Using Weighted Differences in Hazards as Effect Sizes for Survival Data

H. Wang¹ · D. Chen² · Q. Pan³ · M. T. Hueman⁴

Accepted: 24 December 2021 / Published online: 1 February 2022

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2022

Abstract

Sensitive to the change in sample sizes, traditional measures such as values of test statistics or p values can fail to quantify the difference in survival between populations for time-to-event data. We thereby propose to use effect sizes defined as the weighted differences in hazards to evaluate the survival difference between two groups. On the basis of the logrank test statistic, Gehan–Wilcoxon test statistic and Prentice–Wilcoxon test statistic, we developed three effect sizes that compare the survival experiences over the time period of investigation. Estimates of these three effect sizes were provided and their large sample behaviors were studied. In light of the Mann–Whitney parameter, we presented an effect size that compares the survival experiences over the entire possible/hypothetical time period. Two estimates of this effect size were constructed. We compared the proposed effect sizes and illuminated their use by theoretical studies, simulations and real cancer data. The effect sizes proposed in this article can help understand the survival difference in populations and are expected to have promising applications in the field of survival analysis.

Disclaimer The contents, views or opinions expressed in this publication or presentation are those of the authors and do not necessarily reflect official policy or position of Uniformed Services University of the Health Sciences, the Department of Defense (DoD), or Departments of the Army, Navy, or Air Force. Mention of trade names, commercial products, or organizations does not imply endorsement by the US Government.

This work was partially supported by John P. Murtha Cancer Center Research Program (64349-MCC Comprehensive Research).

✉ D. Chen
dechang.chen@usuhs.edu

¹ Department of Biostatistics, The George Washington University, Washington, DC, USA

² Department of Preventive Medicine and Biostatistics, F. Edward Hébert School of Medicine, Uniformed Services University of the Health Sciences, Bethesda, MD, USA

³ Department of Statistics, The George Washington University, Washington, DC, USA

⁴ Department of Surgical Oncology, John P. Murtha Cancer Center, Walter Reed National Military Medical Center, Bethesda, MD, USA

Keywords Effect size · Survival analysis · Test statistic · Mann–Whitney parameter

1 Introduction

To report the difference (with respect to some characteristics, e.g., weight, blood pressure value) between two groups (populations), there is nothing more common than to conduct a hypothesis test and check the “statistical significance.” However, along with a series of criticism [2,44], the conventional, dichotomous way of using the p value (e.g., ≤ 0.05 and > 0.05) has faced an unprecedented debate for its abuse and misuse. To that end, there is a need to find surrogate measures that can genuinely present findings from different studies in the hope of drawing better inferences.

It seems that the p value, even without being dichotomized, could not serve as such a measure of the difference between groups. This is mainly because the p value is influenced by the change in sample sizes and there is no simple relationship between the sample size and the p value that can be used to adjust the result. What measures would better describe the difference between groups?

A recommended alternative would be the effect size [2,44]. An effect size is a quantitative measure of the magnitude of a phenomenon and is unaffected by the change in sample sizes. When analyzing or mining data, effect sizes allow researchers to ensure that statistically significant effects inferred from a sample are substantial enough to be considered practically meaningful [27]. Certain effect sizes have been proposed for data types other than time-to-event, e.g., correlation coefficient, relative risk, Cohen’s d [11], etc. For time-to-event data with censoring involved, there does not seem to have much research focusing on effect sizes that can be used to assess the difference in survival between groups.

An obvious effect size for time-to-event data is the hazard ratio from the Cox model [12]. However, it has long been known that the hazard ratio is a valid measure only under the proportional hazards (PH) assumption. Applying PH analysis to studies where PH does not hold can lead to misleading conclusions [37]. Even worse is that the PH assumption is rarely checked formally. The commonly used tests for examining the PH assumption, such as Grambsch–Therneau test, may not capture non-PH in studies with small sample sizes, but always lead to a rejection when sample sizes are large [42]. Alternatively, the PH assumption can be examined by using residual plots (e.g., plotting martingale or Schoenfeld residuals against time), by checking the predictor–time interaction effect, or by visually looking at the survival curves. Note that these methods usually provide a subjective assessment. Therefore, it is important to find an informative estimate of the difference between two survival curves regardless of the PH assumption. In the current literature, there are two types of estimates toward this goal: the average hazard ratio (AHR) proposed by [38] and the restricted mean survival time (RMST) proposed by [37]. However, the interpretation of AHR can be difficult and misleading due to its complex definition [37]. And because it is difficult to select an appropriate period of time to compute RMST, the use of RMST seems to have been limited.

In this paper, we propose to use the weighted differences in hazards as effect sizes for studying the survival difference between two groups. The use of these proposed

effect sizes do not impose any particular assumptions about the distributions of the failure time. They can be applied with or without the PH assumption. We will show that the commonly used logrank test statistic [30], Gehan–Wilcoxon test statistic [18], Prentice–Wilcoxon test statistic [35], as well as the well known Mann–Whitney parameter [29] can be exploited to develop such effect sizes.

This paper is organized in the following way. Section 2 introduces three test statistics-based effect sizes, estimates of the effect sizes and their large sample results. Section 3 introduces the Mann–Whitney parameter-based effect size and two estimates of the effect size. Section 4 presents a comparison among the proposed effect sizes. Section 5 partitions values of effect sizes to facilitate applications. Section 6 illustrates various aspects, including applications, of the proposed effect sizes by simulations and real cancer data. We conclude in Sect. 7.

2 Effect Sizes Based on Test Statistics

The logrank test is one of the most popular methods for testing the null hypothesis H_0 that there is no difference in the survival between two groups. The test is asymptotically fully efficient for testing equality of survival distributions in a proportional hazard family when no censoring is present or when the censoring distributions are the same for compared groups [28]. The original logrank test assigns an equal weight to each failure time. Gehan [18], Peto and Peto [34], Tharone and Ware [41], Prentice [35], Andersen [3] and Fleming and Harrington [17] modified the logrank test by assigning a different weight at each failure time so that the tests can be more sensitive to the difference in survival distributions in a particular time interval. Assigning varying weights over time to the logrank test produces weighted logrank tests that have been proven to be more efficient when hazards are non-proportional [16].

Suppose that two samples of survival data containing a total of n subjects are available, with sample i coming from group i . Let $t_1, \dots, t_J$ be the distinct failure times in increasing order from the pooled sample. Let D_{ij} and Y_{ij} be, respectively, the number of subjects who failed and were at risk at t_j in sample i ($i = 1, 2; j = 1, 2, \dots, J$). Also let D_j and Y_j be, respectively, the total number of subjects who failed and were at risk at t_j in both samples. Then, the general form of the weighted logrank test statistic, denoted by L_w , is

$$L_w = \frac{\sum_{j=1}^J w_j \left(D_{1j} - D_j \frac{Y_{1j}}{Y_j} \right)}{\left(\sum_{j=1}^J w_j^2 D_j \frac{Y_{1j}}{Y_j} \frac{Y_{2j}}{Y_j} \frac{Y_j - D_j}{Y_j - 1} \right)^{\frac{1}{2}}}, \quad (1)$$

where the weight $w_j = w_n(t_j)$ for a positive weight function $w_n(\cdot)$. L_w includes as special cases many well-known test statistics, such as the logrank test statistic (weights $w_j = 1$), Gehan–Wilcoxon test statistic (weights $w_j = Y_j/n$) and Prentice–Wilcoxon test statistic (weights $w_j = \hat{S}(t_j-) = \prod_{k:k < j} (1 - D_k/Y_k)$) [16].

It is known that L_w is close to zero when two survival functions are almost identical to each other and has a large absolute value when two survival functions are far apart.

Therefore, L_w could serve as a measure of the difference in survival between two groups. However, the absolute value of L_w tends to increase (without bound) as the sample size increases. Consequently, when the data size increases, the p value will decrease to 0, indicating that with a large data size the p value can be non-informative in quantifying differences. Therefore, none of L_w , $|L_w|$, and the p value, computed using big samples, can provide an appropriate measure of the difference in survival. It is then of interest to find a statistic that can measure the survival difference and at the same time is “unaffected” by changes in sample sizes. Such a statistic is an estimate of an effect size and can be derived in light of the numerator of L_w , as shown in this paper. The denominator, an estimate of the standard deviation of the numerator under the null hypothesis H_0 , will not be utilized in constructing estimates of effect sizes.

Along with our development, we will use several notations associated with the groups and samples. Let T denote the random variable of the actual failure time, C the random variable of the censoring time, X the minimum of T and C and Z the indicator of the group number with $Z = i$ indicating group i ($i = 1, 2$). Suppose that in the study, the assignment of a subject to group i has a probability P_i . For group i , let $\lambda_i(t)$, $f_i(t)$, $S_i(t)$ and $S_i^*(t)$ denote, respectively, the hazard function, density function of the failure time, survival function of the failure time and censoring survival function (i.e., the survival function of the censoring time). We assume that in each group the censoring distribution and the survival distribution are continuous and independent of each other. The above quantities can be combined into $\pi(t) = P(X \geq t) = \pi_1(t)P_1 + \pi_2(t)P_2$ with $\pi_i(t) = P(X \geq t | Z = i) = S_i(t)S_i^*(t)$. The quantity $\pi_i(t)$ is the probability of being at risk (at time t) for subjects in group i . The following notations are needed for survival data. For sample i , let n_i denote the total number of subjects under study, and $Q_i(t)$, $N_i(t)$ and $Y_i(t)$ represent, respectively, the number of subjects with observed or unobserved failure times larger than or equal to t , number of observed failures after time 0 up through and including time t and number of subjects at risk at time t . Note that $Q_i(t)$, $N_i(t)$ and $Y_i(t)$ depend on the sample size. To facilitate the calculation, we will use the convention $0/0 = 0$.

Let U denote the numerator of L_w in Eq. (1). We have

$$\frac{n}{n_1 n_2} U = \frac{n}{n_1 n_2} \sum_{j=1}^J w_j \left(\frac{D_{1j} Y_{2j}}{Y_j} - \frac{D_{2j} Y_{1j}}{Y_j} \right) = \sum_{j=1}^J w_j \frac{\frac{Y_{1j}}{n_1} \frac{Y_{2j}}{n_2}}{\frac{Y_j}{n}} \left(\frac{D_{1j}}{Y_{1j}} - \frac{D_{2j}}{Y_{2j}} \right).$$

Clearly, $\frac{Y_{1j}}{n_1}$, $\frac{Y_{2j}}{n_2}$, $\frac{Y_j}{n}$, $\frac{D_{1j}}{Y_{1j}}$, $\frac{D_{2j}}{Y_{2j}}$ are the estimates, respectively, of $\pi_1(t_j)$, $\pi_2(t_j)$, $\pi(t_j)$, $\lambda_1(t_j)\Delta t$, $\lambda_2(t_j)\Delta t$, where Δt is the length of an infinitesimal interval of the time. Below we will show that the limit of $\frac{n}{n_1 n_2} U$ ($n \rightarrow \infty$) is a weighted difference (via integration) between two hazard functions of the two groups. This indicates $\frac{n}{n_1 n_2} U$ could be used to estimate the effect size that measures the difference in survival experiences between group 1 and group 2.

Theorem 1 Suppose that $w_n(t)$ is a predictable process such that $0 \leq |w_n(t)| \leq 1$ for all $t \in [0, \infty]$. Also suppose that for almost all t , $w_n(t)$ converges in probability to a deterministic function $\omega(t)$. Then

$$\frac{n}{n_1 n_2} U \xrightarrow{P} \int_0^\infty \frac{\pi_1(t) \pi_2(t)}{\pi(t)} \omega(t) (\lambda_1(t) - \lambda_2(t)) dt,$$

provided that the integrand is absolutely integrable.

Proof See “Appendix 1”.

The expression of $\int_0^\infty \frac{\pi_1(t) \pi_2(t)}{\pi(t)} \omega(t) (\lambda_1(t) - \lambda_2(t)) dt$ in Theorem 1 is a weighted difference in hazards with weight at time t equal to $\frac{\pi_1(t) \pi_2(t)}{\pi(t)} \omega(t)$ and does not depend on the sample size. Therefore, it can serve as an effect size measuring the survival difference between two groups. In general, this effect size is unknown. Theorem 1 shows that an estimate of the effect size, i.e., $\frac{n}{n_1 n_2} U$, can be computed using the data at hand.

In many cases, it is straightforward to verify the conditions in Theorem 1. Let $\widehat{ES}_L$, $\widehat{ES}_G$, and $\widehat{ES}_P$ denote $\frac{n}{n_1 n_2} U$ corresponding to the weight selection for the logrank, Gehan–Wilcoxon and Prentice–Wilcoxon test statistic, respectively. That is,

$$\widehat{ES}_L \equiv \frac{n}{n_1 n_2} \sum_{j=1}^J \left(\frac{D_{1j} Y_{2j}}{Y_j} - \frac{D_{2j} Y_{1j}}{Y_j} \right), \quad (2)$$

$$\widehat{ES}_G \equiv \frac{n}{n_1 n_2} \sum_{j=1}^J \frac{Y_j}{n} \left(\frac{D_{1j} Y_{2j}}{Y_j} - \frac{D_{2j} Y_{1j}}{Y_j} \right), \quad (3)$$

$$\widehat{ES}_P \equiv \frac{n}{n_1 n_2} \sum_{j=1}^J \prod_{k:k < j} \left(1 - \frac{D_k}{Y_k} \right) \left(\frac{D_{1j} Y_{2j}}{Y_j} - \frac{D_{2j} Y_{1j}}{Y_j} \right). \quad (4)$$

Then it follows from Theorem 1 that the following results hold true.

Corollary 1 (a) If $E_i[\lambda_{3-i}(t)/\lambda_i(t)] < \infty$ ($i = 1, 2$), where the expectation is taken with respect to the failure time in group i , then $\widehat{ES}_L \xrightarrow{P} \int_0^\infty \frac{\pi_1(t) \pi_2(t)}{\pi(t)} (\lambda_1(t) - \lambda_2(t)) dt$, where the integrand is absolutely integrable.

(b) $\widehat{ES}_G \xrightarrow{P} \int_0^\infty \pi_1(t) \pi_2(t) (\lambda_1(t) - \lambda_2(t)) dt$, where the integrand is absolutely integrable.

(c) If $S_1^*(t) = S_2^*(t) = S^*(t)$, then $\widehat{ES}_P \xrightarrow{P} \int_0^\infty S^*(t) S_1(t) S_2(t) (\lambda_1(t) - \lambda_2(t)) dt = \int_0^\infty \frac{\pi_1(t) \pi_2(t)}{S^*(t)} (\lambda_1(t) - \lambda_2(t)) dt$, where the integrand is absolutely integrable.

Proof See “Appendix 2”.

The Corollary indicates that $\widehat{ES}_L$, $\widehat{ES}_G$, and $\widehat{ES}_P$ estimates, respectively, the following three effect sizes:

$$ES_L \stackrel{\text{def}}{=} \int_0^\infty \frac{\pi_1(t) \pi_2(t)}{\pi(t)} (\lambda_1(t) - \lambda_2(t)) dt, \quad (5)$$

$$ES_G \stackrel{\text{def}}{=} \int_0^\infty \pi_1(t) \pi_2(t) (\lambda_1(t) - \lambda_2(t)) dt, \quad (6)$$

and

$$ES_P \stackrel{\text{def}}{=} \int_0^\infty \frac{\pi_1(t)\pi_2(t)}{S^*(t)} (\lambda_1(t) - \lambda_2(t)) dt. \quad (7)$$

Clearly, all three integrals depend on the censoring distribution. The proof of the corollary shows that both ES_G and ES_P belong to the interval $[-2, 2]$. Actually, Sect. 4.2 will show ES_G is a difference between two probabilities and thus lies inside the interval $[-1, 1]$. In comparison, there are no constants a and b such that any ES_L lies inside the interval $[a, b]$. Note that ES_L , ES_G and ES_P from the formulas can be negative and their magnitudes (i.e., the absolute values) are often the main interest in practical use. It follows from their definition that a positive (negative) effect size means a positive (negative) weighted difference in hazards between the two groups, which implies a “higher” (“lower”) hazard in group 1 than in group 2. This information can be useful when comparing two groups. For instance, if group 1 represents the treatment and group 2 represents the control, then a positive ES_L implies an adverse effect from the treatment.

3 Effect Sizes Based on the Mann–Whitney Parameter

The Mann–Whitney parameter $P(T_1 > T_2)$ assesses the stochastic ordering between the distributions of two variables T_1 and T_2 . In this paper, T_i is the variable of actual failure time in group i for $i = 1, 2$. If $P(T_1 > T_2) = 0.5$, then T_1 and T_2 are “statistically indifferent” to each other; if $P(T_1 > T_2) > 0.5$, then T_1 is “statistically preferred” to T_2 [33]. Therefore, $P(T_1 > T_2)$ can serve as a measure of the difference between two distributions of T_1 and T_2 . Based on $P(T_1 > T_2)$, we define

$$ES_{MW} = 1 - 2P(T_1 > T_2) \quad (8)$$

as an effect size of the difference between two distributions. Note that ES_{MW} ranges from -1 to 1 , with its minimum -1 and maximum 1 corresponding to $P(T_1 > T_2) = 1$ and 0 , respectively. As in the test statistic-based effect sizes, often the magnitude of ES_{MW} is of interest for comparisons.

In order to estimate ES_{MW} , it suffices to estimate $P(T_1 > T_2)$. When data are free of censoring, $P(T_1 > T_2)$ can be estimated by the Mann–Whitney test statistic [29]. If censoring exists, $P(T_1 > T_2)$ cannot be estimated by the Mann–Whitney test statistic since some actual failure times are unknown. In this case, an alternative way to estimate $P(T_1 > T_2)$ is needed. Note that

$$P(T_1 > T_2) = \int_{t=0}^\infty \int_{s=t}^\infty f_1(s)f_2(t)dsdt = - \int_0^\infty S_1(t)dS_2(t), \quad (9)$$

where S_i and f_i are, respectively, the survival function and density function for T_i ($i = 1, 2$). Let $\hat{S}_1$ and $\hat{S}_2$ be the Kaplan–Meier estimates of S_1 and S_2 , respectively. Efron proposed to use $\hat{D} = - \int_0^\infty \hat{S}_1(t)d\hat{S}_2(t)$ to estimate the right side of (9), assuming

that the largest observation of each group is uncensored [15]. However, in practice it is very common that this assumption is not met. Without this assumption, using $\widehat{D}$ to estimate $P(T_1 > T_2)$ may cause trouble since the Kaplan–Meier estimates could be undefined for the potentially important tail of a survival function. Below we present two remedies, which overcome the drawback of $\widehat{D}$, to estimate the Mann–Whitney parameter. One approach is based on completing the estimated survival functions using exponential tails, and the other on the conditional estimation.

3.1 Estimating ES_{MW} by Adding Exponential Tails

There are several methods to complete estimated survival functions by adding tails. Let $\widehat{S}(t)$ be the Kaplan–Meier estimate of the survival function $S(t)$. Efron suggested estimating $\widehat{S}(t)$ by 0 beyond the largest observed study time $t_{\max}$ [15]. Gill recommended estimating $\widehat{S}(t)$ by $\widehat{S}(t_{\max})$ beyond $t_{\max}$ [20]. Moeschberger and Klein proposed estimating $\widehat{S}(t)$ by a restricted Weibull survival curve that goes through two points $(0, 1)$ and $(t_{\max}, \widehat{S}(t_{\max}))$ beyond $t_{\max}$ [32]. Brown, Hollander and Kowar suggested estimating $\widehat{S}(t)$ by an exponential curve going through $(0, 1)$ and $(t_{\max}, \widehat{S}(t_{\max}))$ beyond $t_{\max}$ [7]. We note that both methods from Efron and Gill are sensitive to heavy or administrative censoring. Adding Weibull tails could lead to crossing survival functions for some populations where survival functions are not supposed to cross each other. Adding exponential tails does not suffer from these drawbacks and hence is a preferred choice. We start this discussion by replacing the tail of a survival function with an exponential function.

Let τ_i be a time point such that for $t > \tau_i$, $S_i(t)$ is replaced by an exponential function determined by the exponential curve going through two points $(0, 1)$ and $(\tau_i, S_i(\tau_i))$ ($i = 1, 2$). That is, $S_i(t) = e^{-\lambda_i t}$ for $t > \tau_i$ with $\lambda_i = -\log(S_i(\tau_i))/\tau_i$. Then the Mann–Whitney parameter is approximated by $D_e(\tau_1, \tau_2) = A + B + C$ with $A = -\int_0^{\min(\tau_1, \tau_2)} S_1(t) dS_2(t)$; $B = -\int_{\tau_1}^{\tau_2} e^{-\lambda_1 t} dS_2(t)$ if $\tau_1 < \tau_2$, 0 if $\tau_1 = \tau_2$, $\int_{\tau_2}^{\tau_1} S_1(t) \lambda_2 e^{-\lambda_2 t} dt$ if $\tau_1 > \tau_2$; and $C = \frac{\lambda_2}{\lambda_1 + \lambda_2} e^{-(\lambda_1 + \lambda_2) \max(\tau_1, \tau_2)}$. Consequently, ES_{MW} can be approximated by

$$ES_{MWE}(\tau_1, \tau_2) = 1 - 2D_e(\tau_1, \tau_2). \quad (10)$$

It is worth noting that an exponential tail leads to a constant hazard beyond τ_i and two exponential tails do not intersect unless they coincide exactly.

When survival data are available, we can use the Kaplan–Meier estimates $S_i(\cdot)$ and Lebesgue–Stieltjes integration in the calculation of A , B , and C to obtain $\widehat{A}$, $\widehat{B}$, and $\widehat{C}$. Set $\widehat{D}_e(\tau_1, \tau_2) = \widehat{A} + \widehat{B} + \widehat{C}$. Then we can estimate ES_{MW} by using $1 - 2\widehat{D}_e(\tau_1, \tau_2)$. In reality, tied failure times occur due to rounding or common data collection practice/logistics (e.g., failure times are measured in months). As a result, it follows from the law of total probability that $\widehat{D}_e(\tau_1, \tau_2)$ and $\widehat{D}_e(\tau_2, \tau_1)$ based on a real dataset usually slightly underestimate $P(T_1 > T_2)$ and $P(T_2 > T_1)$, respectively. To make the adjustment, we add $\frac{1}{2} (1 - \widehat{D}_e(\tau_2, \tau_1) - \widehat{D}_e(\tau_1, \tau_2))$ to the term $\widehat{D}_e(\tau_1, \tau_2)$ in $1 - 2\widehat{D}_e(\tau_1, \tau_2)$ to obtain an estimate of ES_{MW} , i.e.,

$$\widehat{ES}_{MWE}(\tau_1, \tau_2) = \widehat{D}_e(\tau_2, \tau_1) - \widehat{D}_e(\tau_1, \tau_2). \quad (11)$$

Equation (11) should be used when tied failure times are possible. To be safe, it is recommended that Eq. (11) be used always in practice.

3.2 Estimating ES_{MW} Through Conditioning

Let τ be a preset time point before which we have the knowledge about S_1 and S_2 . Then we may approximate the Mann–Whitney parameter $P(T_1 > T_2)$ by the conditional probability $P(T_1 > T_2 | T_1 \leq \tau \text{ or } T_2 \leq \tau)$, denoted by $D_c(\tau)$, which can be shown to be equal to $\frac{-\int_{t=0}^{\tau} \widehat{S}_1(t) d\widehat{S}_2(t)}{1 - \widehat{S}_1(\tau)\widehat{S}_2(\tau)}$. Consequently, ES_{MW} can be approximated by

$$ES_{MWC}(\tau) = 1 - 2D_c(\tau). \quad (12)$$

When survival data are available, we can estimate $D_c(\tau)$ by $\widehat{D}_c(\tau) = \frac{-\int_{t=0}^{\tau} \widehat{S}_1(t) d\widehat{S}_2(t)}{1 - \widehat{S}_1(\tau)\widehat{S}_2(\tau)}$ using the Kaplan–Meier estimates $\widehat{S}_1(\cdot)$ and $\widehat{S}_2(\cdot)$. Similar to (11), we can obtain another estimate of ES_{MW} with adjustment for ties:

$$\widehat{ES}_{MWC}(\tau) = \frac{-\int_{t=0}^{\tau} \widehat{S}_2(t) d\widehat{S}_1(t)}{1 - \widehat{S}_2(\tau)\widehat{S}_1(\tau)} - \frac{-\int_{t=0}^{\tau} \widehat{S}_1(t) d\widehat{S}_2(t)}{1 - \widehat{S}_1(\tau)\widehat{S}_2(\tau)}. \quad (13)$$

3.3 Selection of τ_1 , τ_2 and τ

We can determine τ_1 , τ_2 for $ES_{MWE}(\tau_1, \tau_2)$ and τ for $ES_{MWC}(\tau)$ by using the data. Let $t_{\max_i}$ be the largest observed time (either censoring or failure time) from group i ($i = 1, 2$). We can set $\tau_1 = \tau_2 = \tau = \tau^*$, where τ^* is less than or equal to $\min(t_{\max_1}, t_{\max_2})$. The time point τ^* should be selected to avoid the unstable end (if any) of the Kaplan–Meier curves and, at the same time, as close as possible to $\min(t_{\max_1}, t_{\max_2})$ to make as much use as possible of the information from the Kaplan–Meier estimates. Such a setting of τ_1 , τ_2 , and τ will allow $\widehat{ES}_{MWE}(\tau_1, \tau_2)$ and $\widehat{ES}_{MWC}(\tau)$ to produce reasonably good estimates of ES_{MW} if the assumption of proportional hazards holds true. This is seen in the following theorem:

Theorem 2 Let $\lambda_1(t)$ and $\lambda_2(t)$ be the hazard functions for T_1 and T_2 , respectively. If there exists a constant r such that $r = \frac{\lambda_1(t)}{\lambda_2(t)}$ for any $t > 0$, then for any $\tau^* > 0$, we have

$$ES_{MW} = ES_{MWE}(\tau_1 = \tau^*, \tau_2 = \tau^*) = ES_{MWC}(\tau = \tau^*) = \frac{r - 1}{r + 1},$$

where ES_{MW} , $ES_{MWE}(\tau_1, \tau_2)$ and $ES_{MWC}(\tau)$ are defined by Eqs. (8), (10) and (12), respectively.

Proof See “Appendix 3”.

Note that in many situations, survival functions can be modeled by distributions with proportional hazards. For example, if $T_1 \sim \text{Weibull}(\alpha, \beta_1)$ and $T_2 \sim \text{Weibull}(\alpha, \beta_2)$,

Fig. 1 The survival functions of the censoring time (gray) and the failure time in group 1 (red) and group 2 (blue)

then it is easy to show the ratio of two hazard functions is a constant equal to $r = \frac{\beta_2^\alpha}{\beta_1^\alpha}$. (Recall that the survival function of a Weibull(α, β) distribution has the expression $S(t) = e^{-(t/\beta)^\alpha}$.) By Theorem 2, we have $ES_{MW} = ES_{MWE}(\tau_1 = \tau^*, \tau_2 = \tau^*) = ES_{MWC}(\tau = \tau^*) = \frac{\beta_2^\alpha - \beta_1^\alpha}{\beta_1^\alpha + \beta_2^\alpha}$.

Also note that the conclusion of the theorem may not hold if the assumption of proportional hazards is violated, as shown below. Let $T_1 \sim \text{Weibull}(2.5, 0.5)$ and $T_2 \sim \text{Weibull}(5, 0.5)$ denote variables of failure times in group 1 and group 2, respectively. The ratio of the two hazard functions ($= 0.5^{3.5}t^{-2.5}$) is not a constant. Assume that in each group, the only censoring is an administrative censoring uniformly distributed between 0 and $\tilde{t} = 0.3$. The survival functions of the failure time and censoring time distributions are shown in Fig. 1. Straightforward computation shows $ES_{MW} = 0.091$, $ES_{MWE}(\tau_1 = 0.3, \tau_2 = 0.3) = 0.570$, and $ES_{MWC}(\tau = 0.3) = 0.584$. Therefore, both ES_{MWE} and ES_{MWC} are far away from ES_{MW} .

In light of Theorem 2, $\widehat{ES}_{MWE}(\tau_1 = \tau^*, \tau_2 = \tau^*)$ and $\widehat{ES}_{MWC}(\tau = \tau^*)$ can estimate ES_{MW} very well when the proportional hazards assumption holds true. It turns out they can also estimate ES_{MW} well under the condition that $\widehat{S}_1(\tau^*)$ and $\widehat{S}_2(\tau^*)$ are close to 0. This is seen in the following theorem.

Theorem 3 If $S_1(\tau^*) \leq \epsilon_1$ and $S_2(\tau^*) \leq \epsilon_2$, then $|ES_{MWE}(\tau_1 = \tau^*, \tau_2 = \tau^*) - ES_{MW}| \leq 2\epsilon_1\epsilon_2$ and $|ES_{MWC}(\tau = \tau^*) - ES_{MW}| \leq 2\epsilon_1\epsilon_2$.

Proof See “Appendix 4”.

Theorem 3 tells us that if the values of survival functions in the two groups are close to 0 at time $t = \tau^*$, then the biases of $ES_{MWE}(\tau_1 = \tau^*, \tau_2 = \tau^*)$ and $ES_{MWC}(\tau = \tau^*)$

will be small. Therefore, we can expect $\widehat{ES}_{MWE}(\tau_1 = \tau^*, \tau_2 = \tau^*)$ and $\widehat{ES}_{MWC}(\tau = \tau^*)$ to perform well if $\widehat{S}_1(\tau^*)$ and $\widehat{S}_2(\tau^*)$ are close to 0.

4 Comparing Effect Sizes

Two types of effect sizes have been developed: the effect sizes ES_L , ES_G and ES_P based on test statistics, and the effect sizes ES_{MW} based on the Mann–Whitney parameter. Various comparisons among these effect sizes are available.

4.1 Effect Sizes as Weighted Differences in Hazards

All four effect sizes represent a weighted difference between two hazard functions and share the same format of definition:

$$\int_0^\infty \xi(t)(\lambda_1(t) - \lambda_2(t))dt. \quad (14)$$

It follows from (5), (6) and (7) that ES_L , ES_G and ES_P (assuming $S_1^*(t) = S_2^*(t) = S^*(t)$) correspond to, respectively,

$$\begin{aligned} \xi_L(t) &= \frac{\pi_1(t)\pi_2(t)}{\pi(t)} = \left(\frac{1}{\pi_1(t)}P_2 + \frac{1}{\pi_2(t)}P_1 \right)^{-1} = \frac{S_1^*(t)S_2^*(t)S_1(t)S_2(t)}{S_1(t)S_1^*(t)P_1 + S_2(t)S_2^*(t)P_2}, \\ \xi_G(t) &= \pi_1(t)\pi_2(t) = S_1^*(t)S_2^*(t)S_1(t)S_2(t), \end{aligned}$$

and

$$\xi_P(t) = \frac{\pi_1(t)\pi_2(t)}{S^*(t)} = S^*(t)S_1(t)S_2(t).$$

The effect size ES_{MW} corresponds to

$$\xi_{MW}(t) = S_1(t)S_2(t),$$

since

$$\begin{aligned} ES_{MW} &= \int_0^\infty S_2(t)f_1(t)dt - \int_0^\infty S_1(t)f_2(t)dt \\ &= \int_0^\infty S_1(t)S_2(t)(\lambda_1(t) - \lambda_2(t))dt. \end{aligned}$$

Formula (14) represents a weighted difference between two hazard functions with the weight at time t equal to $\xi(t)$. The interpretation of the weight for each effect size is clear. For ES_{MW} , the weight $\xi_{MW}(t)$ is the probability that the failure times of both groups are beyond t . For ES_G , the weight $\xi_G(t)$ is the probability that the observed times (either failure time or censoring time) of both groups are beyond t . For ES_L , the

weight $\xi_L(t)$ is the ratio of the probability that the observed times of both groups are beyond t to the probability that the observed time of the combined group is beyond t . Note that $\xi_L(t)$ is essentially the weighted harmonic mean of $\pi_1(t)$ and $\pi_2(t)$. For ES_P where $S_1^*(t) = S_2^*(t) = S^*(t)$, the weight $\xi_P(t)$ is the ratio of the probability that the observed times of both groups are beyond t to the probability that the censoring time is beyond t . Note that $\xi_P(t)$ is in fact the geometric mean of $\xi_{MW}(t)$ and $\xi_G(t)$.

Weights can be compared under various conditions. For any censoring distributions, we have $0 \leq \xi_G(t) \leq \xi_L(t)$ and $0 \leq \xi_G(t) \leq \xi_{MW}(t)$. Assuming the same censoring distribution for both groups, i.e., $S_1^*(t) = S_2^*(t) = S^*(t)$, we obtain $0 \leq \xi_G(t) \leq \xi_P(t) \leq \xi_L(t)$ and $0 \leq \xi_G(t) \leq \xi_P(t) \leq \xi_{MW}(t)$. Furthermore, if assuming the same light censoring distribution for both groups, we see that the weights ξ_G , ξ_P and ξ_{MW} are usually close to each other; if assuming the same heavy censoring distribution for both groups, the weights ξ_G , ξ_P and ξ_L are usually much smaller than the weight ξ_{MW} .

4.2 What Do ES_L , ES_G and ES_P Actually Measure?

The test statistic-based effect sizes ES_L , ES_G and ES_P depend on censoring through the supports and sizes of the censoring survival functions. Supports of the censoring survival functions determine where to compute the weighted differences in hazards. Clearly, the formula (14) shows that these effect sizes are the weighted difference in hazard functions over the time period where censoring is possible. This indicates that these effect sizes capture the difference in hazards during the time period of study which is designed to compare the two groups. Sizes of censoring survival functions impact the magnitudes of the effect sizes. From (14), it is seen that small magnitudes of the effect sizes are produced if either S_1^* or S_2^* or both are small.

To illustrate the above points, consider the scenario where an administrative censoring occurs at time $\tilde{t}$, i.e., the study is terminated at time $\tilde{t}$. We have $ES_G = \int_0^\infty \pi_1(t)\pi_2(t)(\lambda_1(t) - \lambda_2(t))dt = \int_0^{\tilde{t}} \pi_1(t)\pi_2(t)(\lambda_1(t) - \lambda_2(t))dt$. It is seen that ES_G computes the weighted difference in hazard functions before time $\tilde{t}$ and thus ES_G only compares the two groups before time $\tilde{t}$. When $\tilde{t}$ is smaller and thus censoring is heavier, ES_G becomes smaller (assuming $\lambda_1(t) \geq \lambda_2(t)$). As $\tilde{t}$ approaches 0, ES_G goes to 0.

In practice, there are no special restrictions on the use of $\widehat{ES}_G$. Attentions should be paid when using $\widehat{ES}_L$ and $\widehat{ES}_P$. $\widehat{ES}_L$ is sensitive to the change in sample proportions and the use of $\widehat{ES}_P$ requires the equality of two censoring distributions.

To conclude, we note that ES_G is a difference between two special probabilities. Let T_i , C_i and X_i denote, respectively, the variable of failure time, the variable of censoring time and the minimum of T_i and C_i in group i ($i = 1, 2$). Then it is easy to show $ES_G = A - B$, where $A = \int_0^\infty S_2(t)S_1^*(t)S_2^*(t)f_1(t)dt = P(X_2 > X_1, X_1 = T_1)$, $B = \int_0^\infty S_1(t)S_1^*(t)S_2^*(t)f_2(t)dt = P(X_1 > X_2, X_2 = T_2)$. Clearly A is the probability that the observed time of a randomly selected subject from group 2 is longer than that of a randomly selected subject from group 1 who met the event, i.e., the probability that a randomly selected subject from group 2 can be observed to live longer than a randomly selected subject from group 1. A similar statement can be made for B . Therefore, ES_G simply compares the failure times between groups

Fig. 2 The survival functions of the censoring time (gray) and the failure time in group 1 (red) and group 2 (blue)

through comparing observed times. In contrast, $ES_{MW} = P(T_2 > T_1) - P(T_1 > T_2)$ compares the underlying failure times directly.

4.3 Practical Issues in Estimating ES_{MW}

$ES_{MW} = 1 - 2P(T_1 > T_2)$, defined as a simple linear transformation of $P(T_1 > T_2)$, has a clear interpretation. That is, any value of ES_{MW} will immediately translate to the Mann–Whitney parameter $P(T_1 > T_2)$. However, estimating ES_{MW} can be problematic in practice. There may not be enough information to support estimating ES_{MW} . Again consider the scenario where an administrative censoring presents so that the study terminates at time $t = \tilde{t}$. As a result, estimating ES_{MW} could become impossible without additional information on T_1 and T_2 . This is illustrated in the following example. Consider two cases. For case 1 (Fig. 2a), $S_1(t) = (1 - t)\mathbb{I}_{[0,1]}$ and $S_2(t) = (1 - 4t^3)\mathbb{I}_{[0,0.25^{1/3}]}$, so that $ES_{MW} = 1 - 2P(T_1 > T_2) = -0.055$. And for case 2 (Fig. 2b), $S_1(t) = (1 - t)\mathbb{I}_{[0,1]}$ and $S_2(t) = (1 - 4t^3)\mathbb{I}_{[0,0.25]} + (9/8 - 3t/4)\mathbb{I}_{[0.25,1.5]}$, so that $ES_{MW} = 1 - 2P(T_1 > T_2) = 0.477$. Here $\mathbb{I}_{[a,b]}$ equals 1 if t is inside the interval $[a, b]$ and 0 otherwise. Assume that in both cases, the only censoring is the administrative censoring uniformly distributed between 0 and $\tilde{t} = 0.25$. Then in either case, we will only be able to observe the survivals $S_1(t)$ and $S_2(t)$ up to time $\tilde{t}$, which are identical for both cases. Therefore, estimating ES_{MW} using only the observed data in either case becomes problematic. In another word, if we use $\widehat{ES}_{MWE}(\tau_1, \tau_2)$ or $\widehat{ES}_{MWC}(\tau)$ to estimate ES_{MW} , a big error could occur. In general, when estimating ES_{MW} by $\widehat{ES}_{MWE}(\tau_1, \tau_2)$ or $\widehat{ES}_{MWC}(\tau)$, the bias can be quite large if a large portion of the survival function(s) is not observed and we have to choose τ_i or τ such that $\widehat{S}(\tau_i)$ or $\widehat{S}(\tau)$ is much greater than 0.

On the other hand, ES_{MW} can be estimated for several general situations. If the largest observation from each group is a failure time (i.e., the Kaplan–Meier curve drops to 0 at the end), ES_{MW} can be estimated by $1 - 2\widehat{D}$, where $\widehat{D} = -\int_0^\infty \widehat{S}_1(t)d\widehat{S}_2(t)$ is proposed by Efron [15]. If none of the two Kaplan–Meier curves drop to 0 at the end, ES_{MW} can be estimated by both $\widehat{ES}_{MWE}(\tau_1 = \tau^*, \tau_2 = \tau^*)$ and

$\widehat{ES}_{MWC}(\tau = \tau^*)$ when the proportional hazards assumption holds true (Theorem 2) or when $\widehat{S}_1(\tau^*)$ and $\widehat{S}_2(\tau^*)$ are close to 0 (Theorem 3).

If the Kaplan–Meier curve on S_1 drops to 0 at time t_a , the other curve on S_2 stops at time $t_b \geq t_a$ but does not drop to 0, then it is easy to show ES_{MW} can be estimated by $1 + 2 \int_0^{t_a} \widehat{S}_1(t) d\widehat{S}_2(t)$. If the Kaplan–Meier curve on S_1 drops to 0 at time t_a , the other curve on S_2 stops at time $t_b < t_a$ but does not drop to 0, then ES_{MW} can again be estimated by $\widehat{ES}_{MWE}(\tau_1 = \tau^*, \tau_2 = \tau^*)$ and $\widehat{ES}_{MWC}(\tau = \tau^*)$ when the proportional hazards assumption holds true or when $\widehat{S}_1(\tau^*)$ and $\widehat{S}_2(\tau^*)$ are close to 0.

We note that ES_{MW} may have no clinical meaning at all in some situations. For instance, consider two survival functions S_1 and S_2 whose Kaplan–Meier estimates $\widehat{S}_1$ and $\widehat{S}_2$ are close to 1 over a long study time. Then clinically, the difference between S_1 and S_2 could be ignored. However, ES_{MW} between S_1 and S_2 can be large, say, close to 1, since $ES_{MW} = \int_0^\infty S_2(t) S_1(t) (\lambda_1(t) - \lambda_2(t)) dt$ compares the difference in hazards over entire supports of survival functions. As such, a large ES_{MW} could be clinically meaningless. This problem can occur with diseases that have a very optimistic survival, such as thyroid cancer. To resolve, test statistic-based effect sizes should be applied since they only compare the difference over the period of study time.

4.4 Relationship Among ES_L , ES_G , ES_P , ES_{MW} and Hazard Ratio

Examining the functions $\xi(t)$ in (14) leads to the following straightforward comparisons among the values of ES_L , ES_G , ES_P , ES_{MW} and the hazard ratio.

- Assuming $\lambda_1(t) \geq \lambda_2(t)$ for all $t \in [0, \infty]$, we have $0 \leq ES_G \leq ES_L$ and $0 \leq ES_G \leq ES_{MW} \leq 1$.
- Assuming $\lambda_1(t) \geq \lambda_2(t)$ for all $t \in [0, \infty]$ and assuming the same censoring distribution for both groups, we have $0 \leq ES_G \leq ES_P \leq ES_L$ and $0 \leq ES_G \leq ES_P \leq ES_{MW} \leq 1$.
- In the absence of censoring, ES_G and ES_P reduce to ES_{MW} . If censoring is light, then the censoring survival is close to 1 so that ES_G and ES_P are close to ES_{MW} .
- The value of ES_L is influenced by the probabilities by which subjects are assigned to groups. For instance, if $\pi_1(t) \leq \pi_2(t)$ and $\lambda_1(t) \geq \lambda_2(t)$ for all $t \in [0, \infty]$, then ES_L increases as P_1 increases.
- There is a relationship between ES_{MW} and the hazard ratio, a widely used effect size in survival analysis. Assume a constant hazard ratio $r = \frac{\lambda_1(t)}{\lambda_2(t)}$. Then from Theorem 2, we have $ES_{MW} = 1 - 2P(T_1 > T_2) = \frac{r-1}{r+1}$. In particular, $ES_{MW} = 0$ if $r = 1$ and $ES_{MW} \rightarrow 1$ (-1) as $r \rightarrow \infty$ (0). In addition, by Taylor expansion, $ES_{MW} \approx \frac{\theta}{2}$ for θ close to 0, where $\theta = \log r$.

5 Partition of Values of Effect Sizes

The proposed effect sizes can quantify the difference between two populations of time-to-event data. In many cases, one has to assess if an effect size is sufficiently large to be (practically) meaningful. For instance, in a clinical setting, one may need

to evaluate the clinical meaningfulness of an effect size. As such, certain rules are needed in order to determine if an estimate of an effect size is small, medium or large. This involves partitioning the values of effect sizes. For demonstration, below we will develop such rules for effect sizes ES_{MW} and ES_G .

The rule for ES_{MW} can be derived in light of the widely used rule of thumb on the magnitude of Cohen's d . The natural log of the exponentially distributed survival time was considered because the transformed time has the same variance for all exponentially distributed times and thus Cohen's d has a clear interpretation [5]. Under the assumption that the failure time is exponentially distributed in each group, Cohen's d can be written as a function of the hazard ratio $d = -\frac{\log(r)\sqrt{6}}{\pi}$ ("Appendix 5"), and hence small, medium and large hazard ratios correspond to values of 1.29, 1.90 and 2.79, respectively [5]. According to $ES_{MW} = \frac{r-1}{r+1}$, the relationship between the hazard ratio and effect size ES_{MW} stated in Sect. 4.4, we see that the small, medium and large effect sizes ES_{MW} correspond to its values of 0.13, 0.31 and 0.47, respectively. Specifically, using the midpoint 0.22 of 0.13 and 0.31 and midpoint 0.39 of 0.31 and 0.47, we have the following rule of thumb: the effect size ES_{MW} is small if $|ES_{MW}| \in [0, 0.22)$, medium if $|ES_{MW}| \in [0.22, 0.39)$ and large if $|ES_{MW}| \in [0.39, 1]$.

The same idea can be used to derive a rule for ES_G . In addition to the assumption that failure times in the two groups are exponentially distributed, we assume censoring times in the two groups are exponentially distributed. Then ES_G can be written as (see "Appendix 5" for a proof.)

$$ES_G = \frac{e^{-\pi d/\sqrt{6}} - 1}{e^{-\pi d/\sqrt{6}} + 1 + e^{-\pi d/\sqrt{6}} \frac{CR_1}{1-CR_1} + \frac{CR_2}{1-CR_2}}, \quad (15)$$

where d is Cohen's d and $CR_i = P(T_i > C_i)$. Note that CR_i is the censoring rate in group i , which can be estimated by the observed censoring rate $\widehat{CR}_i$. Using the above formula, we can obtain a rule of thumb regarding when the effect size ES_G is small, medium and large. Specifically, for any fixed CR_i , the small, medium and large effect sizes ES_G are computed by using the corresponding values of Cohen's d (small, 0.2; medium, 0.5; large, 0.8). Table 1 shows a list of small, medium and large effect sizes ES_G for selected CR_i . Using the midpoints of numbers in any triplet in the table, intervals for small, medium and large effect sizes can be easily obtained. For instance, for $CR_1 = CR_2 = 10\%$, the triplet consists of 0.11, 0.28, 0.43. Using the midpoint 0.20 of 0.11 and 0.28 and midpoint 0.36 of 0.28 and 0.43, we have the following rule of thumb: with $CR_1 = CR_2 = 10\%$, the effect size ES_G is small if $|ES_G| \in [0, 0.20)$, medium if $|ES_G| \in [0.20, 0.36)$ and large if $|ES_G| \in [0.36, 1]$.

The above rules for ES_{MW} and ES_G are developed under the assumption that actual failure/censoring times follow an exponential distribution. In cases where this assumption does not hold, the rules provide a simple and useful reference.

Table 1 Small, medium and large effect sizes ES_G . The censoring rate in group i is denoted by CR_i ($i = 1, 2$)

CR_1	Norms	CR_2									
		0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
0%	Small	0.13	0.12	0.11	0.10	0.09	0.08	0.07	0.06	0.04	0.02
	Medium	0.31	0.29	0.27	0.24	0.22	0.19	0.16	0.12	0.09	0.04
	Large	0.47	0.44	0.40	0.36	0.32	0.27	0.22	0.17	0.12	0.06
10%	Small	0.12	0.11	0.11	0.10	0.09	0.08	0.07	0.05	0.04	0.02
	Medium	0.30	0.28	0.26	0.24	0.21	0.18	0.15	0.12	0.08	0.04
	Large	0.46	0.43	0.39	0.35	0.31	0.27	0.22	0.17	0.12	0.06
20%	Small	0.12	0.11	0.10	0.09	0.09	0.08	0.07	0.05	0.04	0.02
	Medium	0.29	0.27	0.25	0.23	0.20	0.18	0.15	0.12	0.08	0.04
	Large	0.44	0.41	0.38	0.34	0.30	0.26	0.22	0.17	0.12	0.06
30%	Small	0.11	0.10	0.10	0.09	0.08	0.07	0.06	0.05	0.04	0.02
	Medium	0.27	0.25	0.24	0.22	0.20	0.17	0.15	0.12	0.08	0.04
	Large	0.42	0.40	0.36	0.33	0.29	0.26	0.21	0.17	0.12	0.06
40%	Small	0.10	0.09	0.09	0.08	0.08	0.07	0.06	0.05	0.04	0.02
	Medium	0.25	0.24	0.22	0.21	0.19	0.16	0.14	0.11	0.08	0.04
	Large	0.40	0.38	0.35	0.32	0.28	0.25	0.21	0.16	0.11	0.06
50%	Small	0.09	0.09	0.08	0.08	0.07	0.06	0.06	0.05	0.03	0.02
	Medium	0.23	0.22	0.21	0.19	0.17	0.16	0.13	0.11	0.08	0.04
	Large	0.37	0.35	0.33	0.30	0.27	0.24	0.20	0.16	0.11	0.06
60%	Small	0.08	0.07	0.07	0.07	0.06	0.06	0.05	0.04	0.03	0.02
	Medium	0.20	0.19	0.18	0.17	0.16	0.14	0.12	0.10	0.07	0.04
	Large	0.34	0.32	0.30	0.28	0.25	0.22	0.19	0.15	0.11	0.06
70%	Small	0.06	0.06	0.06	0.06	0.05	0.05	0.04	0.04	0.03	0.02
	Medium	0.17	0.17	0.16	0.15	0.14	0.13	0.11	0.09	0.07	0.04
	Large	0.29	0.28	0.26	0.24	0.22	0.20	0.17	0.14	0.10	0.06
80%	Small	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.03	0.03	0.02
	Medium	0.13	0.13	0.12	0.12	0.11	0.10	0.09	0.08	0.06	0.04
	Large	0.23	0.22	0.21	0.20	0.19	0.17	0.15	0.13	0.09	0.05
90%	Small	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.01
	Medium	0.08	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.05	0.03
	Large	0.14	0.14	0.13	0.13	0.12	0.11	0.11	0.09	0.07	0.05

6 Experiments

Example 1: Simulations

This example presents a simulation study on five effect size estimates: $\widehat{ES}_L$, $\widehat{ES}_G$, $\widehat{ES}_P$, $\widehat{ES}_{MWE}$ and $\widehat{ES}_{MWC}$, defined in (2), (3), (4), (11) and (13), respectively. It

Fig. 3 Survival functions of the failure time and censoring time in Example 1

illustrates various elements developed in previous sections and shows that with large samples, differences in survival between groups can be appropriately described by effect sizes.

The experiment is outlined as follows. First, we chose the two-parameter Weibull distributions as the underlying distributions for the failure and censoring times in groups 1, 2 and 3. We used $T_1 \sim \text{Weibull}(2.5, 0.4)$, $T_2 \sim \text{Weibull}(2.5, 0.5)$ and $T_3 \sim \text{Weibull}(2.5, 0.6)$ for the failure time in group 1, group 2 and group 3, respectively. Then the proportional hazards assumption holds between groups 1 and 2 and between groups 1 and 3, since $\frac{\lambda_1(t)}{\lambda_2(t)} \equiv 1.75$ and $\frac{\lambda_1(t)}{\lambda_3(t)} \equiv 2.76$, where λ_1 , λ_2 and λ_3 are hazard functions in groups 1, 2 and 3, respectively. The censoring times in all three groups were chosen to follow a Weibull(10, 0.6) distribution. The survival functions of the failure and censoring times are shown in Fig. 3. (Note that the censoring survival function has a value smaller than 1×10^{-4} at time $t = 0.75$, so roughly the study was terminated at $t = 0.75$.) Second, survival data were generated, with sample i corresponding to group i ($i = 1, 2, 3$). And we increased the sample size for samples 1 and 2, denoted by n_1 and n_2 , respectively, from 1000 to 20,000 at the same rate, while fixing the sample size for sample 3, denoted by n_3 , at 2000.

Our goal was to compare the survival differences between group 1 and group 2 and between group 1 and group 3. We recorded the values of $\widehat{ES}_L$, $\widehat{ES}_G$, $\widehat{ES}_P$, $\widehat{ES}_{MWE}$ and $\widehat{ES}_{MWC}$ when evaluating the survival difference between group 1 and group 2 and between group 1 and group 3. Here for $\widehat{ES}_{MWE}(\tau_1, \tau_2)$ in (11) and $\widehat{ES}_{MWC}(\tau)$ in (13), we chose $\tau_1 = \tau_2 = \tau$ to be 0.6, since all Kaplan–Meier curves are stable before time $t = 0.6$. The simulation results are summarized in Table 2.

Computation through Eq. (14) shows that values of ES_L , ES_G , ES_P and ES_{MW} between group 1 and 2 are 0.44, 0.26, 0.26 and 0.27, respectively, and that values of

Table 2 Estimates of effect sizes in Example 1

	Group 1 and Group 2						Group 1 and Group 3					
	$n_1 + n_2$	$\widehat{ES}_L$	$\widehat{ES}_G$	$\widehat{ES}_P$	$\widehat{ES}_{MWE}$	$\widehat{ES}_{MWC}$	$n_1 + n_3$	$\widehat{ES}_L$	$\widehat{ES}_G$	$\widehat{ES}_P$	$\widehat{ES}_{MWE}$	$\widehat{ES}_{MWC}$
	2000	0.43	0.25	0.25	0.26	0.26	3000	0.64	0.44	0.45	0.47	0.47
	4000	0.46	0.26	0.27	0.28	0.28	4000	0.69	0.41	0.42	0.45	0.45
	6000	0.47	0.27	0.28	0.29	0.29	5000	0.78	0.46	0.47	0.49	0.49
	8000	0.44	0.25	0.25	0.26	0.26	6000	0.76	0.42	0.44	0.46	0.46
	10,000	0.47	0.27	0.28	0.29	0.29	7000	0.80	0.43	0.44	0.47	0.47
	15,000	0.46	0.26	0.27	0.28	0.28	9500	0.85	0.43	0.44	0.47	0.47
	20,000	0.42	0.25	0.25	0.26	0.26	12,000	0.89	0.45	0.46	0.49	0.49
	25,000	0.46	0.27	0.27	0.28	0.28	14,500	0.84	0.40	0.41	0.44	0.44
	30,000	0.45	0.26	0.26	0.27	0.27	17,000	0.90	0.43	0.44	0.46	0.46
	40,000	0.44	0.25	0.26	0.27	0.27	22,000	0.93	0.43	0.45	0.47	0.47

ES_G , ES_P , and ES_{MW} between group 1 and 3 are 0.43, 0.44 and 0.47, respectively. (Note that $ES_{MW} = \frac{r-1}{r+1}$, the relationship between ES_{MW} and the hazard ratio r stated in Sect. 4.4, can be easily verified between group 1 and group 2 and between group 1 and group 3.) Here in computing ES_L , we used the proportion of subjects in the sample to approximate the probability of assignment of subjects to the corresponding group. ES_L between group 1 and group 3 could not be computed due to the change in sample proportions. The censoring rates in groups 1, 2 and 3 are 11%, 26% and 42%, respectively. Table 2 shows that except for the $\widehat{ES}_L$ column between groups 1 and 3, values of $\widehat{ES}_L$, $\widehat{ES}_G$, $\widehat{ES}_P$, $\widehat{ES}_{MWE}$ and $\widehat{ES}_{MWC}$ are, respectively, close to the values of ES_L , ES_G , ES_P , ES_{MW} and ES_{MW} . This illustrates how estimates of effect sizes developed in Sects. 2 and 3 approximate their corresponding effect sizes. Note that estimates of effect sizes are not much influenced by the change in sample sizes, except for $\widehat{ES}_L$ evaluated between groups 1 and 3.

Table 2 also shows that for each of $\widehat{ES}_L$, $\widehat{ES}_G$, $\widehat{ES}_P$, $\widehat{ES}_{MWE}$, and $\widehat{ES}_{MWC}$, the value obtained between groups 1 and 2 is smaller than that between groups 1 and 3. In particular, all of the values of $\widehat{ES}_{MWE}$, $\widehat{ES}_{MWC}$ and $\widehat{ES}_G$ indicate a medium effect between group 1 and group 2 but a large effect between group 1 and group 3. Therefore, all five types of effect size estimates support the fact that the survival difference between groups 1 and 2 is smaller than the difference between groups 1 and 3 (Fig. 3). These results are consistent with the results of the modified Cohen's d . The modified Cohen's d applies to cases of unequal variances and is defined as the ratio of the difference in means to the square root of the mean of the two variances [11]. Both Cohen's d and modified Cohen's d use the same conventional definition of small, medium and large values of d . The modified Cohen's d for the survival times in this example is -0.516 (medium) between Group 1 and Group 2, and is -0.917 (large) between Group 1 and Group 3.

To conclude this example, we make the following remark. One may wish to employ statistical tests to detect significant differences between groups. When doing this by simulations, we note that as n_1 and n_2 are large (say, above 15,000), p values of tests of logrank, Gehan–Wilcoxon and Prentice–Wilcoxon between group 1 and group 2 will be smaller than those between group 1 and group 3, indicating that the difference between group 1 and group 2 is more significant than the difference between group 1 and group 3. This is contrary to the conclusion from the proposed effect sizes and clearly against the truth shown in Fig. 3. Therefore, reporting only p values, which depend on sample sizes, may not be sufficient when differentiating differences between populations.

Example 2: Gastric Carcinoma Data

The proposed effect sizes can be applied no matter if hazards are proportional. Here we present one example with non-proportional hazards, in which our proposed effect sizes can be used to detect and quantify the difference between two groups, while the approach of the hazard ratio from the Cox regression fails to do so.

The Gastrointestinal Tumor Study Group studied survival of 90 patients with locally advanced, non-resectable gastric carcinoma by comparing two treatments:

Fig. 4 Kaplan–Meier curves of failure times based on the gastric carcinoma data

chemotherapy and radiation, and chemotherapy alone [39]. The aim was to determine if radiation was beneficial to the patients under chemotherapy. 45 and 45 patients were randomly assigned to the groups with and without radiation, respectively. The survival data are publicly available [39]. The Kaplan–Meier curves of the two groups are shown in Fig. 4, which indicates an overall negative impact of radiation on the survival of patients. Note that the observed censoring rates for both groups are low: 18% observed in each group. For convenience, the groups with and without radiation will be referenced as group 1 and group 2, respectively.

The Grambsch–Therneau test for examining the proportional hazard assumption gives a significant p value of 0.0033, suggesting that the hazard ratio would not be an appropriate effect size for analyzing this problem. In fact, the estimate of the hazard ratio (group 1 over group 2) from the Cox regression gives a value of 1.30 with a wide confidence interval (CI) (95% CI: 0.83 to 2.06). These estimates are not very informative when assessing if radiation improves survival.

In comparison, our proposed effect sizes can capture the adverse effect of radiation. Between the groups with and without radiation, $\widehat{ES}_L$, $\widehat{ES}_G$, $\widehat{ES}_P$, $\widehat{ES}_{MWE}$, and $\widehat{ES}_{MWC}$, are, respectively, 0.22 (95% CI, -0.16 to 0.59), 0.27 (95% CI, 0.04 to 0.51), 0.26 (95% CI, 0.03 to 0.50), 0.26 (95% CI, 0.03 to 0.51) and 0.26 (95% CI, 0.03 to 0.51). (Hereafter, the CIs of the effect sizes are the bootstrap confidence intervals [14], based on 100,000 bootstrap samples.) Note that in computing $\widehat{ES}_{MWE}$ and $\widehat{ES}_{MWC}$ we used $\tau_1 = \tau_2 = \tau = \tau^* = \min(t_{\max_1}, t_{\max_2}) = 1472$, since the Kaplan–Meier curves are relatively stable for the entire study period. These two estimates can be used for the example (with non-proportional hazards) since $2\widehat{S}_1(\tau^*) \times \widehat{S}_2(\tau^*) = 2 \times 0.133 \times 0.125 \approx 0.03$, which is small (Theorem 3). From the above estimates, several notes follow below. First, except for $\widehat{ES}_L$, all other four

effect sizes and their CIs show that radiation decreases the survival, which is consistent with the observation in Fig. 4. Second, estimates of ES_{MW} and ES_G suggest a medium effect size (Sect. 5). Third, because of the relatively low censoring rate (18% observed in each group), the estimates of effect sizes based on the test statistics and the Mann–Whitney parameter are close to each other except for $\widehat{ES}_L$ (Sect. 4.4).

Example 3: SEER Thyroid Cancer Data

The tumor, nodal involvement, metastasis (TNM) staging model provides the world-wide standard for cancer prognosis and treatment recommendations [1]. Due to the accumulation of time-to-event data with current advances in cancer research, periodic updates of the TNM are necessary for the improvement of the delivery of patient care. However, such updates often involve expert panels and can be cumbersome when one has to deal with big data. The Ensemble Algorithm for Clustering Cancer Data (EACCD) [8,25] is a machine learning procedure designed specifically to “automatically” update the TNM or its expanded models that involve additional factor(s). This algorithm computes survival differences between groups of patients and utilizes these differences to stratify patients. Below we demonstrate an application of EACCD, integrated with effect sizes, to update and improve the current thyroid cancer staging modeling.

51,291 patients of papillary and follicular thyroid cancer were selected from the SEER 18 databases [40]. Each patient was diagnosed sometime between 2004 and 2010 and was followed up until the year 2015. Cause-specific deaths and survival times (in months) were recorded. In addition, each patient provided information for the following four variables: tumor size (T), regional lymph nodes (N), status of distant metastasis (M) and age (A). All four variables are categorical with 7 levels for T: T0, T1, T2, T2, T3, T4a, T4b; 3 levels for N: N0, N1a, N1b; 2 levels for M: M0 M1; and 2 levels for A: A1 (age < 55), A2 (age ≥ 55). Clinically, it is useful to consider combinations of patients. A combination is defined as a subset of the data corresponding to one level of each factor and is conveniently written using levels of factors (e.g., T2N0M0A2 denotes a subset of patients with T=T2, N=N0, M=M0, A=A2). (Note that the notation of a combination implicitly represents the corresponding underlying population/group.) The dataset contains 39 combinations each including at least 25 patients. Refer to the study by Yang *et al* for specific data management [46].

The goal of applying EACCD is to partition these 39 combinations into g clinically meaningful clusters such that 1) patients from the same cluster have a clinically similar survival while patients from two distinct clusters have clinically different survivals; and 2) g is as small as possible while the set of g clusters can provide the maximum possible accuracy in survival prediction. After such g clusters are obtained, one can study the survival of patients in each cluster and predict the survival for new patients by using survivals from clusters. The merit of applying EACCD is that a large number of original combinations (as often it is) is aggregated into a small number of more manageable clusters that have approximately the same survival prediction accuracy as the original set of combinations [25].

EACCD consists of 3 main steps. Step 1 defines the initial dissimilarity between survival functions of any two groups associated with two combinations. Step 2 computes learned dissimilarities by using initial dissimilarities. Step 3 clusters the combinations by the learned dissimilarities and a linkage method. The clustering result is summarized into a dendrogram which will then be cut to obtain g clusters. Algorithm 1 is one version of the algorithm that utilizes the two-phase Partitioning Around Medoids algorithm (PAM) [26].

Algorithm 1 Ensemble Algorithm for Clustering Cancer Data

Input: A collection of combinations $\text{Com}_1, \dots, \text{Com}_m$ and non-negative weights $w_1, \dots, w_m$ with $\sum_{k=1}^m w_k = 1$.

- 1: Define the initial dissimilarity $\text{dis}_0(\text{Com}_i, \text{Com}_j)$ for any pair Com_i and Com_j .
 - 2: For each k with $1 \leq k \leq m$, apply the two-phase PAM and the initial dissimilarities in Step 1 to partition combinations into k clusters, and for any pair Com_i and Com_j , define $\delta_k(\text{Com}_i, \text{Com}_j) = 1$ if Com_i and Com_j are not assigned into the same cluster and $\delta_k(\text{Com}_i, \text{Com}_j) = 0$ otherwise. Compute the learned dissimilarity $\text{dis}(\text{Com}_i, \text{Com}_j) = \sum_{k=1}^m w_k \delta_k(\text{Com}_i, \text{Com}_j)$ for any pair Com_i and Com_j .
 - 3: Perform hierarchical clustering to cluster the combinations by using $\text{dis}(\text{Com}_i, \text{Com}_j)$ in Step 2.
-

Critical to the application of EACCD is the definition of initial dissimilarities, which can largely affect the output of EACCD [43]. To find an appropriate measure of the initial dissimilarity between two groups associated with two combinations of the thyroid cancer data, we first note that the sample sizes of combinations vary a lot. For instance, T1N0M0A1 has 17,747 cases while T4bN1bM1A2 has only 40 cases. As shown in Example 1, an effect size should serve as a good choice of the initial dissimilarity. In addition, thyroid cancer has a very good prognosis and there are many combinations that have similar and very optimistic survivals before the time of administrative censoring (12 years). According to Sect. 4.3, the test statistic-based effect sizes would serve as better initial dissimilarities than the Mann–Whitney-based effect size.

When applying EACCD to the thyroid dataset (the number of combinations $m = 39$), the equal weights $w_1 = \dots = w_m = 1/m$ and the absolute values of 5 estimates ($\widehat{ES}_L$, $\widehat{ES}_G$, $\widehat{ES}_P$, $\widehat{ES}_{MWE}$ and $\widehat{ES}_{MWC}$) as initial dissimilarities were used in step 1 and the complete linkage [23] was employed in step 3. To compute the Mann–Whitney-based effect sizes, we first note that all Kaplan–Meier curves are stable before $\min_{1 \leq k \leq m} \{t_{\max_k}\}$, where $t_{\max_k}$ is the largest time at which the Kaplan–Meier estimator is defined in Com_k , and therefore, for simplicity, we set $\tau_1 = \tau_2 = \tau = \tau^* = \min_{1 \leq k \leq m} \{t_{\max_k}\}$ for comparing any two combinations.

Figures 5 and 6 show the results based on the test statistic and Mann–Whitney effect sizes, respectively. The left columns of both figures list dendrograms. Each red box represents cutting the dendrogram that corresponds to the optimal C-index [22,25]. Note that g , the number of clusters from cutting, can change when different effect sizes are used. The right columns show the thyroid cancer-specific survival using the Kaplan–Meier estimates. These survival curves allow us to visually evaluate the clusters. Figure 6b and d, based on the effect size $\widehat{ES}_{MW}$, shows that several high curves have a similar survival, indicating the clusters are not well separated and thus

Fig. 5 Dendrograms and survival curves of clusters from EACCD and the test statistic-based effect sizes. A 5-year disease-specific survival rate in percentage is given beneath each combination (leaf of the dendrogram)

form a poor partition of the patients. As explained in Sect. 4.3, ES_{MW} tends to assign a big value of effect size to the difference between two groups (combinations) that share similar and optimistic survival experiences during the time of follow-up, because it compares not only the hazard difference over the period of study time but also the difference after the study. This tendency causes EACCD to partition combinations with similar and high survivals into different clusters, which actually do not have clinically meaningful differences in survival as their survival curves are very similar to each other. By contrast, Fig. 5b, d, and f suggests all ES_L , ES_G , and ES_P lead to a meaningful partition of data.

Fig. 6 Dendrograms and survival curves of clusters from EACCD and the Mann–Whitney-based effect size. A 5-year disease-specific survival rate in percentage is given beneath each combination (leaf of the dendrogram)

The above example shows an application of the effect sizes in upgrading the staging system for thyroid cancer. There exist related works in the literature. The hazard ratio-based effect size and Mann–Whitney parameter-based effect size have been applied to expand the staging system for breast cancer and colorectal cancer [24,25].

7 Conclusion

This paper investigates effect sizes that are used to assess the difference in survival between two groups. It studies three test statistic-based effect sizes ES_L , ES_G and ES_P and one Mann–Whitney-based effect size ES_{MW} , all defined as weighted differences in hazards. The effect sizes ES_L , ES_G and ES_P compare the survival experiences between two groups over the time period of investigation, while ES_{MW} compares the survival experiences over the entire possible time period. Effect sizes ES_L and ES_G , estimated by $\widehat{ES}_L$ and $\widehat{ES}_G$, can be applied to any data, while the use of ES_P , estimated by $\widehat{ES}_P$, assumes equal censoring distributions. Although ES_{MW} has a clear interpretation, it may not be estimable or clinically meaningful. If the Kaplan–Meier curve that stops first drops to 0, ES_{MW} can be estimated directly using the Kaplan–Meier estimates of the survival functions of the two groups. If the Kaplan–Meier curve that stops first does not drop to 0, ES_{MW} can be estimated using both $\widehat{ES}_{MWE}$ and $\widehat{ES}_{MWC}$ under the assumption of proportional hazards between the two groups or under the condition that both curves are low for time close to when they stop.

Effect sizes proposed in this paper have direct applications to big survival data. An estimate of an effect size, computed using two large samples, quantifies the difference between two underlying populations. This is in contrast to checking the p value from a statistical test, which may not provide any insight about the size of the survival difference and could cause misunderstanding.

Effect sizes are particularly useful when repeated comparisons of survival differences are required in a study. For instance, when updating or expanding the TNM staging system for a cancer, assessing the survival difference is needed for many pairs of groups [1]. In this case, effect sizes can be used to differentiate the survival between two groups, so that two groups can be merged if the associated effect size is small. Survival tree modeling [6] represents another scenario that requires a sequential comparison of survival for various pairs of groups. The key element in survival trees is their splitting rules, which can be determined by measuring the dissimilarities between possible nodes and finding the separation that leads to most dissimilar nodes. The largest significant value from test statistics (e.g., logrank, Gehan–Wilcoxon) is often used to suggest the “best” separation, based on the idea that the more extreme the statistic, the more dissimilar the two nodes [9,10]. As seen in previous sections of this article, effect sizes seem to be better choices for measuring dissimilarities between nodes and their use is expected to improve the performance of survival trees.

Hazard ratio, the traditionally used effect size, requires the assumption of proportional hazards. With non-proportional hazards, hazard ratios can fail to detect and quantify the difference between two groups. However, the proposed effect sizes in this paper can be applied no matter if hazards are proportional. This is important in the settings of clinical trials. At the design and planning stages of a study, there is usually no information about whether or not two groups have proportional hazards, so that the assumption of proportional hazards should not be made. The proposed effect sizes, defined as the weighted differences in hazards, would be more reasonable to be used along with the survival time end point for the study than the hazard ratio.

We have defined the effect sizes and studied their estimates for comparing the survival difference between groups. The use of these estimates is straightforward for problems where variable adjustment is not needed. For situations where adjustment needs to be made with important variables, we propose to test the existing adjustment methods, such as matching, stratification and inverse probability weighting [45]. This constitutes an interesting future study. We have not discussed potential analytical formulas for the variances of the proposed estimators. With availability of these formulas, one can investigate not only interval estimation of effect sizes but also the minimum sample sizes required for using the estimators. Finding theoretical formulas for the variances is of interest and deserves further research.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42519-021-00234-8>.

Acknowledgements This study is based upon work supported by John P. Murtha Cancer Center Research Program under the Grant No. 64349-MCC Comprehensive Research.

Availability of Data and Material Data are publicly available.

Code Availability R codes are available as supplementary material.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix 1

Proof of Theorem 1

Before we prove Theorem 1, we need three propositions.

Proposition 1 (Gill's condition) [4,21] *Suppose that for some sequence of process $X_n(t)$ we have*

$$X_n(t) \xrightarrow{P} X(t) \text{ as } n \rightarrow \infty$$

for almost all $t \in [0, \tau]$ with $0 < \tau \leq \infty$, and the deterministic function X satisfies

$$\int_0^\tau |X(t)| dt < \infty.$$

Also suppose that for all $\delta > 0$, there exists k_δ with $\int_0^\tau k_\delta(t) dt < \infty$ such that

$$\liminf_{n \rightarrow \infty} P(|X_n(t)| \leq k_\delta(t) \text{ for all } t \in [0, \tau]) \geq 1 - \delta.$$

Then

$$\sup_{t \in [0, \tau]} \left| \int_0^t X_n(s) ds - \int_0^t X(s) ds \right| \xrightarrow{P} 0.$$

Proposition 2 (Daniels' "in probability linear bound") [4,13] *Let K_n be the empirical distribution function based on a random sample of size n from the continuous (cumulative) distribution function K . Then, for any $\delta \in (0, 1)$,*

$$P(1 - K_n(t-) \leq \delta^{-1}(1 - K(t)) \text{ for all } t \in [0, \infty]) = 1 - \delta.$$

Proposition 3 (Lenglart–Rebolledo inequality) [4,36] *Let N be a counting process, and $M = N - A$ the corresponding local square integrable martingale. Suppose H is a predictable and locally bounded process. Then for any stopping time $T (\leq \infty)$ and any $\epsilon, \eta > 0$,*

$$P \left\{ \sup_{t \leq T} \left\{ \int_0^t H(s) dM(s) \right\}^2 > \epsilon \right\} \leq \frac{\eta}{\epsilon} + P \left\{ \int_0^T H^2(t) d\langle M, M \rangle(t) > \eta \right\},$$

where $\langle M, M \rangle$ is the compensator for M^2 .

With these propositions, we can prove Theorem 1. Rewrite

$$\begin{aligned} U &= \sum_{j=1}^J w_j \left(\frac{Y_{2j}}{Y_{1j} + Y_{2j}} D_{1j} - \frac{Y_{1j}}{Y_{1j} + Y_{2j}} D_{2j} \right) \\ &= \int_0^\infty \frac{Y_2(t) w_n(t)}{Y_1(t) + Y_2(t)} dN_1(t) - \int_0^\infty \frac{Y_1(t) w_n(t)}{Y_1(t) + Y_2(t)} dN_2(t), \end{aligned}$$

where the last equation is an application of the Lebesgue–Stieltjes integration. Letting

$$H_i(t) = \frac{n}{n_1 n_2} \frac{Y_{3-i}(t)}{Y_1(t) + Y_2(t)}, \quad (16)$$

we have $\frac{n}{n_1 n_2} U = \int_0^\infty w_n(t) H_1(t) dN_1(t) - \int_0^\infty w_n(t) H_2(t) dN_2(t)$. According to Doob–Meyer decomposition theorem [31], $N_i(t)$ can be decomposed into a martingale $M_i(t)$ and a compensator $A_i(t) = \int_0^t \lambda_i(s) Y_i(s) ds$, $i = 1, 2$. Therefore,

$$\frac{n}{n_1 n_2} U = \Gamma_n + \Psi_n, \quad (17)$$

where

$$\Gamma_n = \int_0^\infty w_n(t) H_1(t) dM_1(t) - \int_0^\infty w_n(t) H_2(t) dM_2(t)$$

and

$$\Psi_n = \int_0^\infty w_n(t) H_1(t) dA_1(t) - \int_0^\infty w_n(t) H_2(t) dA_2(t).$$

The proof of the theorem consists of two steps.

Step 1 We show $\Gamma_n \xrightarrow{P} 0$ as $n \rightarrow \infty$. We begin with the integral

$$\begin{aligned} \int_0^\infty H_i(t)^2 dA_i(t) &= \int_0^\infty H_i(t)^2 \lambda_i(t) Y_i(t) dt \\ &= \int_0^\infty \left(\frac{n}{n_{3-i}} \frac{Y_{3-i}(t)}{Y_1(t) + Y_2(t)} \right)^2 \frac{\lambda_i(t)}{n_i} \frac{Y_i(t)}{n_i} dt \end{aligned} \quad (18)$$

and will use Proposition 1 to show

$$\int_0^\infty H_i(t)^2 dA_i(t) \xrightarrow{P} 0. \quad (19)$$

First, by the fact that $\frac{n}{n_{3-i}} \xrightarrow{P} \frac{1}{P_{3-i}}$, $\frac{Y_{3-i}(t)}{Y_1(t)+Y_2(t)} \leq 1$ and $\frac{Y_i(t)}{n_i} \leq 1$, we have, for any $0 \leq t < \infty$,

$$\left(\frac{n}{n_{3-i}} \frac{Y_{3-i}(t)}{Y_1(t) + Y_2(t)} \right)^2 \frac{\lambda_i(t)}{n_i} \frac{Y_i(t)}{n_i} = (O_p(1))^2 * o_p(1) * O_p(1) = o_p(1),$$

i.e.,

$$\left(\frac{n}{n_{3-i}} \frac{Y_{3-i}(t)}{Y_1(t) + Y_2(t)} \right)^2 \frac{\lambda_i(t)}{n_i} \frac{Y_i(t)}{n_i} \xrightarrow{P} 0. \quad (20)$$

Second, we note that $(n_i - Q_i(t))/n_i$ is the value, at time $t-$ (time just before t), of the empirical distribution function from the cumulative distribution function of the failure time in group i . According to Proposition 2, for any $\delta \in (0, 1)$, we have,

$$P \left(1 - (n_i - Q_i(t))/n_i \leq \delta^{-1} S_i(t) \text{ for all } t \in [0, \infty] \right) = 1 - \delta.$$

Thus, since $Y_i(t) \leq Q_i(t)$, we have

$$P \left(\frac{Y_i(t)}{n_i} \leq \delta^{-1} S_i(t) \text{ for all } t \in [0, \infty] \right) \geq 1 - \delta. \quad (21)$$

Third, by Bonferroni inequality, we see

$$\begin{aligned} P \left(\frac{n^2}{n_{3-i}^2} \frac{Y_i(t)}{n_i} \leq \left(\frac{1}{P_{3-i}^2} + \delta \right) \delta^{-1} S_i(t) \text{ for all } t \in [0, \infty] \right) \\ \geq P \left(\frac{n^2}{n_{3-i}^2} \leq \frac{1}{P_{3-i}^2} + \delta \right) + P \left(\frac{Y_i(t)}{n_i} \leq \delta^{-1} S_i(t) \text{ for all } t \in [0, \infty] \right) - 1. \end{aligned}$$

Then, from $\frac{n^2}{n_{3-i}^2} \xrightarrow{P} \frac{1}{P_{3-i}^2}$ and Eq. (21), we have

$$\begin{aligned} \liminf_{n \rightarrow \infty} P \left(\frac{n^2}{n_{3-i}^2} \frac{Y_i(t)}{n_i} \leq \left(\frac{1}{P_{3-i}^2} + \delta \right) \delta^{-1} S_i(t) \text{ for all } t \in [0, \infty] \right) \\ \geq \liminf_{n \rightarrow \infty} P \left(\frac{n^2}{n_{3-i}^2} \leq \frac{1}{P_{3-i}^2} + \delta \right) \\ + \liminf_{n \rightarrow \infty} P \left(\frac{Y_i(t)}{n_i} \leq \delta^{-1} S_i(t) \text{ for all } t \in [0, \infty] \right) - 1 \end{aligned}$$

$$\geq 1 - \delta,$$

and correspondingly

$$\liminf_{n \rightarrow \infty} P \left(\frac{n^2}{n_{3-i}^2} \frac{Y_i(t)}{n_i} \lambda_i(t) \leq \left(\frac{1}{P_{3-i}^2} + \delta \right) \delta^{-1} f_i(t) \text{ for all } t \in [0, \infty] \right) \geq 1 - \delta.$$

Hence, from

$$\left(\frac{n}{n_{3-i}} \frac{Y_{3-i}(t)}{Y_1(t) + Y_2(t)} \right)^2 \frac{\lambda_i(t)}{n_i} \frac{Y_i(t)}{n_i} \leq \left(\frac{n}{n_{3-i}} \right)^2 \frac{Y_i(t)}{n_i} \lambda_i(t),$$

we obtain

$$\liminf_{n \rightarrow \infty} P \left(\left(\frac{n}{n_{3-i}} \frac{Y_{3-i}(t)}{Y_1(t) + Y_2(t)} \right)^2 \frac{\lambda_i(t)}{n_i} \frac{Y_i(t)}{n_i} \leq k_\delta(t) \text{ for all } t \in [0, \infty] \right) \geq 1 - \delta, \quad (22)$$

where $k_\delta(t) = \left(\frac{1}{P_{3-i}^2} + \delta \right) \delta^{-1} f_i(t)$ with $\int_0^\infty k_\delta(t) dt = \frac{1}{P_{3-i}^2} \delta^{-1} + 1 < \infty$ for all $\delta > 0$. By Eqs. (18), (20), (22) and Proposition 1, we see (19) is established.

We are now in a position to show $\Gamma_n \xrightarrow{P} 0$. In fact, from Eq. (19) and the fact that $|w_n(t)| \leq 1$ and thus $\int_0^\infty (w_n(t) H_i(t))^2 dA_i(t) \leq \int_0^\infty H_i^2(t) dA_i(t)$, we see $\int_0^\infty (w_n(t) H_i(t))^2 dA_i(t) \xrightarrow{P} 0$. Since $A_i(t)$ is the compensator of the counting process $N_i(t)$ with $A_i(0) = 0$ a.s. and $A_i(t) < \infty$ a.s. for any t , $M_i = N_i - A_i$ is a local square integrable martingale [16]. In addition, since $H_i(t)$ is a left-continuous process and $w_n(t) H_i(t) = \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \frac{w_n(t) Y_{3-i}(t)}{Y_1(t) + Y_2(t)} \leq 2$, $w_n(t) H_i(t)$ is a predictable and bounded process [16]. Moreover, $\langle M_i, M_i \rangle(t) = A_i(t)$ since $A_i(t)$ is continuous [16]. Using Proposition 3, we have, for any $\epsilon, \eta > 0$,

$$\begin{aligned} & P \left\{ \sup_{t \leq \infty} \left\{ \int_0^t w_n(s) H_i(s) dM_i(s) \right\}^2 > \epsilon \right\} \\ & \leq \frac{\eta}{\epsilon} + P \left\{ \int_0^\infty (w_n(t) H_i(t))^2 dA_i(t) > \eta \right\}. \end{aligned}$$

Therefore, $\int_0^\infty w_n(t) H_i(t) dM_i(t) \xrightarrow{P} 0$. Thus, $\Gamma_n = \int_0^\infty w_n(t) H_1(t) dM_1(t) - \int_0^\infty w_n(t) H_2(t) dM_2(t) \xrightarrow{P} 0$.

Step 2 We show $\Psi_n \xrightarrow{P} \int_0^\infty \frac{\pi_1(t)\pi_2(t)}{\pi(t)} \omega(t)(\lambda_1(t) - \lambda_2(t))dt$. Again the proof is done by taking advantage of Proposition 1. Rewrite Ψ_n as

$$\begin{aligned}\Psi_n &= \int_0^\infty \frac{n}{n_1 n_2} \frac{Y_2(t)}{Y_1(t) + Y_2(t)} w_n(t) \lambda_1(t) Y_1(t) dt \\ &\quad - \int_0^\infty \frac{n}{n_1 n_2} \frac{Y_1(t)}{Y_1(t) + Y_2(t)} w_n(t) \lambda_2(t) Y_2(t) dt \\ &= \int_0^\infty \frac{(Y_1(t)/n_1)(Y_2(t)/n_2)}{(Y_1(t) + Y_2(t))/n} w_n(t)(\lambda_1(t) - \lambda_2(t)) dt.\end{aligned}$$

By the law of large numbers, we have $\frac{Y_1(t)}{n_1} \xrightarrow{P} P(X \geq t | Z = 1) = \pi_1(t)$, $\frac{Y_2(t)}{n_2} \xrightarrow{P} P(X \geq t | Z = 2) = \pi_2(t)$, $\frac{Y_1(t) + Y_2(t)}{n} \xrightarrow{P} P(X \geq t) = \pi(t)$, and hence, due to the assumption $w_n(t) \xrightarrow{P} \omega(t)$, the integrand in Ψ_n has the following probability limit:

$$\frac{(Y_1(t)/n_1)(Y_2(t)/n_2)}{(Y_1(t) + Y_2(t))/n} w_n(t)(\lambda_1(t) - \lambda_2(t)) \xrightarrow{P} \frac{\pi_1(t)\pi_2(t)}{\pi(t)} \omega(t)(\lambda_1(t) - \lambda_2(t)). \quad (23)$$

To obtain the lower bound of $\liminf_{n \rightarrow \infty}$, we first observe

$$\left| \frac{(Y_1(t)/n_1)(Y_2(t)/n_2)}{(Y_1(t) + Y_2(t))/n} w_n(t)(\lambda_1(t) - \lambda_2(t)) \right| \leq \frac{n}{n_2} \frac{Y_1(t)}{n_1} \lambda_1(t) + \frac{n}{n_1} \frac{Y_2(t)}{n_2} \lambda_2(t). \quad (24)$$

As before, using Bonferroni inequality, $\frac{n}{n_{3-i}} \xrightarrow{P} \frac{1}{P_{3-i}}$, and Eq. (21), we see

$$\begin{aligned}\liminf_{n \rightarrow \infty} P \left(\frac{n}{n_{3-i}} \frac{Y_i(t)}{n_i} \leq \left(\frac{1}{P_{3-i}} + \frac{\delta}{2} \right) 2\delta^{-1} S_i(t) \text{ for all } t \in [0, \infty] \right) \\ \geq \liminf_{n \rightarrow \infty} P \left(\frac{n}{n_{3-i}} \leq \frac{1}{P_{3-i}} + \frac{\delta}{2} \right) \\ + \liminf_{n \rightarrow \infty} P \left(\frac{Y_i(t)}{n_i} \leq 2\delta^{-1} S_i(t) \text{ for all } t \in [0, \infty] \right) - 1 \\ = \liminf_{n \rightarrow \infty} P \left(\frac{Y_i(t)}{n_i} \leq 2\delta^{-1} S_i(t) \text{ for all } t \in [0, \infty] \right) \\ \geq 1 - \frac{\delta}{2},\end{aligned}$$

and correspondingly

$$\liminf_{n \rightarrow \infty} P \left(\frac{n}{n_{3-i}} \frac{Y_i(t)}{n_i} \lambda_i(t) \leq \left(\frac{1}{P_{3-i}} + \frac{\delta}{2} \right) 2\delta^{-1} f_i(t) \text{ for all } t \in [0, \infty] \right) \geq 1 - \frac{\delta}{2}.$$

Then, using Bonferroni inequality one more time, we obtain

$$\liminf_{n \rightarrow \infty} P \left(\frac{n}{n_2} \frac{Y_1(t)}{n_1} \lambda_1(t) + \frac{n}{n_1} \frac{Y_2(t)}{n_2} \lambda_2(t) \leq k_\delta(t) \text{ for all } t \in [0, \infty] \right) \geq 1 - \delta,$$

where $k_\delta(t) = \left(\frac{1}{P_2} + \frac{\delta}{2}\right) 2\delta^{-1} f_1(t) + \left(\frac{1}{P_1} + \frac{\delta}{2}\right) 2\delta^{-1} f_2(t)$ with $\int_0^\infty k_\delta(t) dt = \frac{2}{P_2} \delta^{-1} + 1 + \frac{2}{P_1} \delta^{-1} + 1 < \infty$. And hence, according to (24),

$$\liminf_{n \rightarrow \infty} P \left(\left| \frac{\frac{Y_1(t)}{n_1} \frac{Y_2(t)}{n_2}}{\frac{Y_1(t) + Y_2(t)}{n}} w_n(t) (\lambda_1(t) - \lambda_2(t)) \right| \leq k_\delta(t) \text{ for all } t \in [0, \infty] \right) \geq 1 - \delta. \quad (25)$$

Finally, by (23), (25), the assumption $\int_0^\infty \left| \frac{\pi_1(t)\pi_2(t)}{\pi(t)} \omega(t) (\lambda_1(t) - \lambda_2(t)) \right| dt < \infty$, and Proposition 1, we have

$$\Psi_n \xrightarrow{p} \int_0^\infty \frac{\pi_1(t)\pi_2(t)}{\pi(t)} \omega(t) (\lambda_1(t) - \lambda_2(t)) dt.$$

Combining results in steps 1 and 2, we see that $\frac{n}{n_1 n_2} U$ converges in probability to $\int_0^\infty \frac{\pi_1(t)\pi_2(t)}{\pi(t)} \omega(t) (\lambda_1(t) - \lambda_2(t)) dt$. $\square$

Appendix 2

Proof of Corollary 1

(a) If $w_j = 1$, then $\omega(t) = 1$. Note the following:

$$\begin{aligned} & \int_0^\infty \left| \frac{\pi_1(t)\pi_2(t)}{\pi(t)} (\lambda_1(t) - \lambda_2(t)) \right| dt \\ & \leq \int_0^\infty \frac{1}{4} \left(\frac{S_2(t)S_2^*(t)}{P_1} + \frac{S_1(t)S_1^*(t)}{P_2} \right) |\lambda_1(t) - \lambda_2(t)| dt \\ & \leq \int_0^\infty \frac{S_2(t)S_2^*(t)}{4P_1} (\lambda_1(t) + \lambda_2(t)) dt + \int_0^\infty \frac{S_1(t)S_1^*(t)}{4P_2} (\lambda_1(t) + \lambda_2(t)) dt \\ & = E_2 \left[\frac{\left(\frac{\lambda_1(t)}{\lambda_2(t)} + 1 \right) S_2^*(t)}{4P_1} \right] + E_1 \left[\frac{\left(\frac{\lambda_2(t)}{\lambda_1(t)} + 1 \right) S_1^*(t)}{4P_2} \right] \\ & < \infty, \end{aligned}$$

where the last inequality is obtained by using the assumption $E_i[\lambda_{3-i}(t)/\lambda_i(t)] < \infty$. Then the result follows immediately from Theorem 1.

(b) If $w_j = Y_j/n$, then $\omega(t) = \pi(t)$. Since

$$\begin{aligned} \int_0^\infty \left| \frac{\pi_1(t)\pi_2(t)}{\pi(t)} \omega(t)(\lambda_1(t) - \lambda_2(t)) \right| dt &= \int_0^\infty |\pi_1(t)\pi_2(t)(\lambda_1(t) - \lambda_2(t))| dt \\ &\leq \int_0^\infty |S_1(t)S_2(t)(\lambda_1(t) - \lambda_2(t))| dt \\ &\leq \int_0^\infty (S_2(t)f_1(t) + S_1(t)f_2(t)) dt \\ &\leq 2, \end{aligned}$$

the result follows immediately from Theorem 1.

(c) Here we have $w_n(t) = \prod_{k:t_k < t} \left(1 - \frac{D_k}{Y_k}\right) = \widehat{S}(t-)$, which is the left-continuous version of the Kaplan–Meier estimator. Let Υ denote the support of $S(t)S^*(t)$, i.e., the set of t such that $S(t)S^*(t) > 0$. We can show that for any $t \in \Upsilon$, $w_n(t) \xrightarrow{P} S(t) = S_1(t)P_1 + S_2(t)P_2$. First, for the pooled population, T is independent of C when $S_1^*(t) = S_2^*(t) = S^*(t)$. This is because $P(T \leq t)P(C \leq t') = P(T \leq t, C \leq t')$, which can be easily seen by conditioning and using the assumption of independence between T and C in each group. Second, for any $\epsilon > 0$ and $\epsilon_1 > 0$,

$$\begin{aligned} P(|\widehat{S}(t-) - S(t)| > \epsilon) &\leq P(|\widehat{S}(t - \epsilon_1) - S(t - \epsilon_1)| > \epsilon/6) \\ &\quad + P(|\widehat{S}(t) - S(t)| > \epsilon/6) \\ &\quad + P(|S(t - \epsilon_1) - S(t)| > \epsilon/6) \\ &\quad + P(|\widehat{S}(t) - S(t)| > \epsilon/2). \end{aligned}$$

Then from the continuity of $S(t)$, the independence between T and C , and the consistency of $\widehat{S}(t)$ [19], we see that for $t \in \Upsilon$,

$$\lim_{n \rightarrow \infty} P(|\widehat{S}(t-) - S(t)| > \epsilon) = 0,$$

showing $w_n(t) \xrightarrow{P} S(t) = S_1(t)P_1 + S_2(t)P_2$ for $t \in \Upsilon$. Therefore,

$$\begin{aligned} \frac{\widehat{S}(t-)(Y_1(t)/n_1)(Y_2(t)/n_2)}{(Y_1(t) + Y_2(t))/n} (\lambda_1(t) - \lambda_2(t)) &\xrightarrow{P} \frac{S(t)\pi_1(t)\pi_2(t)}{\pi(t)} (\lambda_1(t) - \lambda_2(t)) \\ &= S^*(t)S_1(t)S_2(t)(\lambda_1(t) - \lambda_2(t)) \end{aligned}$$

holds for $t \in \Upsilon$. Clearly this limit also holds true for any $t \notin \Upsilon$, since $Y_1(t)$, $Y_2(t)$ and $S^*(t)S_1(t)S_2(t)$ all are 0 for $t \notin \Upsilon$. Hence, Eq. (23) holds. Now the result is obtained by following exactly the proof of Theorem 1 and observing $\int_0^\infty |S^*(t)S_1(t)S_2(t)(\lambda_1(t) - \lambda_2(t))| dt \leq \int_0^\infty (S_2(t)f_1(t) + S_1(t)f_2(t)) dt \leq 2$. $\square$

Appendix 3

Proof of Theorem 2

From $\lambda_1(t) = r\lambda_2(t)$, we have $S_1(t) = (S_2(t))^r$. Then we have the following: $P(T_1 > T_2) = \int_0^\infty S_1(t)f_2(t)dt = \int_0^\infty (S_2(t))^r f_2(t)dt = \frac{1}{r+1}$, so that $ES_{MW} = 1 - 2P(T_1 > T_2) = \frac{r-1}{r+1}$.

Let $\lambda_1^* = -\log(S_1(\tau^*))/\tau^*$ and $\lambda_2^* = -\log(S_2(\tau^*))/\tau^*$. Then it follows from $S_1(t) = (S_2(t))^r$ that we have $\lambda_1^* = r\lambda_2^*$. In light of $D_e(\tau_1 = \tau^*, \tau_2 = \tau^*) = A + B + C$, where $A = \int_0^{\tau^*} S_1(t)f_2(t)dt = \int_0^{\tau^*} (S_2(t))^r f_2(t)dt = \frac{1}{r+1} (1 - (S_2(\tau^*))^{r+1})$, $B = 0$ and $C = \frac{\lambda_2^*}{\lambda_1^* + \lambda_2^*} e^{-(\lambda_1^* + \lambda_2^*)\tau^*} = \frac{1}{r+1} (S_2(\tau^*))^{r+1}$, we see $D_e(\tau_1 = \tau^*, \tau_2 = \tau^*) = \frac{1}{r+1}$ so that $ES_{MWE}(\tau_1 = \tau^*, \tau_2 = \tau^*) = 1 - 2D_e(\tau_1 = \tau^*, \tau_2 = \tau^*) = \frac{r-1}{r+1}$.

Finally, we have $ES_{MWC}(\tau = \tau^*) = 1 - 2D_c(\tau = \tau^*) = \frac{r-1}{r+1}$, since $D_c(\tau = \tau^*) = \frac{1}{1 - S_1(\tau^*)S_2(\tau^*)} \int_0^{\tau^*} S_1(t)f_2(t)dt = \frac{1}{1 - (S_2(\tau^*))^{r+1}} \int_0^{\tau^*} (S_2(t))^r f_2(t)dt = \frac{1}{1 - (S_2(\tau^*))^{r+1}} \frac{1}{r+1} (1 - (S_2(\tau^*))^{r+1}) = \frac{1}{r+1}$. $\square$

Appendix 4

Proof of Theorem 3

To prove the first inequality, recall (Sect. 3.1) $D_e(\tau_1 = \tau^*, \tau_2 = \tau^*) = -\int_0^{\tau^*} S_1(t)dS_2(t) + \frac{\lambda_2^*}{\lambda_1^* + \lambda_2^*} e^{-(\lambda_1^* + \lambda_2^*)\tau^*}$ with $\lambda_i^* = -\log(S_i(\tau^*))/\tau^*$ for $i = 1, 2$ and $P(T_1 > T_2) = -\int_0^{\tau^*} S_1(t)dS_2(t) - \int_{\tau^*}^\infty S_1(t)dS_2(t)$. Therefore, we have

$$\begin{aligned} & |D_e(\tau_1 = \tau^*, \tau_2 = \tau^*) - P(T_1 > T_2)| \\ & \leq \max \left(\left| \frac{\lambda_2^*}{\lambda_1^* + \lambda_2^*} e^{-(\lambda_1^* + \lambda_2^*)\tau^*} \right|, \left| \int_{\tau^*}^\infty S_1(t)f_2(t)dt \right| \right) \\ & \leq \max \left(\frac{\lambda_2^*}{\lambda_1^* + \lambda_2^*} e^{-\lambda_1^*\tau^*} e^{-\lambda_2^*\tau^*}, \int_{\tau^*}^\infty S_1(\tau^*)f_2(t)dt \right) \\ & = \max \left(\frac{\lambda_2^*}{\lambda_1^* + \lambda_2^*} S_1(\tau^*)S_2(\tau^*), S_1(\tau^*)S_2(\tau^*) \right) \\ & \leq \epsilon_1 \epsilon_2. \end{aligned}$$

Hence, $|ES_{MWE}(\tau_1 = \tau^*, \tau_2 = \tau^*) - ES_{MW}| = 2|D_e(\tau_1, \tau_2) - P(T_1 > T_2)| \leq 2\epsilon_1\epsilon_2$.

For the second inequality, recall (Sect. 3.2) $D_c(\tau = \tau^*) = \frac{-\int_0^{\tau^*} S_1(t)dS_2(t)}{1 - S_1(\tau^*)S_2(\tau^*)}$. Therefore,

$$\begin{aligned}
& |D_c(\tau = \tau^*) - P(T_1 > T_2)| \\
&= \left| \left(\frac{1}{1 - S_1(\tau^*)S_2(\tau^*)} - 1 \right) \int_0^{\tau^*} S_1(t)f_2(t)dt - \int_{\tau^*}^{\infty} S_1(t)f_2(t)dt \right| \\
&\leq \max \left(\frac{S_1(\tau^*)S_2(\tau^*)}{1 - S_1(\tau^*)S_2(\tau^*)} \int_0^{\tau^*} S_1(t)f_2(t)dt, \int_{\tau^*}^{\infty} S_1(t)f_2(t)dt \right) \\
&\leq \max \left(\frac{S_1(\tau^*)S_2(\tau^*)}{1 - S_1(\tau^*)S_2(\tau^*)} (1 - S_2(\tau^*)), \int_{\tau^*}^{\infty} S_1(t)f_2(t)dt \right) \\
&\leq \max \left(\frac{S_1(\tau^*)S_2(\tau^*)}{1 - S_1(\tau^*)S_2(\tau^*)} (1 - S_1(\tau^*)S_2(\tau^*)), \int_{\tau^*}^{\infty} S_1(\tau^*)f_2(t)dt \right) \\
&= \max (S_1(\tau^*)S_2(\tau^*), S_1(\tau^*)S_2(\tau^*)) \\
&\leq \epsilon_1 \epsilon_2.
\end{aligned}$$

Hence,

$$|ES_{MWC}(\tau = \tau^*) - ES_{MW}| = 2 |D_c(\tau) - P(T_1 > T_2)| \leq 2\epsilon_1 \epsilon_2.$$

□

Appendix 5

Proof of (15)

Assume $S_1(t)$, $S_2(t)$, $S_1^*(t)$ and $S_2^*(t)$ are all exponentially distributed with $S_1(t) = e^{-\lambda_1 t}$, $S_1^*(t) = e^{-\lambda_1^* t}$, $S_2(t) = e^{-\lambda_2 t}$, $S_2^*(t) = e^{-\lambda_2^* t}$ and $\lambda_1 = r\lambda_2$ (i.e., hazard ratio is r). Since $CR_1 = 1 - P(C_1 > T_1) = 1 - \int_0^{\infty} S_1^*(t)f_1(t)dt = \frac{\lambda_1^*}{\lambda_1 + \lambda_1^*}$, we have $\lambda_1^* = \frac{\lambda_1 CR_1}{1 - CR_1}$. Similarly, $\lambda_2^* = \frac{\lambda_2 CR_2}{1 - CR_2}$. Then we have $ES_G = \int_0^{\infty} e^{-\lambda_1 t} e^{-\lambda_2 t} e^{-\lambda_1^* t} e^{-\lambda_2^* t} (\lambda_1 - \lambda_2) dt = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2 + \lambda_1^* + \lambda_2^*} = \frac{r - 1}{r + 1 + r \frac{CR_1}{1 - CR_1} + \frac{CR_2}{1 - CR_2}}$.

In addition, it is known that the natural log of the exponentially distributed survival time has mean $-\log(\lambda_i) - \gamma$ (γ is the Euler–Mascheroni constant) and variance $\pi^2/6$ [5]. Therefore, the Cohen's d based on the differences of the log failure times between the two groups is

$$d = \frac{(-\log(\lambda_1) - \gamma) - (-\log(\lambda_2) - \gamma)}{\sqrt{\pi^2/6}} = -\frac{\log(r)\sqrt{6}}{\pi},$$

which is a function of r . Using this relationship between r and d , we see that ES_G can be represented as a function of Cohen's d and censoring rates CR_i , as shown in formula (15). □

References

1. Amin MB, Edge S, Greene FL, Byrd DR, Brookland RK, Washington MK et al (2017) AJCC cancer staging manual, 8th edn. Springer, Berlin
2. Amrhein V, Greenland S, McShane B (2019) Scientists rise up against statistical significance. *Nature* 567(7748):305–307. <https://doi.org/10.1038/d41586-019-00857-9>
3. Andersen PK, Gill RD (1982) Cox's regression model for counting processes: a large sample study. *Ann Stat* 10(4):1100–1120
4. Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York
5. Azuero A (2016) A note on the magnitude of hazard ratios. *Cancer* 122(8):1298–1299. <https://doi.org/10.1002/cncr.29924>
6. Bou-Hamad I, Larocque D, Ben-Ameur H (2011) A review of survival trees. *Stat Surv* 5:44–71. <https://doi.org/10.1214/09-SS047>
7. Brown BW Jr, Hollander M, Korwar RM (1974) Nonparametric tests of independence for censored data, with application to heart transplant studies. In: Proshan F, Serfling RJ (eds) Reliability and biometry. SIAM, Philadelphia, pp 327–354
8. Chen D, Xing K, Henson D, Sheng L, Schwartz AM, Cheng X (2009) Developing prognostic systems of cancer patients by ensemble clustering. *J Biomed Biotechnol* 2009:632786. <https://doi.org/10.1155/2009/632786>
9. Ciampi A, Thiffault J, Nakache JP, Asselain B (1986) Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Comput Stat Data Anal* 4(3):185–204. [https://doi.org/10.1016/0167-9473\(86\)90033-2](https://doi.org/10.1016/0167-9473(86)90033-2)
10. Ciampi A, Hogg SA, McKinney S, Thiffault J (1988) RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. Methods and program features. *Comput Methods Progr Biomed* 26(3):239–256
11. Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale
12. Cox DR (1972) Regression models and life-tables. *J R Stat Soc Ser B Stat Methodol* 34:187–220. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>
13. Daniels HE (1945) The statistical theory of the strength of bundles of threads. I. *Proc R Soc Lond A* 183(995):405–435. <https://doi.org/10.1098/rspa.1945.0011>
14. Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge
15. Efron B (1967) The two sample problem with censored data. In: Proceedings of 5th Berkeley symposium, vol 4, pp 831–853
16. Fleming TR, Harrington DP (1991) Counting processes and survival analysis. Wiley, New York
17. Fleming TR, Harrington DP, O'sullivan M (1987) Supremum versions of the log-rank and generalized Wilcoxon statistics. *J Am Stat Assoc* 82(397):312–320. <https://doi.org/10.1080/01621459.1987.10478435>
18. Gehan EA (1965) A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 52(1–2):203–224. <https://doi.org/10.2307/2333825>
19. Gill RD et al (1983) Large sample behaviour of the product-limit estimator on the whole line. *Ann Stat* 11(1):49–58
20. Gill RD (1980) Censoring and stochastic integrals. *Stat Neerl* 34(2):124. <https://doi.org/10.1111/j.1467-9574.1980.tb00692.x>
21. Gill RD (1983) Discussion of the papers by Helland and Kurtz. *Bull Int Stat Inst* 50(3):239–243
22. Harrell FE, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 15:361–387
23. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York
24. Hueman M, Wang H, Henson D, Chen D (2019) Expanding the TNM for cancers of the colon and rectum using machine learning: a demonstration. *ESMO Open* 4(3):e000518. <https://doi.org/10.1136/esmoopen-2019-000518>
25. Hueman MT, Wang H, Yang CQ, Sheng L, Henson DE, Schwartz AM, Chen D (2018) Creating prognostic systems for cancer patients: a demonstration using breast cancer. *Cancer Med* 7(8):3611–3621. <https://doi.org/10.1002/cam4.1629>

26. Kaufman L, Rousseeuw PJ (2009) Finding groups in data: an introduction to cluster analysis. Wiley, Hoboken
27. Khalilzadeh J, Tasci AD (2017) Large sample size, significance level, and the effect size: solutions to perils of using big data for academic research. *Tour Manag* 62:89–96. <https://doi.org/10.1016/j.tourman.2017.03.026>
28. Lawless JF (2003) Statistical models and methods for lifetime data. Wiley, Hoboken
29. Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 18(1):50–60
30. Mantel N (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 50(3):163–170
31. Meyer PA (1962) A decomposition theorem for supermartingales. *Ill J Math* 6(2):193–205
32. Moeschberger M, Klein JP (1985) A comparison of several methods of estimating the survival function when there is extreme right censoring. *Biometrics* 41(1):253–259
33. Montes I, Miranda E, Montes S (2014) Decision making with imprecise probabilities and utilities by means of statistical preference and stochastic dominance. *Eur J Oper Res* 234(1):209–220
34. Peto R, Peto J (1972) Asymptotically efficient rank invariant test procedures. *J R Stat Soc Ser A* 135(2):185–206. <https://doi.org/10.2307/2344317>
35. Prentice RL (1978) Linear rank tests with right censored data. *Biometrika* 65(1):167–179. <https://doi.org/10.2307/2335292>
36. Rebollo R (1980) Central limit theorems for local martingales. *Probab Theory Relat Fields* 51(3):269–286. <https://doi.org/10.1007/BF00587353>
37. Royston P, Parmar MK (2011) The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med* 30(19):2409–2421
38. Schemper M, Wakounig S, Heinze G (2009) The estimation of average hazard ratios by weighted Cox regression. *Stat Med* 28(19):2473–2489
39. Stablein DM, Carter WH Jr, Novak JW (1981) Analysis of survival data with nonproportional hazard functions. *Control Clin Trials* 2(2):149–159
40. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973–2015) (Nov 2017) sub (2018) National Cancer Institute. DCCPS, Surveillance Research Program
41. Tarone RE, Ware J (1977) On distribution-free tests for equality of survival distributions. *Biometrika* 64(1):156–160. <https://doi.org/10.1093/biomet/64.1.156>
42. Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T et al (2014) Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol* 32(22):2380–2385. <https://doi.org/10.1200/JCO.2014.55.2208>
43. Wang H, Chen D, Hueman MT, Sheng L, Henson DE (2017) Clustering big cancer data by effect sizes. In: *IEEE/ACM second CHASE conference*. IEEE, pp 58–63 <https://doi.org/10.1109/CHASE.2017.60>
44. Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a world beyond “ $p < 0.05$ ”. *Am Stat* 73:1–19 <https://doi.org/10.1080/00031305.2019.1583913>
45. Xie J, Liu C (2005) Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Stat Med* 24(20):3089–3110. <https://doi.org/10.1002/sim.2174>
46. Yang C, Gardiner L, Wang H, Hueman M, Chen D (2019) Creating prognostic systems for well differentiated thyroid cancer using machine learning. *Front Endocrinol* 10:288. <https://doi.org/10.3389/fendo.2019.00288>