} < 1$.

Moreover for any $0 < t_0 < t_{\mathrm{sup}}$,

$$\begin{aligned} \frac{\partial}{\partial t_0}\left\{\frac{\int I_{\{t\geq t_0\}}F(t)V(t)\,dt}{\int I_{\{t\geq t_0\}}V(t)\,dt}\right\} &= \frac{-V\left(t_0\right)F\left(t_0\right)\int_{t_0}^{\infty}V(t)\,dt + V\left(t_0\right)\int_{t_0}^{\infty}F(t)V(t)\,dt}{\left\{\int_{t_0}^{\infty}V(t)\,dt\right\}^2} \\ &= \frac{V\left(t_0\right)\int_{t_0}^{\infty}\left\{F(t) - F\left(t_0\right)\right\}V(t)\,dt}{\left\{\int_{t_0}^{\infty}V(t)\,dt\right\}^2} \geq 0. \end{aligned}$$

Therefore $\xi_1^2(t_0)/\xi_F^2(t_0)$ is a decreasing function of t_0 on $(0, t_{\sup})$. Thus there exists $t^* \in (0, t_{\sup})$, such that $\xi_1^2(t_0)/\xi_F^2(t_0) = 1$, where t^* also depends on the survival and enrollment assumptions. Thus for $t_0 < t^*$, $\xi_1(t_0)/\xi_F(t_0) > 1$; for $t^* < t_0 < t_{\sup}$, $\xi_1(t_0)/\xi_F(t_0) < 1$.

Similar to the proof of Theorem 1, we have that the LR is asymptotically less efficient than the FH under delayed treatment effect in the neighborhood of H_0 when separation time t_0 is large enough. ■

Under $H_{a,n}^{\text{Delay}}$, we let $g_0(t) = I_{\{t \geq t_0\}}$. Then for any g,

$$\left|\frac{\xi_{g_0}(t_0)}{\xi_g(t_0)}\right| = \frac{\int I_{\{t\geq t_0\}} V(t)\,dt}{\sqrt{\int I_{\{t\geq t_0\}} V(t)\,dt}} \frac{\sqrt{\int g^2(t) V(t)\,dt}}{\int g(t) V(t)\,dt}$$
$$= \frac{\sqrt{\int I_{\{t\geq t_0\}} V(t)\,dt \int g^2(t) V(t)\,dt}}{\int g(t) V(t)\,dt} \geq 1$$

by Cauchy-Schwartz inequality. The equality holds if and only if $I_{\{t\geq t_0\}} V(t) = g^2(t) V(t)$. That is $g(t) = g_0(t)$. Therefore $\left|\xi_{g_0}(t_0)\right| > \left|\xi_g(t_0)\right|$ for any $g \neq g_0$. This shows that in Equation (1), if we set the weight $w_{n,j} = I_{t_j \geq t_0}$ and denote the new test as $U_{t_0,n}$, then

Proposition A.3**.** Suppose the Assumptions (A1) and (A3) in the appendix hold. For testing $H_0 : \lambda_1(t) = \lambda_0(t)$ versus $H_{a,n}^{\text{PH}} : \lambda_1(t)/\lambda_0(t) = e^{\theta_n}$, where $\theta_n = \delta/\sqrt{n}$ and $\delta < 0$, $U_{t_0,n}$ test is the locally most efficient test in the family given by Equation (1) in the neighborhood of H_0.

$H_{a,n}^{\text{PH}} : \lambda_1(t)/\lambda_0(t) = e^{\theta_n}$ is a special case of the lag model in Zucker and Lakatos (1990). Proposition A.3 can also be derived from Zucker and Lakatos (1990). Although $U_{t_0,n}$ is the locally most efficient test under $H_{a,n}$, it is not feasible in applications because the separation time is often unknown to us. It thus only serves as a reference line for other tests in theory.

A.4 | PROOF OF THEOREMS 3 AND 4

The property of $Z_{\max}$ is based on the asymptotic joint distribution of $(U_{1,n}, U_{F,n})$ in Equation (2). Note that according to Fleming and Harrington (1991), the asymptotic correlation between the LR and FH is given as

$$\rho = \rho_F = \frac{\int F(t) V(t)\,dt}{\sqrt{\int V(t)\,dt}\sqrt{\int F^2(t) V(t)\,dt}}, \tag{A.2}$$

which depends on the survival and censoring assumptions. By definition in Equation (A.2), ρ is always positive. We first introduce a lemma for bivariate normal distributions.

Lemma A.1. Let $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left\{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right\}$ with $\mu_1 < \mu_2 < 0$ and $\rho > 0$. For $\alpha \in (0,1)$, $z_{\alpha/2}$ and $c_{\alpha,\rho}$ satisfy $P\left(|X_1 - \mu_1| < z_{\alpha/2}\right) = P\left\{\max\left(|X_1 - \mu_1|, |X_2 - \mu_2|\right) < c_{\alpha,\rho}\right\} = 1 - \alpha/2$. Denote $q_1 = P\left\{\max\left(|X_1|, |X_2|\right) < c_{\alpha,\rho}\right\}$, $q_2 = P\left(|X_2| < z_{\alpha/2}\right)$. Then $q_1 < q_2$ if $\mu_1 < \Delta_\alpha$, where $\Delta_\alpha = -\frac{\left(4c_{\alpha,\rho}^2 - 14c_{\alpha,\rho}z_{\alpha/2} + 10z_{\alpha/2}^2\right)\rho + \left(c_{\alpha,\rho}^2 + 4c_{\alpha,\rho}z_{\alpha/2} - 5z_{\alpha/2}^2\right)}{(c_{\alpha,\rho} - 3z_{\alpha/2})\rho + 2z_{\alpha/2}}$.

Proof: Assume $Z_1, Z_2 \overset{iid}{\sim} N(0,1)$, then we can write $X_1 = \sqrt{1-\rho^2}Z_1 + \rho Z_2 + \mu_1$, $X_2 = Z_2 + \mu_2$. q_1 and q_2 can be written as

$$q_1 = \Phi\left(-\mu_2 + z_{\alpha/2}\right) - \Phi\left(-\mu_2 - z_{\alpha/2}\right),$$
$$q_2 = P\left(-c_{\alpha,\rho} - \mu_1 < \sqrt{1-\rho^2}Z_1 + \rho Z_2 < c_{\alpha,\rho} - \mu_1, -c_{\alpha,\rho} - \mu_2 < Z_2 < c_{\alpha,\rho} - \mu_2\right).$$

Furthermore $q_1 - q_2 = (I + II) - (III + IV)$, where

$$I = \int_{-\mu_2+z_{\alpha/2}}^{-\mu_2+c_{\alpha,\rho}} dz_2 \int_{-c_{\alpha,\rho}-\mu_1<\sqrt{1-\rho^2}z_1+\rho z_2<c_{\alpha,\rho}-\mu_1} \phi(z_1)\,\phi(z_2)\,dz_1,$$

$$II = \int_{-\mu_2-c_{\alpha,\rho}}^{-\mu_2-z_{\alpha/2}} dz_2 \int_{-c_{\alpha,\rho}-\mu_1<\sqrt{1-\rho^2}z_1+\rho z_2<c_{\alpha,\rho}-\mu_1} \phi(z_1)\,\phi(z_2)\,dz_1,$$

$$III = \int_{-\mu_2-z_{\alpha/2}}^{-\mu_2+z_{\alpha/2}} dz_2 \int_{\sqrt{1-\rho^2}z_1+\rho z_2<c_{\alpha,\rho}-\mu_1} \phi(z_1)\,\phi(z_2)\,dz_1,$$

$$IV = \int_{-\mu_2-z_{\alpha/2}}^{-\mu_2+z_{\alpha/2}} dz_2 \int_{-c_{\alpha,\rho}-\mu_1<\sqrt{1-\rho^2}z_1+\rho z_2} \phi(z_1)\,\phi(z_2)\,dz_1,$$

with $\phi(\cdot)$ the pdf of standard normal. Since I, II, III, IV can all be written as the form of $\int\int_\Omega \phi(z_1)\,\phi(z_2)\,dz_1 dz_2$, which depends on how far the region Ω is away from the origin, we have

$$II \le \int_{-\mu_1-z_{\alpha/2}}^{-\mu_1+c_{\alpha,\rho}-2z_{\alpha/2}} dz_2 \int_{-c_{\alpha,\rho}-\mu_2<\sqrt{1-\rho^2}z_1+\rho z_2<c_{\alpha,\rho}-\mu_2} \phi(z_1)\,\phi(z_2)\,dz_1. \tag{A.3}$$

We divide I into two parts $I = I_1 + I_2$ with

$$I_1 = \int_{-\mu_2+z_{\alpha/2}}^{-\mu_2+c_{\alpha,\rho}} dz_2 \int_{-c_{\alpha,\rho}-\mu_1<\sqrt{1-\rho^2}z_1+\rho z_2<-c_{\alpha,\rho}-\mu_1+\rho(3z_{\alpha/2}-c_{\alpha,\rho})} \phi(z_1)\,\phi(z_2)\,dz_1,$$

$$I_2 = \int_{-\mu_2+z_{\alpha/2}}^{-\mu_2+c_{\alpha,\rho}} dz_2 \int_{-c_{\alpha,\rho}-\mu_1+\rho(3z_{\alpha/2}-c_{\alpha,\rho})<\sqrt{1-\rho^2}z_1+\rho z_2<c_{\alpha,\rho}-\mu_1} \phi(z_1)\,\phi(z_2)\,dz_1.$$

It's easy to see that

$$I_1 \le \int_{-\mu_2+c_{\alpha,\rho}-2z_{\alpha/2}}^{-\mu_2+z_{\alpha/2}} dz_2 \int_{z_{\alpha/2}-2c_{\alpha,\rho}-\mu_1<\sqrt{1-\rho^2}z_1+\rho z_2<-c_{\alpha,\rho}-\mu_1} \phi(z_1)\,\phi(z_2)\,dz_1. \tag{A.4}$$

Denote A as the common point shared by the integration regions on the right hand sides in Equations (A.3), (A.4), and B as the point closest to the origin in the integration region of I_2. Denote $\|A\|, \|B\|$ as the distances of A, B to the origin, respectively. When $\|A\| \le \|B\|$, one can map the integration region of I_2 to an area in that of III. We have

$$\begin{aligned}\|A\|^2 - \|B\|^2 = \frac{1}{1-\rho^2}\Big[&2\left\{\left(c_{\alpha,\rho} - 3z_{\alpha/2}\right)\rho + 2z_{\alpha/2}\right\}\mu_1 + 2\left(-2z_{\alpha/2}\rho - c_{\alpha,\rho} + 3z_{\alpha/2}\right)\mu_2 \\ &+ \left(4c_{\alpha,\rho}^2 - 14c_{\alpha,\rho}z_{\alpha/2} + 10z_{\alpha/2}^2\right)\rho + \left(c_{\alpha,\rho}^2 + 4c_{\alpha,\rho}z_{\alpha/2} - 5z_{\alpha/2}^2\right)\Big].\end{aligned}$$

Since $\mu_1 < \mu_2 < 0$, when $\mu_1 < \Delta_\alpha$, we have $\|A\| \le \|B\|$, and furthermore,

$$I_2 \le \int_{-\mu_2+c_{\alpha,\rho}-2z_{\alpha/2}}^{-\mu_2+2c_{\alpha,\rho}-3z_{\alpha/2}} dz_2 \int_{\sqrt{1-\rho^2}z_1+\rho z_2<z_{\alpha/2}-2c_{\alpha,\rho}-\mu_1} \phi(z_1)\,\phi(z_2)\,dz_1. \tag{A.5}$$

Note that the right hand sides in Equations (A.3), (A.4), (A.5) all have their integration regions as a subset of III, and these regions have no overlapping. Therefore we have $I + II \le III$. Thus $q_1 - q_2 = (I + II) - (III + IV) < 0$.■

Based on the above lemma, we can compare the efficiency of $Z_{\max}$ with the LR and FH under different alternatives within the framework of local asymptotics. The following assumptions are needed to develop the theoretical results.

(A3). Assume $\delta\sqrt{P_0P_1}\sqrt{\int V(t)\,dt} < \Delta_\alpha$.

(A4). Assume $\delta\sqrt{P_0P_1}\int I_{\{t\geq t_0\}}F(t)V(t)\,dt/\sqrt{\int F^2(t)V(t)\,dt} < \Delta_\alpha$.

Assumptions (A3), (A4) add constraints on the asymptotical means of the LR and FH tests under the PH and delayed alternatives. They are basically assuming the distributions of the LR and FH tests under shrinking alternatives are not too close to the null distribution $N(0,1)$, which means the limiting values of their power functions are not too small.

A.4.1 | Proof of Theorem 3

Under the PH alternative $H_{a,n}^{\mathrm{PH}}: \lambda_1(t)/\lambda_0(t) = e^{\theta_n}$ with $\theta_n = \delta/\sqrt{n}$ and $\delta < 0$, we have already proven $\lim_{n\to\infty}\sqrt{n}\tau_1(\theta_n) = \xi_1 < \lim_{n\to\infty}\sqrt{n}\tau_F(\theta_n) = \xi_F < 0$. Assumption (A3) ensures $\xi_1 < \Delta_\alpha$.

Note that $1-\lim_{n\to\infty}\psi_{Z_{\max}}(\theta_n) = P\left\{\max\left(\left|U_{1,n}\right|,\left|U_{F,n}\right|\right) < c_{\alpha,\rho}\right\}$, $1-\lim_{n\to\infty}\psi_{U_{F,n}}(\theta_n) = P\left(\left|U_{F,n}\right| < z_{\alpha/2}\right)$. Since $\begin{pmatrix}U_{1,n}\\U_{F,n}\end{pmatrix} \xrightarrow{d} N\left\{\begin{pmatrix}\xi_1\\\xi_F\end{pmatrix},\begin{pmatrix}1&\rho\\\rho&1\end{pmatrix}\right\}$, by the above Lemma A.1, combined with Assumption (**A3**): one has $\left\{1-\lim_{n\to\infty}\psi_{Z_{\max}}(\theta_n)\right\} - \left\{1-\lim_{n\to\infty}\psi_{U_{F,n}}(\theta_n)\right\} < 0$. That is, $\lim_{n\to\infty}\psi_{U_{F,n}}(\theta_n) < \lim_{n\to\infty}\psi_{Z_{\max}}(\theta_n)$ .Therefore $Z_{\max}$ is asymptotically more efficient than the FH under the PH alternative in the neighborhood of H_0. ■

A.4.2 | Proof of Theorem 4

Under delayed treatment effect $H_{a,n}^{\mathrm{Delay}}: \frac{\lambda_1(t)}{\lambda_0(t)} = 1-\left(1-e^{\theta_n}\right)I_{(t\geq t_0)}$ with t_0 large enough such that $\xi_F(t_0) < \xi_1(t_0)$. Similarly under local asymptotics with $\theta_n = \delta/\sqrt{n}$ and $\delta < 0$, Assumption (A4) ensures $\xi_F < \Delta_\alpha$.

Since $1-\lim_{n\to\infty}\psi_{Z_{\max}}(\theta_n) = P\left\{\max\left(\left|U_{1,n}\right|,\left|U_{F,n}\right|\right) < c_{\alpha,\rho}\right\}$, $1-\lim_{n\to\infty}\psi_{U_{1,n}}(\theta_n) = P\left(\left|U_{1,n}\right| < z_{\alpha/2}\right)$, similar to the proof of Theorem 3, one has $\left\{1-\lim_{n\to\infty}\psi_{Z_{\max}}(\theta_n)\right\} - \left\{1-\lim_{n\to\infty}\psi_{U_{1,n}}(\theta_n)\right\} < 0$.That is, $\lim_{n\to\infty}\psi_{U_{1,n}}(\theta_n) < \lim_{n\to\infty}\psi_{Z_{\max}}(\theta_n)$ .Therefore $Z_{\max}$ is asymptotically more efficient than the LR in the neighborhood of H_0 when separation time is large enough. ■

REFERENCES

1. Chen T. T. (2013). Statistical issues and challenges in immuno-oncology. *Journal for immunotherapy of cancer*, 1(1), 18.
2. Cohen, E.E., Harrington, K.J., Le Tourneau, C., Dinis, J., Licitra, L., Ahn, M.J., Soria, A., Machiels, J.P., Mach, N., Mehra, R. and Burtness, B. (2017). LBA45_PRPembrolizumab (pembro) vs standard of care (SOC) for recurrent or metastatic head and neck squamous cell carcinoma (R/M HNSCC): Phase 3 KEYNOTE-040 trial. *Annals of Oncology*, 28.
3. Cox, D.R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological),* 34(2), 187-220.
4. Ferris, R.L., Blumenschein Jr, G., Fayette, J., Guigay, J., Colevas, A.D., Licitra, L., Harrington, K., Kasper, S., Vokes, E.E., Even, C. and Worden, F. (2016). Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *New England Journal of Medicine*, 375(19), 1856-1867.
5. Fine, G.D. (2007). Consequences of delayed treatment effects on analysis of time-to-event endpoints. *Drug Information Journal*, 41(4), 535-539.
6. Fleming T.R. and Harrington D.P. (1991). *Counting processes and survival analysis.* New York: John Wiley & Sons.
7. Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52, 203-224.
8. Gill, R.D. (1980). Censoring and stochastic integrals. *Statistica Neerlandica*, 34(2), 124-124.
9. Guyot P., Ades A.E., Ouwens M.J., and Welton N.J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*, 12(1), 9.
10. Harrington, D.P. and Fleming, T.R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69(3), 553-566.
11. Hasegawa T. (2014). Sample size determination for the weighted log-rank test with the Fleming-Harrington class of weights in cancer vaccine studies. *Pharmaceutical Statistics*, 13(2), 128-135.

12. Herbst, R.S., Baas, P., Kim, D.W., Felip, E., Pérez-Gracia, J.L., Han, J.Y., Molina, J., Kim, J.H., Arvis, C.D., Ahn, M.J. and Majem, M. (2016). Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *The Lancet*, 387(10027), 1540-1550.
13. Hoos, A. (2012). Evolution of end points for cancer immunotherapy trials. *Annals of oncology*, *23*(8), 47-52.
14. Hoos, A., Eggermont, A.M., Janetzki, S., Hodi, F.S., Ibrahim, R., Anderson, A., Humphrey, R., Blumenstein, B., Old, L. and Wolchok, J. (2010). Improved endpoints for cancer immunotherapy trials. *Journal of the National Cancer Institute*, 102(18), 1388-1397.
15. Kalbfleisch, J.D. (1978). Likelihood methods and nonparametric tests. *Journal of the American Statistical Association*, 73(361), 167-170.
16. Kalbfleisch, J.D. and Prentice, R.L. (1981). Estimation of the average hazard ratio. *Biometrika*, 68(1), 105-112.
17. Karrison, T. G. (2016). Versatile tests for comparing survival curves based on weighted log-rank statistics. *Stata Journal*, 16(3), 678-690.
18. Kaufman, H., Schwartz, L.H., William, W.N., Sznol, M., del Aguila, M., Whittington, C., Fahrbach, K., Xu, Y., Masson, E., Dempster, S. and Vergara-Silva, A.L. (2017). Evaluation of clinical endpoints as surrogates for overall survival in patients treated with immunotherapies.
19. Lachin, J. M., Foulkes, M.A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*, 42(3), 507-519.
20. Lakatos E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*, 44, 229-241.
21. Lee, J. W. (1996). Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics*, 721-725.
22. Lee, S. H. (2007). On the versatility of the combination of the weighted log-rank statistics. *Computational Statistics & Data Analysis*, 51(12), 6557-6564.
23. Lin, R.S. and León, L.F. (2017). Estimation of treatment effects in weighted log-rank tests. *Contemporary clinical trials communications*, 8, 147-155.
24. Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports,* 50(3), 163-170.
25. Motzer, R.J., Escudier, B., McDermott, D.F., George, S., Hammers, H.J., Srinivas, S., Tykodi, S.S., Sosman, J.A., Procopio, G., Plimack, E.R. and Castellano, D. (2015). Nivolumab versus everolimus in advanced renal-cell carcinoma. *New England Journal of Medicine*, 373(19), 1803-1813.
26. Peto, R. (1972). Rank tests of maximal power against Lehmann-type alternatives. *Biometrika*, 59(2), 472-5.
27. Peto, R., Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A (General),* 135(2), 185-207.
28. Rao, C.R., 2001. *Linear Statistical Inference and its Applications.* New York: John Wiley & Sons.
29. Self, S., Prentice, R., Iverson, D., Henderson, M., Thompson, D., Byar, D., Insull, W., Gorbach, S.L., Clifford, C., Goldman, S. and Urban, N. (1988). Statistical design of the women's health trial. *Controlled Clinical Trials*, 9(2), 119-136.
30. Schemper, M., Wakounig, S. and Heinze, G. (2009). The estimation of average hazard ratios by weighted Cox regression. *Statistics in medicine*, 28(19), 2473-2489.
31. Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68, 316-318.
32. Serfling, R.J., (2009). Approximation theorems of mathematical statistics (Vol. 162). New York: John Wiley & Sons.
33. Sit T., Liu M., Shnaidman M., and Ying Z. (2016). Design and analysis of clinical trials in the presence of delayed treatment effect. *Statistics in Medicine*, 35(11), 1774-1779.
34. Tarone, R.E. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika*, 64(1), 156-160.
35. Xu Z., Zhen B.G., Park Y., and Zhu B. (2017). Designing therapeutic cancer vaccine trials with delayed treatment effect. *Statistics in Medicine*, *36***,** 592-605.
36. Yang, M., Hua, Z., and Vardhanabhuti, S. (2018+). Sample size determination under non-proportional hazards.
37. Yang, S., Hsu, L. and Zhao, L. (2005). Combining asymptotically normal tests: case studies in comparison of two groups. *Journal of statistical planning and inference*, 133(1), 139-158.
38. Zhang, D. and Quan, H. (2009). Power and sample size calculation for log?rank test with a time lag in treatment effect. *Statistics in medicine*, 28(5), 864-879.
39. Zucker, D. M. and Lakatos, E. (1990). Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika*, 77, 853-864.

How to cite this article: M. Yang, Z. Hua, L. Xue, and M. Hu (2018), $Z_{\max}$ test for delayed effect in immuno-oncology, *Pharmaceutical statistics*, *2018;00:1–6*.