On Effect Sizes for Nonparametric Comparison of Censored Survival Outcomes

Yongzhao Shao[*†] Zhaoyin Zhu[*]

Abstract

Survival outcomes are frequently randomly censored with unknown censoring distributions. Due to complexities caused by censoring, useful effect sizes for nonparametric comparison of censored survival outcomes have not been systematically investigated despite existence of several well known nonparametric tests such as the log-rank and Wilcoxon tests. Effect size generally emphasizes the magnitude of the difference between the studied survival endpoints rather than confounding this with sample size as in the case of p-value. This paper investigates weakness and advantages of existing and newly proposed effect sizes for the nonparametric comparison of time to event outcomes.

Key Words: censored survival outcome, effect size, non-parametric test, time-to-event endpoint sample size, clinical trials

1. Introduction

When comparing two treatments in terms of survival outcomes, it is of great interest to describe the results in terms of magnitude of difference in efficacies not just whether they are equally efficacious. It is well known that effect size is an important way of quantifying the magnitude of difference between two groups that has many advantages over the use of p-value alone or merely reporting a statistical significance (yes or no) alone (Cohen 1990; Sullivan and Feinn 2012). Survival outcomes are frequently randomly censored with unknown censoring distributions. Due to complexities caused by censoring, useful effect sizes for nonparametric comparison of censored survival outcomes have not been systematically investigated despite existence of several well known nonparametric tests such as the logrank and Wilcoxon tests. Effect size emphasizes the magnitude of the difference between the survival endpoints. While a small p-value is important as an indication of statistical significance or indication that the observed difference is not likely due to chance alone. However, p-value is confounded with sample size, thus statistical significance alone does not imply the difference is of clinical significance or practical significance no matter how small the p-values is. Thus, instead of reporting p-value alone, p-value should go together with effect size whenever possible. However, for censored survival outcomes, neither statistical textbooks nor computer packages routinely specify an effect size for the censored survival endpoints, e.g., when the Wilcoxon test is conducted.

Appropriate selection of effect size depends on the specific needs in practical applications. In particular, specifying an effect size is of great importance in deciding sample size and evaluating statistical power in clinical trial and study design. Technically speaking, effect size selection in a hypothesis-driven study has to account for the specific test to be used in the planned data analysis. Moreover, after data has been collected upon finishing a clinical trial or study, effect size is useful in summarizing and reporting the observed difference or improvement in efficacy in addition to the p-value. Significance from the p-value and confidence intervals in terms of effect size are naturally complement each other. Unfortunately, existing statistical textbooks and commonly used softwares do not offer clear

[*] Division of Biostatistics, NYU School of Medicine, New York, NY 10016, USA

[†] E-mail: shaoy01@nyu.edu

guidelines on how to select effect size for the common non-parametric tests on comparing survival probabilities. Discussion and guidelines on effect size selection are of practical importance for both the design and the analysis of clinical study/trial with censored survival endpoints. However, there is a lack of literature on systematic discussion of this important issue. This paper investigates weakness and advantages of existing effect sizes relative to newly proposed effect sizes for comparing time-to-event outcomes.

2. Methods

2.1 Set up. In survival studies, one of the meaningful criteria for a new medical treatment is whether it can significantly enlong the survival time of a patient compared with the existing treatment which maybe called standard care in general. The survival time can be any time-to-event such as overall survival or cancer progression-free survival. The nonparametric hypothesis might be formulated as $H_0 : P(X \geq Y) = 1/2$ where X and Y are the survival times of randomly selected patients from with the new treatment and the standard care, respectively. The alternative hypothesis is $H_a : P(X \geq Y) \neq 1/2)$ and desired alternative hypothesis is $P(X \geq Y)$ is 1 or much larger than 1/2. Wilcoxon test (or Gehan's test) or log-rank test might be used to test this non-parametric hypothesis. For brevity, in the rest of this paper we will subsequently consider the nonparametric test of $H_0 : P(X \geq Y) = 1/2$ and assume X and Y have continuous CDFs F and G, respectively, and $P(X = Y) = 0$.

2.2 Effect size. While it is of importance and interest to conduct formal nonparametric hypothesis tests on survival endpoints and obtain the corresponding p-values. As is well known, we also want to report some suitable effect size to quantify the strength or the magnitude of the difference between two comparison groups. Effect size is useful in the study design well as data analysis of clinical trials and other biomedical studies. Effect size estimation is as important as reporting a p-value. For example, even when p-values are not significantly small as in some early phase clinical trials, effect size estimates are fundamental in planning validation studies or other future studies (e.g. in power estimation). Unfortunately, existing statistical textbooks and commonly used softwares do not offer clear guidelines on how to select effect size for the common non-parametric tests on comparing survival endpoints. In particular, log-rank test is commonly suggested and widely used to compare two independent survival curves, and hazard ratio (HR) is a widely used measure of effect size following a significant p-value of the log rank test. An important limitation for using hazard ratio (HR) or $\log$(HR) as an effect size is that, in practice, the HR often changes with time (or time-dependent). HR or log(OR) is not generally a valid measure of effect size and is valid only when HR is independent of time, e.g. under Cox's proportional hazard models. Even under the proportional hazard assumption, the use of hazard ratio as an effect size can still be problematic in some practical applications [Heman 2010]. In fact, despite the recommendations in many texts and its wide use in practice, the log-rank test is not a consistent non-parametric test. In fact, frequently, for two different survival curves with $P(X \geq Y) \neq 1/2$, when F and G cross each other, the log-rank test may have no power at all no matter how large the sample size is. Similarly, there is no systematic guideline on suitable effect size for Wilcoxon test although specifying an effect size is of great importance in deciding sample size and statistical power in trial design where Wilcoxon test is planned test. Technically speaking, effect size selection in a hypothesis-driven study has to account for the operating characteristics of the planned test as part of the planned data analysis. Also, when we report a small p-value or statistical significance, e.g. based on the Wilcoxon test, we want to report an effect size and corresponding confidence interval and would like to have certain degree of consistency or concordance between the p-value and confidence interval. There still exist some knowledge gaps and lack of literature concerning these fundamental questions in nonparametric comparison of survival endpoints.

In practice, following Wilcoxon test (or Gehan's test) the difference in median survival time is often used as a measure of effect size following a significant p-value. Median survival time is conceptually easy to understand. However, as is well known, in many cases when $P(X \geq Y) \neq 1/2$, the Wilcoxon test has high power to detect the difference between groups, but the median survival time can be identical. Thus, the difference between median survival time is considerer as an useful effect size only in very limited situations.

Similar to median survival time, differences between mean survival time or restricted mean survival time (RMST) are also useful effect sizes but with non-negligible discordance with the Wilcoxon test. The mean survival time is conceptually easy to understand. First, as is well known, mean is an average. Second, the mean of a survival time X with CDF F is also the area under the corresponding survival curves, that is

$$\mu_X = \int_0^\infty [1 - F(t)]dt.$$

However, many survival curves do not have a well defined finite mean , i.e., $\mu_X = \infty$. The restricted mean survival time (RMST) is another effect size of interest [Chen and Tsiatis 2001, Andersen,et al. 2004, Tian et al 2013, Uno et al 2014]. For two numbers a and b, we denote $a \wedge b = \min(a, b)$. On the time interval $[0, \tau]$, the RMST of a survival time X is defined as

$$\mathrm{RMST}(\tau) = E(X \wedge \tau).$$

Even when $\mu_X = \infty$, $\mathrm{RMST}(\tau)$ is always finite for any $\tau < \infty$. The difference in mean or RMST reflects the difference in areas under survival curves. Such differences are meaningful when the two survival curves do not cross at any finite positive time. However, when two survival curves cross each other, the difference between mean survival time or RSMT are not necessarily informative as measures of effect size for the Wilcoxon test. For many situations where $P(X \geq Y) \neq 1/2$ and thus the Wilcoxon test has high power when sample size is large, the difference in mean or in RMST reflects the difference in areas under survival curves of F and G may be close to 0, and a confidence interval may contain zero when the Wilcoxon test has a very small p-value. Thus there often exist discordance between the p-value and confidence intervals for the difference in median survival time and difference in mean survival time or RSMT.

Note that Gehan's test statistic or the Mann-Whitney-Wilcoxon test statistic is generally a consistent estimate of $P(X > Y)$. Thus one may use $C = P(X > Y)$ as a measure of effect size. Indeed, Efron (1967) has considered the effect measure $C = P(X > Y)$ and provided a consistent estimator based on the KM estimators of F and G [Koziol and Jia 2009]. However, $P(X > Y)$ is not identifiable with heavy censoring in tail (thus, it cannot not always be consistently estimated). Also, $P(X > Y)$ is not capable of providing effect size measures at different time points or time-intervals.

2.3 New effect sizes. For two numbers a and b, we denote $a \wedge b = \min(a, b)$. We propose a class of effect size measures of the form:

$$C(a, \tau) = C_{a,\tau} = P(a \leq Y \wedge \tau \leq X \wedge \tau), \tag{1}$$

where a and τ are pre-selected non-negative constants. Naturally, τ can be the end of study or any pre-selected landmark time of particular interest (e.g., $\tau = 3, 5$ or 10 years). In particular, when the interval $[a, \tau) = [0, \infty)$, it is clear that $C(a, \tau) = P(X \geq Y)$ is identical to Efron's C-index. $C_{a,\tau} = 0$ or1means one treatment almost surely yields longer survival than the other treatment within the time interval $[a, \tau]$. $C_{a,\tau} = 1/2$ means the two treatment yields equivalent chance of survival. In general, we might be interested in some time intervals $[a, \tau) \neq [0, \infty)$, e.g. when the survival curves have one known crossing point at a. In this case, it might be more informative to consider $C(a, \tau) = C_{a,\tau} = P(a \leq Y \wedge \tau \leq X \wedge \tau)$, rather than consider the overall $C = P(X > Y)$ in Efron (1967), Newcombe (2006a, b), Koziol and Jia (2009). For brevity,]in this paper, we focus on discussion of the case $a = 0$. That is, we focus on the case $C_\tau = C_{0,\tau} = P(Y \wedge \tau \leq X \wedge \tau)$.

The C_τ and RMST are quite different, although they might complement each other. First of all, $C_\tau \in [0, 1]$ while RMST can be unbounded as $\tau \to \infty$. C_τ has a well defined limit as $\tau \to \infty$ which has been described by Efron (1967). The RMST might diverge as τ increases because not all survival times have well defined finite mean values. Secondly, the C_τ is the mean or limit of the Wilcoxon tests statistic, thus is a natural effect size measure. There is no natural collection between the behavior of Wilcoxon tests statistic and that of RMST. Another difference between C_τ and other effect sizes including RMST is in the following. Suppose X corresponding to a new treatment that improves the survival Y of an existing treatment by adoption of improved care e.g. better personalized patient management that would definitely improve clinical outcome [Reisberg et al 2017], although not sure by how much it will help. That is, both treatments have the same clinical intervention (i.e. same drug and/or same therapy) but the X group has additionally enhanced personalized care and patient management. Then we would expect X is at least as good as Y, that is, $P(X \geq Y) \approx 1$ thus Wilcoxon test has high power, and large effect size in C_τ or $K\tau$. However, for majority of the patients, the improvement in survival might be tiny and thus the difference in median survival time or RMST might be negligible.

When we are interested in comparing two treatments for patients at "high risk of dying' before some time τ, the following measure of effect size is of practical interest

$$K_\tau = P(X \geq Y | X \leq \tau, Y \leq \tau). \tag{2}$$

It is clear that C_τ is related to K_τ. It is easy to verify that

$$C_\tau = P(Y \leq X | X \leq \tau, Y \leq \tau) F(\tau) G(\tau) + P(X > \tau).$$

Thus,

$$C_\tau = K_\tau F(\tau) G(\tau) + [1 - F(\tau)]. \tag{3}$$

For nonparametric maximum likelihood estimation, we can estimate $F(\tau)$ and $G(\tau)$ using the Kaplan-Meier estimate. If we can constently estimate K_τ, then by equation (3), we also can obtain consistent estimate of C_τ, and vice versus. Indeed from equation (3), we have

$$K_\tau = [C_\tau + F(\tau) - 1]/[F(\tau) G(\tau)].$$

The new effect sizes C_τ or K_τ proposed here are in the forms of concordance probabilities that are known to be useful as measures of discriminative power in distingushing two groups [Harrell et al 1982; Gönen and Heller 2005, Koziol and Jia 2009, Uno et al 2011, Zhang and Shao 2017, Han et al 2017]. In other words, as overall predictive accuracy, based on $C_\tau = 1$ or 0, one can predict that one treatment almost surely yields longer survival than the other treatment.

2.4 Inference on effect sizes. We consider the problem of comparing two survival outcomes X and Y as in the setting of a randomized clinical trial. Say X is the survival time for the experimental treatment arm and Y denotes the survival time of the control arm. Let $X_1^*, \cdots, X_n^*$ and $Y_1^*, \cdots, Y_m^*$ denote the true independent survival time with cumulative distribution functions $F(s)$ and $G(s)$, respectively. Let $U_1, \cdots, U_n$ and $V_1, \cdots, V_m$ be censoring time for the two arms, respectively. Suppose $X_i = \min(X_i^*, U_i), i = 1, \cdots, n$ and $Y_j = \min(Y_j^*, V_j), j = 1, \cdots, m$ are the observed survival time in the two arms. We use $\delta_i = I(X_i^* < U_i), i = 1, \cdots, n$ and $\psi_j = I(Y_j^* < V_j), j = 1, \cdots, m$ to denote censoring indicators. Without much loss of generality, we assume tied observations are unlikely to happen ($P(X_i^* = X_j^*) = 0$), otherwise tied observations can be adjusted.

In common practice, when given censored data (X, Y, δ, ψ), it is customary to plot the two Kaplan-Meier survival curves for visdual comparison followed by nonparametric hypothesis tests for $H_0 : F = G$ and the corresponding p-values [Collett 2014, 3rd Edith]. The most widely used non-parametric tests for the comparison of two survival outcomes or two survival curves are the log-rank test and the Wilcoxon test [Collett 2014]. Both tests are straghtforward to conduct using commonly available statistical packages including SAS and R which routinely provide p-values of these tests. Then, the question of interset if how can we estimate the effect sizes C_τ or K_τ? How to conduct inference about C_τ or K_τ?

We are given censored data (X, Y, δ, ψ) to estimate $K_\tau = P(X^* > Y^* | X^* \leq \tau, Y^* \leq \tau)$ and $C_\tau = K_\tau F(\tau)G(\tau) + 1 - F(\tau)$. Note that $P(Y^* < X^* \leq \tau) = \int_0^\tau [G(s) - 1] dF(s)$, $P(X^* \leq \tau) = 1 - F(\tau)$, $P(Y^* \leq \tau) = 1 - G(\tau)$.

$$K_\tau = P(X^* > Y^* | X^* \leq \tau, Y^* \leq \tau) = \frac{P(Y^* < X^* \leq \tau)}{P(X^* \leq \tau) P(Y^* \leq \tau)}.$$

Given censored data (X, Y, δ, ψ), the nonparametric Kaplan-Meier method provides consistent estimator $\hat{F}$ and $\hat{G}$, thus also consistent estimator of K_τ:

$$\hat{K}_\tau = \frac{\int_0^\tau [\hat{G}(s) - 1] d\hat{F}(s)}{(1 - \hat{F}(\tau))(1 - \hat{G}(\tau))} \tag{4}$$

and

$$\hat{C}_\tau = \hat{K}_\tau \hat{F}(\tau) \hat{G}(\tau) + 1 - \hat{F}(\tau). \tag{5}$$

The consistency and asymptotic normality of $\hat{C}_\tau$ follow directly from the consistency and asymptotic normality of KM estimates. Moreover, based on the explicit formulas for KM estimates, we obtain explicit formulas for $\hat{K}_\tau$ and $\hat{C}_\tau$. Moreover, these explicit formulas enable efficient numerical computation of these estimates. The asymptotic normality and the easy computation allow us to obtain variance estimates and standard errors using bootstrap. We have developed an R package to compute $\hat{K}_\tau$ and $\hat{C}_\tau$ for any given τ. In a special case of interest, when $\tau = \infty$, our estimates in equation (4) and (5) and the R algorithm can compute Efron's estimate for $P(Y < X)$. In particular, based on our bootstrapped confidence interval, we can also conduct bootstrap test for $H_0 : P(X^* > Y^* | X^* \leq \tau, Y^* \leq \tau) = 0.5$, or $H_0 : K_\tau = 0.5$ as demonstrated in the next section.

3. Simulation Studies

For numerical illustration, we only report numerical study of K_τ because the results of C_τ are similar. We generated survival time from exponential distributions for the two groups with parameters (hazard rates) ϕ and $\rho\phi$ respectively. That is

$$X^* \sim \exp(\rho\phi), \qquad Y^* \sim \exp(\phi).$$

The censoring time was generated independently from a uniform distribution from 0 to 2, i.e. $U \sim U[0, 2]$, $V \sim U[0, 2]$. The true value of our proposed effect size $K_\tau = P(X^* > Y^* | X^* \leq \tau, Y^* \leq \tau)$ can be calculated via

$$K_\tau = \frac{1/(1 + \rho) - \exp(-\rho\phi\tau) + \rho \exp(-\phi(1 + \rho)\tau)/(1 + \rho)}{(1 - \exp(-\phi\rho\tau))(1 - \exp(-\phi\tau))}.$$

$X_i = \min(X_i^*, U_i), i = 1, \cdots, n$ and $Y_j = \min(Y_j^*, V_j), j = 1, \cdots, m$ are observed survival times. $\delta_i = I(X_i^* < U_i), i = 1, \cdots, n$ and $\psi_j = I(Y_j^* < V_j), j = 1, \cdots, m$ are censoring indicators. After sample the censored data (X, Y, δ, ψ), we then estimate $K_\tau = P(X^* > Y^* | X^* \leq \tau, Y^* \leq \tau)$. To evaluate the performance of our estimator under different scenarios, we set $\phi = 2$, $\rho = 1, 2$ and $\tau = 0.5, 1, 2$. For comparison, we set sample size for the two groups $n = m = 100, 200$ and the number of bootstrap replications to estimate variance equals 500. Findings based on 500 simulations in each instance are given in Table 1 and Table 2.

From Table 1, under different scenarios, our proposed estimators for the effect size measure K_τ are very close to the true value. The type I error or coverage probability for 95% confidence interval is getting closer to the nominal level as sample size increases. Table 2 demonstrates that when the true probability is away from 0.5, our proposed test has adequate power to detect the difference between two groups significantly.

Table 1: Type I error or coverage probability for 95% bootstrap confidence interval

Sample Size	(ρ, ϕ, τ)	True	Mean	SD	95% CP
m=n=100	(1,2,0.5)	0.5	0.496	0.055	0.958
m=n=200	(1,2,0.5)	0.5	0.497	0.039	0.944
m=n=100	(2,2,0.5)	0.423	0.421	0.050	0.956
m=n=200	(2,2,0.5)	0.423	0.422	0.028	0.946
m=n=100	(2,2,1)	0.373	0.369	0.044	0.946
m=n=200	(2,2,1)	0.373	0.374	0.031	0.948
m=n=100	(2,2,2)	0.339	0.337	0.042	0.942
m=n=200	(2,2,2)	0.339	0.338	0.029	0.952

Table 2: Power of bootstrap tests $H_0 : P(X^* > Y^* | X^* \leq \tau, Y^* \leq \tau) = 0.5$

Sample Size	(ρ, ϕ, τ)	Mean	Power (%)
m=n=100	(2,2,0.5)	0.421	40.2
m=n=200	(2,2,0.5)	0.422	65.0
m=n=100	(2,2,1)	0.369	84.8
m=n=200	(2,2,1)	0.374	99.4

4. Discussion

When comparing two treatments in terms of survival outcomes, it is of great interest to describe the results in terms of measures of magnitude difference in efficacies not just whether they are equally efficacious. It is well known that effect size is an important way of quantifying the magnitude of difference between two groups that has many advantages over the use of p-value alone or merely reporting of a statistical significance alone (Cohen 1990; Sullivan and Feinn 2012). Survival outcomes are frequently randomly censored with unknown censoring distributions. Due to complexities caused by censoring, useful effect sizes for nonparametric comparison of censored survival outcomes have not been systematically investigated despite existence of several well known nonparametric tests such as the logrank and Wilcoxon tests. Effect size emphasizes the magnitude of the difference between the survival endpoints rather than confounding this with sample size as in the case of a reported p-value. However, for censored survival outcomes, neither statistical textbooks nor computer packages routinely specify an effect size for the censored survival endpoints, e.g., when the Wilcoxon test is conducted. Discussion on effect size selection is of practical importance for study/trial designs in comparison of censored survival endpoints. Also, when the clinical trial or study is successfully conducted, effect size is useful in summarizing and reporting study findings. This paper investigates weakness and advantages of existing effect sizes for comparing studies with time to event outcomes and discuss new measures of effect size for Wilcoxon test. We developed bootstrap based confidence intervals for the effect sizes and bootstrap based hypothesis testing for the given effect sizes. Numerical study indicates that the bootstrap confidence intervals have desired coverage probability and the bootstrap hypothesis tests have desired type I errors. We also have developed an R package to implement the proposed effect sizes with confidence intervals.

5. Acknowledgements

This research was supported in part by NIH grants P30 AG0851, 2P30 CA16087 and 5P30 ES00260.

REFERENCES

Andersen, P. K., Hansen, M. G., and Klein, J. P. (2004), "Regression analysis of restricted mean survival time based on pseudo-observations", *Lifetime Data Analysis,* 10(4), 335-350.

Chen, P. Y., and Tsiatis, A. A. (2001), "Causal inference on the difference of the restricted mean lifetime between two groups," *Biometrics,* 57(4), 1030-1038.

Cohen, J. (1990), "Things I have learned (so far)", *American Psychologist,* 45(12), 1304.

Collett, D. (2014) "*Modelling Survival Data in Medical Research*", Third Edition. Chapman and Hall/CRC.

Efron, B. (1967), "The two-sample problem with censored data," *Proceedings of the Fifth Berkley Symposium on Mathematical Statistics and Probability*, 4, 831-853.

Gönen, Mithat and Heller, Glenn (2005), "Concordance probability and discriminatory power in proportional hazards regression", *Biometrika*, 92, 965-970.

Han, X, and Zhang, Y, and Shao, Y. (2017) "On comparing 2 correlated C indices with censored survival data." *Stat Med.* 2017 Jul 31. doi: 10.1002/sim.7414. [Epub ahead of print]

Harrell, Frank E and Califf, Robert M and Pryor, David B and Lee, Kerry L and Rosati, Robert A. (1982) "Evaluating the yield of medical tests", JAMA, 247, 2543–2546.

Heman, M. A. (2010) "The Hazards of Hazard Ratios", *Epidemiology*, 21, 13-15.

Koziol, J.A. and Jia, Z. (2009), "The concordance index C and the Mann-Whitney parameter $Pr(X > Y)$ with randomly censored data," *Biometrical Journal*, 51, 467-474.

Newcombe, R.G. (2006a), "Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1, general issues and tail-area-based methods," *Statistics in Medicine,* 25, 543-557.

Newcombe, R.G. (2006b), "Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: asymptotic methods and evaluation," *Statistics in Medicine,* 25, 559-573.

Reisberg, B. et al (2017), "Comprehensive, Individualized, Person-Centered Management of Community-Residing Persons with Moderate-to-Severe Alzheimer Disease: A Randomized Controlled Trial," *Dementia and Geriatric Cognitive Disorders*, 43, 100-17.

Sullivan, G. M. and Feinn, R. (2012), "Using effect sizeor why the P value is not enough", *Journal of Graduate Medical Education,* 4(3), 279-282.

Tian, L., Zhao, L., and Wei, L. J. (2013), "Predicting the restricted mean event time with the subject's baseline covariates in survival analysis," *Biostatistics,* 15(2), 222-233.

Uno, Hajime and Cai, Tianxi and Pencina, Michael J and D'Agostino, Ralph B and Wei, LJ (2011), "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data", *Statistics in medicine*, 30, 1105–1117.

Uno, H., Claggett, B., Tian, L. et al (2014), "Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis," *Journal of clinical Oncology,* 32(22), 2380-2385.

Zhang, Y and Shao Y. (2017), "Concordance measure and discriminatory accuracy in transformation cure models." *Biostatistics*. 2017 May 5. doi: 10.1093/biostatistics/kxx016. [Epub ahead of print]