Menemui Matematik (Discovering Mathematics)
Vol. 42, No. 2: 86-91 (2020)

Non-Parametric Tests for Comparing More than Two Survivor Functions with Interval Censored Data via Midpoint Imputation

Syaza Diyana Nadhirah Mohamad Suhaini and Jayanthi Arasan[a)]
Department of Mathematics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor.
[a)] jayanthi@upm.edu.my.

ABSTRACT

The objective of this research is to compare the performances of the Log-rank and Wilcoxon non-parametric tests to test differences in the survival of 3 groups of patients when data is complete, right and interval censored. For interval censored data, midpoint imputation is used to estimate the survival time of each individual. A thorough power study was carried out to assess the performance of the tests with various distributions, censoring proportions (cp), sample sizes (n), significance levels (α) and effect sizes (Λ). The results clearly indicate that the Log-rank test performs slightly better than the Wilcoxon test when the proportional hazard (PH) assumption is fulfilled. However, the performance of both tests are almost the same, regardless of the PH assumption at larger n and Λ. Finally, the tests were applied to real data from an AIDS clinical trial, which involve three treatment groups with both right and interval censored data.

Keywords: Log-rank test, Wilcoxon test, Interval Censored, Midpoint Imputation, g-sample

INTRODUCTION

In a clinical trial, it is quite impossible to get all information from patients that are recruited in the study and data is often interval censored. Interval censored data consists of failure time that lies between two endpoints. Two commonly known non-parametric procedures, known as Log-rank and Wilcoxon tests, are designed to compare two or more survival curves when observations are censored. There are many literatures available from previous research on these tests such as Gehan [1], Prentice [2], Mantel [3], Cox [4], and Peto & Peto [5]. In the absence of censoring, Log-rank and Wilcoxon procedures will be reduced to the Mann-Whitney test for two treatment groups and the Kruskal-Wallis test for more than two treatment groups [6]. Since both tests are non-parametric in nature, no assumption about distribution of survival data is needed.

The assessment on performance of Log-rank and Wilcoxon tests begins when Lee et al. [7] compare the size and power of the tests using small samples from several types of survival distributions with and without censoring. Following that, the assessment has been studied extensively by considering various conditions such as different types of distribution, sample size, censoring percentage, Proportional Hazard (PH) assumption and early-or-late hazard differences in survival curves. Authors who have made a contribution include Latta [8], Beltangady and Frankowski [9], Leton and Zuluaga [10], Martinez and Naranjo [11], Darilay and Naranjo [12], Ramakrishman and Ravanan [13] and Karadeniz and Ercan [14] [15].

Most researchers conducted their studies by focusing only on two treatment groups and right censored data in assessing the performance of Log-rank and Wilcoxon tests. The main goal of this research is to assess the performance of both tests in comparing more than two treatment groups when dealing with interval censored data.

Section 2 provides a brief description on survivor function estimator and two non-parametric tests assessed in this study. A brief discussion on several conditions tested on assessing the performance of Log-rank and Wilcoxon tests can be found in Section 3. Section 4 discusses the power study of both non-parametric tests, as well as the real data analysis that is used for illustration purposes on the usage of both non-parametric tests in real life situation. Finally, the performance of both tests will be concluded in the last section.

KAPLAN-MEIER ESTIMATOR AND NON-PARAMETRIC TESTS

A survivor function, $S(t)$, is the probability of a random variable Texceeding the specified time, t. It is the probability that an individual will survives greater than or equal to time t. In practical, this function can only be estimated from a sample of survival time, since it is impossible to get all information that is related to a study within the specified period of time. The most popular proposed non-parametric methods of estimating a survivor is Kaplan-Meier Product Limit (KMPL) estimator which is defined as,

$$S(t)=\prod_{k;t_k\le t}\left(1-\frac{d_k}{n_k}\right). \tag{1}$$

Let say death is the event of interest in the study and $k = 1,2,\dots,r$ is the number of distinct failure time. Then, t_k is a time when at least one death happened, d_j is the number of death that happened at time t_k and n_k is the number of individuals who still alive at time t_k.

The test statistic for Log-rank and Wilcoxon are almost the same in general. However, Wilcoxon test tend to have more weight on the early-stage of the survivor function. In general, the test statistic for both tests is made up of a division between two terms. The first term is a vector of $g-1$ components of U_h, denoted as $\boldsymbol{U}$ and the second term is the variance-covariance matrix of $V_{hh'}$, denoted as $\boldsymbol{V}$. Eq. 2 shows the U_{Lh} components used to obtained the vector $\boldsymbol{U}_L$ of Log-rank test while eq. 3 shows the covariance formula of U_h and $U_{h'}$ components that is used to obtained the variance-covariance matrix of $\boldsymbol{V}_L$,

$$U_{Lh}=\sum_{k=1}^{r}\left(d_{hk}-e_{hk}\right), \tag{2}$$

$$V_{Lhh'}=\sum_{k=1}^{r}\frac{n_{hk}d_k\left(n_k-d_k\right)}{n_k\left(n_k-1\right)}\left(\delta_{hh'}-\frac{n_{h'k}}{n_k}\right). \tag{3}$$

Note that e_{hk} is the expected number of death which is computed as $e_{hk}=n_{hk}d_k/n_k$ and h is the number of treatment groups involve in the study where $h,h'=1,2,\dots,g-1$, which gives $\delta_{h,h'}=I\left[h-h'\right]$. The, the test statistic for Log-rank test is defined as,

$$W_L=U_L'V_LU_L\sim\chi^2_{g-1}. \tag{4}$$

Similarly, the test statistic for Wilcoxon test are made up of two terms constructed with the same component as in eq. 2 and 3. The only difference is that there exist a weight of n_k inside both components as stated in eq. 5 and 6,

$$U_{Wh} = \sum_{k=1}^{r} n_k \left(d_{hk} - e_{hk} \right), \tag{5}$$

$$V_{Whh'} = \sum_{k=1}^{r} n_k^2 \frac{n_{hk} d_k \left(n_k - d_k \right)}{n_k \left(n_k - 1 \right)} \left(\delta_{hh'} - \frac{n_{h'k}}{n_k} \right). \tag{6}$$

Similar to Log-rank test, the test statistic for Wilcoxon test will be

$$W_W = U_W' V_W U_W \sim \chi^2_{g-1}. \tag{7}$$

DESIGN OF EXPERIMENT

Many imputation methods have been employed to deal with interval censored data. Among them are left, interval and right imputation method. This study only implements the midpoint imputation since this method has been proven to provide a better result compared to the other two methods [16]. Let say the exact failure time lies between an interval of $\left(L_i, R_i\right]$ where L_i and R_i are the left and right endpoints of survival times, respectively. By using midpoint imputation, the exact failure time can be estimated by,

$$t_i = \frac{L_i + R_i}{2}. \tag{8}$$

In this research, the performance of Log-rank and Wilcoxon tests are assessed using three different treatment groups only with uncensored data, followed by right censored data and finally with right and interval censored data. There are five conditions considered in the assessments which are believed to affect the performance of Log-rank and Wilcoxon tests. Among them are type of sample distribution, sample sizes (n), censoring proportions (cp), significance levels (α) and effect sizes (Λ). These data were simulated from three different distributions namely Exponential, Weibull and Gompertz. The survival times, t_i, were generated using the inverse transform method where $t_i = F^{-1}(u_i)$ and u_i is a pseudo-random number generated from a Uniform distribution. The data is simulated such that the PH assumption is always fulfilled.

In this research, treatment group with equal size of n of 20, 30, 40 and 60 and significance levels of $\alpha = 0.05$ and $\alpha = 0.10$ is used. For right and interval censoring cases, the assessment of both tests is carried out under censoring percentage of 10%, 15%, 20%, and 25% in each treatment groups. Interestingly for this case, the censoring percentage of both right and interval censored observations will be equal. If survival data consists of 10% censored observations, 5% of them will be right censored and another 5% will be interval censored.

An effect size is determined by transforming the hazard ratio (HR) into a standardized mean difference and takes the absolute value of it [17]. It is defined as small, medium or large when $\Lambda = 0.2$, $\Lambda = 0.5$ and $\Lambda = 0.8$. These values are used based on the correction made by Azuero [17] who simplified some algebra on HR so that it is comparable with Cohen's d difference. Let say $h_1(t)$ and $h_2(t)$ be the hazard rate of Group 1 and Group 2, respectively. Then the hazard ratio of Group 2 over Group 1 is defined as

$$HR = \frac{h_2(t)}{h_1(t)}, \tag{9}$$

and the effect size, Λ, can be obtained by

$$\Lambda = \left| \frac{\sqrt{6}\ln(HR)}{\pi} \right|. \tag{10}$$

The performance of Log-rank and Wilcoxon tests are assessed by using empirical power which ranges from 0 to 1. It is the probability that the test will reject a false H_0 where under H_0, there is no difference in survival between two or more independent groups. An empirical power is used to measure the ability of both tests to detect any differences in probability of surviving among treatments groups. Higher empirical power indicates that the test is more reliable under certain conditions. General simulation study procedures are provided as follows,

1. Generate survival times of 3 treatment groups, independently, from a desired continuous distribution, n, cp, andΛ.
2. Specify the desired α, then, perform the Log-rank and Wilcoxon tests on generated survival data.
3. Check either the tests are significant or not by comparing the p-value and α.
4. Replicate the procedure 5000 times and compute the empirical power.

RESULTS

Power Study of Log-rank and Wilcoxon Tests

In this section, the results of empirical power for Log-rank and Wilcoxon tests are divided into two parts. The first part discusses the performance of both test when incorporating with complete and right censored data while the second part discusses the performance of tests when dealing with right and interval censored data. The assessment is carried out under several conditions as mentioned in section 3.

Results for Complete and Right Censored Data

Figure 1 shows empirical power graph of Log-rank and Wilcoxon tests for survival data with uncensored and right censored observations under different distribution, sample size, level of significance, censoring percentage, and effect size. Based on graph (a) and (d) for Exponential distributions, the power of both tests tend to converge to a value of 0.05 and 0.10 respectively as the sample size increase regardless of their censoring percentage. This is because the effect size used is small, which make the tests less powerful to detect any differences between the survivor functions. In other words, both tests provide large Type I error when the effect size used is small. Similar discussion can be implemented for Weibull and Gompertz distributions.

Meanwhile, graph (b) and (e) show that when the effect size between survivor functions is medium, the Log-rank provides more power under different censoring percentage compared to Wilcoxon since most of the Log-rank lines situated above the Wilcoxon lines. However, the performances for both tests are much better when survival data consists of complete observations rather than censored data for all type of distributions.

As the sample size used is 40 and above, the performances of both tests are almost similar, where they provide good empirical powers. So, regardless of the satisfactory for PH assumption, the performance of both tests will be similarly great under medium effect size if only if the sample size used is large enough. Other than that, the empirical power of these tests will be higher if the

significance level used is large. This claim can be supported by the lines of both tests in graph (e) under all conditions where these lines plotted at higher empirical power compared to lines in graph (b). Unlike in Exponential and Weibull distribution, the power of Log-rank is far way better compared to Wilcoxon when analyzing survival data that follow Gompertz distribution.

Graph (c) and (f) for Exponential and Gompertz distributions show that both tests provide similar power regardless of the sample size, significance level and censoring percentage as long as the effect size between survivor functions are large enough. Contradict to these two distributions, graph (c) and (f) for Weibull distribution show that Log-rank and Wilcoxon tests have similar performance if and only if more than 40 samples are used in each treatment group.

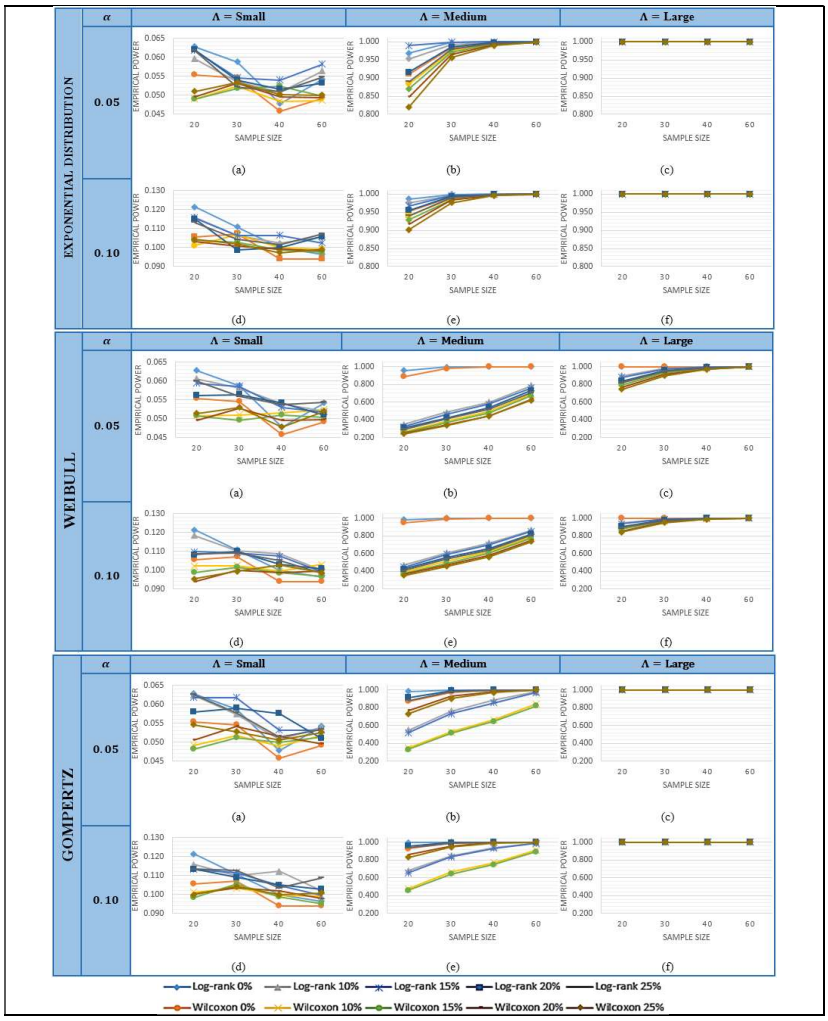

Figure 1: Empirical Power Graph of Log-rank & Wilcoxon test for 3 different distributions with uncensored & right censored data under different $\boldsymbol{n}$, $\boldsymbol{\alpha}$, $\boldsymbol{cp}$ and $\boldsymbol{\Lambda}$.

Results for Right and Interval Censored Data

Figure 2 shows the empirical power graph of Log-rank and Wilcoxon tests for survival data with right and interval censored observations which follow Exponential, Weibull and Gompertz distribution under different sample sizes, level of significance, censoring percentage, and effect

size. Based on graph (a) and (d) for all distributions, the power of both tests tend to converge to a value of 0.05 and 0.10 respectively as the sample size increase regardless of their censoring percentage. Log-rank still performs better than Wilcoxon just like in the previous discussion for complete and right censored data because the PH assumption is satisfied for all distributions. Graph (b) and (e) in Exponential and Gompertz distributions show almost similar empirical power pattern. However, both tests perform better when survival data follow Exponential distribution compared to Gompertz. Surprisingly, Log-rank and Wilcoxon have low performance if the sample size used is less than 30 although the effect size between the survivor functions is medium when survival data follow Weibull distribution. This claim is supported by comparing graph (b) and (e) in Weibull with those in Exponential and Gompertz distributions. Nevertheless, as the sample size used is 40 and above with medium and large effect sizes, the performance of both tests are almost the same, where both tests provide good empirical powers under different censoring percentage, significance level and type of distribution. Similar to the previous discussion in section A, when larger significance level and effect size are used, the performance of both tests will be better similarly great.

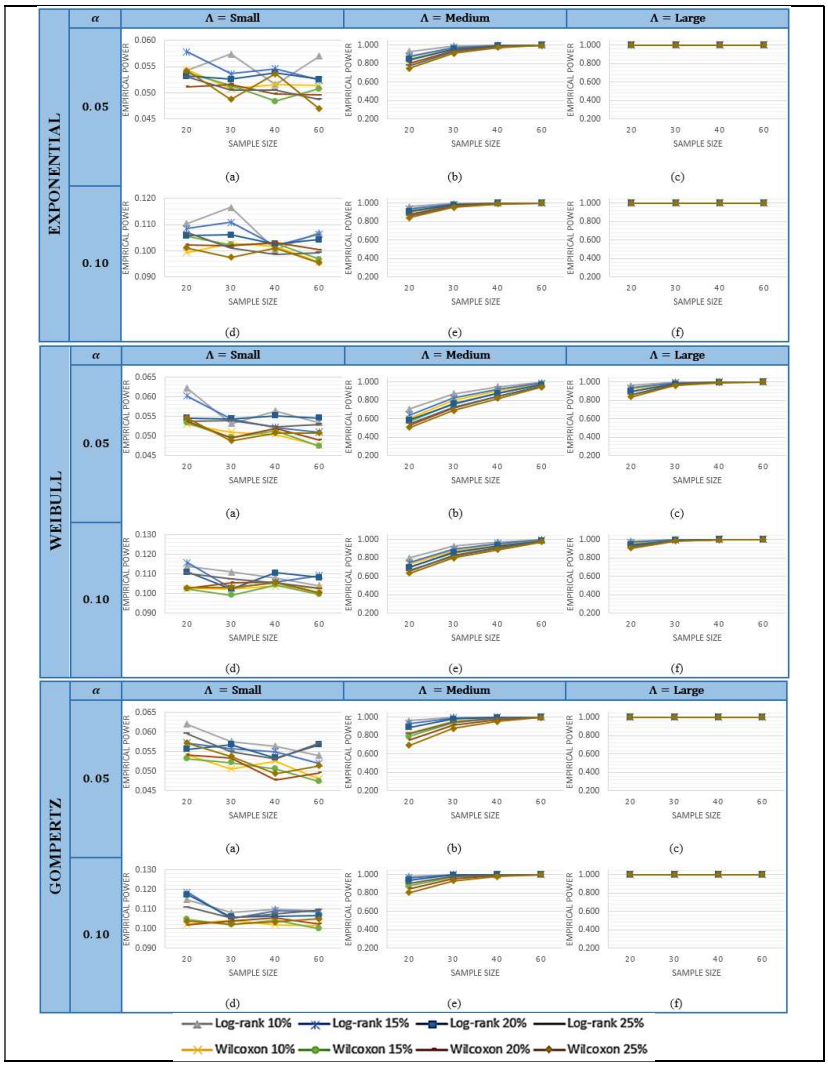

Figure 2: Empirical Power Graph of Log-rank & Wilcoxon test for 3 different distributions with right & interval censored data under different $\boldsymbol{n}$, $\boldsymbol{\alpha}$, $\boldsymbol{cp}$ and $\boldsymbol{\Lambda}$.

Analysis of Real Life Data

A real-life data analysis was conducted for illustration purposes of Log-rank and Wilcoxon tests usage. Datasets of an AIDS clinical trial designed to study the benefits of Zidovudine therapy in patients in the early stage of HIV infection (*illut3* obtained from *FHtest* package in R software) consists of 1607 observations with three different treatment groups known as Deferred Therapy (G1), 500mg/day Dosage (G2) and 1500mg/day Dosage (G3). In general, there are 541 patients in G1, 538 patients in G2 and 528 patients in G3. About 46.58% of observations in G1, 50.37% in G2 and 56.44% in G3 are censored where both right and interval censoring involve. A goodness-of fit test was conducted to determine the type of distribution that fits the AIDS survival data well. The accuracy criterion used in this project is Akaike Criterion Information (AIC). Results in Table 1 show that Gompertz distribution with the shape and scale parameters of $\lambda = 0.0377$ and $k = -0.0188$ is a good model to describe the data compared to the other two distributions as it provides the smallest AIC.

Table 1: AIC value for Exponential, Weibull and Gompertz distribution.

Distribution	Parameters	Estimate	SE	95% LCI	95% UCI	AIC
Exponential	λ	0.0267	0.0016	0.0238	0.0299	7608.446
Weibull	k	0.8759	0.0257	0.8268	0.9278	7588.901
	λ	38.2728	2.5812	33.5339	43.6814	
Gompertz	k	-0.0188	0.0023	-0.0234	-0.0143	7537.207
	λ	0.0377	0.0026	0.0329	0.0432	

Kaplan-Meier estimator is used to estimate and plot the survivor function for each treatment group as in Figure 3. Based on the plots, the probability of surviving for patients in G3 is the highest, followed by G1 and G2. Also, the black dashes lines which indicate the median survival time shows that G3 has the highest median survival time compared to G2 and G1 which means that patients who receive 1500mg dosage per day has better chances of surviving than patients who received 500mg/day Dosage and Deferred Therapy. However, the medians of G1 and G2 are quite far compared to G2 and G3. This means that those who received 500mg/day Dosage also have better chances of surviving compared to patients who just received Deferred Therapy. Further evidence as in Figure 4 shows the hazard rate of G1 is higher compared to G2 and G3, which means that patients who received Deferred Therapy has higher risk of death compared to those who received 500mg and 1500mg dosage per day.

Before performing the Log-rank and Wilcoxon tests, the PH assumption is assessed first by using *cox.zph* function from *survival* package to obtain both individual and global p-value. G1 is used as the reference category for this assessment. Figure 5 shows the graphs of the scaled Schoenfeld residuals against the survival time. When the proportional hazards assumption holds, the Schoenfeld residuals will be close to zero. The plots show that the smoothed point wise confidence bands are all around 0, which confirms that there is no obvious evidence against the PH assumption.

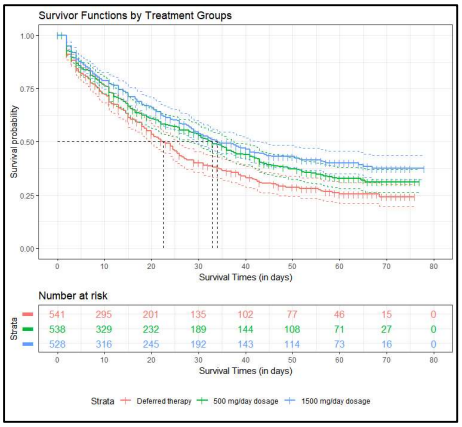

Figure 3: Survival plot and risk table for Deferred Therapy, 500mg/day Dosage and 1500mg/day Dosage.

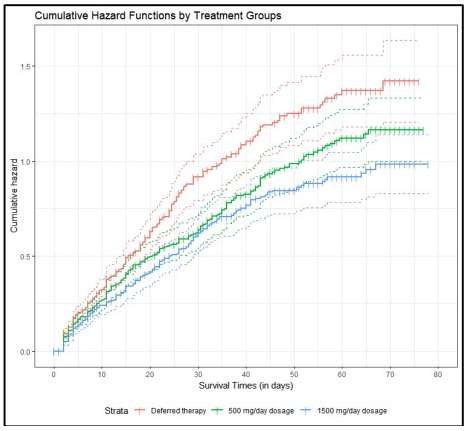

Figure 4: Cumulative hazard plot for Deferred Therapy, 500mg/day Dosage and 1500mg/day Dosage.

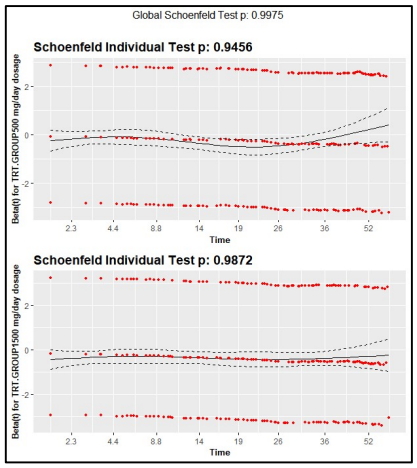

Figure 5. Schoenfeld residuals for treatment groups against survival time plot.

Log-rank and Wilcoxon tests are used to assess the significance difference between several survivor functions. Under null hypothesis, there is no difference in probability of surviving between all three treatment groups. Based on the results in Table 2, Log-rank provide smaller p-value compared to Wilcoxon because these survival data satisfy the PH assumption as mentioned previously. Log-rank test will perform better on detecting the significance different between survivor functions when PH assumption is satisfied compared to Wilcoxon. Hence, for this data, there is sufficient evidence to conclude that there is a significance difference in probability of surviving between patients who received Deferred Therapy, 500mg dosage/day and 1500mg dosage/day.

Table 2: Non-parametric tests for comparing three survivor functions.

Non-parametric test	χ_2^2	p-value
Log-rank	18.6	0.00009
Wilcoxon	17.3	0.00020

5. CONCLUSION & RECOMMENDATION

Overall, there is not much difference in the performance of Log-rank and Wilcoxon tests when dealing with survival data having right censored data alone, or right and interval censored data. This is because both censored data still contribute some information to the analysis although the exact failure time cannot be observed precisely. The only difference between these two data is in interval censored data, the two endpoint where the failure time lies is known, whereas for right censored, there is only the left endpoint exists.

Based on this study, Log-rank and Wilcoxon tests performs well regardless of type of distributions, censoring percentage, and significance level, as long as the effect size between the survivor functions is large and the sample size used is more than 40 samples for each group. The analysis of real data shows that Gompertz distribution is a good parametric model to describe the AIDS survival data. The real data also satisfied the PH assumption. Thus, it was expected that Log-rank would perform brighter compared to Wilcoxon test.

Future works may be employed by extending this power study in analyzing other non-parametric tests for survivor functions comparison. Other than that, this analysis can be extended by allowing for unequal sample size, and censoring percentage in each treatment group. Stratified tests are also suggested if there exist variables which may affect the survival times. Instead of comparing whether there is significance difference between the survivor functions, a trend test for more than two survivor functions may be performed too. This is because when dealing with an ordinal grouping variable, a trend test may be, actually, of greater relevance than the g-sample test.

ACKNOWLEDGMENTS

This work was supported by the Putra Grant, VOT 9595300, Universiti Putra Malaysia.

REFERENCES

[1] E. Gehan, "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored," *Biometrika,,* vol. 52, no. 1/2, pp. 203-223, 1965.

[2] R. Prentice, "Linear Rank Tests with Right Censored Data," *Biometrika,* vol. 65, pp. 167-179, 1978.

[3] N. Mantel, "Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration," *Cancer Chemotherapy Reports,* vol. 50, pp. 163-170, 1966.

[4] D. Cox, "Regression Models and Life-tables (with discussion)," *Journal of the Royal Statistical,* vol. 34, no. 2, pp. 187-220, 1972.

[5] R. Peto and J. Peto, "Asymptotically Efficient Rank Invariant Test Procedures," *Journal of the Royal Statistical Society. Series A (General),* vol. 135, no. 2, pp. 185-207, 1972.

[6] C. Le, Applied Survival Analysis, New York: John Wiley & Sons, 1997.

[7] E. Lee, M. Desu and E. Gehan, "A Monte Carlo Study of the Power of Some Two-Sample Tests," *Biometrika,* vol. 62, no. 2, p. 425–432, 1975.

[8] R. B. Latta, "A Monte Carlo Study of Some Two-Sample Rank Tests with Censored Data," *Journal of American Statistical Association,* vol. 76, no. 375, p. 713–719, 1981.

[9] M. Beltangady and R. Frankowski, "Effect of Unequal Censoring on the Size and Power of the LogRank and Wilcoxon Types of Tests for Survival Data," *Statistics in Medicine,* vol. 8, no. 8, pp. 937-945, 1989.

[10] E. León and P. Zuluaga, "Equivalence between Score and Weighted Tests for Survival Curves," *Communications in Statistics - Theory and Methods,* vol. 30, no. 4, p. 591–608, 2001.

[11] R. Martinez and J. Naranjo, "A Pretest for Choosing Between Logrank and Wilcoxon Tests in The Two-sample Problem," *International Journal of Statistics,* vol. 68, no. 2, pp. 111-125, 2010.

[12] A. Darilay and J. Naranjo, "A Pretest for Using Logrank or Wilcoxon in The Two-Sample Problem," *Computational Statistics & Data Analysis,* vol. 55, no. 7, pp. 2400-2409, 2011.

[13] M. Ramakrishman and R. Ravanan, "Non-parametric Methods for Comparing Two Survival Distributions," *Journal of Arts, Science & Commerce,* vol. IV, no. 2, pp. 121-125, April 2013.

[14] P. Karadeniz and I. Ercan, "Examining Tests for Comparing Survival Curves with Right Censored Data," *Statistics in Transition New Series,* vol. 8, no. 2, p. 311–328, June 2017.

[15] P. Karadeniz and I. Ercan, "On Comparing Survival Curves with Right-Censored Data According to the Events Occurs at the Beginning, in the Middle and at the End of Study Period," *International Journa of Statistics in Medical Research,* vol. 7, pp. 117-128, 2018.

[16] A. Zyoud, F. A. Elfaki and M. Hrairi, "Parametric Model Based On Imputations Techniques for Partly Interval Censored Data," *Journal of Physics: Conference Series,* p. 949, 2017.

[17] A. Azuero, "A Note on the Magnitude of Hazard Ratios," *Cancer,* vol. 122, no. 8, pp. 1298-1299, 2016.