Effective Comparison of Two Potentially Crossing Hazard Rate Curves

Xiaoxi Zhang[a], Somnath Datta[a] and Peihua Qiu[a]

[a]Department of Biostatistics, University of Florida, Gainesville, FL, United States

ARTICLE HISTORY
Compiled March 25, 2024

ABSTRACT
In survival data analysis, comparison of two hazard rate curves is critically important for evaluating a treatment effect. In many applications, the two hazard curves could potentially cross each other, violating the proportional hazards assumption in the Cox's model. In such cases, the traditional tests like the log-rank test and the Peto-Peto test that were developed based on that assumption would be ineffective. There have been some discussions in the literature on comparison of two potentially crossing hazard curves, based on either parametric modeling or nonparametric testing approaches. However, the assumed models of the existing parametric methods are often difficult to justify in practice. On the other hand, the existing nonparametric tests are usually based on the maximization with respect to an unknown crossing point, leading to complex null distributions for the corresponding test statistics. We suggest a novel method in this paper for comparing two hazard curves based on a nonparametric testing procedure. Its test statistic avoids the maximization mentioned above and consequently has the desirable asymptotic normality property under some regularity conditions. We show that the new method is effective for comparing two potentially crossing hazard curves.

KEYWORDS
Additive tests; Asymptotic normality; Crossing point; Hazard rate functions; Survival analysis; Weighted log-rank test

1. Introduction

Comparison of two hazard rate functions is critically important for the purpose of evaluating treatment effects when analyzing survival data [cf., 14,15]. To this end, the log-rank test is the most widely used test whose performance is optimal when the two hazard rate functions satisfy the Cox proportional hazard model assumption. Many modified versions of the log-rank test, including the Gehan test and the Peto-Peto test, have been proposed in the literature to place more emphasis on earlier failure times [14, Chapter 7]. However, it has been well demonstrated that all these tests could have low power when the two related hazard curves cross each other so that the Cox proportional hazard assumption is violated [e.g., 2,18,19,21,22]. This paper suggests a general approach for effective comparison of two potentially crossing hazard curves.

The crossing hazards phenomenon is common in applications where treatment effects are quite different in different time periods. For instance, surgeries can usually

CONTACT Peihua Qiu. Email: pqiu@ufl.edu

improve a patient's long-term health. But, in a short term, they may cause high mortality due to infections or other short-term risks [24]. In the literature, there have been many existing methods for proper comparison of two potentially crossing hazard curves. Some early methods employ the modeling approach by including the crossing structure of the hazard rate functions explicitly in a parametric model [e.g., 1,2,19]. However, their assumed parametric models are often difficult to justify in practice. Therefore, some methods based on nonparametric tests have also been developed based on the following observation about the log-rank test. When two hazard rate functions cross each other, early differences between the two functions would be canceled out by late differences of opposite sign in the log-rank test statistic, which explains why that test would be ineffective in such cases. To avoid this cancelation, many existing methods for comparing two crossing hazard rate functions define their test statistics using the absolute or squared differences between the two estimated hazard rate functions [cf., 9,18], or adopt the weighted log-rank testing framework by choosing special weights that change signs before and after a potential crossing point [cf., 20,21]. Some recent methods suggested combining several weighted log-rank tests for comparing two potentially crossing hazard curves [e.g., 5,16]. [10] suggested two tests based on the Pearson chi-squared test and the log-likelihood ratio test for comparing multiple non-proportional hazard rate functions. For comprehensive numerical comparisons among various existing methods, see [6,17].

In many existing methods mentioned above, the test statistics are derived specifically for the alternative hypothesis that the related hazard curves cross each other at an unknown crossing point. Such a problem formulation excludes some important cases when two hazard curves are different but not crossing. To overcome this limitation, [24] suggested a two-stage additive testing procedure in which the log-rank test was used in the first stage to detect non-crossing difference between the two hazard curves and a specific weighted log-rank test was used in the second stage to detect any crossing difference. In the original two-stage procedure, the p-value was computed using the method suggested for additive tests in [26]. [4] showed that the testing procedure could be improved by using the Fisher's combined probability test in order to compute the p-value. The two-stage method was generalized for comparing multiple hazard rate functions in [3].

In some existing methods [e.g., 20,24] including the two-stage method, the test statistics for comparing two potentially crossing hazard curves are constructed based on certain metrics measuring the difference between the two estimated hazard curves that are maximized with respect to an unknown crossing point. As studied originally by [23] and confirmed by [24], such test statistics have bimodal asymptotic null distributions, and therefore their p-values are difficult to compute accurately. This is one major reason why the related methods are ineffective in certain cases. Regarding the existing weighted log-rank tests designed for comparing two potentially crossing hazard curves, [24] pointed out that it was inappropriate to use the constant weights -1 and 1 before and after the potential crossing point, as done in [1] and [20]. So, in their suggested weighted log-rank test statistic, two different constants with opposite signs were used as weights. Their weighting scheme, however, still has the following two limitations. First, non-constant weights might be more reasonable to use since intuitively observed data closer to the crossing point would contribute less to testing the difference between the two hazard curves because the difference between the two hazard curves is smaller at such places. Second, the weight suggested in [24] is discontinuous at the potential crossing point, making the distribution of the related test statistic analytically complex to study.

In this paper, we propose a novel weighted log-rank test for comparing two potentially crossing hazard curves. Its test statistic uses a continuous weighting scheme that takes larger values at places farther away from the potential crossing point. It also avoids the maximization with respect to the unknown crossing point when defining its test statistic. Consequently, the null distribution of its test statistic is asymptotically normal, which is preferable compared to the bimodal asymptotic distributions of certain existing methods discussed above. This novel weighted log-rank test is then used in the two-stage additive testing framework for detecting any difference between the two hazard curves, including the crossing or non-crossing (e.g., parallel) differences. To properly define the overall p-value of the two-stage additive testing procedure, the test statistics used in the two stages are designed to be asymptotically independent of each other. Then, the method by [26] and the Fisher-test method [7] are combined properly to compute the overall p-value of the two-stage additive test. The proposed method is shown to be effective in many cases, compared to some state-of-the-art competing methods.

The rest of the paper is organized as follows. Our suggested method is described in detail in Section 2. The asymptotic normality of the proposed weighted log-rank test for comparing two potentially crossing hazard curves is established in Section 3. Some simulation results for evaluating the numerical performance of our proposed method in comparison with some competing methods are given in Section 4. Section 5 demonstrates a real data analysis by using our proposed method. Some remarks conclude the paper in Section 6. Proofs of two theorems are provided in Appendix.

2. The Proposed Method

Our proposed method is described in several parts. The problem formulation in the two-stage additive testing framework is introduced in Subsection 2.1. The proposed weighted log-rank test for comparing two potentially crossing hazard curves is discussed in Subsection 2.2. The proposed method to determine the overall p-value of the two-stage testing procedure is described in Subsection 2.3.

2.1. Problem formulation and the two-stage additive tests

In most applications for comparing two hazard rate functions, we are interested in testing whether the two functions are the same or not in a study time period. To be more specific, let $h_0(t)$ and $h_1(t)$ be the hazard rate functions of the survival times of the subjects in the control and treatment groups, respectively. Then, we are interested in the following hypothesis:

$$\begin{aligned} &H_0 : h_1(t) = h_0(t), \text{ for all } t \in [0, \mathcal{T}] \text{ versus} \\ &H_1 : h_1(t) \neq h_0(t), \text{ for some } t \in [0, \mathcal{T}], \end{aligned} \tag{1}$$

where $[0, \mathcal{T}]$ is the study time period. The alternative hypothesis H_1 in (1) contains cases when the two hazard curves are different but not crossing (denoted as $H_1^{(1)}$) and the cases when they cross each other in $[0, \mathcal{T}]$ (denoted as $H_1^{(2)}$).

In the literature, many existing methods for comparing two potentially crossing hazard curves have been deveoped for testing H_0 versus $H_1^{(2)}$ (e.g., [19]). These methods cannot effectively detect the non-crossing difference between the two hazard curves.

To overcome this limitation, [24] suggested to handle the testing problem (1) using a two-stage additive testing procedure with the following two stages:

First Stage: Test for hypotheses H_0 versus $H_1^{(1)}$ by the conventional log-rank test, and

Second Stage: Test for hypotheses H_0 versus $H_1^{(2)}$ by a testing procedure designed specifically for detecting a crossing pattern of the two hazard curves.

The entire two-stage additive test rejects H_0 when either the Stage-I test rejects H_0 or the Stage-I test fails to reject H_0 but the Stage-II test rejects H_0. For this two-stage additive testing procedure, it should be reasonable to use the conventional log-rank test as the Stage-I test since it would be optimal or close to optimal for detecting non-crossing difference between the two hazard curves. But, the weighted log-rank test suggested in [24] as a Stage-II test would have several fundamental limitations, as pointed out in Section 1.

2.2. Proposed two-stage additive testing procedure

Let n_j be the number of subjects in group j, for $j = 1, 2$, and $\{t_1,\ t_2,\ \ldots,\ t_D\}$ be the set of D distinct ordered event times in the pooled sample. For the jth group at time t_i, d_{ij} denotes the observed number of events and Y_{ij} denotes the number of individuals at risk, for $i = 1, 2, \cdots, D$, and $j = 1, 2$. Let $d_i = d_{i1} + d_{i2}$ and $Y_i = Y_{i1} + Y_{i2}$, for each i. Then, the test statistic of the conventional log-rank test used in Stage-I of the two-stage additive testing procedure is defined to be

$$U = \frac{\sum_{i=1}^{D} w_{i1} \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^{D} w_{i1}^2 \frac{Y_{i1}}{Y_i} \frac{Y_{i2}}{Y_i} \frac{Y_i - d_i}{Y_i - 1} d_i}}, \tag{2}$$

where the weights w_{i1} are all equal to 1 in the log-rank test. It has been well discussed in the literature that the asymptotic null distribution of U is standard normal under some regularity conditions [25].

For the second stage of the two-stage additive testing procedure, we propose a new weighted log-rank test for detecting a possible crossing pattern of the two hazard rate functions. To be more specific, for $j = 1, 2$, let F_j and G_j represent the cumulative distribution functions (cdf) of the event time and the censoring time, respectively, of the jth group, and S_j and L_j represent the survival functions of the event time and the censoring time, respectively. Then,

$$S_j(t) = 1 - F_j(t), \quad L_j(t) = 1 - G_j(t), \quad \text{for } t \in [0, \mathcal{T}].$$

Under H_0 in (1), we have $S_1(t) = S_2(t) = S(t)$ and $F_1(t) = F_2(t) = F(t)$, for any $t \in [0, \mathcal{T}]$. Then, the new weighted log-rank test statistic has the same expression as that of U in (2), except that the weight at time t is defined to be

$$w_2(t) = -1 + c(t - t_D), \tag{3}$$

where $c \leq 0$ is a constant and t_D is the largest observed event time among all subjects in the pooled sample. The new weighting function is shown in Figure 1, from which it can be seen that it is a linear function that changes signs at $t = t_D + 1/c$. Because of this property of the weighting function $w_2(t)$, early differences between the two estimated hazard rate functions would be avoided to be mostly cancelled out by late differences in the related weighted log-rank test statistic in cases when the two hazard curves cross each other. Thus, the resulting weighted log-rank test could detect a potential crossing pattern of the two hazard rate functions.

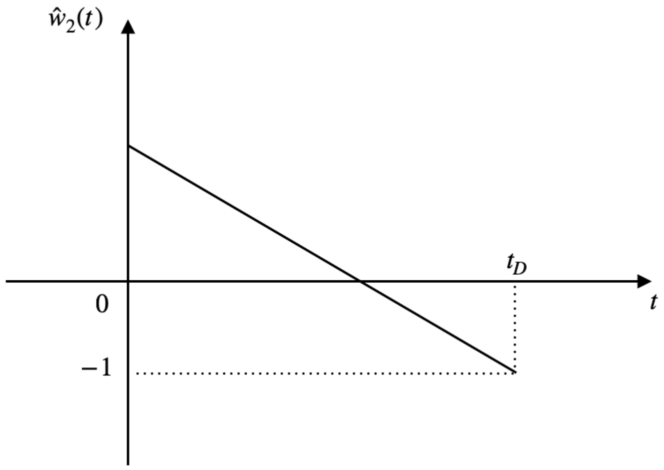

Figure 1. Proposed weighting function used in the weighted log-rank test for a Stage-II test in the two-stage additive testing procedure.

To calculate the overall p-value of the two-stage additive testing procedure properly, [26] suggested that the test statistics used in its two stages should be asymptotically independent of each other. To make the new weighted log-rank test statistic with the weighting function $w_2(t)$ in (3) uncorrelated with the Stage-I test statistic U defined in (2), it can be checked that c in (3) should be estimated by

$$\widehat{c} = \frac{\sum_{i=1}^{D} \frac{\widehat{L}_1(t_i)\widehat{L}_2(t_i)}{(n_1/n)\widehat{L}_1(t_i)+(n_2/n)\widehat{L}_2(t_i)} \Delta \widehat{S}(t_i)}{\sum_{i=1}^{D} (t_i - t_D) \frac{\widehat{L}_1(t_i)\widehat{L}_2(t_i)}{(n_1/n)\widehat{L}_1(t_i)+(n_2/n)\widehat{L}_2(t_i)} \Delta \widehat{S}(t_i)}, \tag{4}$$

where $\widehat{L}_j(t)$, for $j = 1, 2$, are the Kaplan-Meier estimates of the survival functions of the censoring times of the two groups, and $\widehat{S}(t)$ is the Kaplan-Meier estimate of the survival function of the event times computed from the pooled sample. Furthermore, the resulting test statistic used in the second stage becomes

$$V = \frac{\sum_{i=1}^{D} \widehat{w}_2(t_i) \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^{D} \widehat{w}_2^2(t_i) \frac{Y_{i1}}{Y_i} \frac{Y_{i2}}{Y_i} \frac{Y_i - d_i}{Y_i - 1} d_i}}, \tag{5}$$

where $\widehat{w}_2(t_i)$ is defined in (3) after c is replaced by $\widehat{c}$ in (4). In the next section, it will be shown that U and V are indeed asymptotically independent of each other under H_0 and some regularity conditions.

From (5), it can be seen that our proposed test statistic V for detecting the crossing difference between the two hazard curves avoids the maximization with respect to the unknown crossing point that many existing methods require [cf., 24]. That is because $\widehat{c}$ is a data-driven constant that can be adjusted automatically by the observed data to meet the requirement of asymptotic independence between the test statistics in the two stages. See the proof of Theorem 2 in Appendix B for details. Because of this property, it will be shown in the next section that its null distribution would be asymptotically normal under some regularity conditions, instead of the complicated bimodal asymptotic distributions that many existing test statistics for comparing two potentially crossing hazard curves have [cf., 23]. This asymptotic normality property makes the calculation of the p-value of the test using V much easier, and the test becomes more effective as well because i) the bootstrap procedure that is routinely used for computing the p-value related to a bimodal asymptotic distribution can be avoided, and ii) the boundary problem of the maximization procedure mentioned above that the crossing point cannot be in the boundary regions of the study time peroid $[0, \mathcal{T}]$ is avoided as well. Numerical results presented in Section 4 will confirm these conclusions.

2.3. Calculation of the p-value for the proposed two-stage additive test

As discussed in Subsection 2.1, the two-stage additive testing procedure rejects H_0 if and only if the Stage-I test rejects H_0 or the Stage-I test fails to reject H_0 but the Stage-II test rejects H_0. It fails to reject H_0 if and only if both the Stage-I and Stage-II tests fail to reject H_0. Let α_1, α_2, and α be the significance levels of the Stage-I test, the Stage-II test, and the entire two-stage additive test, respectively. Then, based on the asymptotic independence between the test statistics U and V used in the two stages that will be confirmed in Section 3 below, it can be checked that the following equation is asymptotically valid:

$$\alpha_1 + \alpha_2(1 - \alpha_1) = \alpha. \tag{6}$$

By this result, Sheng and Qiu [26] defined the overall p-value of the two-stage additive testing procedure to be

$$p\text{-value}_{SQ} = \begin{cases} p_1, & \text{if } p_1 \leq \alpha_1 \\ \alpha_1 + p_2(1 - \alpha_1), & \text{otherwise,} \end{cases} \tag{7}$$

where p_1 and p_2 denoted the p-values of the Stage-I and Stage-II tests, respectively. The quantity p-value$_{SQ}$ in (7) depends on α_1. [26] suggested choosing $\alpha_1 = \alpha_2 = 1 - \sqrt{1 - \alpha}$ by the result (6) in cases when there is no prior information about the crossing pattern of the two hazard curves. This selection, however, treats the two stages equally, which may result in a less effective testing procedure.

When the two test statistics used in a two-stage additive testing procedure are independent, another popular method to compute the overall p-value of the two-stage additive test is the Fisher-test method [4,7]. Under H_0 in (1), both p_1 and p_2 would follow a uniform distribution on [0,1]. Therefore, both $-2\log(p_1)$ and $-2\log(p_2)$ would follow a chi-square distribution with degrees of freedom 2. So, $-2\log(p_1 p_2)$ would follow a chi-square distribution with degrees of freedom 4, and the overall p-value of the two-stage additive test can be defined to be

$$p\text{-value}_F = H[-2\log(p_1 p_2)], \tag{8}$$

where $H(\cdot)$ is the survival function of the chi-square distribution with degrees of freedom 4.

Based on a large numerical study, [4] has shown that the Fisher-test method would be more robust than the method by [26] in cases when both p_1 and p_2 are small, in the sense that the two-stage additive test with its p-value calculated by the former method would be more powerful than the test with its p-value calculated by the later method in such cases. In other cases considered in their numerical study, the two-stage additive test with its p-value calculated by the method of [26] could be more robust. To make use of the strength of both methods, we suggest calculating the overall p-value of the two-stage additive testing procedure by the following formula:

$$\begin{aligned} p\text{-value} \;=\; & \min\Big\{\Big[p\text{-value}_{SQ(\alpha_1=0)} + p\text{-value}_{SQ(2\alpha_1=\alpha_2)} + p\text{-value}_{SQ(\alpha_1=\alpha_2)} + \\ & p\text{-value}_{SQ(\alpha_1=2\alpha_2)} + p\text{-value}_{SQ(\alpha_1=\alpha)}\Big] \Big/ (5c_1), p\text{-value}_F\Big\} \Big/ c_2, \end{aligned} \tag{9}$$

where $c_1 > 0$ and $c_2 > 0$ are two constants chosen such that the type-I error probability of the two-stage additive testing procedure is the pre-specified value α. In Expression (9), we first calculate 5 p-values by the method (7) in cases when α_1 and α_2 are chosen such that Equation (6) holds and i) $\alpha_1 = 0$, ii) $2\alpha_1 = \alpha_2$, iii) $\alpha_1 = \alpha_2$, iv) $\alpha_1 = 2\alpha_2$, and v) $\alpha_1 = \alpha$, respectively. These five cases are considered to accommodate major crossing and non-crossing patterns of the two hazard curves. Under H_0, each of the five p-values would have a uniform distribution on [0,1]. Their average, however, would not usually be uniformly distributed [cf., 11]. So, the constant c_1 is chosen such that their average divided by c_1, which is the first element in "$\min\{\cdot,\cdot\}$" of (9), would have the property that the event of "the first element is less than or equal to α" has the probablity of α under H_0. As pointed out earlier, the second element in "$\min\{\cdot,\cdot\}$" of (9) would have a uniform distribution on [0,1] under H_0. Then, our defined overall p-value of the two-stage additive testing procedure is the minimum of the two elements in "$\min\{\cdot,\cdot\}$" of (9), and the adjustment constant c_2 is used to make sure that the type-I error probability of the test is α. It has been confirmed numerically that for all α values in $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.2\}$,

$$c_1 = 1.37, \qquad c_2 = 0.76.$$

In (9), instead of setting $\alpha_1 = \alpha_2$ as done in [24], we have considered five cases when α_1 changes from 0 to α when using the method (7) to compute the overall p-value of the two-stage additive testing procedure, which represent different degrees of importance of the Stage-I test in calculating the overall p-value by the method (7). The proposed overall p-value is the minimum of the average of the five p-values computed by the method (7) and the p-value computed by the method (8), after proper adjustments made by the two constants c_1 and c_2. So, by using this approach, major crossing and non-crossing patterns of the two hazard curves have been accommodated in calculating the overall p-value by the method (7), and the observed data have been used to determine whether the method (7) or the method (8) should be used in calculating the overall p-value of the two-stage additive testing procedure. Instead of the average of the five values of $p\text{-value}_{SQ}$ used in (9), we have also considered using their minimum. Based on a large numerical study, it turns out that the method using the average would perform better in most cases considered.

3. Statistical Properties

In this section, we derive some statistical properties of the test statistics U and V (cf., (2) and (5)) used in the proposed two-stage additive testing procedure.

Theorem 1 *For $j = 1, 2$, assume that the event time in the jth group has the cdf F_j with a continuous probability density function (pdf), the censoring time has the cdf G_j, observations in the treatment and control groups are independent of each other, and the censoring times are independent of the event times in each group. Then, under H_0 in (1), the asymptotic null distribution of V is $N(0, 1)$.*

Theorem 2 *Under the assumptions in Theorem 1, the two statistics U and V defined in (2) and (5) are asymptotically independent of each other.*

Proofs of Theorems 1 and 2 are given in Appendix.

4. Simulation Study

In this section, we evaluate the numerical performance of the proposed two-stage additive testing procedure discussed in Sections 2 and 3 by Monte Carlo simulations. First, we investigate the finite-sample distributional properties of the test statistics U and V defined in (2) and (5). For this simulation, we assume that the treatment and control groups have the same hazard rate functions 1 (i.e., $h_0(t) = h_1(t) = 1$, for all t). We consider the sample size of both groups being 100 (i.e., $n_1 = n_2 = 100$), and generate the censoring times from a uniform distribution on the interval $[0, 1.6]$. The procedure is repeated for 5,000 times, from which 5,000 values of U and V are computed. The two plots in the first row of Figure 2 show the density histograms of the 5,000 values of U and the 5,000 values of V, respectively, where the solid curve in each plot is the density curve of the standard normal distribution. It can be seen from these two plots that both U and V follow approximately the standard normal distribution under the null hypothesis H_0. To check the asymptotic independence between U and V, we can compare the joint density histogram of the 5,000 values of (U, V) and their joint density histogram constructed under the assumption that U and V are independent of each other. Under the assumption of independence, the joint density of (U, V) equals the product of two individual densities of U and V. The two plots in the second row of Figure 2 show the joint density histograms of (U, V) with and without the assumption of independence, respectively. It can be seen that the two joint density histograms are almost identical, which is consistent with the result in Theorem 2 that U and V are asymptotically independent under the null hypothesis H_0.

Next, we evaluate the numerical performance of the proposed two-stage additive testing procedure, denoted as NP representing "new procedure", in comparison with some existng competing methods. In the simulation study, the sample sizes of both the treatment and control groups are fixed at 100, and the following 7 cases under 3 different censoring schemes are considered, where the censoring times are generated from the uniform distributions on the intervals [0, 1], [0, 1.6] and [0, 2.6], respectively, in the three censoring schemes. Under each censoring scheme, the hazard rate function of the control group is set to be $h_0(t) = 1$, and that of the treatment group is set to be $h_1(t) = 1$ in Case 1 and $h_1(t) = a(t - b) + 1$ in Cases 2-7, where a is the slope taking the values of 2.0, 2.0, 2.0, 1.2, 1.2, and 1.2, and b is the crossing time taking the values of 0.2, 0.3, 0.4, 0.4, 0.5, and 0.6, respectively, in Cases 2-7. The two hazard rate functions in Cases 1-7 are shown in Figure 3.

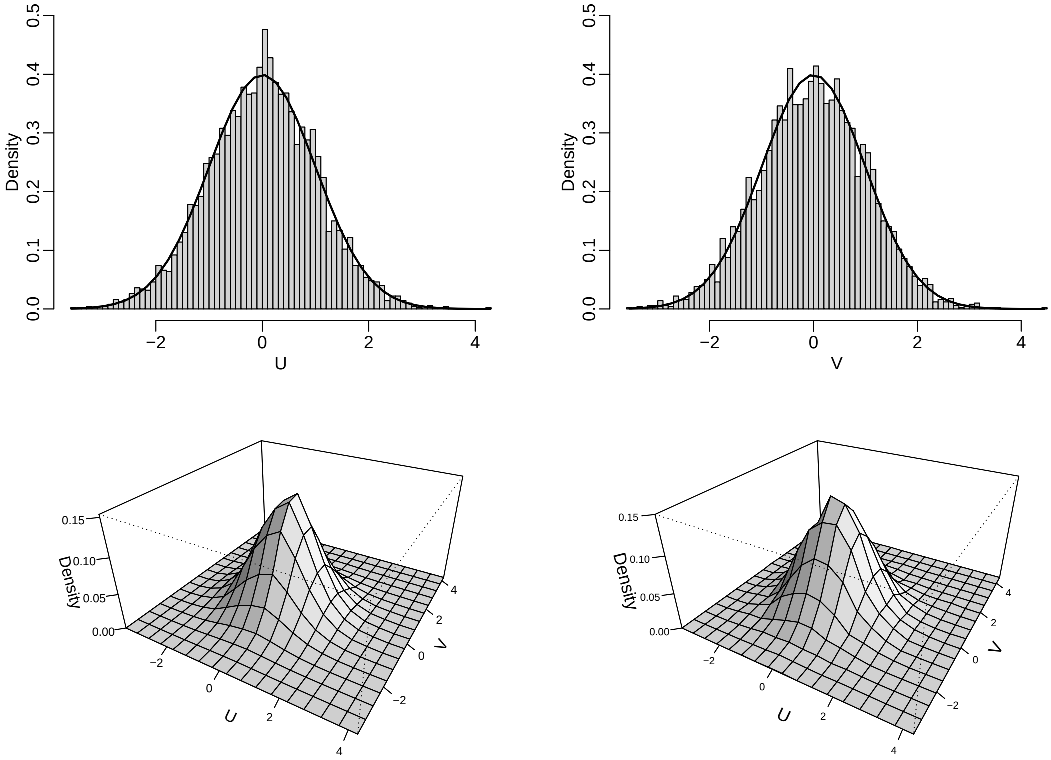

Figure 2. Individual density histograms of U and V based on 5,000 replicated simulations (left and right panels in the first row), and their joint density histograms with and without the assumption of independence between U and V (left and right panels in the second row).

Based on the comparative studies in [6,17], the two-stage test TS suggested by [24], the tests KONGC and KONGL by [10] that are based on the Pearson's chi-squared test and the log-likelihood ratio test, respectively, the MDIR test by [5] that combines multiple weighted log-rank tests, and the MAXC test by [16] that combines multiple Fleming-Harrington weighted log-rank tests [8] have good overall performance compared to many other competing methods. Therefore, in this paper we evaluate the numerical performance of NP in comparison with these alternative methods that were designed for comparing two potentially crossing hazard rate curves, plus the traditional log-rank (LR) test and Peto-Peto (PP) test. The LR and PP tests are constructed under the Cox proportional hazard assumption. Thus, they are powerful in cases when two hazard curves are different but do not cross each other. In comparison, the TS, KONPC, KONPL, MDIR, MAXC, and NP tests are constructed for detecting arbitrary difference between two hazard rate curves, including the ones with crossing patterns.

In our simulation study, the overall significance level α of each method is fixed at 0.05, and all results are based on 1,000 replications. For the TS test, the bootstrap sample size for computing its p-value is fixed at 1,000. For the KONPC, KONPL and MDIR tests, the number of permutations for computing their p-values is also set to be 1,000. For the MDIR test, two different versions considering two and four directions, denoted as MDIR_2 and MDIR_4, respectively, are considered as suggested in [5]. For

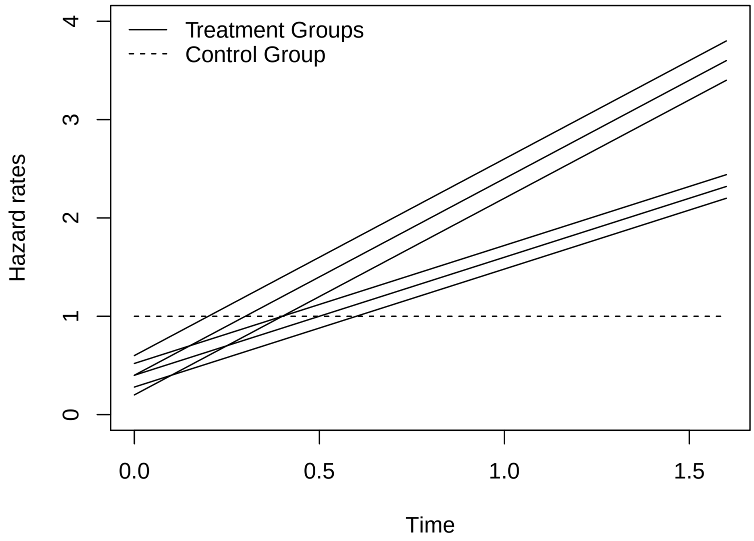

Figure 3. The dashed line represents $h_1(t) = h_0(t) = 1$ in Case 1, and the solid lines denote $h_1(t)$ in Cases 2-7.

the MAXC test, different Fleming-Harrington weighted log-rank tests are combined in the way as suggested in [16]. For the NP test, its Stage-I test (i.e., the one using the test statistic U in (2)) is denoted as LR and the Stage-II test (i.e., the one using the test statistic V in (5)) is denoted as WLR. When the overall p-value of the NP test is computed by (7) with $2\alpha_1 = \alpha_2$, $\alpha_1 = \alpha_2$, or $\alpha_1 = 2\alpha_2$, the related NP test is denoted as NPSQ($2\alpha_1 = \alpha_2$), NPSQ($\alpha_1 = \alpha_2$), and NPSQ($\alpha_1 = 2\alpha_2$), respectively. When the overall p-value of the NP test is computed by (8), the NP test is denoted as NPF. The NP test with its overall p-value computed by (9) is denoted as NPSQF. These are considered here to demonstrate the overall strength of NPSQF in comparison with its variants.

The censoring rates of the control and treatment groups in the seven cases described above under the three different censoring schemes are presented in Table 1. From the table, it can be seen that the censoring rates are between 58%-75% under the censoring scheme 1, between 40%-56% under the censoring scheme 2, and between 25%-37% under the censoring scheme 3. Thus, the three censoring schemes can represent the high, medium and low censoring levels, respectively.

Table 2 tabulates the crossing patterns in different simulation settings under Censoring Schemes I-III. In the simulation study, two different slopes for the treatment hazard rate function are considered (cf., Figure 3). For each slope, three different crossing point locations are considered. However, a crossing point can be considered as early or late also depends on the censoring rates. Under all three censoring schemes, crossing points in various different cases are roughly classified as Early, Middle, and Late in Table 2 for convenience of discussions later.

The empirical sizes and powers of the related testing methods are presented in Tables 3-5 under the censoring schemes I-III, respectively, where the sizes of the tests are given in the first columns of the tables corresponding to Case 1 (i.e., $h_0(t) = h_2(t) = 1$).

Table 1. Censoring rates in the seven cases under the three censoring schemes considered in the simulation study.

	Censoring Scheme I		Censoring Scheme II		Censoring Scheme III	
Cases	Control	Treatment	Control	Treatment	Control	Treatment
1	0.632	0.632	0.499	0.498	0.357	0.355
2	0.632	0.587	0.499	0.403	0.357	0.250
3	0.632	0.634	0.499	0.441	0.357	0.275
4	0.632	0.687	0.499	0.487	0.357	0.305
5	0.632	0.664	0.499	0.487	0.357	0.312
6	0.632	0.698	0.499	0.519	0.357	0.335
7	0.632	0.734	0.499	0.556	0.357	0.362

Table 2. Crossing patterns in various cases considered under Censoring Schemes I-III.

			Censoring Scheme		
Cases	Slope	Crossing Time	I	II	III
2	2.0	0.2	Middle	Early	Early
3	2.0	0.3	Late	Middle	Early
4	2.0	0.4	Late	Late	Middle
5	1.2	0.4	Middle	Early	Early
6	1.2	0.5	Late	Middle	Early
7	1.2	0.6	Late	Late	Middle

In each table, the three largest power values in each column corresponding to Cases 2-7 are highlighted by bold numbers. From these tables, we can have the following conclusions. First, the sizes of all tests are close to the nominal significance level of $\alpha = 0.05$. Second, compared to the existing tests including LR, PP, TS, KONPC, KONPL, MDIR_2, MDIR_4 and MAXC tests, our proposed method NPSQF has larger power in all cases considered, except a small number of cases when the two hazard curves cross early or late in which the performance of NPSQF is close to the best ones of the existing tests. Third, the traditional tests LR and PP perform poorly in most cases considered. Fourth, MDIR_2 and MAXC perform well in some cases when the crossing time is small (e.g., Case 2 in Tables 4 and 5).

Next we focus on the performances of the various versions of NP as presented in Tables 3-5. First, NPF is more powerful than NPSQ when both of the Stage-I test (i.e., LR) and the Stage-II test (i.e., WLR) have relatively large powers to detect the crossing difference between the two hazard curves, which usually happens when the two hazard curves cross at an early or late time (e.g., Cases 4 and 7 in Table 3, Cases 2 and 7 in Table 4, and Cases 2 and 3 in Table 5). Second, among the five versions of NPSQ (note: LR is the same as NPSQ with $\alpha_1 = \alpha$ and WLR is same as NPSQ with $\alpha_1 = 0$), the one with a smaller α_1 value would perform better in most cases considered, because the Stage-II test (WLR) would be more focused in such a two-stage test which is favorable to compare two crossing hazard curves. However, in some cases when the crossing point is small or large (e.g., Case 7 in Table 3), the above conclusion may not be true because the crossing pattern is not obvious in the observed data in such cases. Third, NPSQF performs well in all cases considered. Therefore, NPSQF is recommended if there is no prior information about the crossing pattern of the two hazard curves.

We also conduct some simulations in cases when $h_1(t)$ is non-monotonic linear, cubic polynomial, and exponential under the censoring scheme III and the same simulation setups as before. The figure of $h_0(t)$ and $h_1(t)$ are shown in Figure 5 and the results are presented in Table 7 in Appendix C. From the results, it can be seen that our proposed methods WLR, NPSQ($2\alpha_1 = \alpha_2$), NPSQ($\alpha_1 = \alpha_2$), NPSQ($\alpha_1 = 2\alpha_2$), NPF, and NPSQF have larger powers, compared to alternative methods LR, PP, TS, KONPC, KONPL, MDIR_2, MDIR_4, and MAXC in all cases considered.

Table 3. Sizes and powers of different methods for comparing two hazard curves in various cases under the censoring scheme I.

	Cases						
Methods	1	2	3	4	5	6	7
PP	0.047	0.053	0.084	0.322	0.141	0.305	**0.552**
TS	0.046	0.252	0.289	0.497	0.166	0.275	0.460
KONPC	0.054	0.157	0.131	0.275	0.119	0.242	0.418
KONPL	0.054	0.155	0.135	0.282	0.118	0.240	0.421
MDIR_2	0.048	0.283	0.326	0.552	0.187	**0.326**	0.543
MDIR_4	0.049	0.219	0.240	0.454	0.149	0.262	0.430
MAXC	0.039	0.198	0.125	0.261	0.097	0.219	0.429
LR	0.042	0.079	0.050	0.203	0.093	0.234	0.476
WLR	0.049	**0.330**	**0.404**	**0.568**	**0.207**	0.258	0.320
NPSQ($2\alpha_1 = \alpha_2$)	0.050	**0.292**	**0.348**	0.541	0.191	0.293	0.468
NPSQ($\alpha_1 = \alpha_2$)	0.048	0.262	0.318	0.520	0.183	0.295	0.484
NPSQ($\alpha_1 = 2\alpha_2$)	0.045	0.232	0.274	0.474	0.166	0.292	0.496
NPF	0.054	0.273	0.313	**0.559**	**0.194**	**0.337**	**0.577**
NPSQF	0.051	**0.294**	**0.354**	**0.575**	**0.203**	**0.327**	**0.546**

Table 4. Sizes and powers of different methods for comparing two hazard curves in various cases under the censoring scheme II.

	Cases						
Methods	1	2	3	4	5	6	7
PP	0.048	0.114	0.053	0.176	0.088	0.212	0.447
TS	0.045	0.586	0.621	0.768	0.304	0.404	0.564
KONPC	0.049	0.546	0.467	0.467	0.179	0.216	0.358
KONPL	0.048	0.548	0.462	0.473	0.179	0.218	0.362
MDIR_2	0.044	**0.664**	0.660	0.767	0.330	0.443	0.619
MDIR_4	0.041	0.567	0.553	0.679	0.258	0.339	0.506
MAXC	0.055	0.618	0.445	0.351	0.139	0.181	0.358
LR	0.051	0.323	0.101	0.052	0.051	0.096	0.241
WLR	0.050	0.556	**0.708**	**0.834**	**0.455**	**0.530**	**0.627**
NPSQ($2\alpha_1 = \alpha_2$)	0.045	0.605	**0.673**	**0.799**	**0.382**	**0.485**	0.621
NPSQ($\alpha_1 = \alpha_2$)	0.049	0.601	0.631	0.774	0.341	0.455	0.603
NPSQ($\alpha_1 = 2\alpha_2$)	0.052	0.577	0.583	0.740	0.281	0.403	0.556
NPF	0.510	**0.661**	0.631	0.753	0.322	0.443	**0.628**
NPSQF	0.048	**0.662**	**0.681**	**0.800**	**0.373**	**0.493**	**0.658**

Table 5. Sizes and powers of different methods for comparing two hazard curves in various cases under the censoring scheme III.

Methods	Cases 1	2	3	4	5	6	7
PP	0.054	0.203	0.061	0.096	0.061	0.147	0.330
TS	0.041	0.840	0.877	0.935	0.543	0.645	0.775
KONPC	0.040	0.849	0.822	0.855	0.421	0.469	0.571
KONPL	0.040	0.855	0.821	0.857	0.430	0.474	0.576
MDIR_2	0.045	**0.922**	**0.916**	**0.951**	0.599	0.659	0.805
MDIR_4	0.050	0.848	0.846	0.910	0.481	0.525	0.667
MAXC	0.044	**0.907**	0.806	0.706	0.393	0.324	0.385
LR	0.048	0.633	0.353	0.121	0.098	0.051	0.088
WLR	0.041	0.693	0.867	**0.961**	**0.662**	**0.753**	**0.846**
NPSQ($2\alpha_1 = \alpha_2$)	0.040	0.844	0.892	0.950	**0.605**	**0.718**	**0.818**
NPSQ($\alpha_1 = \alpha_2$)	0.042	0.847	0.865	0.940	0.576	0.673	0.795
NPSQ($\alpha_1 = 2\alpha_2$)	0.045	0.839	0.830	0.918	0.519	0.615	0.754
NPF	0.050	**0.915**	**0.907**	0.933	0.577	0.637	0.787
NPSQF	0.041	0.906	**0.906**	**0.952**	**0.625**	**0.705**	**0.819**

5. A Case Study

In this section, we demonstrate the proposed method using a real dataset from the Veterans' Administration Lung Cancer study discussed in [13] that aimed to compare the effects of a standard therapy (control group) with a test therapy (treatment group) in the treatment of advanced inoperable lung cancer. Among 130 patients under the age of 70 in the study, 67 of them were randomized to the control group and 63 to the treatment group. Time to death for each patient was recorded as the primary outcome measure. There were 5 censored observations in the control group and 4 censored observations in the treatment group. This dataset can be obtained from the *R*-package **survival**. The estimated hazard rate functions of the control and treatment groups, using kernel-based methods [12], are shown in Figure 4. From the figure, it can be seen that the two hazard curves cross around the 100th day after the study started.

Next, the alternative methods LR, PP, TS, KONPC, KONPL, MDIR_2, MDIR_4, MAXC, as well as WLR, NPSQ($2\alpha_1 = \alpha_2$), NPSQ($\alpha_1 = \alpha_2$), NPSQ($\alpha_1 = 2\alpha_2$), NPF, and NPSQF are applied to the dataset to compare the two hazard curves. All these methods are set up in the same way as that in the simulation studies presented in Section 4 with the overall significance level of each method being $\alpha = 0.05$. Their p-values are given in Table 6. From the table, it can be seen that the crossing pattern between the two hazard curves in this example can only be detected by WLR, NPSQ($2\alpha_1 = \alpha_2$), NPSQ($\alpha_1 = \alpha_2$), and NPSQF, although the p-values of TS, KONPC, KONPL, MDIR_2, NPSQ($\alpha_1 = 2\alpha_2$) and NPF are also quite small. It is reasonable in this example that WLR has the smallest p-value since WLR is developed specially for detecting a crossing pattern of the two hazard curves, and the crossing pattern is quite obvious in Figure 4. It can be seen that NPSQF can also detect such a crossing difference between the two hazard curves, while the alternative existing methods LR, PP, TS, KONPC, KONPL, MDIR_2, MDIR_4, and MAXC cannot.

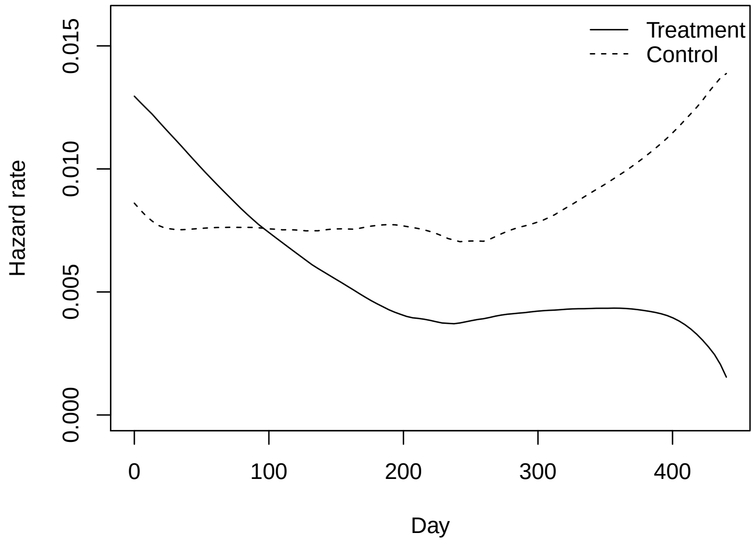

Figure 4. Estimated hazard curves of the treatment and control groups of the Veterans' Administration Lung Cancer study.

Table 6. Calculated p-values of various methods for comparing two hazard curves in the Veterans' Administration Lung Cancer study. The numbers in bold denote the four smallest p-values.

Method	PP	TS	KONPC	KONPL	MDIR_2	MDIR_4	MAXC
p-value	0.329	0.092	0.107	0.106	0.104	0.155	0.452
Method	LR	WLR	NPSQ ($2\alpha_1 = \alpha_2$)	NPSQ ($\alpha_1 = \alpha_2$)	NPSQ ($\alpha_1 = 2\alpha_2$)	NPF	NPSQF
p-value	0.991	**0.023**	**0.040**	**0.048**	0.056	0.072	**0.046**

6. Concluding Remarks

We have presented a new two-stage additive testing procedure to compare two hazard curves that may or may not cross each other. In the new testing procedure, the traditional log-rank test is used in its first stage, and a special weighted log-rank test with a linear weighting function (cf., (3)) is used in its second stage. Compared to the existing tests designed for comparing two potentially crossing hazard curves, the new test used in the second stage avoids the maximization with respect to the unknown crossing point. Thus, its test statistic has the preferable asymptotic normality under the null hypothesis, instead of the complex bimodal null distribution. Consequently, calculation of its p-value becomes more convenient and accurate, and the resulting test becomes more powerful, which has been confirmed by the numerical studies presented in Sections 4 and 5. There are still some issues with the proposed method that need to be addressed in future research. For instance, the current version of our proposed method cannot estimate the crossing point well since it is constructed mainly for comparing the two hazard curves, instead of estimation of the crossing point. It cannot accommodate potential impact of covariates either. In addition, the proposed new method can only handle cases with one crossing point between the two hazard curves. In reality, the two hazard curves could have multiple crossing points. There are applications where we need to compare more than two hazard curves as well. All these research problems will be pursued elsewhere.

Acknowledgments

The authors thank the editor, the associate editor and two referees for many constructive comments and suggestions which improved the quality of the paper greatly.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] Anderson JA, Senthilselvan A. A two-step regression model for hazard functions. Appl Stat. 1982;31:44–51.
[2] Breslow NE, Edler L, Berger J. A two–sample censored-data rank test for acceleration. Biometrics. 1984;40:1049–1062.
[3] Chen Z, Huang H, Qiu P. Comparison of multiple hazard rate functions. Biometrics. 2016;72:39–45.
[4] Chen Z, Huang H, Qiu P. An improved two-stage procedure to compare hazard curves. J Stat Comput Simul. 2017;87:1877–1886.
[5] Ditzhaus M, Friedrich S. More powerful logrank permutation tests for two-sample survival data. J Stat Comput Simul. 2020;90:2209–2227.
[6] Dormuth I, Liu T, Xu J, Yu M, Pauly M, Ditzhaus M. Which test for crossing survival curves? A user's guideline. BMC Med Res Methodol. 2022;22:34.
[7] Fisher RA. Statistical Methods for Research Workers. Edinburgh, Scotland: Oliver and Boyd; 1925.
[8] Fleming TR, Harrington PD. Counting Processes and Survival Analysis. New York: John Wiley & Sons; 1991.
[9] Fleming TR, O'Fallon JR, O'Brien PC, Harrington DP. Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. Biometrics. 1980;36:607–625.
[10] Gorfine M, Schlesinger M, Hsu L. k-sample omnibus non-proportional hazards tests based on right-censored data. Stat Methods Med Res. 2020;29:2830–2850.
[11] Hall P. The distribution of means for samples of size n drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. Biometrika. 1927;19:240–245.
[12] Hess KR, Serachitopol DM, Brown BW. Hazard function estimators: a simulation study. Stat Med. 1999;18(22):3075–3088.
[13] Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. New York: John Wiley & Sons; 2002.
[14] Klein JP, Moeschberger ML. Survival Analysis (2nd edition). New York: Springer-Verlag; 2003.
[15] Lawless JF. Statistical Models and Methods for Lifetime Data (2nd edition). New York: John Wiley & Sons; 2002.
[16] Lee SH. On the versatility of the combination of the weighted log-rank statistics. Comput Stat Data Anal. 2007;51:6557–6564.

[17] Li H, Han D, Hou Y, Chen H, Chen Z. Statistical inference methods for two crossing survival curves: A comparison of methods. PLoS ONE. 2015;10:e0116774.
[18] Lin X, Wang H. A new testing approach for comparing the overall homogeneity of survival curves. Biometrical Journal. 2004;46:489–496.
[19] Liu K, Qiu P, Sheng J. Comparing two crossing hazard rates by Cox proportional hazards modeling. Stat Med. 2007;26:375–391.
[20] Mantel N, Stablein DM. The crossing hazard function problem. Statistician. 1988;37:59–64.
[21] Moreau T, Maccario J, Lellouch J, Huber C. Weighted log rank statistics for comparing two distributions. Biometrika. 1992;79:195–198.
[22] O'Quigley J. On a two-sided test for crossing hazard rates. Statistician. 1994;43:563–569.
[23] O'Quigley J, Pessione F. The problem of a covariate-time qualitative interaction in a survival study. Biometrics. 1991;47:101–115.
[24] Qiu P, Sheng J. A two-stage procedure for comparing hazard rate functions. J R Stat Soc Ser B. 2008;70:191–208.
[25] Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika. 1981;68:316–319.
[26] Sheng J, Qiu P. p-value calculation for multi-stage additive tests. J Stat Comput Simul. 2007;77:1057–1064.

7. Appendices

Appendix A. Proof of Theorem 1

For $j = 1, 2$ and $k = 1,\ 2,\ \dots,\ n_j$, let T_{kj} be the event time of the kth subject in group j with c.d.f. F_j, C_{kj} be the censoring time with c.d.f. G_j, and

$$S_j(s) = 1 - F_j(s),\ L_j(s) = 1 - G_j(s),\ X_{kj} = \min(T_{kj},\ C_{kj}),$$
$$\delta_{kj} = I_{\{T_{kj} < C_{kj}\}},\ \pi_j(s) = P(X_{kj} > s) = S_j(s)L_j(s).$$

In the above expression for $\pi_j(s)$, we have made a conventional assumption that the event times T_{kj} and censoring times C_{kj} are independent of each other. Also, under H_0, $S_1 = S_2 = S$.

Let $\boldsymbol{w} = (w_1, w_2, \cdots, w_D)^T$ denote a vector of weights used in either U or V. Then, we define the test statistic $Z(\boldsymbol{w})$ and its estimated variance $\widehat{\sigma}(\boldsymbol{w})$ as follows:

$$Z(\boldsymbol{w}) = h \sum_{i=1}^{D} w_i \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right),\ \widehat{\sigma}(\boldsymbol{w}) = h^2 \sum_{i=1}^{D} w_i^2 \frac{Y_{i1}}{Y_i} \frac{Y_{i2}}{Y_i} \frac{Y_i - d_i}{Y_i - 1},$$

where $h = \sqrt{n/(n_1 n_2)}$. We also define the following counting processes: for j=1,2,

$$\bar{Y}_j(s) = \sum_{k=1}^{n_j} I_{\{X_{kj} \geq s\}},\ \bar{N}_j(s) = \sum_{k=1}^{n_j} I_{\{X_{kj} \leq s, \delta_{kj}=1\}}.$$

For group j, $\bar{Y}_j(s)$ defined above is the at-risk process which is left continuous, and $\bar{N}_j(s)$ is the event process which is right continous. Let $\widehat{S}(s)$ be the Kaplan-Meier estimator of the survival function $S(s)$, and $W(s)$ be a bounded predictable function of $\widehat{S}(s-)$ having the property that $(W(t_1), W(t_2), \cdots, W(t_D))^T = \boldsymbol{w}$. Then, $Z(\boldsymbol{w})$ can be written as

$$Z(\boldsymbol{w}) = h \int_0^u W(s) \frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}_1(s)+\bar{Y}_2(s)} \left\{ \frac{d\bar{N}_1(s)}{\bar{Y}_1(s)} - \frac{d\bar{N}_2(s)}{\bar{Y}_2(s)} \right\}, \tag{A.1}$$

where $u = \min\{s : \min(\pi_1(s), \pi_2(s)) = 0\}$.

We can explore the properties of V by regarding it as a statistic of the class K discussed in Section 3.3 of Fleming and Harrington (1991), where K is defined as

$$K(s) = hW(s) \frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}_1(s)+\bar{Y}_2(s)}. \tag{A.2}$$

Next, let us define $\widehat{\pi}(s) = (n_1+n_2)^{-1}(\bar{Y}_1(s)+\bar{Y}_2(s))$ to be the pooled sample estimator of $p_1\pi_1(s) + p_2\pi_2(s)$, where $\pi_j(s)$ is the proportions of subjects who are still at risk at time s, for $j = 1, 2$, $\pi(s)$ is the proportion of subjects in the pooled sample who are still at risk at time s, $p_1 = n_1/(n_1 + n_2)$ and $p_2 = n_2/(n_1 + n_2)$. Then, $\widehat{\pi}(s)$, $\widehat{\pi}_1(s) = \bar{Y}_1(s)/n_1$ and $\widehat{\pi}_2(s) = \bar{Y}_2(s)/n_2$ are all consistent estimators of $\pi(s)$, $\pi_1(s)$ and $\pi_2(s)$, respectively. The asmptotic normality of V can be confirmed by checking the three regularity conditions of Corollary 7.2.1 in Fleming and Harrington (1991) below.

The first regularity condition of Corollary 7.2.1 in Fleming and Harrington (1991) is that: for $j = 1, 2$,

$$\frac{K^2(s)}{\bar{Y}_j(s)} \xrightarrow{p} \xi_j(s), \text{ as } n \to \infty,$$

where the convergence is uniform on $[0, t]$ for any $t \in I = \{t : \pi_1(t)\pi_2(t) > 0\}$, $\xi_j(s)$ is a nonnegative, left-continuous function with right-hand limits such that $\xi_j(t) < \infty$, $\xi_j^+(s)$ has bounded variation on each closed subinterval of I, and $\xi_j(s) = 0$ for any $t \notin I$. This condition is satisfied here if we define $\xi_j(s) = W^2(s)\frac{p_1p_2\pi_1^2(s)\pi_2^2(s)}{p_j\pi_j(p_1\pi_1(s)+p_2\pi_2(s))^2}$ based on Equation (A.2).

To discuss the second regularity condition, let us define

$$\sigma^2(\boldsymbol{w}) = \int_0^u [h_1(s) + h_2(s)][1 - \Delta\Lambda(s)]d\Lambda(s), \text{ for } u \notin I,$$

where $\Lambda(s) = \int_0^s \{1 - F(s-)\}^{-1}dF(s)$ is the common cumulative hazard function of the event time under H_0 which is continous because the common c.d.f. is assumed to have a continous density function. Then, the second regularity condition of Corollary 7.2.1 in Fleming and Harrington (1991) is that for any $\epsilon > 0$,

$$\lim_{t\uparrow u} \limsup_{n\to\infty} P\left\{ \int_t^u K^2 \frac{\bar{Y}_1 + \bar{Y}_2}{\bar{Y}_1\bar{Y}_2} d\Lambda > \epsilon \right\}, \text{ for any } \epsilon > 0. \tag{A.3}$$

To check (A.3), we first notice that

$$\begin{aligned}\sigma^2(\boldsymbol{w}) &= \int_0^u W^2(s)\frac{\pi_1(s)\pi_2(s)}{p_1\pi_1(s)+p_2\pi_2(s)}(1-\Delta\Lambda(s))d\Lambda(s)\\ &= \int_0^u W^2(s)\frac{\pi_1(s)\pi_2(s)}{p_1\pi_1(s)+p_2\pi_2(s)}\frac{1}{S(s)}dF(s)\\ &= \int_0^u W^2(s)\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s).\end{aligned} \tag{A.4}$$

Equation (A.3) is valid because

$$\begin{aligned}&\lim_{t\uparrow u}\limsup_{n\to\infty} P\left\{\int_t^u K^2\frac{\bar{Y}_1+\bar{Y}_2}{\bar{Y}_1\bar{Y}_2}d\Lambda>\epsilon\right\}\\ =&\lim_{t\uparrow u}\limsup_{n\to\infty} P\left\{\int_t^u W^2(s)\frac{\widehat{\pi}_1(s)\widehat{\pi}_2(s)}{\widehat{\pi}(s)}d\Lambda>\epsilon\right\}\\ =&0.\end{aligned}$$

Finally, the third regularity condition of Corollary 7.2.1 in Fleming and Harrington (1991) is that for any $u<\infty$ and $\epsilon>0$,

$$\lim_{n\to\infty} P\left\{\int_u^\infty K^2\frac{\bar{Y}_1+\bar{Y}_2}{\bar{Y}_1\bar{Y}_2}d\Lambda>\epsilon\right\}=0.$$

This regularity condition is valid here because

$$\begin{aligned}&\lim_{n\to\infty} P\left\{\int_u^\infty K^2\frac{\bar{Y}_1+\bar{Y}_2}{\bar{Y}_1\bar{Y}_2}d\Lambda>\epsilon\right\}\\ =&\lim_{n\to\infty} P\left\{\int_u^\infty W^2(s)\frac{\widehat{\pi}_1(s)\widehat{\pi}_2(s)}{\widehat{\pi}(s)}d\Lambda>\epsilon\right\}\\ =&0.\end{aligned}$$

Therefore, by Corollary 7.2.1 in Fleming and Harrington (1991), we have

$$Z(\boldsymbol{w})/\sigma(\boldsymbol{w}) \xrightarrow{D} N(0,1), \text{ as } n\to\infty. \tag{A.5}$$

In addition, by Corollary 7.2.1 in Fleming and Harrington (1991), we also have $\widehat{\sigma}^2(\boldsymbol{w}_2) \xrightarrow{P} \sigma^2(\boldsymbol{w}_2)$ where $\boldsymbol{w}_2$ is the vector of weights used in V. So, by the Slutsky's theorem, we have

$$Z(\boldsymbol{w}_2)/\widehat{\sigma}(\boldsymbol{w}_2) \xrightarrow{D} N(0,1).$$

Thus, the conclusion in Theorem 1 is true.

Appendix B. Proof of Theorem 2

First, we can always find a constant $w_0>0$ such that

$$\sigma^2(\boldsymbol{w}_1)=\sigma^2(\boldsymbol{w}_2), \tag{A.6}$$

where $\boldsymbol{w}_1 = (W_1(t_1), \cdots, W_1(t_D))^T = w_0 \mathbf{1}_D$ and $\boldsymbol{w}_2 = (W_2(t_1), \cdots, W_2(t_D))^T$. Define

$$U^* = \frac{Z(\boldsymbol{w}_1)}{\sigma(\boldsymbol{w}_1)}, \ \ V^* = \frac{Z(\boldsymbol{w}_2)}{\sigma(\boldsymbol{w}_2)}.$$

The asymptotic independence between U^* and V^* can be obtained if we can show that

$$\begin{pmatrix} U^* \\ V^* \end{pmatrix} \xrightarrow{D} N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{I}_2\right), \ \text{as } n \to \infty. \tag{A.7}$$

To prove (A.7), let us consider the following linear combination

$$aU^* + bV^* = \frac{Z(a\boldsymbol{w}_1 + b\boldsymbol{w}_2)}{\sigma(\boldsymbol{w}_1)}, \tag{A.8}$$

where a and b are two arbitrary constants and the equation is based on Equation (A.6). Similar to Equation (A.5), by Corollary 7.2.1 in Fleming and Harrington (1991), we have

$$Z(a\boldsymbol{w}_1 + b\boldsymbol{w}_2)/\sigma(a\boldsymbol{w}_1 + b\boldsymbol{w}_2) \xrightarrow{D} N(0,1), \ \text{as } n \to \infty. \tag{A.9}$$

Thus, if we can prove

$$\frac{\sigma^2(a\boldsymbol{w}_1 + b\boldsymbol{w}_2)}{\sigma^2(\boldsymbol{w}_1)} \xrightarrow{Pr} a^2 + b^2, \ \text{as } n \to \infty, \tag{A.10}$$

then by the Slutsky's theorem and the results in (A.8) and (A.9), we have

$$aU + bV \xrightarrow{D} N(0, a^2 + b^2), \ \text{as } n \to \infty.$$

To prove (A.10), by the results in (A.4) and (A.6), we have

$$\begin{aligned}
&\frac{\sigma^2(a\boldsymbol{w}_1 + b\boldsymbol{w}_2)}{\sigma^2(\boldsymbol{w}_1)} \\
=&\frac{1}{\sigma^2(\boldsymbol{w}_1)} \int_0^u [aW_1(s) + bW_2(s)]^2 \frac{L_1(s)L_2(s)}{p_1 L_1(s) + p_2 L_2(s)} dF(s) \\
=&\frac{1}{\sigma^2(\boldsymbol{w}_1)} \int_0^u [a^2 W_1^2(s) + b^2 W_2^2(s) + 2abW_1(s)W_2(s)] \frac{L_1(s)L_2(s)}{p_1 L_1(s) + p_2 L_2(s)} dF(s) \\
=&\frac{1}{\sigma^2(\boldsymbol{w}_1)} [(a^2 + b^2)\sigma^2(\boldsymbol{w}_1) + 2abw_0 \int_0^u W_2(s) \frac{L_1(s)L_2(s)}{p_1 L_1(s) + p_2 L_2(s)} dF(s)].
\end{aligned} \tag{A.11}$$

Therefore, by (A.11), if we can prove that

$$\int_0^u W_2(s) \frac{L_1(s)L_2(s)}{p_1 L_1(s) + p_2 L_2(s)} dF(s) \xrightarrow{Pr} 0, \ \text{as } n \to \infty,$$

then Equation (A.10) will follow. To this end, first we notice that

$$
\begin{aligned}
&\int_0^u W_2(s)\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s) \\
&= -\int_0^u \frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s) + \widehat{c}\int_0^u (s-u)\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s),
\end{aligned} \tag{A.12}
$$

where

$$
\widehat{c} = \frac{\sum_{i=1}^D \frac{\widehat{L}_1(t_i)\widehat{L}_2(t_i)}{(n_1/n)\widehat{L}_1(t_i)+(n_2/n)\widehat{L}_2(t_i)}\Delta\widehat{S}(t_i)}{\sum_{i=1}^D (t_i-t_D)\frac{\widehat{L}_1(t_i)\widehat{L}_2(t_i)}{(n_1/n)\widehat{L}_1(t_i)+(n_2/n)\widehat{L}_2(t_i)}\Delta\widehat{S}(t_i)}.
$$

Next, we want to show that

$$
\widehat{c} \xrightarrow{Pr} k_r, \text{ as } n\to\infty, \tag{A.13}
$$

where

$$
k_r = \frac{\int_0^u \frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s)}{\int_0^u [s-u]\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s)}.
$$

To show (A.13), we have

$$
\begin{aligned}
&\sum_{i=1}^D \frac{\widehat{L}_1(t_i)\widehat{L}_2(t_i)}{(n_1/n)\widehat{L}_1(t_i)+(n_2/n)\widehat{L}_2(t_i)}\Delta\widehat{S}(t_i) \\
=\;& \sum_{i=1}^D \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}\Delta\widehat{S}(s) \\
=\;& \int_0^{t_D} \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}\Delta\widehat{S}(s) \\
=\;& -\int_0^{t_D} \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}\Delta\widehat{F}(s),
\end{aligned} \tag{A.14}
$$

where $\widehat{F}(s) = 1 - \widehat{S}(s)$. From (A.14), we have

$$
\begin{aligned}
&\left| \int_0^{t_D} \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s) + (n_2/n)\widehat{L}_2(s)} d\widehat{F}(s) - \int_0^u \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} dF(s) \right| \\
\leq &\left| \int_0^{t_D} \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s) + (n_2/n)\widehat{L}_2(s} d\widehat{F}(s) - \int_0^u \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s) + (n_2/n)\widehat{L}_2(s)} d\widehat{F}(s) \right| + \\
&\left| \int_0^u \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s) + (n_2/n)\widehat{L}_2(s)} d\widehat{F}(s) - \int_0^u \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} d\widehat{F}(s) \right| + \\
&\left| \int_0^u \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} d\widehat{F}(s) - \int_0^u \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} dF(s) \right| \\
:= &A_1 + A_2 + A_3.
\end{aligned}
$$

The right-hand side of the last inequality has three parts. We evaluate each of them separately. For A_1, we have

$$
\begin{aligned}
&\left| \int_0^{t_D} \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s) + (n_2/n)\widehat{L}_2(s)} d\widehat{F}(s) - \int_0^u \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s) + (n_2/n)\widehat{L}_2(s)} d\widehat{F}(s) \right| \\
= &\left| \int_u^{t_D} \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s) + (n_2/n)\widehat{L}_2(s)} d\widehat{F}(s) \right| \\
\leq &\left| \int_u^{t_D} 1 d\widehat{F}(s) \right| \\
= &|t_D - u| \xrightarrow{Pr} 0, \text{ as } n \to \infty.
\end{aligned}
$$

For A_2, based on the Taylor Polynomial of a function of two variables, we have

$$
\begin{aligned}
&\left| \int_0^u \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s) + (n_2/n)\widehat{L}_2(s)} d\widehat{F}(s) - \int_0^u \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} d\widehat{F}(s) \right| \\
= &\left| \int_0^u [\frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s) + (n_2/n)\widehat{L}_2(s)} - \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)}] d\widehat{F}(s) \right| \\
\leq &\int_0^u \left[\frac{1}{p_2} \sup_s |\widehat{L}_1(s) - L_1(s)| + \frac{1}{p_1} \sup_s |\widehat{L}_2(s) - L_2(s)| \right] d\widehat{F}(s) \xrightarrow{Pr} 0, \text{ as } n \to \infty.
\end{aligned}
$$

Finally, for A_3, we have

$$
\begin{aligned}
&\left| \int_0^u \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} d\widehat{F}(s) - \int_0^u \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} dF(s) \right| \\
= &\left| \int_0^u \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} d[\widehat{F}(s) - F(s)] \right| \xrightarrow{Pr} 0, \text{ as } n \to \infty.
\end{aligned}
$$

Therefore, we have

$$
\left| \int_0^{t_D} \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s) + (n_2/n)\widehat{L}_2(s)} d\widehat{F}(s) - \int_0^u \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} dF(s) \right| \xrightarrow{Pr} 0, \text{ as } n \to \infty
$$

Similarly, we can obtain the results

$$\sum_{i=1}^{D}(t_i-t_D)\frac{\widehat{L}_1(t_i)\widehat{L}_2(t_i)}{(n_1/n)\widehat{L}_1(t_i)+(n_2/n)\widehat{L}_2(t_i)}\Delta\widehat{S}(t_i)=-\int_0^{t_D}(s-u)\frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}d\widehat{F}(s)$$

and

$$\left|\int_0^{t_D}(s-u)\frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}d\widehat{F}(s)-\int_0^{u}(s-u)\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s)\right|\xrightarrow{Pr}0,\ \text{as}\ n\to\infty.$$

Then, Equation (A.13) follows. By (A.12) and (A.13), we have the result in (A.10). Consequently, the result in Equation (A.7) is proved. Therefore, U^* and V^* are asymptotically independent.

Now, the test statistics U and V can be written as

$$U=U^*\frac{\sigma(\boldsymbol{w}_1)}{\widehat{\sigma}(\boldsymbol{w}_1)},\quad V=V^*\frac{\sigma(\boldsymbol{w}_2)}{\widehat{\sigma}(\boldsymbol{w}_2)}.$$

Since $\widehat{\sigma}^2(\boldsymbol{w}_j)\xrightarrow{P}\sigma^2(\boldsymbol{w}_j)$, for $j=1,2$, we have the result that U and V are asymptotically independent.

Appendix C. Additional Simulation Results

Some additional simulation results discussed at the end of Section 4 are given in Table 7 and Figure 5 here.

Table 7. Sizes and powers of different methods for comparing two hazard curves in cases when $h_1(t)$ is non-monotonic linear (Non-Mono), cubic polynomial (Poly), and exponential (Exp) under the censoring scheme III.

	Cases		
Methods	Non-Mono	Poly	Exp
PP	0.192	0.114	0.090
TS	0.615	0.868	0.911
KONPC	0.822	0.721	0.795
KONPL	0.824	0.732	0.802
MDIR_2	0.050	0.080	0.122
MDIR_4	0.859	0.798	0.879
MAXC	0.583	0.571	0.694
LR	0.053	0.083	0.120
WLR	**0.962**	**0.938**	**0.966**
NPSQ($2\alpha_1=\alpha_2$)	**0.944**	**0.922**	**0.956**
NPSQ($\alpha_1=\alpha_2$)	0.928	0.906	0.947
NPSQ($\alpha_1=2\alpha_2$)	0.905	0.882	0.930
NPF	0.916	0.896	0.940
NPSQF	**0.937**	**0.921**	**0.957**

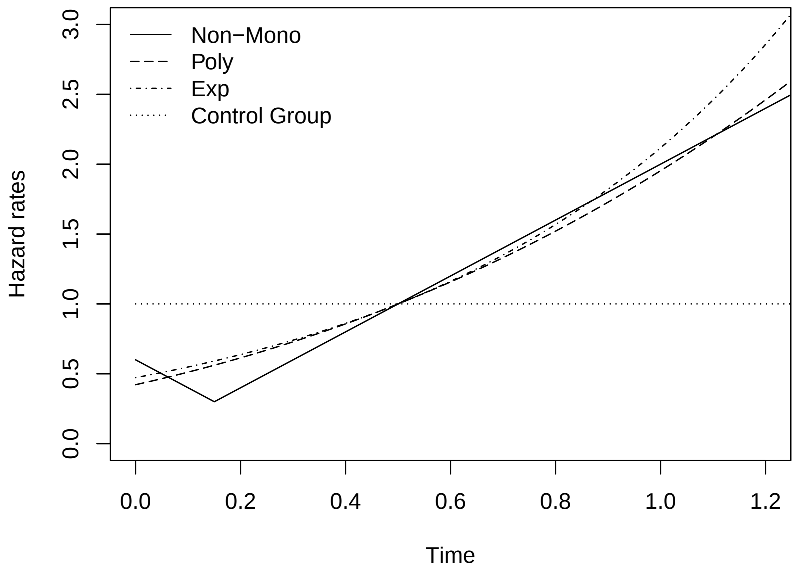

Figure 5. The dotted line represents $h_0(t) = 1$, the solid line denotes $h_1(t) = 2(t-0.15)(1-2I(t \leq 0.15))+0.3$ which is a non-monotonic linear hazard (Non-Mono), the long-dashed line denotes $h_1(t) = (0.5t+0.75)^3$ which is a cubic polynomial hazard (Poly), and the dot-dashed line denotes $h_1(t) = \exp[1.5(t-0.5)]$ which is a exponential hazard (Exp).