Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ORIGINAL ARTICLE

Comparing Two Hazard Curves When There Is a Treatment Time-Lag Effect

Xiaoxi Zhang, Somnath Datta and Peihua Qiu

Department of Biostatistics, University of Florida, Florida, USA

Correspondence
Peihua Qiu
Department of Biostatistics
University of Florida
Gainesville FL, 32611, USA
Email: pqiu@ufl.edu

Abstract

In cancer and other medical studies, time-to-event (e.g., death) data are common. One major task to analyze time-to-event (or survival) data is usually to compare two medical interventions (e.g., a treatment and a control) regarding their effect on patients' hazard to have the event in concern. In such cases, we need to compare two hazard curves of the two related patient groups. In practice, a medical treatment often has a time-lag effect, i.e., the treatment effect can only be observed after a time period since the treatment is applied. In such cases, the two hazard curves would be similar in an initial time period, and the traditional testing procedures, such as the log-rank test, would be ineffective in detecting the treatment effect because the similarity between the two hazard curves in the initial time period would attenuate the difference between the two hazard curves that is reflected in the related testing statistics. In this paper, we suggest a new method for comparing two hazard curves when there is a potential treatment time-lag effect based on a weighted log-rank test with a flexible weighting scheme. The new method is shown to be more effective than some representative existing methods in various cases when a treatment time-lag effect is present.

KEYWORDS:
Box-Cox transformation, hazard curve, survival data, time-lag, treatment effect, weighted log-rank test

1 | INTRODUCTION

Survival analysis is widely used in medical studies and other applications for analyzing time-to-event data[1,2,3]. In medical studies, when assessing treatment effects for some life-threatening diseases like cancers, heart diseases, and HIV infections, the primary outcome is often a patient's time-to-death after a treatment is applied. In some other applications to treat milder diseases like influenza, researchers usually focus on a patient's time-to-recovery from such diseases. Besides medical studies,

survival analysis is also essential in many other fields, such as credit risk analysis, production adoption analysis, and quality control in medical systems. When assessing treatment effect, the most commonly used approach in practice is to compare two hazard curves of the treatment and control groups. This paper develops a new method for comparing two hazard curves in cases when there is a treatment time-lag effect (i.e., there would be a delay in time the treatment effect can be observed).

Because of its importance, there are already many methods developed in the literature for effective comparison of two hazard curves[3,4]. Traditional methods, such as the Log-rank, Gehan-Wilcoxon, and Peto-Peto tests, are developed under the assumption that the Cox proportional hazards model is appropriate for describing the hazard curves[2]. In practice, however, there are many cases when this assumption is invalid. One important case is when the two hazard curves cross each other. There have been some methods developed for comparing two crossing hazard curves. See, for instance, the papers[5,6,7,8,9,10,11,12,13] and the references cited therein.

In practice, a medical treatment often has a time-lag effect. In some applications, a treatment does not show a benefit early on, but it has a long-term advantage. For example, immuno-oncology is a rapidly evolving area in the development of anti-cancer drugs. The effect of Immuno-oncology is not typically directed to the tumor itself; it instead boosts or releases the brake from a patient's immune system, and this positive effect may not be observed immediately[14]. See Section 4 for another real-data example. In such cases, the Cox proportional hazards assumption would be violated, and the two hazard curves would be similar in an initial time period of a study and become different afterward. If a traditional method (e.g., the Log-rank test) is used to compare such hazard curves, then the treatment effect shown after the initial time period would be difficult to detect, because the similarity between the two curves in the initial time period would attenuate the overall difference between the two hazard curves in the entire study period that is reflected in the related testing statistic.

To address the treatment time-lag effect, there have been some discussions in the literature. For instance, Dinse et al[15] proposed a method for estimating the length of the time-lag period based on the Kaplan-Meier estimates of the survival functions. The resulting estimate has been confirmed positively biased[16]. Gierz et al[17] developed an alternative method for estimating the length of the time-lag period by using an empirical divergence measure. Although it has been shown that this method performs better than the one by Dinse et al in various cases, both methods cannot test whether the treatment effect is significant or not. In cases when the time-lag period can be pre-specified, Zucker and Lakatos[18] proposed two weighted Log-rank tests for comparing the related hazard curves. In practice, however, a precise specification of the time-lag period could be challenging. Park and Qiu[16] proposed a generalized Cox proportional hazards model that can accommodate the time-lag effect. However, this method assumed the log hazard ratio to be a linear function of time, which could be too restrictive for some applications.

In this paper, a novel weighted Log-rank test is suggested for comparing two hazard curves when there is a potential treatment time-lag effect. The new weighted Log-rank test employs a flexible weighting scheme based on the Box-Cox transformation.

It does not require any prior knowledge about the length of the potential time-lag period, and is not confined to any particular parametric forms for the hazard ratio. Numerical studies show that it is effective in various cases considered.

The remainder of the paper is organized as follows. Section 2 provides a formulation of the research problem and a detailed description of our proposed method. Section 3 presents some simulation results about the numerical performance of the proposed method in comparison with some representative existing methods. Section 4 demonstrates the proposed method using two real-world datasets. Section 5 gives some concluding remarks.

2 | PROPOSED METHOD

In this paper, our major goal is to test whether there is a treatment time-lag effect based on the observed survival data. Let h_1 and h_0 represent the hazard rate functions of the treatment and control groups, respectively. Then, we are interested in testing the following hypotheses:

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

where $[0,\mathcal{T}]$ denotes the entire study period, $[0,\tau]$ is the time-lag period, and τ is the length of this period and also called the time-lag point hereafter.

To test the hypotheses in (1), we suggest a weighted log-rank testing procedure described below. 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 weighted log-rank test[3] is defined to be

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

where $\{w(t_i), i = 1, 2, \cdots, D\}$ are the pre-specified weights.

For the test statistic U defined in (2), the difference between two estimated hazard rate functions is calculated at each time point t_i, and U is just a standardized weighted average of all such differences at $\{t_i, i = 1, 2, \ldots, D\}$. Intuitively, for testing the hypotheses in (1), the weight $w(t_i)$ should be 0 when $t_i \le \tau$ (i.e., before the time-lag point) because the inclusion of the terms corresponding to $\{t_i, t_i \le \tau\}$ in U cannot improve its power to detect the treatment effect after τ. On the other hand, when $t_i > \tau$ and t_i increases, the weight $w(t_i)$ should get larger since the related terms in U could provide more information about the

treatment effect. Based on these intuitions, we suggest using the following weighting function in (2):

$$w(t;\alpha,\tau) = \left[\mathrm{BC}_\alpha(t) - \mathrm{BC}_\alpha(\tau)\right] I(t > \tau), \tag{3}$$

where $I(t > \tau)$ is an indicator function that equals 1 when $t > \tau$ and 0 otherwise, $\mathrm{BC}_\alpha(t)$ is the modified Box-Cox transformation defined by

$$\mathrm{BC}_\alpha(t) = \begin{cases} \log(t), & \text{if } \alpha = 0 \\ t^\alpha, & \text{otherwise,} \end{cases} \tag{4}$$

and $\alpha > 0$ is a coefficient. It should be pointed out that the regular Box-Cox transformation is defined to be $\log(t)$ when $\alpha = 0$ and $(t^\alpha - 1)/\alpha$ otherwise. Because the weights are used in both the numerator and denominator of (2) and the difference $\mathrm{BC}_\alpha(t) - \mathrm{BC}_\alpha(\tau)$ is used in defining the weight function, it can be easily checked that the test statistic U would not change if the regular Box-Cox transformation is replaced by its simplified version in (4). See the paper[9] for some related discussions. By changing the value of α, the weighting function $w(t;\alpha,\tau)$ can have many different patterns. So, it is flexible. In this paper, we only consider non-negative values for α to make sure that $w(t;\alpha,\tau)$ is an increasing function of t after the time-lag point τ. This restriction can be lifted easily if it is unreasonable in some applications, making the weighting function even more flexible. Figure 1 shows the weighting function $w(t;\alpha,\tau)$ in cases when $\mathcal{T} = 3$, $\tau = 1$, and α changes its value among $\{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$. From the plot, it can be seen that the increasing rate is larger when α is larger. When $\alpha = 0$, the weighting function increases slowly with t in a logarithm rate. When $\alpha = 2$, it increases quite fast in a quadratic rate.

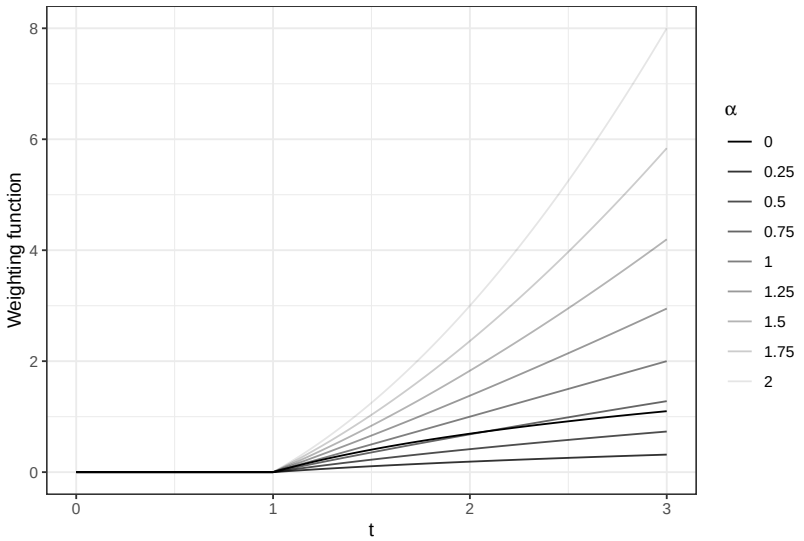

Figure 1 The weighting function $w(t;\alpha;\tau)$ in cases when $\mathcal{T} = 3, \tau = 1$, and α changes its value among $\{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$.

In the case when the time-lag point τ is known, the test statistic of the weighted Log-rank test using the weighting function defined in (3) is then

$$U_{\alpha,\tau} = \frac{\sum_{i=1}^{D} w(t_i;\alpha,\tau)\left(d_{i1} - Y_{i1}\frac{d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{D} w^2(t_i;\alpha,\tau)\frac{Y_{i1}}{Y_i}\frac{Y_{i2}}{Y_i}\frac{Y_i-d_i}{Y_i-1}d_i}}. \tag{5}$$

In practice, however, τ is usually unknown, and the parameter α should be chosen properly too. To address these issues, the test statistic of our proposed weighted Log-rank test is defined to be

$$\widetilde{U} = \max_{\alpha\geq 0}\max_{\tau\in[0,\mathcal{T}]}\left|U_{\alpha,\tau}\right|. \tag{6}$$

In (6), the value of τ is first searched for a given α so that the magnitude of the weighted Log-rank test statistic $U_{\alpha,\tau}$ defined in (5) is maximized. The maximizer, denoted as $\widetilde{\tau}(\alpha)$, should be a good estimate of the time-lag point if the given α value is a good choice for the weighting function $w(t;\alpha,\tau)$. Then, the value of α is also searched to maximize the value of $\left|U_{\alpha,\widetilde{\tau}(\alpha)}\right|$, and $\widetilde{U}$ is defined to be the maximum.

In the survival analysis literature, it has been shown that the weighted Log-rank test statistic $U_{\alpha,\tau}$, for given values of α and τ, would have a normal asymptotic distribution under some regularity conditions[4,19]. However, the proposed test statistic $\widetilde{U}$ involves two maximizations of $\left|U_{\alpha,\tau}\right|$ with respect to α and τ. In a different case involving a maximization of a weighted Log-rank test statistic, O'Quigley and Pessione[20] showed that the distribution of the resulting maximum would be bimodal, and they introduced a method called the Direct Bootstrap for computing the p-value for the related testing procedure. Next, we will show that the distribution of the proposed test statistic $\widetilde{U}$ is also bimodel. Thus, the Direct Bootstrap procedure is also considered here for computing the p-value of the testing procedure using $\widetilde{U}$ defined in (6).

In cases when $h_0(t) = h_1(t) = 1$ (i.e., H_0 in (1) is true), $n_1 = n_2 = 200$, $[0, \mathcal{T}] = [\text{⌀}, \text{⌀}.]$ and the censoring times are generated from uniform distributions on [0; 1.8], the density histogram of $\widetilde{U}$ based on 8,000 replicated simulations is shown in Figure 2(a). In each simulation, $n_1 = n_2 = 200$ survival times are generated as specified above for both the treatment and control groups, and then the value of $\widetilde{U}$ is computed. So, Figure 2(a) is made based on 8,000 values of $\widetilde{U}$. From the plot, it can be seen that the null distribution of $\widetilde{U}$ is indeed bimodal and symmetric about 0. Figure 2(b) and 2(c) show the density histograms of $\widetilde{U}$ in cases when $h_1(t)$ changes to

$$h_1(t) = \{\exp[1.5(t-1.2)]\}\, I(t > 1.2) + I(t \leq 1.2)$$

and

$$h_1(t) = \{\exp[1.5(t-0.9)]\}\, I(t > 0.9) + I(t \leq 0.9),$$

respectively, and other settings are kept unchanged. Obviously, the time-lag point is respectively 1.2 and 0.9 in these two cases when H_0 is violated and there is a treatment time-lag effect. From the two plots, it can be seen that i) the bimodal shape of the density histogram is retained, but the distribution of $\widetilde{U}$ becomes skewed towards the direction of the alternative hypothesis, and

ii) when the time-lag point τ is larger (i.e., $\tau = 1.2$ in plot (b)), the distribution is less skewed, implying that the alternative hypothesis is more difficult to confirm by the test due to the fact that a longer time-lag period would attenuate the overall difference between the two related hazard curves more seriously.

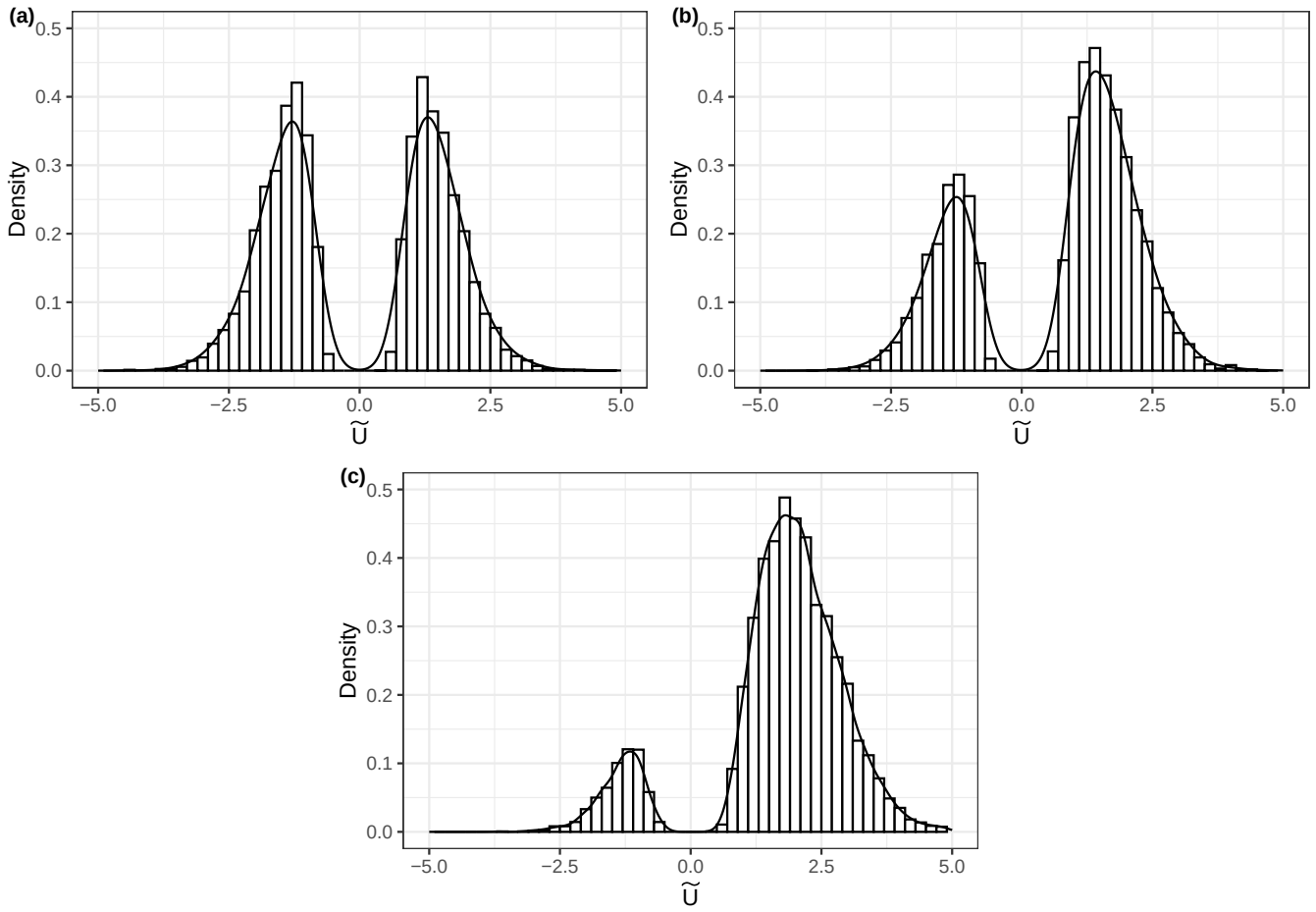

Figure 2 Panels (a)-(c) show the density histograms of $\widetilde{U}$ in cases when $h_0(t) = 1$, and $h_1(t) = 1$ (plot (a)), $h_1(t) = \{\exp[1.5(t - 1.2)]\}I(t > 1.2) + I(t \leq 1.2)$ (plot (b)), or $h_1(t) = \{\exp[1.5(t - 0.9)]\}I(t > 0.9) + I(t \leq 0.9)$ (plot (c)).

To calculate the p-value of the test using $\widetilde{U}$ by the Direct Bootstrap procedure, a bootstrap sample of size n_j is first drawn with replacement from the observed data of group j, for $j = 1, 2$, and then the value of $\widetilde{U}$ is computed by (6). This bootstrap resampling process is repeated for B times. Among the B values of $\widetilde{U}$, let B^+ be the number of positive values and B^- be the number of negative values. Then, the p-value of the test using $\widetilde{U}$ for testing the hypotheses in (1) is defined to be

$$p\text{-value} = 2\min(B^+, B^-)/B.$$

See the paper[20] for a related discussion.

3 | SIMULATION STUDY

In this section, we evaluate the numerical performance of our proposed method in detecting the treatment time-lag effect using simulation studies in comparison with several representative existing methods. The existing methods considered in the simulation studies include the Log-rank (LR), Gehan-Wilcoxon (GE), Tarone-Ware (TW), Peto-Peto (PE), Fleming-Harrington (FH), and Zucker and Lakatos (ZC) testing procedures. In the Fleming-Harrington testing procedure, there are two parameters ρ and λ used in its weighting function. As in Park and Qiu's paper[16], the following three combinations for these two parameters are considered here: (ρ=0, λ=1), (ρ=1, λ=0), and (ρ=1, λ=1), and the related testing procedures are denoted as FH^{01}, FH^{10}, and FH^{11}, respectively. To use the Zucker and Lakatos testing procedure, the potential time-lag point τ^* should be specified in advance. In this section, we set τ^* to be the following three values: τ -0.05, τ, and τ+0.05, where τ denotes the true time-lag point. The three related testing procedures are denoted as ZC^-, ZC, and ZC^+, respectively, representing the testing approach when the pre-specified time-lag point is smaller than, equal to, or larger than the true time-lag point. Since our proposed method is a special weighted Log-rank test that incorporates the modified Box-Cox transformation (4), it is denoted as BC hereafter.

The simulation is performed in the following cases. In Case 1, it is assumed that $h_0(t) = h_1(t) = 1$ (i.e., H_0 in (1) is true). The other cases are described below when H_0 is invalid, $h_0(t) = 1$, and $h_1(t)$ takes one of the following patterns:

Exponential patterns: $h_1(t) = \{\exp[1.5(t-\tau)]\}\, I(t > \tau) + I(t \le \tau)$, where τ changes among $\{0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2\}$, and the related cases are denoted as Cases E2– E8, respectively.

Linear patterns: $h_1(t) = [5(t-\tau)+1]\, I(t > \tau) + I(t \le \tau)$, where τ changes among $\{0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2\}$, and the related cases are denoted as Cases L2–L8, respectively.

Quadratic patterns: $h_1(t) = \left[(t-\tau+1)^2\right] I(t > \tau) + I(t \le \tau)$, where τ changes among $\{0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2\}$, and the related cases are denoted as Cases Q2–Q8, respectively.

These cases are shown in Figure 3, from which it can be seen that various cases with different time-lag points and different patterns of $h_1(t)$ have been considered here.

In each case considered above, n_1 (n_2) observations of survival times are generated randomly from a survival time distribution with the hazard function $h_0(t)$ ($h_1(t)$), where $n_1 = n_2$ and n_1 can take the value of 100, 150, or 200. Two different censoring schemes are considered, which are referred to as censoring scheme I and censoring scheme II. Under the censoring scheme I, the censoring times are generated randomly from the uniform distribution on the interval [0, 3.6], while the censoring times are generated randomly from the uniform distribution on the interval [0, 1.8] under the censoring scheme II. Therefore, the maximum observed survival time in the simulated data is 3.6 and 1.8, respectively, under these two censoring schemes. Tables 1-2 summarize the censoring rates of the observed survival data in the control and treatment groups in each case considered under

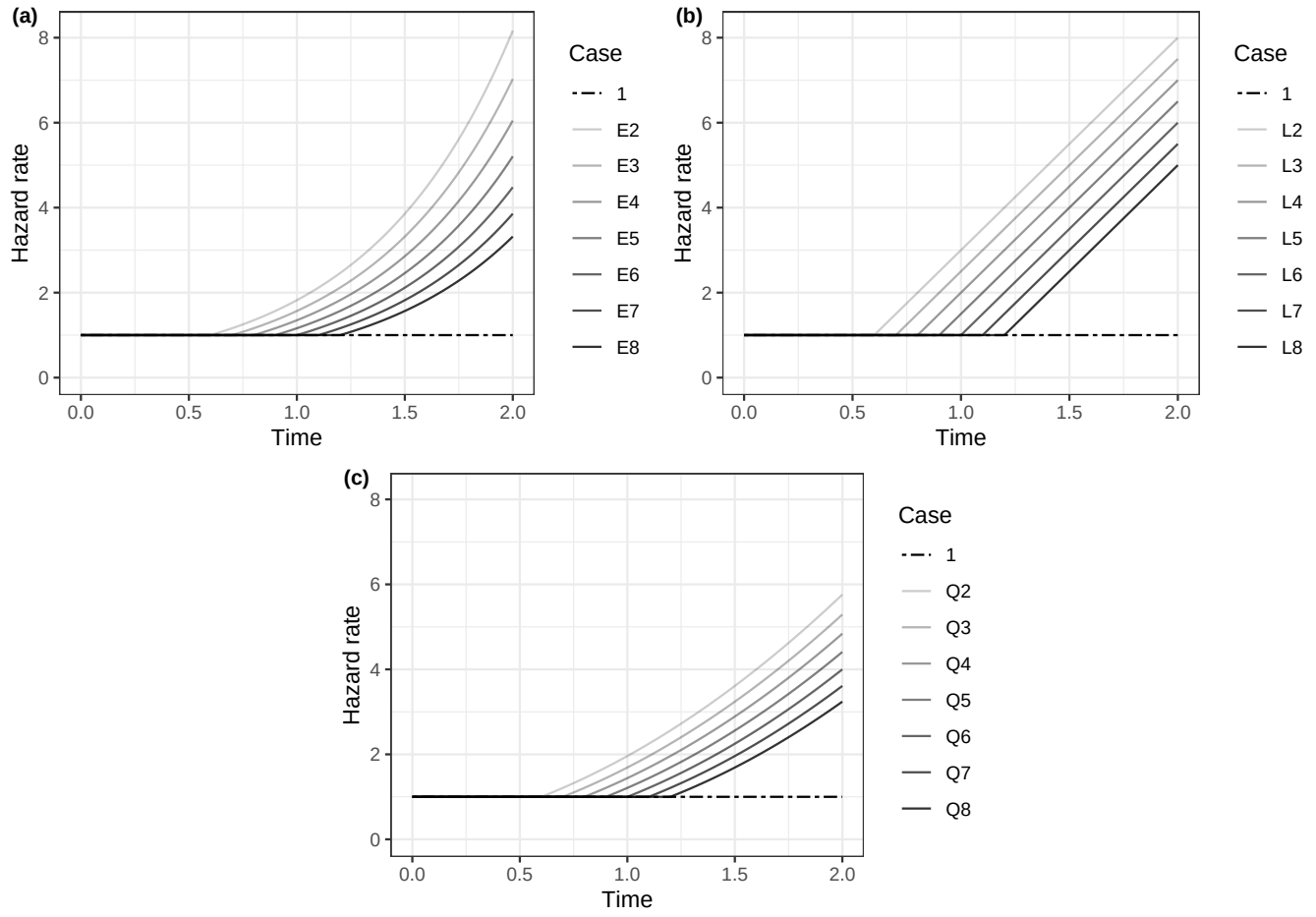

Figure 3 Panels (a)-(c) present the hazard rate functions in cases when $h_0(t) = 1$, and $h_1(t)$ has the exponential, linear, and quadratic patterns, respectively, with different time-lag periods between the two hazard functions.

the two censoring schemes. It can be seen that the censoring rates range from 18% to 28% under the censoring scheme I, and from 37% to 47% under the censoring scheme II. Thus, the two censoring schemes can represent a low and a high censoring level, respectively.

[Tables 1-2 about here]

Next, we study the empirical sizes and powers for all the following tests: LR, GE, TW, PE, FH^{01}, FH^{01}, FH^{01}, ZC^{-}, ZC, ZC^{+}, and BC, under different censoring schemes and sample sizes. To determine the critical value of a given test, a Monte Carlo simulation with 8,000 replications is performed under the null hypothesis that both hazard rate functions of the control and treatment groups are the constant 1. Then, the critical values of the test are selected to be the 2.5% and 97.5% quantiles of the empirical distribution of the test statistic obtained from the 8,000 replicated simulations, given the significance level α of the test to be 0.05. Then, the size and power of the test are obtained by performing another 1,000 replicated simulations under the null

(i.e., $h_1(t) = 1$) and an alternative (i.e., $h_1(t)$ takes one of the alternative patterns shown in Figure 3) hypothesis, respectively. The results under the two censoring schemes and three sample sizes considered are presented in Tables 3-8, respectively.

[Tables 3-8 about here]

From Tables 3-8, it can be seen that the sizes of all testing methods are quite close to the nominal significance level of 0.05 in all cases considered. The proposed method BC has the largest power in all cases considered, except the seven cases E7, E8, L8, Q7, and Q8 when $n_1 = n_2 = 100$ and E8 and Q8 when $n_1 = n_2 = 150$ under the censoring scheme II. In these 7 out of a total 126 cases considered in the tables, the time-lag point τ is quite large, and the data censoring rate is quite high. In these cases, the methods ZC and/or ZC^+ perform slightly better than BC, with the price that the time-lag point needs to be pre-specified. If the time-lag point is not pre-specified properly, then the method ZC may not perform well since BC performs better than one of ZC^-, ZC, and ZC^+ in 4 out of the 7 cases mentioned above. From the tables, we can also observe that the traditional methods LR, GE, TW, and PE for comparing the two hazard rate functions are not effective in cases when there is a treatment time-lag effect, since their power values are much smaller than those of other methods in all cases considered. Therefore, if a treatment time-lag effect is present, they should be avoided. For the methods FH and ZC, their performance is reasonably good only when their parameters (i.e., ρ and λ in FH and τ^* in ZC) are chosen properly, which may not be easy in practice. From the tables, it can be seen that the power of each method decreases when the time-lag point increases, which is intuitively reasonable because the overall difference between the two hazard rate functions becomes smaller when the time-lag point increases. When other simulation settings remain unchanged, all methods seem to have larger powers when the censoring rate is lower and/or the sample sizes are larger, which is intuitively reasonable as well. As a summary, this example confirms that the proposed method BC provides a powerful tool for comparing two hazard rate functions when there is a treatment time-lag effect.

4 | CASE STUDY

In this section, we apply our proposed method BC and other comparative approaches discussed in Section 3 to two real-data examples. The first example was originally discussed in Mantel et al's paper[21] and became a popular example in the survival analysis literature. The data were from a study to investigate whether a drug is useful in preventing the formation of tumors in rats. To this end, 100 distinct litters were used, and each litter contained three rats. One randomly selected rat from each litter received the drug treatment; the other two were treated with a placebo. The event of interest was the tumor incidence, and the study was conducted over a period from day 0 to day 100 after the treatment. The observed survival data are available in the *R*-package **survival**. Due to the limited number of male rats with tumors (only 2 out of 150), our analysis focuses only on female rats. Consequently, there were 31 rats in the drug group (treatment group) with a censoring rate of 48% and 50 rats in the placebo group (control group) with a censoring rate of 66%. By using kernel-based methods[22], the estimated hazard rate functions for

both the control and treatment groups are shown in Figure 4. From the figure, it can be seen that the two hazard curves are almost the same until about 80 days, and the hazard curve of the drug group tends to be above that of the placebo group afterward.

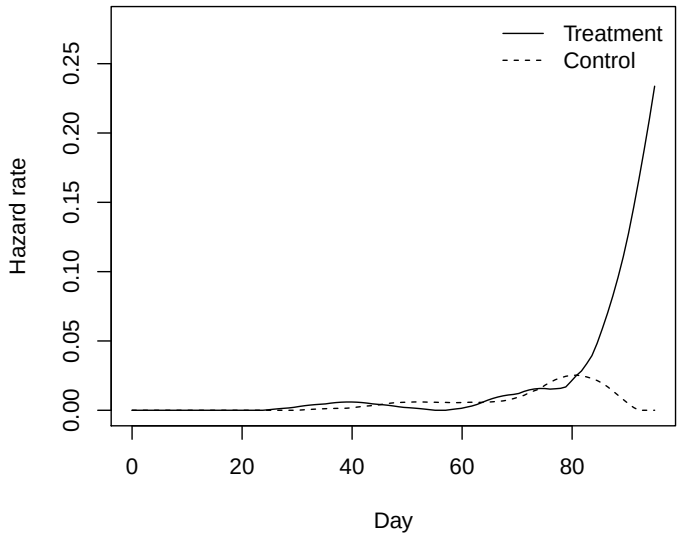

Figure 4 Estimated hazard curves of the treatment and control groups of the Rats dataset.

Next, we apply the methods LR, GE, TW, PE, FH^{01}, FH^{01}, FH^{01}, ZC^{-}, ZC, ZC^{+}, and BC to this dataset to compare the two hazard curves. For each method, the conventional significance level of 0.05 is used. To compute the *p*-value of the proposed method BC, the Direct Bootstrap procedure discussed at the end of Section 2 is used, in which the bootstrap sample size is chosen to be 2,000. For the ZC method, the time-lag point is pre-specified to be 77, 81 (the point estimate of τ by BC), or 85, and the three versions of the method are denoted as ZC^{-}, ZC, and ZC^{+}, respectively. Table 9 presents the *p*-values of all the testing methods considered. From the results in the table, it can be seen that the four conventional testing methods LR, GE, TW, and PE cannot detect the treatment effect in this example, as expected, since the treatment time-lag effect would attenuate the overall difference between the two hazard curves so that their test statistic values become small in magnitude. The methods FH and ZC can detect the treatment effect only when their parameters are pre-specified properly (i.e., ρ=0 and λ=1 in FH and τ = 85 in ZC). As a comparison, the proposed testing method BC is very significant in this example and its *p*-value is much smaller than those of the other methods. Therefore, this example confirms that BC is a powerful tool for comparing two hazard curves when the treatment time-lag effect is present.

[Table 9 about here]

Additionally, we apply the related methods to another example with the Veteran dataset that was discussed in the paper[23]. The data are included in the *R*-package **survival** and can be downloaded there. This Veteran study included 106 lung cancer patients over 50 years old. Among them, 55 patients were randomly assigned to receive a test therapy (treatment group), and another 51 patients were assigned to receive standard therapy (control group). The primary outcome for each patient was the time to death. During the study, two censored observations were recorded in the treatment group and five in the control group. Figure 5 illustrates the estimated hazard rate functions for both groups. The treatment time-lag effect can be seen from Figure 5. Specifically, there is a slight variation in the difference between the two hazard rate functions of the treatment and control groups before day 360. After day 360, the hazard rate function of the control group is significantly higher than the hazard rate function of the treatment group. Then, we apply all related methods to the dataset using the same setups as those in the previous example. Specifically, we employ a bootstrap sample size of 2000 for the BC method, and the pre-specify time-lag point is set at 355, 370, and 385 for the methods ZC^{-}, ZC, and ZC^{+}, respectively. The testing results of all methods are presented in Table 10. From the table, it can be seen that only ZC^{+} and BC can successfully detect the time-lag difference between the two hazard rate functions. The ZC method can detect the difference when the pre-specified time-lag point is relatively large and our proposed BC method has the smallest *p*-value among all methods in this example. So, this example further confirms the effectiveness of the proposed method for comparing two hazard curves when there is a treatment time-lag effect.

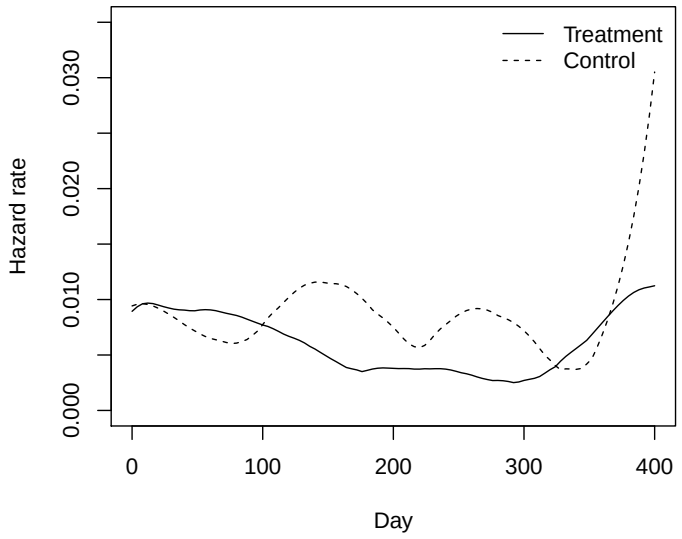

Figure 5 Estimated hazard curves of the treatment and control groups of the Veteran dataset.

[Table 10 about here]

5 | CONCLUDING REMARKS

In the previous several sections, we have discussed our proposed method for comparing two hazard rate functions in cases when there is a treatment time-lag effect. This method is constructed under a weighted log-rank testing framework with a flexible weighting scheme that incorporates the modified Box-Cox transformation (4). Numerical studies in Sections 3 and 4 show that it provides an effective tool for detecting treatment effect when there is a treatment time-lag effect. However, there are still some issues about the proposed method. For instance, our study focuses on hypothesis testing of a treatment time-lag effect rather than point estimation of the time-lag point. It is still unknown yet how to define a proper estimate for the time-lag point based on the proposed testing procedure. Also, although the weighting function defined in (3)-(4) is already very flexible, it is still unknown whether it is possible to derive a data-driven nonparametric weighting scheme for the proposed testing procedure. In addition, observed survival times are often associated with some covariates (e.g., gender, age) in practice. It is still unknown how to accommodate such covariates in the proposed method. All these issues need much future research.

ACKNOWLEDGMENTS:

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

CONFLICT OF INTEREST

The authors have no conflict of interest.

DATA AVAILABILITY STATEMENT

The survival data used in the case study can be downloaded in the open-source *R*-package **survival**.

References

1. Lawless JF. *Statistical models and methods for lifetime data*. John Wiley & Sons, 1982.
2. Bain LJ, Englehardt M. *Statistical Analysis of Reliability and Life-testing Models: Theory and Methods (2nd ed.)*. Marcel Dekker, 1991.
3. Klein JP, Moeschberger ML, others . *Survival analysis: techniques for censored and truncated data*. 1230. Springer, 2003.

4. Fleming TR, Harrington DP. *Counting processes and survival analysis*. John Wiley & Sons, 1991.

5. Chen Z, Huang H, Qiu P. Comparison of multiple hazard rate functions. *Biometrics.* 2016;72(1):39–45.

6. Cheng MY, Qiu P, Tan X, Tu D. Confidence intervals for the first crossing point of two hazard functions. *Lifetime Data Analysis.* 2009;15:441–454.

7. Ditzhaus M, Friedrich S. More powerful logrank permutation tests for two-sample survival data. *Journal of Statistical Computation and Simulation.* 2020;90(12):2209–2227.

8. Lin X, Wang H. A new testing approach for comparing the overall homogeneity of survival curves. *Biometrical Journal: Journal of Mathematical Methods in Biosciences.* 2004;46(5):489–496.

9. Liu K, Qiu P, Sheng J. Comparing two crossing hazard rates by Cox proportional hazards modelling. *Statistics in Medicine.* 2007;26(2):375–391.

10. O'Quigley J. On a two-sided test for crossing hazards. *Journal of the Royal Statistical Society: Series D (The Statistician).* 1994;43(4):563–569.

11. Park KY, Qiu P. Model selection and diagnostics for joint modeling of survival and longitudinal data with crossing hazard rate functions. *Statistics in Medicine.* 2014;33(26):4532–4546.

12. Qiu P, Sheng J. A two-stage procedure for comparing hazard rate functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 2008;70(1):191–208.

13. 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(1):e0116774.

14. Chen TT. Statistical issues and challenges in immuno-oncology. *Journal for Immunotherapy of Cancer.* 2013;1:1–9.

15. Dinse GE, Piegorsch WW, Boos DD. Confidence statements about the time range over which survival curves differ. *Journal of the Royal Statistical Society Series C: Applied Statistics.* 1993;42(1):21–30.

16. Park K, Qiu P. Evaluation of the treatment time-lag effect for survival data. *Lifetime Data Analysis.* 2018;24:310–327.

17. Gierz K, Park K, Qiu P. Non-parametric treatment time-lag effect estimation. *Statistical Methods in Medical Research.* 2022;31(1):62–75.

18. Zucker DM, Lakatos E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika.* 1990;77(4):853–864.

19. Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika.* 1977;64(2):247–254.

20. O'Quigley J, Pessione F. The problem of a covariate-time qualitative interaction in a survival study. *Biometrics.* 1991:101–115.

21. Mantel N, Bohidar NR, Ciminera JL. Mantel-Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information. *Cancer Research.* 1977;37(11):3863–3868.

22. Hess KR, Serachitopol DM, Brown BW. Hazard function estimators: a simulation study. *Statistics in Medicine.* 1999;18(22):3075–3088.

23. Kalbfleisch J, Prentice R. The statistical analysis of failure time data. *Wiley series in probability and mathematical statistics Show all parts in this series.* 1980.

[99]

□

Table 1 Censoring rates of the control and treatment groups in various cases under the censoring scheme I.

	$n_1 = n_2 = 100$		150		200	
Case	Control	Treatment	Control	Treatment	Control	Treatment
1	0.2704	0.2697	0.2697	0.2707	0.2701	0.2700
E2	0.2704	0.2034	0.2697	0.2043	0.2701	0.2037
E3	0.2704	0.2105	0.2697	0.2111	0.2701	0.2107
E4	0.2704	0.2168	0.2697	0.2175	0.2701	0.2170
E5	0.2704	0.2225	0.2697	0.2232	0.2701	0.2227
E6	0.2704	0.2277	0.2697	0.2284	0.2701	0.2279
E7	0.2704	0.2323	0.2697	0.2330	0.2701	0.2325
E8	0.2704	0.2368	0.2697	0.2374	0.2701	0.2368
L2	0.2704	0.1866	0.2697	0.1874	0.2701	0.1870
L3	0.2704	0.1950	0.2697	0.1959	0.2701	0.1954
L4	0.2704	0.2027	0.2697	0.2036	0.2701	0.2032
L5	0.2704	0.2103	0.2697	0.2106	0.2701	0.2100
L6	0.2704	0.2164	0.2697	0.2169	0.2701	0.2164
L7	0.2704	0.2219	0.2697	0.2228	0.2701	0.2224
L8	0.2704	0.2271	0.2697	0.2281	0.2701	0.2275
Q2	0.2704	0.2019	0.2697	0.2028	0.2701	0.2023
Q3	0.2704	0.2094	0.2697	0.2100	0.2701	0.2095
Q4	0.2704	0.2158	0.2697	0.2165	0.2701	0.2159
Q5	0.2704	0.2214	0.2697	0.2222	0.2701	0.2217
Q6	0.2704	0.2269	0.2697	0.2275	0.2701	0.2270
Q7	0.2704	0.2316	0.2697	0.2321	0.2701	0.2316
Q8	0.2704	0.2359	0.2697	0.2366	0.2701	0.2361

Table 2 Censoring rates of the control and treatment groups in various cases under the censoring scheme II.

	$n_1 = n_2 = 100$		150		200	
Case	Control	Treatment	Control	Treatment	Control	Treatment
1	0.4641	0.4634	0.4638	0.4640	0.4638	0.4643
E2	0.4641	0.4062	0.4638	0.4070	0.4638	0.4067
E3	0.4641	0.4194	0.4638	0.4198	0.4638	0.4196
E4	0.4641	0.4300	0.4638	0.4309	0.4638	0.4308
E5	0.4641	0.4393	0.4638	0.4404	0.4638	0.4404
E6	0.4641	0.4467	0.4638	0.4479	0.4638	0.4479
E7	0.4641	0.4521	0.4638	0.4533	0.4638	0.4534
E8	0.4641	0.4564	0.4638	0.4575	0.4638	0.4579
L2	0.4641	0.3738	0.4638	0.3741	0.4638	0.3739
L3	0.4641	0.3903	0.4638	0.3907	0.4638	0.3908
L4	0.4641	0.4057	0.4638	0.4061	0.4638	0.4058
L5	0.4641	0.4190	0.4638	0.4196	0.4638	0.4193
L6	0.4641	0.4302	0.4638	0.4311	0.4638	0.4311
L7	0.4641	0.4398	0.4638	0.4410	0.4638	0.4410
L8	0.4641	0.4480	0.4638	0.4491	0.4638	0.4490
Q2	0.4641	0.4025	0.4638	0.4033	0.4638	0.4030
Q3	0.4641	0.4161	0.4638	0.4165	0.4638	0.4161
Q4	0.4641	0.4270	0.4638	0.4276	0.4638	0.4277
Q5	0.4641	0.4362	0.4638	0.4375	0.4638	0.4375
Q6	0.4641	0.4444	0.4638	0.4455	0.4638	0.4454
Q7	0.4641	0.4505	0.4638	0.4518	0.4638	0.4517
Q8	0.4641	0.4550	0.4638	0.4562	0.4638	0.4565

Table 3 Sizes and powers of different testing procedures for detecting the treatment time-lag effect in various cases under the censoring scheme I and the sample size $n_1 = n_2 = 100$. The number in bold in each row is the largest value in that row.

	Methods										
Cases	LR	GE	TW	PE	FH^{01}	FH^{10}	FH^{11}	ZC^{-}	ZC	ZC^{+}	BC
1	0.049	0.054	0.052	0.050	0.035	0.055	0.043	0.039	0.043	0.038	0.037
E2	0.426	0.081	0.159	0.098	0.830	0.113	0.384	0.691	0.710	0.729	**0.909**
E3	0.320	0.066	0.125	0.083	0.708	0.085	0.271	0.605	0.614	0.645	**0.850**
E4	0.246	0.059	0.095	0.071	0.574	0.072	0.197	0.512	0.539	0.557	**0.784**
E5	0.194	0.057	0.079	0.062	0.457	0.061	0.144	0.433	0.453	0.471	**0.698**
E6	0.155	0.058	0.069	0.056	0.362	0.059	0.102	0.367	0.383	0.404	**0.596**
E7	0.119	0.056	0.063	0.058	0.285	0.056	0.083	0.309	0.325	0.348	**0.500**
E8	0.093	0.054	0.058	0.057	0.211	0.053	0.062	0.263	0.274	0.288	**0.427**
L2	0.700	0.135	0.303	0.192	0.980	0.191	0.679	0.942	0.949	0.958	**0.990**
L3	0.550	0.101	0.222	0.141	0.937	0.147	0.526	0.881	0.905	0.916	**0.974**
L4	0.444	0.081	0.159	0.097	0.859	0.108	0.390	0.810	0.840	0.864	**0.936**
L5	0.335	0.063	0.122	0.081	0.758	0.082	0.270	0.742	0.772	0.790	**0.893**
L6	0.256	0.055	0.095	0.069	0.615	0.068	0.194	0.649	0.686	0.699	**0.835**
L7	0.198	0.057	0.078	0.062	0.493	0.062	0.142	0.557	0.586	0.613	**0.764**
L8	0.155	0.058	0.069	0.057	0.384	0.057	0.103	0.472	0.500	0.528	**0.672**
Q2	0.451	0.084	0.178	0.111	0.836	0.123	0.413	0.706	0.733	0.752	**0.883**
Q3	0.338	0.070	0.126	0.086	0.724	0.090	0.294	0.619	0.645	0.672	**0.821**
Q4	0.258	0.061	0.100	0.072	0.586	0.072	0.211	0.530	0.567	0.577	**0.755**
Q5	0.201	0.055	0.080	0.063	0.465	0.063	0.153	0.447	0.466	0.487	**0.669**
Q6	0.161	0.058	0.069	0.057	0.375	0.060	0.113	0.389	0.402	0.420	**0.565**
Q7	0.123	0.057	0.067	0.058	0.299	0.054	0.089	0.323	0.335	0.360	**0.478**
Q8	0.099	0.055	0.059	0.057	0.223	0.055	0.067	0.280	0.287	0.309	**0.419**

Table 4 Sizes and powers of different testing procedures for detecting the treatment time-lag effect in various cases under the censoring scheme II and the sample size $n_1 = n_2 = 100$. The number in bold in each row is the largest value in that row.

	Methods										
Cases	LR	GE	TW	PE	FH^{01}	FH^{10}	FH^{11}	ZC^{-}	ZC	ZC^{+}	BC
1	0.050	0.047	0.052	0.051	0.043	0.053	0.045	0.041	0.041	0.044	0.049
E2	0.132	0.052	0.061	0.061	0.299	0.066	0.152	0.249	0.264	0.287	**0.390**
E3	0.089	0.049	0.058	0.056	0.179	0.062	0.096	0.175	0.192	0.199	**0.276**
E4	0.068	0.049	0.056	0.053	0.107	0.052	0.066	0.116	0.123	0.143	**0.166**
E5	0.062	0.049	0.055	0.056	0.080	0.049	0.050	0.100	0.109	0.111	**0.130**
E6	0.054	0.047	0.055	0.054	0.060	0.048	0.046	0.081	0.088	0.085	**0.092**
E7	0.056	0.047	0.053	0.054	0.048	0.048	0.046	0.066	**0.073**	0.071	0.060
E8	0.053	0.047	0.052	0.054	0.044	0.050	0.045	0.060	0.063	**0.065**	0.060
L2	0.303	0.069	0.118	0.104	0.657	0.112	0.373	0.573	0.611	0.646	**0.752**
L3	0.205	0.058	0.086	0.081	0.493	0.077	0.233	0.450	0.487	0.512	**0.616**
L4	0.125	0.050	0.063	0.061	0.318	0.064	0.144	0.340	0.357	0.374	**0.457**
L5	0.090	0.047	0.057	0.052	0.189	0.059	0.092	0.228	0.253	0.275	**0.331**
L6	0.069	0.048	0.057	0.055	0.119	0.052	0.061	0.161	0.169	0.183	**0.207**
L7	0.062	0.048	0.054	0.053	0.079	0.050	0.042	0.121	0.134	0.139	**0.141**
L8	0.053	0.047	0.054	0.054	0.056	0.049	0.046	0.092	0.095	**0.096**	0.090
Q2	0.153	0.056	0.070	0.069	0.330	0.067	0.170	0.281	0.298	0.322	**0.400**
Q3	0.103	0.049	0.057	0.056	0.211	0.063	0.106	0.200	0.218	0.233	**0.294**
Q4	0.072	0.048	0.058	0.052	0.126	0.053	0.073	0.137	0.146	0.169	**0.189**
Q5	0.067	0.048	0.057	0.054	0.088	0.049	0.053	0.108	0.122	0.128	**0.139**
Q6	0.057	0.047	0.055	0.054	0.066	0.048	0.047	0.090	0.090	0.090	**0.105**
Q7	0.052	0.047	0.055	0.054	0.053	0.049	0.045	0.069	0.077	**0.083**	0.082
Q8	0.053	0.047	0.053	0.054	0.045	0.050	0.043	0.064	0.066	**0.070**	0.066

Table 5 Sizes and powers of different testing procedures for detecting the treatment time-lag effect in various cases under the censoring scheme I and the sample size $n_1 = n_2 = 150$. The number in bold in each row is the largest value in that row.

	Methods										
Cases	LR	GE	TW	PE	FH^{01}	FH^{10}	FH^{11}	ZC^{-}	ZC	ZC^{+}	BC
1	0.050	0.050	0.049	0.050	0.047	0.047	0.065	0.050	0.051	0.049	0.045
E2	0.594	0.087	0.234	0.125	0.944	0.125	0.530	0.859	0.881	0.888	**0.988**
E3	0.487	0.071	0.161	0.094	0.876	0.094	0.383	0.780	0.811	0.829	**0.969**
E4	0.365	0.062	0.114	0.076	0.783	0.080	0.261	0.711	0.731	0.755	**0.935**
E5	0.271	0.054	0.092	0.064	0.660	0.065	0.185	0.622	0.643	0.674	**0.884**
E6	0.212	0.052	0.074	0.059	0.544	0.058	0.139	0.541	0.564	0.586	**0.811**
E7	0.162	0.048	0.067	0.054	0.433	0.055	0.104	0.469	0.490	0.509	**0.730**
E8	0.124	0.048	0.060	0.051	0.308	0.055	0.091	0.382	0.405	0.429	**0.638**
L2	0.860	0.192	0.452	0.279	0.997	0.283	0.845	0.987	0.990	0.994	**0.999**
L3	0.741	0.117	0.320	0.186	0.988	0.188	0.694	0.970	0.977	0.980	**0.997**
L4	0.610	0.082	0.227	0.118	0.963	0.122	0.537	0.942	0.959	0.966	**0.993**
L5	0.503	0.069	0.161	0.092	0.903	0.088	0.378	0.903	0.919	0.931	**0.981**
L6	0.387	0.062	0.112	0.073	0.814	0.077	0.260	0.835	0.856	0.871	**0.957**
L7	0.286	0.054	0.090	0.063	0.695	0.065	0.181	0.757	0.783	0.808	**0.915**
L8	0.220	0.051	0.076	0.057	0.575	0.058	0.141	0.678	0.704	0.722	**0.861**
Q2	0.621	0.096	0.250	0.141	0.946	0.145	0.567	0.876	0.892	0.905	**0.977**
Q3	0.509	0.074	0.176	0.101	0.887	0.100	0.413	0.801	0.821	0.843	**0.949**
Q4	0.389	0.064	0.123	0.078	0.791	0.083	0.281	0.730	0.749	0.770	**0.912**
Q5	0.284	0.057	0.096	0.068	0.674	0.069	0.197	0.636	0.669	0.701	**0.850**
Q6	0.227	0.052	0.079	0.061	0.558	0.064	0.153	0.570	0.590	0.609	**0.784**
Q7	0.177	0.048	0.069	0.056	0.442	0.054	0.109	0.495	0.517	0.534	**0.705**
Q8	0.135	0.049	0.061	0.052	0.321	0.056	0.092	0.405	0.422	0.451	**0.624**

Table 6 Sizes and powers of different testing procedures for detecting the treatment time-lag effect in various cases under the censoring scheme II and the sample size $n_1 = n_2 = 150$. The number in bold in each row is the largest value in that row.

	Methods										
Cases	LR	GE	TW	PE	FH^{01}	FH^{10}	FH^{11}	ZC^{-}	ZC	ZC^{+}	BC
1	0.056	0.046	0.053	0.052	0.053	0.047	0.053	0.054	0.055	0.048	0.052
E2	0.185	0.051	0.069	0.063	0.438	0.064	0.208	0.355	0.380	0.401	**0.561**
E3	0.117	0.049	0.055	0.053	0.277	0.055	0.125	0.258	0.272	0.296	**0.404**
E4	0.085	0.046	0.053	0.051	0.178	0.057	0.091	0.175	0.179	0.203	**0.275**
E5	0.065	0.045	0.053	0.055	0.110	0.051	0.073	0.125	0.130	0.130	**0.181**
E6	0.057	0.043	0.054	0.054	0.075	0.050	0.060	0.100	0.107	0.101	**0.121**
E7	0.057	0.044	0.054	0.052	0.064	0.047	0.054	0.084	0.078	0.073	**0.096**
E8	0.054	0.045	0.052	0.053	0.054	0.047	0.055	0.067	0.075	**0.082**	0.075
L2	0.451	0.068	0.146	0.128	0.847	0.132	0.526	0.772	0.792	0.813	**0.910**
L3	0.294	0.054	0.091	0.082	0.665	0.080	0.330	0.632	0.657	0.691	**0.815**
L4	0.178	0.052	0.060	0.058	0.484	0.061	0.197	0.469	0.496	0.525	**0.660**
L5	0.116	0.046	0.053	0.052	0.286	0.051	0.121	0.335	0.350	0.376	**0.485**
L6	0.081	0.044	0.053	0.053	0.171	0.056	0.087	0.228	0.242	0.261	**0.335**
L7	0.063	0.045	0.054	0.054	0.100	0.051	0.066	0.149	0.161	0.172	**0.208**
L8	0.059	0.043	0.054	0.050	0.069	0.048	0.053	0.088	0.103	0.116	**0.135**
Q2	0.204	0.055	0.075	0.070	0.477	0.067	0.242	0.394	0.426	0.448	**0.584**
Q3	0.129	0.049	0.056	0.055	0.324	0.055	0.144	0.293	0.311	0.335	**0.439**
Q4	0.094	0.047	0.055	0.053	0.205	0.055	0.101	0.208	0.228	0.239	**0.306**
Q5	0.071	0.044	0.054	0.054	0.121	0.056	0.076	0.143	0.159	0.158	**0.210**
Q6	0.061	0.044	0.054	0.054	0.085	0.051	0.064	0.105	0.111	0.116	**0.139**
Q7	0.057	0.043	0.054	0.051	0.068	0.047	0.054	0.086	0.091	0.081	**0.106**
Q8	0.054	0.046	0.053	0.053	0.057	0.048	0.055	0.071	0.075	**0.085**	0.080

Table 7 Sizes and powers of different testing procedures for detecting the treatment time-lag effect in various cases under the censoring scheme I and the sample size $n_1 = n_2 = 200$. The number in bold in each row is the largest value in that row.

	Methods										
Cases	LR	GE	TW	PE	FH^{01}	FH^{10}	FH^{11}	ZC^{-}	ZC	ZC^{+}	BC
1	0.058	0.0530	0.055	0.058	0.058	0.057	0.066	0.056	0.062	0.061	0.510
E2	0.728	0.104	0.272	0.159	0.988	0.158	0.630	0.935	0.956	0.965	**0.996**
E3	0.590	0.079	0.194	0.109	0.952	0.115	0.461	0.890	0.904	0.923	**0.991**
E4	0.466	0.062	0.134	0.088	0.883	0.083	0.315	0.823	0.844	0.859	**0.983**
E5	0.349	0.063	0.106	0.066	0.783	0.071	0.230	0.755	0.777	0.792	**0.962**
E6	0.268	0.060	0.086	0.059	0.660	0.063	0.164	0.664	0.686	0.716	**0.933**
E7	0.191	0.060	0.072	0.062	0.538	0.063	0.127	0.562	0.607	0.625	**0.869**
E8	0.150	0.057	0.064	0.062	0.419	0.056	0.098	0.492	0.515	0.545	**0.795**
L2	0.938	0.218	0.532	0.329	1.000	0.335	0.928	0.999	0.999	0.999	**1.000**
L3	0.857	0.141	0.396	0.218	1.000	0.218	0.811	0.995	0.997	0.998	**1.000**
L4	0.749	0.095	0.273	0.148	0.992	0.154	0.635	0.985	0.987	0.991	**0.998**
L5	0.611	0.075	0.195	0.104	0.965	0.106	0.467	0.965	0.973	0.977	**0.996**
L6	0.487	0.061	0.131	0.085	0.906	0.083	0.315	0.918	0.930	0.943	**0.987**
L7	0.361	0.062	0.103	0.065	0.810	0.071	0.221	0.848	0.874	0.894	**0.980**
L8	0.275	0.059	0.085	0.059	0.689	0.063	0.161	0.785	0.811	0.833	**0.952**
Q2	0.745	0.113	0.301	0.181	0.990	0.180	0.671	0.947	0.960	0.968	**0.994**
Q3	0.611	0.085	0.212	0.115	0.956	0.119	0.504	0.904	0.922	0.941	**0.986**
Q4	0.485	0.067	0.146	0.093	0.894	0.092	0.340	0.840	0.858	0.873	**0.973**
Q5	0.363	0.062	0.113	0.070	0.795	0.074	0.247	0.775	0.799	0.812	**0.945**
Q6	0.277	0.059	0.086	0.061	0.675	0.067	0.177	0.695	0.711	0.731	**0.912**
Q7	0.202	0.059	0.076	0.062	0.556	0.063	0.135	0.591	0.625	0.652	**0.849**
Q8	0.159	0.055	0.067	0.061	0.435	0.057	0.105	0.522	0.536	0.567	**0.779**

Table 8 Sizes and powers of different testing procedures for detecting the treatment time-lag effect in various cases under the censoring scheme II and the sample size $n_1 = n_2 = 200$. The number in bold in each row is the largest value in that row.

	Methods										
Cases	LR	GE	TW	PE	FH^{01}	FH^{10}	FH^{11}	ZC^{-}	ZC	ZC^{+}	BC
1	0.059	0.051	0.049	0.052	0.064	0.056	0.052	0.064	0.065	0.057	0.060
E2	0.217	0.051	0.080	0.075	0.534	0.072	0.240	0.434	0.470	0.500	**0.656**
E3	0.142	0.052	0.056	0.053	0.332	0.061	0.142	0.305	0.320	0.346	**0.485**
E4	0.093	0.054	0.056	0.055	0.210	0.053	0.096	0.210	0.225	0.244	**0.340**
E5	0.064	0.052	0.055	0.055	0.132	0.052	0.077	0.140	0.156	0.164	**0.228**
E6	0.055	0.053	0.054	0.057	0.087	0.050	0.060	0.096	0.112	0.118	**0.151**
E7	0.055	0.050	0.051	0.052	0.069	0.053	0.052	0.082	0.087	0.087	**0.098**
E8	0.055	0.051	0.050	0.050	0.069	0.054	0.052	0.072	0.075	0.075	**0.077**
L2	0.556	0.076	0.183	0.148	0.940	0.151	0.639	0.876	0.901	0.918	**0.974**
L3	0.379	0.058	0.100	0.093	0.798	0.097	0.397	0.752	0.783	0.813	**0.916**
L4	0.223	0.051	0.074	0.069	0.576	0.065	0.231	0.581	0.613	0.648	**0.787**
L5	0.142	0.051	0.055	0.055	0.362	0.056	0.132	0.414	0.446	0.484	**0.604**
L6	0.087	0.053	0.055	0.055	0.216	0.053	0.090	0.284	0.301	0.335	**0.423**
L7	0.062	0.051	0.052	0.054	0.123	0.052	0.073	0.179	0.197	0.214	**0.265**
L8	0.056	0.050	0.051	0.055	0.085	0.052	0.058	0.129	0.129	0.136	**0.163**
Q2	0.253	0.054	0.087	0.078	0.584	0.080	0.281	0.483	0.520	0.551	**0.679**
Q3	0.153	0.052	0.063	0.058	0.374	0.062	0.161	0.353	0.372	0.411	**0.527**
Q4	0.106	0.054	0.056	0.056	0.240	0.053	0.107	0.254	0.275	0.282	**0.382**
Q5	0.068	0.053	0.055	0.055	0.151	0.056	0.079	0.169	0.186	0.195	**0.253**
Q6	0.058	0.051	0.053	0.055	0.101	0.051	0.066	0.118	0.136	0.139	**0.280**
Q7	0.057	0.049	0.051	0.056	0.076	0.051	0.056	0.092	0.091	0.098	**0.112**
Q8	0.055	0.051	0.050	0.050	0.069	0.053	0.051	0.077	0.081	0.078	**0.084**

Table 9 Calculated *p*-values of various methods for detecting treatment effect in the Rats data example. The number in bold denotes the smallest *p*-values.

LR	GE	TW	PE	FH^{01}	FH^{10}	FH^{11}	ZC^{-}	ZC	ZC^{+}	BC
0.231	0.686	0.512	0.454	0.038	0.584	0.110	0.149	0.050	0.012	**0.003**

Table 10 Calculated *p*-values of various methods for detecting the delayed effect in the Veteran data example. The numbers in bold denote the smallest *p*-values.

LR	GE	TW	PE	FH^{01}	FH^{10}	FH^{11}	ZC^{-}	ZC	ZC^{+}	BC
0.518	0.890	0.903	0.933	0.169	0.936	0.725	0.141	0.095	0.042	**0.039**