Hindawi
Journal of Probability and Statistics
Volume 2023, Article ID 9945446, 10 pages
https://doi.org/10.1155/2023/9945446

Research Article

New Test for the Comparison of Survival Curves to Detect Late Differences

Ildephonse Nizeyimana,[1] **Samuel Mwalili**,[2] **and George Orwa**[2]

[1]*Pan African University Insitute for Basic Sciences, Technology and Innovation, Nairobi, Kenya*
[2]*Jomo Kenyatta University of Agriculture and Technology, Juja, Kenya*

Correspondence should be addressed to Ildephonse Nizeyimana; ildenize01@gmail.com

Received 28 December 2022; Revised 3 February 2023; Accepted 14 February 2023; Published 1 March 2023

Academic Editor: Chin-Shang Li

Background. Survival analysis attracted the attention of different scientists from various domains such as engineering, health, and social sciences. It has been widely exploited in clinical trials when comparing different treatments looking at their survival probabilities. Kaplan–Meier curves plotted from the Kaplan–Meier estimates of survival probabilities are used to depict the general image for such situations. *Methods*. The weighted log-rank test has been dealt with by suggesting different weight functions which give specific strength in specific situations. In this work, we proposed a new weight function comprising all numbers at risk, i.e., the overall number at risk and the separate numbers at risk in the groups under study, to detect late differences between survival curves. *Results*. The new test has been found to be a good alternative after the FH (0, 1) test in detecting late differences, and it outperformed all tests in case of small samples and heavy censoring rates according to the simulation studies. The new test kept the same strength when applied to real data where it showed itself to be among the powerful ones or even outperforms all other tests under consideration. *Conclusion*. As the new test stays stronger in the case of small samples and heavy censoring rates, it may be a better choice whenever targeting the detection of late differences between the survival curves.

1. Introduction

Survival analysis has so many applications in the real world such as engineering like testing the lifetime of life bulbs, medicine like testing the efficiency of different treatments, and it finds even its role in social sciences. In medical research studies, the comparison of two medical treatments is of crucial importance because they help to decide on which treatment works better than another. This is where the comparison of survival curves has its role.

The comparison of survival curves is done when two or more samples are submitted to different treatments or drugs. When comparing drugs, they test them on parallel groups and they decide which one is more efficient. Efficiency may be referred to as the time it takes to cause positive effect if any and at which percentage. For the comparison of survival curves, we consider and record the survival probabilities at each instant of interest for the groups or samples under consideration and we draw the Kaplan–Meier curves and compare them using different techniques. Different scenarios are explored and some tests are more powerful in specific scenarios accordingly. Such scenarios are proportional hazards, early differences, and late differences. Some also include middle differences even though they do not attract the attention of many and this may probably be due to the fact that it rarely happens. The test that is explored in this research is appropriate while investigating the late differences between curves.

2. Materials and Methods

2.1. Weighted Log-Rank Test. The weighted log-rank test is sometimes used in testing the equality of survival distributions. Taking the case of two groups or two treatments, the type of hypotheses that are being tested is of the following form:

H_0: $S_1(t) = S_2(t)$ for all t, against

H_1: $S_1(\mathrm{t}) \neq S_2(t)$ for some t, where $S_i(t)$ is the survival in group i at time t.

In the case of nonproportional hazard rates, the comparison of survival curves is preferably done using different weighted log-rank tests. The weight function is of crucial role, and its misspecification leads to inaccurate results and will cause the loss of power of the test.

The weighted log-rank statistic is written in a stochastic integral form by the following quantity:

$$L_w = \int_0^\tau w(t) \frac{R_1(t) R_2(t)}{R(t)} \left[\frac{dN_1(t)}{R_1(t)} - \frac{dN_2(t)}{R_2(t)} \right], \tag{1}$$

where τ is the total time of the study, w (t) is the weight function at time t, $R_i(t)$ is the number of items/individuals at risk at time t in the i^{th} group, $R(t)$ is the overall number of items/individuals at risk at time t, and $N_i(t)$ is the number of items/individuals which underwent the event of interest by time t in the i^{th} group [1]-[2].

The variance of this weighted log-rank statistic is estimated by the quantity:

$$\widehat{\sigma}^2_{L_w} = \int_0^\tau w^2(t) \frac{R_1(t) R_2(t)}{R(t)} \frac{dN(t)}{R(t)}, \tag{2}$$

where $N(t) = N_1(t) + N_2(t)$.

Computationally, the weighted log-rank statistic is written as follows:

$$U = \sum_{j=1}^{k} w_j \left(d_{ij} - d_j \frac{r_{ij}}{r_j} \right), \tag{3}$$

where w_j is the weight at time t_j, r_j is the overall number of items/individuals at risk at time t_j, r_{ij} is the number of items/individuals at risk at time t_j in the i^{th} group, d_{ij} is the number of events of interest at time t_j in the i^{th} group, and d_j is the overall number of events of interest at time t_j.

The statistic U is such that its expected value is E [U] = 0 and $\mathrm{Var}(U) = \sum_{j=1}^k w_j^2 (r_{1j} r_{2j} d_j (r_j - d_j))/(r_j^2 (r_j - 1))$, and hence, the statistic to be computed becomes

$$\chi^2_{wl} = \frac{\left(\sum_{j=1}^k w_j \left(d_{ij} - d_j \left(r_{ij}/r_j \right) \right) \right)^2}{\sum_{j=1}^k w_j^2 \left(r_{1j} r_{2j} d_j \left(r_j - d_j \right) / r_j^2 \left(r_j - 1 \right) \right)}, \tag{4}$$

where r_j is the overall number at risk in both groups at time t_j and r_{ij} is the number at risk in the i^{th} group at time t_j. We recall that the statistic mentioned above is asymptotically chi-square distributed ($\chi^2_{wl} \sim \chi^2(1)$) and can be reduced to a normal distributed statistic as follows:

$$T = \frac{\sum_{j=1}^k w_j \left(d_{ij} - d_j \left(r_{ij}/r_j \right) \right)}{\sqrt{\sum_{j=1}^k w_j^2 \left(r_{1j} r_{2j} d_j \left(r_j - d_j \right) / r_j^2 \left(r_j - 1 \right) \right)}} \sim N(0,1). \tag{5}$$

The weighted log-rank test statistic contains all three quantities, while the weights considered by different researchers were based on r_j transformed differently or the overall survival probability [3]. Even the survival probabilities considered were the overall ones for the overall sample.

One of the famous weight functions is displayed in Table 1.

Various modifications and improvements have been made to get more powerful weight functions. For example, Garès et al. [4] used the $G^{\rho,\gamma}$ family of tests which was proposed by Fleming and Harrington [5] to investigate the late effects in controlled trials. There exists another test statistic found from a given number of FH statistics tests and it is called Max-combo test statistic [6]-[7]. This test is calculated as the maximum (linear) combination of a selected set of FH tests ($G^{1,1}$), ($G^{1,0}$), ($G^{0,1}$), and ($G^{0,0}$). This technique was introduced because nearly each test statistic has high power in a specific situation, and it would be more helpful to know the situation before.

However, it is not easy to know if in the situation under study, there are early or late effects. FH (0, 1) is more powerful in the case of late effects or late separation of survival curves, while FH (1, 0) becomes more powerful in the case of early effects or early separation of the survival curves. The lack of prior knowledge about the (location of) effects is the cause of using the combination of two or more tests in order to capture every feature [7].

According to the work done by Rückbeil et al. [6], they dealt with the Max-Combo test statistic from three standardized FH tests which are ($G^{1,0}$), ($G^{0,1}$), and ($G^{0,0}$) under five different randomization procedures. They compared the separate FH tests and Max-Combo test, and it was found that the Max-Combo test in each case was the second in power where the highest power of Max-Combo of 83% was observed when they were assessing late treatment effects.

The study Lee [8] has dealt with the standardization of the weighted log-rank test statistics and the Max-combo test statistics Lin et al. [9]. This is the statistic divided by the square root of its variance estimate. Three cases were considered for multiple standardized weighted log-rank test statistics. Considering the corresponding Z statistics Z_1 and Z_2 from ($G^{1,0}$) and ($G^{0,1}$), respectively, as studied by [8]; the three cases are as follows:

(i) The average of the absolute values. This is, $(|Z_1| + |Z_2|)/2$.

(ii) The absolute value of the average. This is, $|Z_1 + Z_2|/2$.

(iii) The maximum of the absolute values. This is, $Max(|Z_1|, |Z_2|)$.

Lee [10] evaluated the maximum and average of ($G^{0,0}$), ($G^{0,2}$), ($G^{2,0}$), and ($G^{2,2}$). Karrison [11] considered Max $(|Z_1|, |Z_2|, \|Z_3|)$, where the Z statistics Z_1, Z_2, and Z_3 were from ($G^{0,0}$), ($G^{0,1}$), and ($G^{1,0}$). This combination covers a good range of possibilities including early differences or late ones and proportional hazards features.

Abou-Shaara [12] studied the similarities between the Kaplan–Meier and ANOVA in his work, and he finally found that the two methods lead to the same conclusion.

There can be a need of estimating the confidence interval of the estimated probability [13], and it is found as follows:

Table 1: Some famous weight functions.

Tests	Weight functions
Log-rank	1
Gehan–Wilcoxon	r_j
Tarone–Ware	$\sqrt{r_j}$
Peto-Peto	$\widetilde{S}(t)$
Modified Peto-Peto	$(\widetilde{S}(t))/(r_j+1)$
Fleming–Harrington	$\widehat{S}(t_{j-1})^{\rho}[1-\widehat{S}(t_{j-1})]^{\gamma}$

$$\widehat{S}(t) \pm 1.96\sqrt{\widehat{\mathrm{Var}}[\widehat{S}(t)]}, \tag{6}$$

where $\widehat{\mathrm{Var}}[\widehat{S}(t)]$ is computed according to Greenwood's formula as follows:

$$\widehat{\mathrm{Var}}[\widehat{S}(t)] = \widehat{S}(t)^2 \sum_{j:\, t_j \le t} \frac{d_j}{r_j\left(r_j - d_j\right)}. \tag{7}$$

Klein et al. [14] proposed a test called a naive test of the null hypothesis for some fixed time points. Such test might be obtained from cumulative hazards $\widehat{H_i}(t)$ or survival probabilities $\widehat{S_1}(t)$.

Qian and Zhou [15] proposed a family of hazard rate functions of hyperbolic-cosine-shaped (CH) type and the deduced CH class weight functions generated good statistic tests for the late differences detection.

2.2. New Weight Function. The existing weight functions are built-in functions of r_j and, hence, vary in function of the total remaining number of individuals at risk in general. The use of r_j transformed in different ways shows that only the size of the total number of individuals at risk in general is taken into account. However, the separate numbers r_{1j} and r_{2j} of individuals at risk in each of the groups would be involved and may probably help to capture more features. The involvement of r_{1j} and r_{2j} separately in the weight will help to detect the difference in the occurrence of the event interest in the two groups at each time point depending on the relation between the two numbers. There is, therefore, a need of a new weight function comprising simultaneously r_j, r_{1j}, and r_{2j} which will change in function of the three variables and hence probably take into account the variations between r_{1j} and r_{2j}. This new weight is thought of being more adaptive since it captures, to some extent, the difference in variations between r_{1j} and r_{2j} by itself and it will be relatively small (big) for small (big) differences in the two quantities. In other words, if the occurrences are likely equal in both the groups, the weight will be relatively less heavy than when the occurrences will be higher in one group than another. While r_j was considering the overall change (and hence general occurrences), separate changes in numbers of individuals at risk in the respective groups are needed for the search of more accuracy and precision of the test.

The new weight function that has been proposed in this study is of the following form:

$$w_j = w_j\left(r_j, r_{1j}, r_{2j}\right) = \frac{r_j}{r_{1j}r_{2j}}, \tag{8}$$

and according to its form, this weight function is monotone increasing. For different couples (r_{1j}, r_{2j}) whose sum is r_j, the new weight will be relatively higher as the difference between r_{1j} and r_{2j} increases compared to when the two numbers are nearly equal.

The stochastic form of the first statistic will be reduced to

$$L_{wNew} = \int_0^{\tau}\left[\frac{dD_1(t)}{R_1(t)} - \frac{dD_2(t)}{R_2(t)}\right], \tag{9}$$

with its corresponding variance which is as follows:

$$\widehat{\sigma}^2_{L_{wNew}} = \int_0^{\tau}\frac{R(t)}{R_1(t)R_2(t)}\frac{dD(t)}{R(t)}. \tag{10}$$

From the direct observation, it can be seen that this statistic depends on the variations in numbers of events in the respective groups, which may lead to the probable predicted sensitivity.

Substituting the new weight function in the general weighted log-rank statistic, we obtain the new statistic which is as follows:

$$\chi^2_{wl} = \frac{\left(\sum_{j=1}^{k}\left(r_j/r_{1j}r_{2j}\right)\left(d_{ij} - d_j\left(r_{ij}/r_j\right)\right)\right)^2}{\sum_{j=1}^{k}\left(r_j/r_{1j}r_{2j}\right)^2\left(r_{1j}r_{2j}d_j\left(r_j - d_j\right)/r_j^2\left(r_j - 1\right)\right)}, \tag{11}$$

or simply

$$\chi^2_{wl} = \frac{\left(\sum_{j=1}^{k}\left(r_j/r_{1j}r_{2j}\right)\left(d_{ij} - d_j\left(r_{ij}/r_j\right)\right)\right)^2}{\sum_{j=1}^{k}\left(d_j\left(r_j - d_j\right)/r_{1j}r_{2j}\left(r_j - 1\right)\right)}. \tag{12}$$

2.3. Power and Relative Efficiency of a Test. The power of the test statistic is by default expressed as follows: $1-\beta$, where β is the probability of type two error. With the statistic of the weighted log-rank test, we have quantities which help to get the power. Assuming the quantity $U = \sum_{j=1}^{k}\widetilde{w}_j(d_{ij} - d_j(r_{ij}/r_j))$ found in the numerator, we have the corresponding variance $V = \sum_{j=1}^{k}\widetilde{w}_j^2(r_{1j}r_{2j}d_j(r_j - d_j))/(r_j^2(r_j - 1))$ on the denominator, and they are such that $(U/\sqrt{V}) \sim N(0,1)$ [16]. The power of the test statistic is then computed as follows:

$$Power = p\left(\frac{U}{\sqrt{V}} > \Phi^{-1}\left(\frac{1-\alpha}{2}\right)\right). \tag{13}$$

Since the p value is also one among the methods of testing the hypothesis, it is good to recall how it is found from the two statistics. With U and V, the one-side$p$ value is calculated as follows:

P – value $= \Phi(U/\sqrt{V})$ [17].

Having two weighted log-rank statistics T_w and T_w, the ARE of T_w relative to T_l as proposed by Jiménez et al. [18] is given by

$$RE\left(T_w, T_l\right) = \left(\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}\left(Power\left(T_w\right)\right)}{\Phi^{-1}(1-\ \alpha) + \Phi^{-1}\left(Power\left(T_l\right)\right)}\right)^2, \tag{14}$$

where Φ^{-1} is the quantile function of the standard normal distribution and $\alpha = 0.05$.

Computationally, the power of the Z statistic obtained from the log-rank test is found as follows:

$$Power\left(T_w\right) = \frac{1}{M}\sum_{1}^{M} 1\left(Z_w > \Phi^{-1}(1-\alpha)\right), \tag{15}$$

where M is the number of simulations which were performed (example: 10,000, 5,000, 1,000, ...), and in our computations, we used $M = 5000$.

3. Data Analysis

3.1. Simulation Study Scenario. The ideal illustration of late separation is depicted in Figure 1. To carry out the simulation study, we used the simsurv *R* package which helped to simulate survival times from standard parametric distributions. In our case, we used the Weibull distribution to simulate the survival times. For one group, we generated the survival times using the Weib (1.2, 3.6), while for the second group, the survival times were generated from Weib (2.9, 5.4) (60% of the survival times for this group) and the remaining (40%) were generated from Weib (1.5, 3.6).

For any case, we performed 5,000 simulations, and the analysis was done by *R*. We considered the cases of equal sample sizes in all our simulations. The notation ($n1$, $n2$) (c) has been used, where $n1 = n2$ represents the sample size under consideration and $n1 = n2$ is the number of individuals in each group and c is the overall censoring rate. The censoring rates taken into account are 20%, 40%, and 60% and $c = 0$ means that there has been no censoring. There are therefore four simulation cases for each sample size. The used sample sizes per group are 20, 50, 80, and 100.

3.2. Simulation Results. To make it more visible and separate, we look at the following plot in Figure 2 which shows graphically the variations in power as obtained in Table 2. To read the plots well, NoCens100 stands for the case of no censoring in the case of a sample size of 100 individuals per group. It is the same for 80, 50, and 20. Cens10020 stands for the case of 100 individuals per group with the overall censoring rate of 20% and the same analogy applies to others.

The new test may be recommended as an alternative of test while aiming at the detection of late differences between treatments. It imposes itself as a good choice when the sample size becomes smaller. In other words, the new test outperforms the existing ones for small sample sizes ($n \leq 50$) . To see this more clearly, we used the relative efficiencies of all tests (in power) compared to the standard log-rank test. We will mainly look at FH (1, 1) and FH (0, 1), and the new test looks to be relatively more efficient. In regard to the efficiency of the tests, we evaluate them relatively to the standard log-rank test. This last is known to perform better in the case of proportional hazards but still keeps some level of power in other scenarios. Even in our case of late differences detection, it was the third choice after being outperformed by our newly proposed test. Table 3 shows the heatmap of the relative efficiencies of other tests at all levels of censoring under consideration with respect to the LR test.

As seen on Figure 3, the graph at the left side is a random simulation for sample size $n = 100$, while the right one is for $n = 20$, and the censoring rate is 20% in both cases. As it can be seen, the FH (0, 1) weight in dashed red increases gradually and this justifies its high power for late differences. The separation of curves usually happens gradually, and hence, as the difference becomes higher, the FH (0, 1) weight becomes higher too.

For the new weight in solid blue, there is only a brutal increase in a very small number of time points at the end, while it is relatively very small since the beginning of the study. This behavior can help us to justify its efficiency for the case of small sample size because in such case, the late separation does not take longer, and hence, the new weight will not lose many event times of the separation. Apart from this, the new weight could be powerful in case of brutal separation in the very last few event times, and this may not happen often practically. However, again, in the few cases of strength, the new weight can reach to numbers above 1 as seen on the graph at the right where it even reached 2 at one last point. It is clear that the relative weakness of the new weight for large sample sizes resides in that fact of failing to capture some event times at the beginning of the separation which may normally start around the middle time of the study and remain sensitive to a very limited number of last event times as both graphs show. In contrast, FH (0, 1) captures gradually all separation since their occurrence as shown by its gradual shape or gradual increase. We recall that where the new weight drops to 0 is when the number at risk in one of the groups becomes 0 because there is no comparison at such points and onward. To make it well understood, assume the separation happened at the time point 30 (graph at the left). We can see how much is the difference between the two weights since then and hence the loss of power for the new weight. For the graph at the right, if the separation started from time point 25, for example, we notice that the difference between the two weights is not that high as at the left side case. But again, we may highlight that the new weight is very strict on the very late few event times with exceptionally higher weights. The lower loss of power for the new weight in the case of censoring resides in the fact that this last reduces the number of event times, and because the new weight needs just the very last few event times, it does not lose too much power as the FH (0, 1) which might have benefited from many event times since the beginning of the separation. This is why small sample size cases and heavy censoring cases are the favoring ones for the new weight which needs just fewer last event times than FH (0, 1). This is not strange because every weight function has some circumstances when it excels in power but fails in others. Our newly suggested weight is then powerful in heavy censoring and/or small sample size cases.

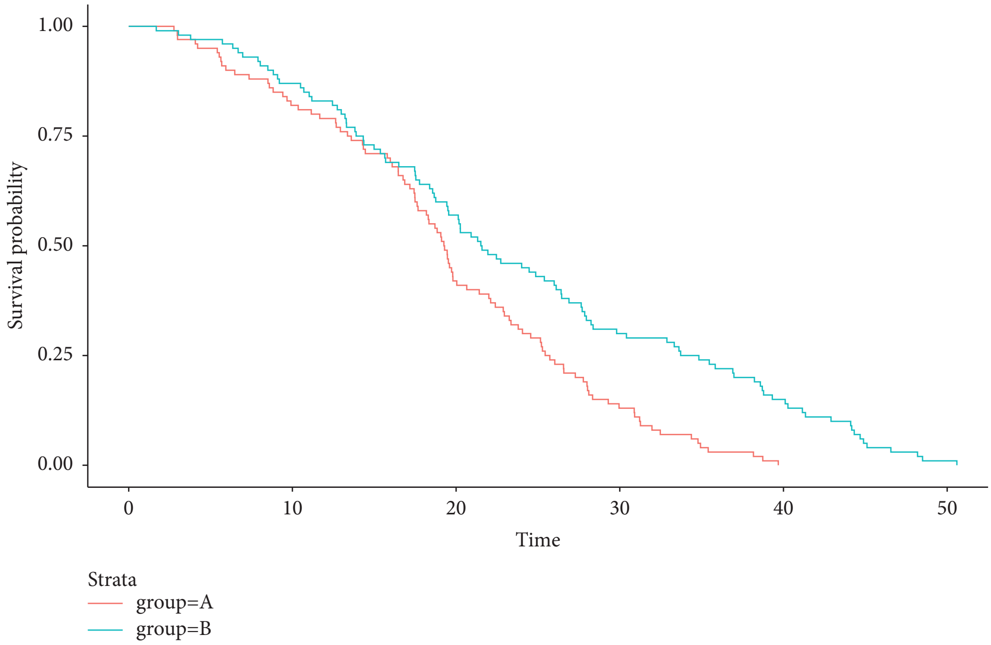
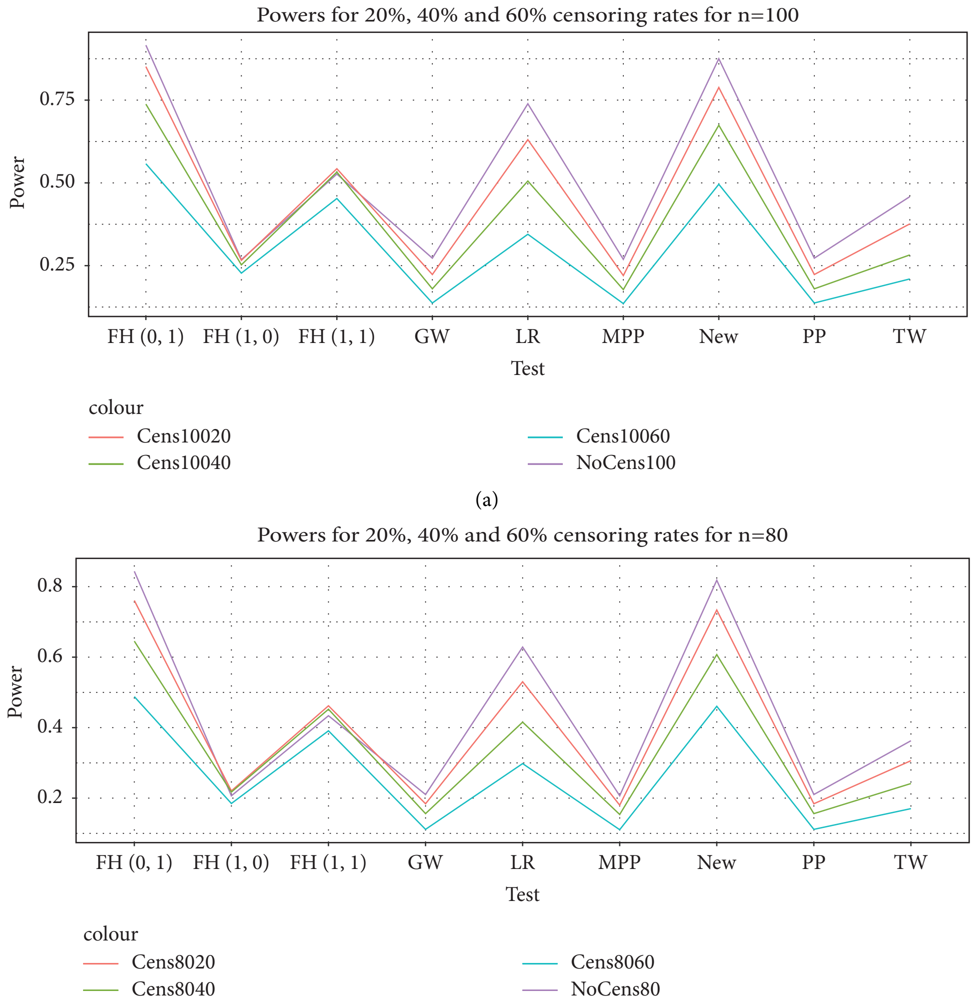


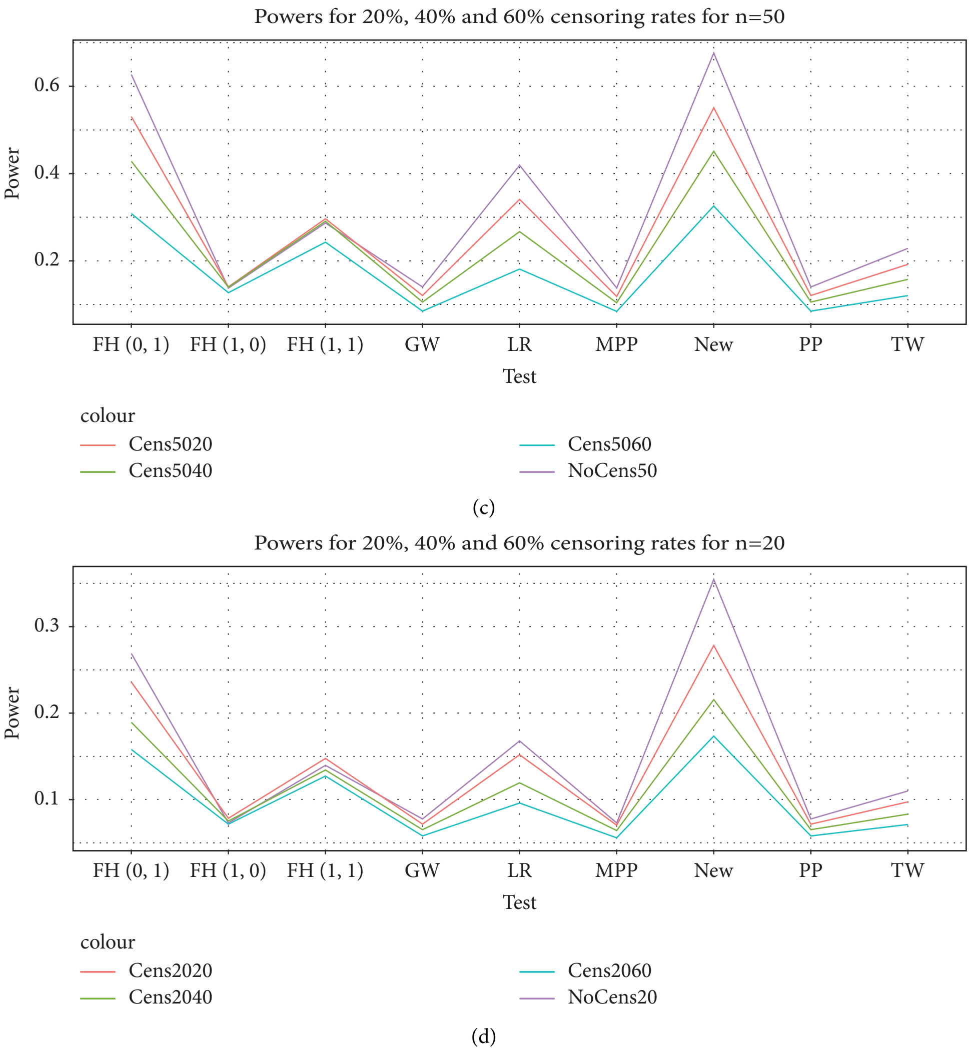
Figure 1: Illustration for late separation between KM curves.



(a)

(b)

Figure 2: Continued.



(c)

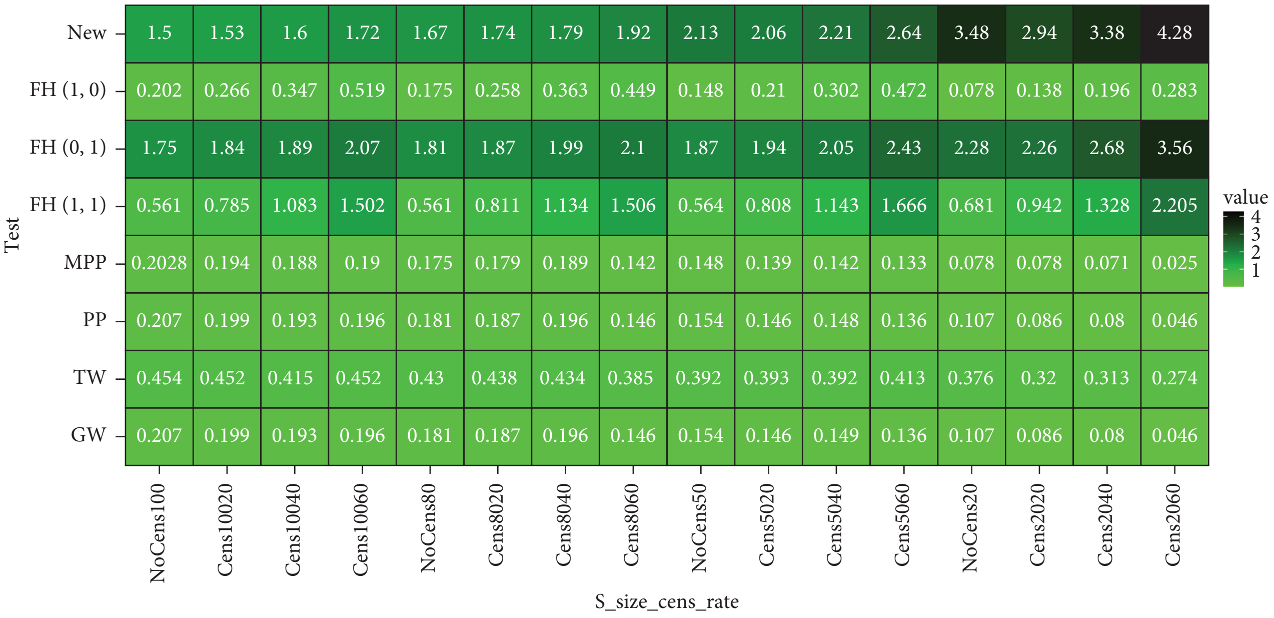
(d)

Figure 2: Power plots for different sample sizes and censoring rates. (a) Powers for 20%, 40%, and 60% censoring rates for $n = 100$. (b) Powers for 20%, 40%, and 60% censoring rates for $n = 80$. (c) Powers for 20%, 40%, and 60% censoring rates for $n = 50$. (d) Powers for 20%, 40%, and 60% censoring rates for $n = 20$.

Table 2: Estimated powers of the weighted log-rank test statistics for late differences.

(n1, n2) (censoring percent)	LR	GW	TW	PP	MPP	FH (1, 1)	FH (0, 1)	FH (1, 0)	New
(100, 100) (0)	0.7390	0.2726	0.4584	0.2726	0.2690	0.5266	0.9160	0.2682	0.8752
(100, 100) (20)	0.6306	0.2232	0.3762	0.2232	0.2200	0.5428	0.8512	0.2660	0.7882
(100, 100) (40)	0.5056	0.1800	0.2824	0.1800	0.1776	0.5326	0.7380	0.2524	0.6742
(100, 100) (60)	0.3448	0.1372	0.2098	0.1372	0.1354	0.4528	0.5578	0.2274	0.4958
(80, 80) (0)	0.6286	0.2102	0.3628	0.2102	0.2068	0.4338	0.8436	0.2066	0.8182
(80, 80) (20)	0.5302	0.1840	0.3062	0.1840	0.1800	0.4620	0.7608	0.2206	0.7338
(80, 80) (40)	0.4156	0.1560	0.2414	0.1560	0.1536	0.4522	0.6454	0.2172	0.6072
(80, 80) (60)	0.2980	0.1114	0.1702	0.1114	0.1104	0.3910	0.4880	0.1848	0.4606
(50, 50) (0)	0.4190	0.1402	0.2286	0.1402	0.1378	0.2868	0.6272	0.1376	0.6764
(50, 50) (20)	0.3410	0.1206	0.1920	0.1206	0.1184	0.2964	0.5306	0.1404	0.5508
(50, 50) (40)	0.2670	0.1056	0.1576	0.1056	0.1040	0.2908	0.4284	0.1396	0.4512
(50, 50) (60)	0.1812	0.0848	0.1204	0.0848	0.0842	0.2428	0.3084	0.1270	0.3258
(20, 20) (0)	0.1678	0.0776	0.1100	0.0776	0.0730	0.1396	0.2690	0.0730	0.3546
(20, 20) (20)	0.1518	0.0716	0.0974	0.0716	0.0704	0.1476	0.2360	0.0784	0.2782
(20, 20) (40)	0.1194	0.0652	0.0832	0.0652	0.0642	0.1342	0.1894	0.0752	0.2156
(20, 20) (60)	0.0960	0.0580	0.0712	0.0580	0.0558	0.1272	0.1580	0.0716	0.1734

Table 3: Heatmap for relative efficiencies compared to the standard log-rank test.



Test	NoCens100	Cens10020	Cens10040	Cens10060	NoCens80	Cens8020	Cens8040	Cens8060	NoCens50	Cens5020	Cens5040	Cens5060	NoCens20	Cens2020	Cens2040	Cens2060
New	1.5	1.53	1.6	1.72	1.67	1.74	1.79	1.92	2.13	2.06	2.21	2.64	3.48	2.94	3.38	4.28
FH (1, 0)	0.202	0.266	0.347	0.519	0.175	0.258	0.363	0.449	0.148	0.21	0.302	0.472	0.078	0.138	0.196	0.283
FH (0, 1)	1.75	1.84	1.89	2.07	1.81	1.87	1.99	2.1	1.87	1.94	2.05	2.43	2.28	2.26	2.68	3.56
FH (1, 1)	0.561	0.785	1.083	1.502	0.561	0.811	1.134	1.506	0.564	0.808	1.143	1.666	0.681	0.942	1.328	2.205
MPP	0.2028	0.194	0.188	0.19	0.175	0.179	0.189	0.142	0.148	0.139	0.142	0.133	0.078	0.078	0.071	0.025
PP	0.207	0.199	0.193	0.196	0.181	0.187	0.196	0.146	0.154	0.146	0.148	0.136	0.107	0.086	0.08	0.046
TW	0.454	0.452	0.415	0.452	0.43	0.438	0.434	0.385	0.392	0.393	0.392	0.413	0.376	0.32	0.313	0.274
GW	0.207	0.199	0.193	0.196	0.181	0.187	0.196	0.146	0.154	0.146	0.149	0.136	0.107	0.086	0.08	0.046

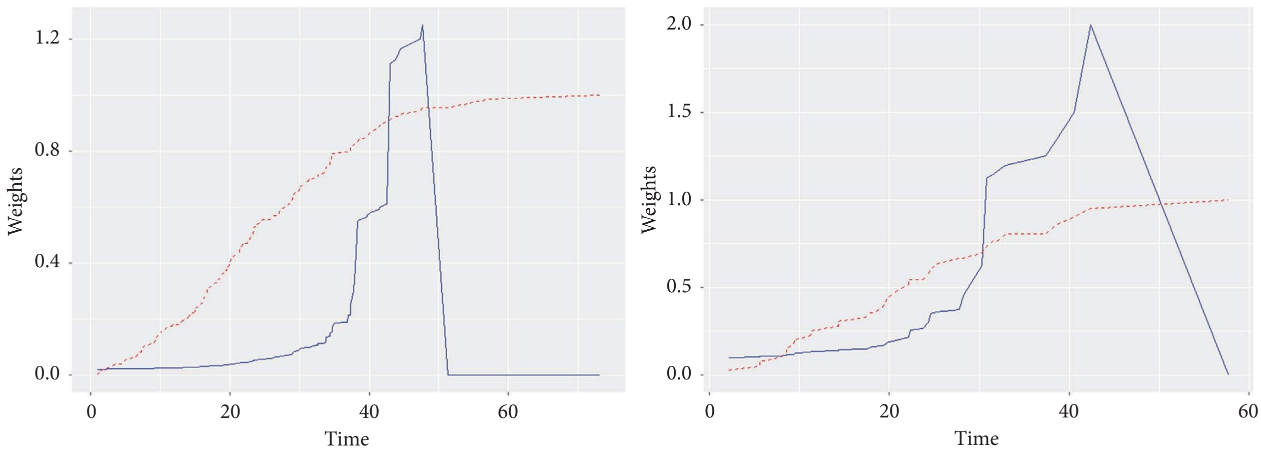


Figure 3: Graphical illustration of the two competing weights: FH (0, 1) and the new weight.

3.3. Discussion. As shown by the heatmap, we have three relatively powerful tests when compared to the standard log-rank test. Those are FH (1, 1), FH (0, 1), and the new test. FH (1, 1) is more powerful than LR when the censoring rate is higher than 50% since the relative efficiency has been more than 150% only in the case where the censoring was 60%. For the cases of no censoring and for those of censoring of 20%, the test has not been more efficient than the standard LR test irrespective of the sample size under consideration.

The FH (0, 1) test, which is usually known to be the most powerful for late differences, still keeps its power, but it becomes outperformed by the newly proposed test for small sample sizes, that is, for $n \leq 50$. We can take two extreme points for the two tests. For $n = 100$ with no censoring, the RE of FH (0, 1) was 175% while it was 150% for the new test. This implies that the difference in relative efficiency is 25% (or we can say that FH (0, 1) is 25% more relatively powerful than the new test when both are compared to the LR for $n = 100$.)

For $n = 20$ with the censoring rate of 60%, the RE of FH (0, 1) is 356%, while it is 428% for the new test, and this implies that the new test is 72% relatively more powerful than FH (0, 1) when both tests are compared to the LR test.

So, we can see that the new test will make a higher difference in relative efficiency where it is relatively powerful than what FH (0, 1) does in its favorable conditions. Noting the importance of sample size, the new test may be a good recommendation due to its behavior in case of small samples and heavy censoring.

To get a more general recommendation between the two tests, we can do an unweighted sum of differences of relative efficiencies in all cases under study and see the result. That is, we take the relative efficiencies for FH (0, 1) minus those of the new test (RE (FH (0, 1))—RE (new test)) in each case and

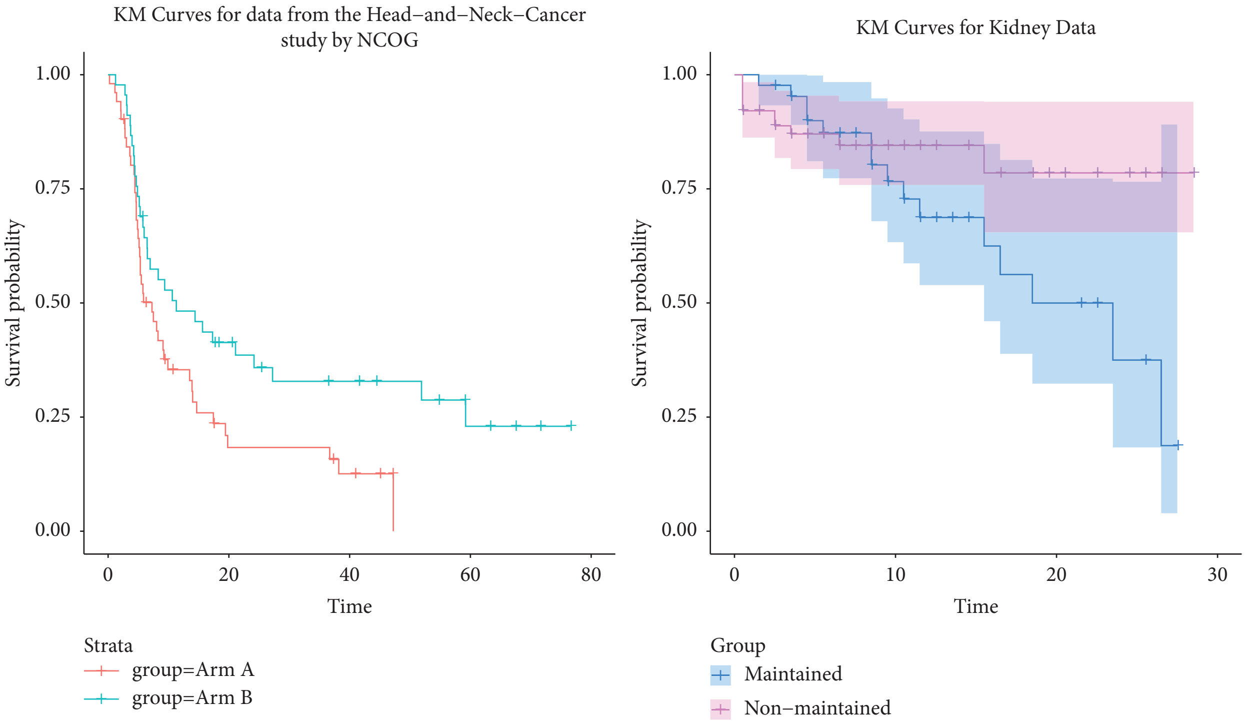


Figure 4: Graphs for real data application cases.

we sum up to see which one is generally relatively more efficient. Operating on the data in the heatmap, we obtain −220% in total, which shows that the new test is relatively more efficient than FH (0, 1) in general. This is immediately linked to the fact that where the new test is relatively more efficient, it makes bigger differences.

3.4. Application to Real Data. To check the reliability of the new test, we preferred using two real datasets to be sure of the comparison. Those datasets are as follows:

(i) Head-and-Neck-Cancer Study by the Northern Oncology Group (NCOG)

(ii) Time to infection of Kidney Dialysis Patients Data

The data from the Head-and-Neck-Cancer Study which was done by Northern Oncology Group (NCOG) are found in Efron [19] and have been reused by many other authors including Qian and Zhou [15] recently. Arm A represents patients who underwent radiation therapy and those who underwent radiation plus chemotherapy were put in Arm B.

For the second dataset of time to infection of kidney dialysis patients, it is a (built-in) dataset found in *R* under KMsurv package. The group was formed referring to the methods for placing catheters in kidney dialysis patients. Surgically placed catheter made group 1 and percutaneously placed catheter made group 2. The plot of Kaplan–Meier curves for both datasets is shown.

From Figure 4, we notice that for NCOG data, the curves are closer to each other at the beginning but separate later where Arm B appears to have higher survival probabilities than Arm A. The two-sided$p$ values for the nine tests have been computed and are given in table.

As seen from the p values in Table 4, the newly proposed test showed itself as stronger than any others as it has the smallest p value of 0.0129, followed by the Fleming–Harrington (FH (0, 1)) with p value = 0.0223 and lastly by the standard log-rank test with p value = 0.047. This is in accordance with the simulation results even though the new test seems to outperform the existing stronger test for late differences, FH (0, 1).

However, this is not strange because even the difference in the powers observed in the simulation was not that high enough that one may not hesitate to recommend this new test as a good choice. The other tests got p values greater than 0.05 because they are usually known to be weak in the detection of late differences, and this is no surprising based on the shape of the two curves. Their failure or weakness to detect such difference might be from their nature. However, since the difference seems to be significant by an immediate look at the graph, if one-sided$p$ values are under consideration, the majority of all these tests could have their p values to be less than 0.05, and hence, the difference might be detected. In such a case, only GW and FH (1, 0) might be the only ones to fail detecting such difference. The general observation which will remain intact is that the new test performed better than any other test in this case.

As it can be immediately observed from the KM curves for kidney data, the two survival curves crossed each other at the early stages where they were even close to each other. After crossing each other, they separated quickly, and this will lead us to the justification of the p value obtained for FH (1, 1) in Table 5. It has been obtained that in addition to the two tests which were expected to detect such differences, we got another one (FH (1, 1)) which is stronger in the detection

Table 4: *P* values of the tests for head-and-neck-cancer data by NCOG.

Tests	*P* value
LR	0.047
GW	0.110
TW	0.0766
PP	0.0846
MPP	0.086
FH (1, 1)	0.060
FH (0, 1)	0.0223
FH (1, 0)	0.151
New	0.0129

Table 5: *P* values of the tests for the time to infection of kidney dialysis patients' data.

Tests	*P* values
LR	0.112
GW	0.963
TW	0.525
PP	0.259
MPP	0.278
FH (1, 1)	0.005
FH (0, 1)	0.0046
FH (1, 0)	0.243
New	0.021

of middle differences. In other words, because it gives heavier weights to middle events and reduces as they go farther from the median time, it detected those differences in this case because in the middle of the study period, the curves had already been separated as it can be immediately seen on the graph. It is to be highlighted that this test has been surprising since it was at the point of outperforming both expected tests with the p value of 0.005. However, FH (0, 1) remained the first among the three tests with the p value of 0.0046 and the new test was the third with p value of 0.021. Contrary to the first NCOG data, even if we had taken one-sided$p$ values, no change might have been observed on the tests with significant p values.

4. Conclusion

The newly proposed test is a good alternative for the detection of late differences between survival curves. It shares the same positive behavior with FH (0, 1) of being relatively more efficient and powerful than the LR, and even though the reduction of power as the censoring rate increase is common, this reduction is relatively small for the new test compared to the remaining others (including the LR test and FH (0, 1)). The new test may, therefore, be the first choice in cases of small sample sizes and heavy censoring rates. The same strength has been observed while dealing with real datasets when the new test remains still sensitive for late differences in survival. Based on the fact that the small size of the sample and censoring are the major threats in survival analysis studies, referring to the power and higher relative efficiency of the new test in such cases, one may consider it as a better choice for late differences detection between survival curves.

Data Availability

The data used to support the findings of this study are publicly and freely available. One dataset is accessed through R software, and another is in the cited research.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank Dr. Bayowa Teniola Babalola for his assistance in conceptual understanding and Marcin Kosiński for his guidance on the programming part.

References

[1] D. M. Zucker and E. Lakatos, "Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment," *Biometrika*, vol. 77, no. 4, pp. 853–864, 1990.

[2] B. R. Logan and S. Mo, "Group sequential tests for long-term survival comparisons," *Lifetime Data Analysis*, vol. 21, no. 2, pp. 218–240, 2015.

[3] P. G. Karadeniz and I. Ercan, "Examining tests for comparing survival curves with right censored data," *Statistics in Transition New Series*, vol. 18, no. 2, pp. 311–328, 2017.

[4] V. Garès, S. Andrieu, J.-F. Dupuy, and N. Savy, "On the fleming—harrington test for late effects in prevention randomized controlled trials," *Journal of Statistical Theory and Practice*, vol. 11, no. 3, pp. 418–435, 2017.

[5] T. R. Fleming and D. P. Harrington, "Counting processes and survival analysis john wiley and sons," *Wiley Inc. New York*, 1991.

[6] M. V. Rückbeil, M. Manolov, and R.-D. Hilgers, "The choice of a randomization procedure in survival studies with non-proportional hazards," *Statistics in Biopharmaceutical Research*, pp. 1–9, 2021.

[7] T. J. Prior, "Group sequential monitoring based on the maximum of weighted log-rank statistics with the fleming–harrington class of weights in oncology clinical trials," *Statistical Methods in Medical Research*, vol. 29, no. 12, pp. 3525–3532, 2020.

[8] S.-H. Lee, "On the versatility of the combination of the weighted log-rank statistics," *Computational Statistics and Data Analysis*, vol. 51, no. 12, pp. 6557–6564, 2007.

[9] R. S. Lin, J. Lin, S. Roychoudhury et al., "Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis," *Statistics in Biopharmaceutical Research*, vol. 12, no. 2, pp. 187–198, 2020.

[10] J. W. Lee, "Some versatile tests based on the simultaneous use of weighted log-rank statistics," *Biometrics*, vol. 52, no. 2, pp. 721–725, 1996.

[11] T. G. Karrison, "Versatile tests for comparing survival curves based on weighted log-rank statistics," *STATA Journal*, vol. 16, no. 3, pp. 678–690, 2016.

[12] H. F. Abou-Shaara, "Scientific note: similarities between survival analysis using kaplan-meier and anova," *Thailand Statistician*, vol. 16, no. 2, pp. 221–229, 2018.

[13] D. G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text*, Springer, Berlin Heidelberg, 2012.

[14] J. P. Klein, B. Logan, M. Harhoff, and P. K. Andersen, "Analyzing survival curves at a fixed point in time," *Statistics in Medicine*, vol. 26, no. 24, pp. 4505–4519, 2007.

[15] K. Qian and X. Zhou, "Weighted log-rank test for clinical trials with delayed treatment effect based on a novel hazard function family," *Mathematics*, vol. 10, no. 15, p. 2573, 2022.

[16] D. Magirr and C.-F. Burman, "Modestly weighted logrank tests," *Statistics in Medicine*, vol. 38, no. 20, pp. 3782–3790, 2019.

[17] D. Magirr, "Non-proportional hazards in immuno-oncology: is an old perspective needed?" *Pharmaceutical Statistics*, vol. 20, no. 3, pp. 512–527, 2021.

[18] J. L. Jiménez, J. Niewczas, A. Bore, and C.-F. Burman, "A modified weighted log-rank test for confirmatory trials with a high proportion of treatment switching," *PLoS One*, vol. 16, no. 11, Article ID e0259178, 2021.

[19] B. Efron, "Logistic regression, survival analysis, and the kaplan-meier curve," *Journal of the American Statistical Association*, vol. 83, no. 402, pp. 414–425, 1988.