Harvard University

Harvard University Biostatistics Working Paper Series

Year 2016 *Paper* 210

Efficiency of Two Sample Tests via the t-Mean Survival Time for Analyzing Event Time Observations

Lu Tian[*] Haoda Fu[†] Stephen J. Ruberg[‡]

Hajime Uno[**] LJ Wei[††]

[*] Stanford University, lutian@stanford.edu

[†] Eli Lilly, haoda-fu@lilly.com

[‡] Eli Lilly, ruberg-stephen-j@lilly.com

[**] Dana-Farber Cancer Institute, huno@jimmy.harvard.edu

[††] Harvard University, wei@sdac.harvard.edu

Efficiency of Two Sample Tests via the t-Mean Survival Time for Analyzing Event Time Observations

Lu Tian, Haoda Fu, Stephen J. Ruberg, Hajime Uno, and LJ Wei

Abstract

In comparing two treatments with the event time observations, the hazard ratio (HR) estimate is routinely used to quantify the treatment difference. However, this model dependent estimate may be difficult to interpret clinically especially when the proportional hazards (PH) assumption is violated. An alternative estimation procedure for treatment efficacy based on the restricted means survival time or t-year mean survival time (t-MST) has been discussed extensively in the statistical and clinical literature. On the other hand, a statistical test 1 via the HR or its asymptotically equivalent counterpart, the logrank test, is asymptotically distribution-free. In this paper, we assess the relative efficiency of the hazard ratio and t-MST tests with respect to the statistical power using various PH and non-PH models under theoretical and practical settings. When the PH assumption is valid, the t-MST test performs almost as well as the HR test. For non-PH models, the t-MST test can substantially outperform its HR counter- part. On the other hand, the HR test can be powerful when the true difference of two survival functions is quite large at end of the study. Unfortunately, for this case, the HR estimate may not have a simple clinical interpretation for the treatment effect due to the violation of the PH assumption.

Efficiency of Two Sample Tests via the t-Mean Survival Time for Analyzing Event Time Observations

Lu Tian[1,*], Haoda Fu[2], Stephen J. Ruberg[3],
Hajime Uno[4], and LJ. Wei[5]

[1]Department of Biomedical Data Science, Stanford University
lutian@stanford.edu.
[2]Eli Lilly and Company.
haoda-fu@lilly.com
[3]Eli Lilly and Company.
ruberg-stephen-j@lilly.com
[4] Dana-Faber/Harvard Cancer Institute.
huno@jimmy.harvard.edu
[5]Department of Biostatistics, Harvard University.
wei@sdac.harvard.edu

November 30, 2016

Abstract

In comparing two treatments with the event time observations, the hazard ratio (HR) estimate is routinely used to quantify the treatment difference. However, this model dependent estimate may be difficult to interpret clinically especially when the proportional hazards (PH) assumption is violated. An alternative estimation procedure for treatment efficacy based on the restricted means survival time or t-year mean survival time (t-MST) has been discussed extensively in the statistical and clinical literature. On the other hand, a statistical test via the HR or its asymptotically equivalent counterpart, the logrank test, is asymptotically distribution-free. In this paper, we assess the relative efficiency of the hazard ratio and t-MST tests with respect to the statistical power using various PH and non-PH models under theoretical and practical settings. When the PH assumption is valid, the t-MST test performs almost as well as the HR test. For non-PH models, the t-MST test can substantially outperform its HR counterpart. On the other hand, the HR test can be powerful when the true difference of two survival functions is quite large at end of the study. Unfortunately, for this case, the HR estimate may not have a simple clinical interpretation for the treatment effect due to the violation of the PH assumption.

Keyword *Asymptotic relative efficiency; Hazard ratio; Proportional hazards model; Survival analysis; t-year mean survival time*

1 Introduction

In a randomized, comparative clinical trial with an event time as the study end point, treatment difference is often summarized by the hazard ratio (HR) by assuming a proportional hazards (PH) model, namely, the ratios of two hazard functions between two treatment groups, labeled as 1 and 0, are approximately constant over time (Cox, 1972). Under the PH assumption, the HR can be consistently estimated by maximizing a partial likelihood function from the Cox model. However, a HR of, for instance, 0.7 for treatment 1 versus treatment 0 cannot be interpreted as a 30% risk reduction in favor of treatment 1 since the hazard is not a probability measure. Moreover, owing to the lack of a simple estimator of the hazard function for treatment 0, it is not clear whether a HR of 0.7 is clinically meaningful. For a "low" hazard function for treatment group 0, a 30% reduction of hazard in treatment group 1 may not be clinically important. Furthermore, when the PH assumption is violated, the HR estimator from the Cox model converges to a parameter which is difficult to interpret (Kalbfleisch and Prentice, 1981; Lin and Wei, 1989). For this case, one often considers the estimated HR as an approximation to a weighted average of the HR's over time. Unfortunately, the weights depend on the censoring distributions. This adds another complexity of translating HR for effective clinical decision making. Other model-based summary measures for the group difference have similar issues as the HR (Wei, 1992).

There are several alternatives to summarize a survival distribution. For example, the median survival time, the t-year event rate and the t-year mean survival time, t-MST. The t-year mean survival time is also coined as the restricted mean survival time. In this paper we are interested in studying properties of using a summary measure as a group contrast based on the t-MST. The inference procedures for t-MST and the function thereof have been studied extensively, for example, by Karrison (1987); Zucker (1998); Royston and Parmar (2011); Zhao et al. (2012); Tian et al. (2014) and Uno et al. (2014). The t-MST has a clear physical and clinical interpretation. For example, an observed 2.5 year of the 3-year mean survival time means if a patient is followed up to 3 years, on average, he or she would survive 2.5 years. The t-MST can be readily estimated via the area under the corresponding Kaplan-Meier (KM) curve up to t year. Contrary to the model-dependent nature of the HR, the treatment effect can be quantified by, for example, the difference in t-MST between two treatment groups, which is purely nonparametric without requiring any model assumption. This group difference measure with a reference value of t-MST from the control arm in a comparative study is more informative than the HR.

Since the HR estimate is routinely used for making inference about the treatment effect, it would be interesting to compare it with the t-MST based inference procedure. However, since these two estimation procedures empirically quantify different parameters, it seems difficult, if not impossible, to study their relative merits from an estimation point of view. On the other hand, under the testing hypothesis paradigm, the HR based test is asymptotically distribution-free. Therefore, it is possible to compare the t-MST and HR based tests with respect to, for example, their conventional statistical power profiles. It is interesting to note that recently Trinquart et al. (2016) utilized the data from a large number of clinical trials to show empirically these two types of tests are quite concordant in terms of the conventional statistical significance interpretation. In this paper, we formally compare these two types of tests under a more theoretical setting. In general, we find that the test based on t-MST performs well compared with its HR counterpart when the PH assumption is valid. This is similar to the comparison of two treatments where the response measure is a continuous variable. In that situation, the Wilcoxon rank sum test has comparable efficiency to the t-test when the underlying data model is a normal distribution, but can be considerably more powerful when the underlying data model deviates from normality. For certain non-PH models, for instance, when the two survival functions are quite large at the end of the study followup time, the HR test is quite powerful. Unfortunately, the corresponding HR estimate is difficult to interpret clinically. With this additional finding, the robust, easily interpretable t-MST based inference procedure provides a useful alternative tool to the HR based counterpart in survival analysis. We use the data from a recent clinical trial and extensive numerical study with practical settings to illustrate our findings.

Note that there are quite a few two sample tests available in the literature to handle non-PH cases (Fleming et al., 1987; Pepe and Fleming, 1989; Kosorok and Lin, 1999; Uno et al., 2015). Unfortunately they don't necessarily have the corresponding estimation procedures available. It is ideal if we have a clinically interpretable, nonparametric estimator for the treatment effect, which can also serve as a test statistic for testing the equivalence of two survival functions.

2 The HR and t-MST based tests for the equivalence of two survival functions

Let T_1 and T_0 be the event times in treatment groups 1 and 0, respectively. The group difference in t-MST up to the time point t is defined as

$$D = E(T_1 \wedge t) - E(T_0 \wedge t),$$

which is the area between two survival curves over the time interval $[0, t]$:

$$\int_0^t \{S_1(u) - S_0(u)\} du,$$

where $S_j(\cdot)$ is the survival function of the failure time T_j in arm $j, j = 0, 1$. In order to make D identifiable based on observed data, $\pi_j(t) = \mathrm{pr}\,(T_j \wedge C_j \geq t)$ needs to be bounded below from zero, where C_j denotes the censoring time in treatment group $j, j = 0, 1$. In practice, t can be chosen, for example, as the minimum of the 95th percentiles of observed $T_j \wedge C_j$ from two treatment groups. The difference in t-MST can then be estimated by

$$\hat{D} = \int_0^t \left\{ \hat{S}_1(u) - \hat{S}_0(u) \right\} du,$$

where $\hat{S}_j(t)$ is the KM estimator for $S_j(t)$. As the sample size $n \to \infty$, $\sqrt{n}(\hat{D} - D)$ converges weakly to a mean zero Gaussian distribution whose variance can be consistently estimated by

$$\hat{\sigma}_D^2 = \int_0^t \left\{ \int_v^t \hat{S}_0(u) du \right\}^2 \frac{d\hat{\Lambda}_0(v)}{p_0 \hat{\pi}_0(v)} + \int_0^t \left\{ \int_v^t \hat{S}_1(u) du \right\}^2 \frac{d\hat{\Lambda}_1(v)}{p_1 \hat{\pi}_1(v)}, \quad (1)$$

where p_j is the proportion of patients randomized into group j, $\hat{\pi}_j(v)$ is the empirical counterpart of $\pi_j(v)$ and $\hat{\Lambda}_j(t)$ is the Nelsen-Aalan estimator for the cumulative hazard function in group $j, j = 0, 1$. Under the null hypothesis that there is no difference between two survival functions, the distribution of $\sqrt{n}\hat{D}/\hat{\sigma}_D$ is approximately the standard normal for large n.

Similarly, let $\hat{\theta}$ be the estimator of θ =log(HR) by maximizing the partial likelihood function over the time interval $[0, t]$. Under the PH model, as $n \to \infty$, $\sqrt{n}(\hat{\theta} - \theta)$ converges weakly to a mean zero Gaussian distribution , whose variance can be consistently estimated by

$$\hat{\sigma}_\theta^2 = \left[\int_0^t \frac{e^{\hat{\theta}} p_0 p_1 \hat{\pi}_0(u) \hat{\pi}_1(u)}{p_0 \hat{\pi}_0(u) + e^{\hat{\theta}} p_1 \hat{\pi}_1(u)} d\tilde{\Lambda}_0(u) \right]^{-1}, \quad (2)$$

where $\tilde{\Lambda}_0(u)$ is the Breslow estimator for the cumulative hazard function at the treatment group 0. Note that similar to t-MST, in theory the large sample normal approximation to the distribution of HR estimator is only valid with observations within a finite time interval, say $[0, t]$, where $\pi_0(t)\pi_1(t) > 0$ (p289-290, Chapter 8, Fleming and Harrington (2011)). There is a misconception that one may use all the observed and censored data to make inference about the treatment effect via the HR estimate. Under the null hypothesis, $\theta = 0$ and the distribution of $\sqrt{n}\hat{\theta}/\hat{\sigma}_\theta$ is approximately standard normal for large n. Therefore, tests based on both $\hat{D}$ and $\hat{\theta}$ are asymptotical valid in retaining the appropriate type I error rate.

3 Asymptotical efficiency of the test based on $\hat{D}$ and $\hat{\theta}$

In this section, we derive the asymptotic efficiencies of the tests based on $\hat{D}$ and $\hat{\theta}$ under a sequence of contiguous alternatives (Hoeffding and Rosenblatt, 1955; Hodges and Lehmann, 1956). Specifically, for the current case, these alternatives are defined as

$$\log\left\{\frac{\lambda_{1n}(u)}{\lambda_0(u)}\right\} = \frac{\alpha(u)}{\sqrt{n}},$$

where $\lambda_0(t)$ and $\lambda_{1n}(t)$ are the hazard functions of the failure time T_0 from the treatment 0 and T_{1n} from the treatment 1, whose distribution depends on the sample size n, respectively (Lagakos, 1988; Slud, 1991). Here, $\alpha(\cdot)$ is a deterministic function over time providing a specific alternative hypothesis of interest. For any given $\alpha(\cdot)$, we can derive the asymptotic efficiency of the test under the mild regularity conditions given in p230-232, Chapter 6, Fleming and Harrington (2011).

For study size n, it follows from similar arguments in Schoenfeld (1981) and Harrington and Fleming (1982) that under the above contiguous alternatives $\sqrt{n}\hat{D}$ is asymptotic normal with mean

$$\int_0^t \left\{\int_0^v \alpha(u)\lambda_0(u)du\right\} S_0(v)dv$$

and variance

$$\sigma_D^2 = \int_0^t \left\{\int_v^t S_0(u)du\right\}^2 \frac{\lambda_0(v)}{w(v)}dv, \tag{3}$$

which is the limit of $\hat{\sigma}_D^2$ defined in (1), where

$$w(v) = \frac{p_0 p_1 \pi_0(v)\pi_1(v)}{p_0\pi_0(v) + p_1\pi_1(v)}.$$

Thus, the asymptotic efficiency of the test based $\hat{D}$ is

$$\xi_1 = \frac{\left[\int_0^t \{\int_v^t S_0(u)du\}\alpha(v)\lambda_0(v)dv\right]^2}{\int_0^t \left\{\int_v^t S_0(u)du\right\}^2 w(v)^{-1}\lambda_0(v)dv}.$$

Similarly, it follows from Lin and Wei (1989) that under above contiguous alternatives the test statistic $\sqrt{n}\hat{\theta}$ is also asymptotic normal with mean

$$\frac{\int_0^t w(u)\alpha(u)\lambda_0(u)du}{\int_0^t w(u)\lambda_0(u)du}$$

and variance

$$\sigma_\theta^2 = \left(\int_0^t w(u)\lambda_0(u)dt \right)^{-1}, \tag{4}$$

which is the limit of $\hat{\sigma}_\theta^2$ in (2). The asymptotic efficiency for this test is

$$\xi_2 = \frac{\left[\int_0^t w(u)\alpha(u)\lambda_0(u)du \right]^2}{\int_0^t w(u)\lambda_0(u)du}.$$

It follows that the asymptotical relative efficiency (ARE) of the t-MST based test relative to the HR-based test is ξ_1/ξ_2, which can be interpreted approximately as the inverse of the ratio of the sample sizes needed for two tests to have the same power to detect the alternative $\lambda_{1n}(u) = \exp\{\alpha(u)/\sqrt{n}\}\lambda_0(u)$.

Note that the standard logrank test, a score test based on the Cox partial likelihood function for testing the equivalence of two survival functions, is asymptotically equivalent to the Wald-type of test based on the HR and therefore ξ_1/ξ_2 is also the ARE of the t-MST based test relative to the logrank test.

For comparing two tests, the ARE may not be informative since there is no reference value to interpret the ratio. Since ARE is approximately a ratio of the sample size of two tests to have similar power for testing the same alternative hypothesis, this connection provides a clear practical interpretation of ARE (Schoenfeld and Richter, 1982; Zhang and Quan, 2009). Specifically, assuming that the alternative hypothesis consists of two fixed unequal hazard functions, say, $\lambda_0(u)$ and $\lambda_1(u)$, the sample size needed for the t-MST based test can be approximated by $(z_{1-\alpha/2} + z_{1-\beta})^2/\xi_1$, if this alternative is "close" to the null hypothesis, where z_q is qth quantile of the standard normal and α and β are the Type I and II error rates, respectively. More generally, a more accurate sample size estimator for $100(1-\beta)\%$ power at the two-sided significance level of α is

$$(z_{1-\alpha/2} + z_{1-\beta})^2 \frac{\int_0^t \left[\left\{ \int_v^t S_0(u)du \right\}^2 \frac{\lambda_0(v)}{p_0\pi_0(v)} + \left\{ \int_v^t S_1(u)du \right\}^2 \frac{\lambda_1(v)}{p_1\pi_1(v)} \right] dv}{\left[\int_0^t \{S_1(u) - S_0(u)\}du \right]^2}, \tag{5}$$

for any pair of $\{\lambda_0(u), \lambda_1(u)\}$. Similarly, the corresponding sample size estimate analogy for HR based test is

$$(z_{1-\alpha/2}+z_{1-\beta})^2 \frac{\int_0^t w(u)^2 \left\{ \frac{\lambda_0(u)}{p_0\pi_0(u)} + \frac{\lambda_1(u)}{p_1\pi_1(u)} \right\} du}{\left[\int_0^t w(u) \left\{ \lambda_1(u) - \lambda_0(u) \right\} du \right]^2}. \tag{6}$$

When $\lambda_1(u)$ is close to $\lambda_0(u)$, the ratio of the two sample size estimates is approximately equal to the inverse of ARE. The ARE, coupled with these two sample size estimates, may provide a meaningful comparison of the two tests. However, when $\lambda_1(u)$ is quite different from $\lambda_0(u)$, ARE may not be a good approximation to the ratio of the sample sizes needed for two tests of interest. In such a case, one may consider the empirical relative efficiency (ERE), E_D^2/E_θ^2, where

$$E_D = \frac{E(\hat{D})}{se(\hat{D})} \text{ and } E_\theta = \frac{E(\hat{\theta})}{se(\hat{\theta})}$$

are the effect sizes standardized by their standard errors for tests based on $\hat{D}$ and $\hat{\theta}$, respectively. Unlike ARE, ERE often does not have a closed-form expression and can only be approximated by numerical simulations for given sample sizes. Although computing ERE is in general more difficult than ARE, it can approximate the ratio of (5) and (6) fairly well and be used to evaluate the relative merits of two tests under any alternative. An example is given in the next section to illustrate this point.

4 Example and Numerical Study

In this section, we first use a study recently conducted by the ECOG-ACRIN Cancer Research Group to compare low- and high-dose dexamethasone for treating newly diagnosed multiple myeloma (Rajkumar et al., 2010) to illustrate how to interpret the ARE (t-MST vs. HR based tests). This study randomized 222 patients to the low-dose group and 223 patients to the high-dose group. Figure 2a shows the resulting KM curves of overall survival by treatment groups. Visually it seems that the patients in the low-dose group survived longer than those in the high dose group. The p-values from the tests based on HR and t-MST are 0.47 and 0.04, respectively. The discrepancy is likely due to the presence of crossing hazards. In this study, the nonparametric KM curves are fairly similar to their parametric counterparts based on a Weibull model (Figure 1b). Now, suppose that we are interested in designing a new study using the results from this ECOG-ACRIN study. That is, assume that the observed pattern of the two observed hazard functions from this study is the alternative hypothesis for a new study. The question is how the t-MST test would perform compared with the HR based test with respect to the ARE and how to interpret this ratio measuring the relative merit of these two tests. To this end, we assume that the underlying hazard functions follow Weibull distributions as shown in Figure 1b. Furthermore, we assume this new study would have similar patient's accrual and follow up time patterns, that is the entire study duration is about 43 months with a 19-month enrollment period. Furthermore, we let the time of loss of follow-up follows an exponential distribution with an annual drop-off rate of 1% for both arms without accounting for the administrative censoring due to the staggered entry of study patients. We also let $t = 40$ months for defining the follow-up period of interest for t-MST and HR. Under these assumptions, the required sample size for 80% power at the significance level of 0.05 is $n_\theta = 2392$ per arm for the HR-based test and $n_D = 564$ per arm for the t-MST based test. The ARE is 2.09, which also strongly favors t-MST based test but quantitatively different n_θ/n_D. In this case, we also can use simulation to directly estimate the ERE, the finite sample analogy of ARE. To this end, we simulate 5000 sets of data consisting of 1000 patients each under the assumed alternatives and obtain the corresponding $\hat{D}$ and $\hat{\theta}$ from each simulated data set. We then approximate the expectation and variance of $\hat{D}$ (and $\hat{\theta}$) by their empirical counterparts. The resulting ERE is 4.14, which is pretty close to n_θ/n_D. It suggests that the "alternative" of interest is not adequately close to null and ARE may not accurately reflect the ratio of the needed sample sizes of the two tests. With the above sample size estimates, one still can assess the practical gain from the t-MST test over the HR test. We have performed the additional simulation studies to empirically estimate the true power of HR and t-MST based tests with the estimated sample sizes. The empirical powers based on 5000 simulations are 79.5% and 80.5% for t-MST and HR based tests, respectively, which confirms the adequacy of the estimated sample sizes based on (5) and (6).

Now, we use the above setting, but modify the underlying Weibull distribution of the low dose group to follow the PH assumption with a true HR of 0.70 against the high-dose arm. The survival curves of this alternative are plotted in Figure 1c. With this specific PH alternative, the required sample sizes are $n_D = 641$ for t-MST based test and $n_\theta = 593$ for HR based test.

Both ARE and ERE are 0.92, which is pretty close to n_θ/n_D. This example suggests that even when the PH assumption is valid, the t-MST test is essentially as powerful as the HR test.

We further explore extensively the relative merits between the two tests under various scenarios. Firstly, we consider the case where the PH assumption holds true, i.e., $\alpha(u) = 1$. For this case, the ARE is

$$\frac{\left\{\int_0^t \{\int_v^t S_0(u)du\}\lambda_0(v)dv\right\}^2}{\int_0^t \left\{\int_v^t S_0(u)du\right\}^2 w(v)^{-1}\lambda_0(v)dv \times \int_0^t w(u)\lambda_0(u)du},$$

which is always ≤ 1 by Cauchy inequality, suggesting that the HR-based test is more powerful under the PH assumption. To quantify the efficiency loss of the t-MST based test under practical settings, we assume that the study has a recruitment period of 19 months and additional follow-up of 24 months after the last patient entered the trial as in the aforementioned ECOG-ACRIN study. Specifically, the censoring time is assumed to be the minimum of $C_L \sim \text{EXP}(\lambda_c)$ and $C_A \sim U(\tau_L, \tau_U)$, reflecting the loss of follow-up and administrative censoring due to staggered entry, respectively. It follows that the survival function for the censoring time is

$$S_C(u) = e^{-\lambda_c u}\frac{\{\tau_U - \max(u, \tau_L)\}}{(\tau_U - \tau_L)} I(u \leq \tau_U),$$

where $[\tau_L, \tau_U] = [24, 43]$. Here, we let $t = 40$ months. Furthermore, we set $\lambda_c = 0.06\%$ matching the annual loss of follow-up rate observed in the ECOG-ACRIN study and $\lambda_c = 2.5\%$, an annual loss of follow-up rate of 26%, to represent heavier censoring caused by, for example, drop-out. Lastly, we also assume that the survival time in the treatment group 0 follows a Weibull distribution with the shape parameter being the maximum likelihood estimator in the high-dose group of the ECOG-ACRIN study, that is, having a decreasing hazard function $\lambda_0(u) = a_0 u^{-0.174}$. With this setup, ARE is determined by a_0, the scale parameter of the Weibull distribution in the treatment group 0. We adjusted the scale parameter so that $S_0(t) = 0.1, \cdots, 0.9$. The detailed results on ARE are summarized in Table 1. Under the PH assumption, the HR-based test is slightly more powerful as anticipated. Furthermore, when the cumulative event rate is relatively high, for example, $\text{pr}(T_0 < t) \geq 0.5$, as for studies of serious diseases with high event rates, the efficiency loss of t-MST based test relative to the HR based optimal test is almost negligible.

Table 1: Asymptotical relative efficiency ARE) and empirical relative efficiency (ERE) under PH alternatives with a HR of 0.7; EREs are estimated based on 5000 sets simulated data.

	ARE(ERE)	
Censoring	light	heavy
$S_0(t) = 0.90$	0.90(0.92)	0.99(1.02)
$S_0(t) = 0.80$	0.91(0.93)	0.99(1.01)
$S_0(t) = 0.70$	0.92(0.94)	0.99(1.01)
$S_0(t) = 0.60$	0.93(0.92)	1.00(1.02)
$S_0(t) = 0.50$	0.94(0.95)	1.00(1.02)
$S_0(t) = 0.40$	0.95(0.96)	1.00(1.02)
$S_0(t) = 0.30$	0.97(0.99)	1.00(1.02)
$S_0(t) = 0.20$	0.98(1.00)	0.99(1.01)
$S_0(t) = 0.10$	0.99(1.01)	0.98(1.01)

To confirm these theoretical AREs, we have also performed numerical simulations to obtain the ERE by estimating E_D and E_R based on results from 5000 simulated data sets. The true HR for two treatment groups (treatment group 1 vs. 0) is 0.70. The resulting EREs are all very close to their asymptotical counterparts (Table 1). For example, when $\lambda_c = 0.06\%$ and $S_0(t) = 0.50$, the ARE and ERE are 0.94 and 0.95, respectively.

Next, we consider three non-PH settings: (i) HR is monotone increasing from < 1 at time 0 to > 1 at time t; (ii) HR is monotone increasing from < 1 at time 0 to 1 at time t; and (iii) HR is monotone decreasing from 1 at time 0 to < 1 at time t. The first setting corresponds to the worst violation of the PH assumption: crossing hazards. In this case, we let $\alpha(u) = \log(0.46) + (u/t)^s$, i.e., the HR is 0.46 at time zero, crosses 1 and eventually increases to 1.25 at time t. In this case, the parameter s dictates how fast the HR increases with time and ARE. Although HR crosses one, two survival functions here don't cross within the interval $[0, t]$. The second setting corresponds to the case where the treatment benefit is large at time 0 but gradually diminishes. To characterize this pattern, we let $\alpha(u) = -c(t-u)^s$, where the constant c is chosen such that the average HR, $t^{-1}\int_0^t \exp\{\alpha(u)\}du$, is approximately 0.7. The last setting corresponds to the case, where the treatment benefit is small at the beginning but grows gradually during the follow-up. Specifically, we let $\alpha(u) = -cu^s$, where c is a constant such that the average HR is approximately 0.7. For all three settings, both the analytic ARE and ERE based on 5000 simulations are evaluated for all combinations of $s \in \{0.5, 1, 2\}$, $\lambda_c \in \{0.06\%, 2.5\%\}$ and survival rates $S_0(t) \in \{0.30, 0.50, 0.70\}$ and reported in Table 2.

When HR is monotone increasing and crosses one during the follow-up (Case i), the t-MST based test can be substantially more efficient than the HR-based test; when the HR is less than one at time 0 and increases to 1 at the end the study (Case ii), the t-MST based test is also more efficient than the HR-based test but the difference is modest; and when the HR is 1 at time 0 and decreases over time (Case iii), the t-MST based test can be less efficient than the HR-based test. Similar to the case where the PH assumption holds, the EREs are in general consistent with their asymptotical counterparts in all settings, supporting the use of AREs for evaluating the finite sample performances of two tests.

It is interesting to note that under the contiguous alternative, the t-MST based test is asymptotically equivalent to the test based on

$$\int_0^t \frac{\int_v^t S_0(u)du}{\int_0^t S_0(u)du} d\left\{\hat{\Lambda}_1(v) - \hat{\Lambda}_0(v)\right\},$$

while the HR-based test is equivalent to the test based on

$$\int_0^t S_C(v)S_0(v)d\left\{\hat{\Lambda}_1(v) - \hat{\Lambda}_0(v)\right\}.$$

The difference between these two tests thus only depends on the choice of weight function, which may explain the results of the above numerical studies. Here we assume that the censoring distributions at two groups are identical and thus $w(v) \propto S_C(v)S_0(v)$. To be specific, let

$$w_D(v) = \frac{\int_v^t S_0(u)du}{\int_0^t S_0(v)dv} \quad \text{and} \quad w_\theta(v) = S_C(v)S_0(v)$$

be the weight functions of the t-MST and HR based tests, respectively. $w_D(v)$ is independent of the censoring distribution and monotone decreasing from 1 to 0, while $w_\theta(v)$ depends on the censoring distribution and monotone decreasing from 1 to $S_C(t)S_0(t) > 0$. In Figure 2, we plot the weight functions for both tests with following survival functions for the censoring distribution

$$S_C(v) = e^{-\lambda_c v}\frac{\{\tau_u - \max(v, \tau_l)\}}{(\tau_u - \tau_l)}I(v \leq \tau_u),$$

Table 2: Asymptotical relative efficiency (ARE) and empirical relative efficiency (ERE) under nonproportional hazards alternatives; EREs are estimated based on 5000 sets simulated data.

	$S_0(t) = 0.30$		$S_0(t) = 0.50$		$S_0(t) = 0.70$	
	ARE(ERE)		ARE(ERE)		ARE(ERE)	
Censoring	light	heavy	light	heavy	light	heavy
			$\alpha(u) = \log(0.46) + (u/t)^s$			
$s = 0.5$	1.99(1.81)	1.13(1.10)	2.48(2.17)	1.24(1.19)	2.98(2.57)	1.36(1.29)
$s = 1.0$	1.40(1.32)	1.09(1.05)	1.50(1.41)	1.13(1.09)	1.59(1.49)	1.17(1.14)
$s = 2.0$	1.20(1.14)	1.08(1.02)	1.24(1.18)	1.09(1.05)	1.29(1.21)	1.12(1.04)
			$\alpha(u) = -c(\alpha)(t-u)^s$			
$s = 0.5$	1.07(1.05)	1.05(1.01)	1.09(1.06)	1.06(1.03)	1.09(1.06)	1.06(1.04)
$s = 1.0$	1.19(1.15)	1.07(1.03)	1.22(1.18)	1.09(1.05)	1.25(1.20)	1.10(1.07)
$s = 2.0$	1.38(1.32)	1.10(1.05)	1.45(1.39)	1.13(1.09)	1.50(1.44)	1.17(1.13)
			$\alpha(u) = -c(\alpha)u^s$			
$s = 0.5$	0.81(0.80)	0.99(0.97)	0.78(0.77)	0.97(0.95)	0.76(0.75)	0.96(0.93)
$s = 1.0$	0.72(0.71)	0.97(0.95)	0.69(0.68)	0.94(0.92)	0.66(0.65)	0.90(0.89)
$s = 2.0$	0.60(0.59)	0.95(0.93)	0.57(0.55)	0.89(0.87)	0.54(0.52)	0.84(0.82)

where $\lambda_c = 0.06\%$ or 2.5% and $[\tau_l, \tau_u] = [24, 43]$. We also assume that the survival function of the failure time is $S_0(v) = \exp(-0.016v^{0.826})$, i.e., $S_0(40) \approx 0.7$. It is clear that $w_D(v)$ assigns less weights to later difference in hazard function compared with $w_\theta(v)$ when the censoring is light. It suggests that when the treatment difference is anticipated at later study time points, the t-MST based test tends to be less powerful than the HR based test. On the other hand, $w_D(v)$ is very similar to $w_\theta(v)$, when the drop-out rate is high, which explains that AREs and EREs of the t-MST based test compared to HR based test are fairly close to one in the heavy censoring settings of Table 2.

5 Remarks

In this paper, we compared the inference procedures based on HR and the t-MST difference under a hypothesis testing paradigm. Through an extensive numerical study with various treatment difference profiles, the AREs for the t-MST and HR tests are similar under the PH-models. For the non-PH models, unless the true difference of two survival functions is quite large near the end of the study and the censoring is light during the study follow-up, the t-MST test generally outperforms the HR test. Referring again to the two sample test for continuous data, there are approaches for measuring the degree of deviation from normality in order to construct highly efficient tests for shift in location relative to the t-test (Ruberg, 1986). Perhaps a line of future research is to measure deviations from the PH model and construct a family of optimal tests that account for the degree of deviation from PH. Although it is important to obtain a p-value to assess statistical significance for a comparative study, estimation of the treatment effect is more informative for clinical decision makings. To make coherent evaluation on the group difference, we suggest using the estimation procedure to conduct hypothesis testing. There are quite a few novel statistical tests for non-proportional hazards alternatives proposed and discussed in the literature (Fleming et al., 1987; Pepe and Fleming, 1989; Kosorok and Lin, 1999; Uno et al., 2015). However, there are no coherent corresponding estimation procedures to quantify the treatment difference. A more powerful test than the HR test for a non-PH case without a companion estimation procedure would have limited value clinically.

There are very few two-sample model-free estimating procedures for treatment effect, which can also be used for testing the equivalence of two survival functions in survival analysis. For example, one may use the difference or ratio of two t-year event rates, median survival times, and t-MSTs. Often the median event time is not estimable due to a short study followup time. Moreover, the standard error of the estimated difference of two median event times can be quite large. Using the t-year event rate difference as the primary parameter of interest would ignore the temporal profile of the treatment effect before time point t. The estimate for the t-MST difference or ratio appears to be a useful tool for analyzing event time observations.

References

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statisitcal Society-B*.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*, volume 169. John Wiley & Sons.

Fleming, T. R., Harrington, D. P., and O'sullivan, M. (1987). Supremum versions of the log-rank and generalized wilcoxon statistics. *Journal of the American Statistical Association*, 82(397):312–320.

Harrington, D. and Fleming, T. (1982). A class of rank test procedures for censored survival data. *Biometrika*.

Hodges, J. and Lehmann, E. (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical statistics*, 27:324–335.

Hoeffding, W. and Rosenblatt, J. (1955). The efficiency of tests. *Annals of Mathematical Statistics*, 26:52–63.

Kalbfleisch, J. and Prentice, R. (1981). Estimation of the average hazard ratio. *Biometrika*, 68:105–112.

Karrison, T. (1987). Restricted mean life with adjustment for covariates. *Journal of American Statistical Association*, pages 1169–1176.

Kosorok, M. R. and Lin, C.-Y. (1999). The versatility of function-indexed weighted log-rank statistics. *Journal of the American Statistical Association*, 94(445):320–332.

Lagakos, S. (1988). The loss in efficiency from misspecifying covariates in proportional hazards regression models. *Biometrika*, 75:156–160.

Lin, D. and Wei, L. (1989). The robust inference for the Cox proportional hazards model. *Journal of American Statistical Association*, 84:1074–1078.

Pepe, M. S. and Fleming, T. R. (1989). Weighted kaplan-meier statistics: a class of distance tests for censored survival data. *Biometrics*, pages 497–507.

Rajkumar, S., Jacobus, S., Callander, N., Fonseca, R., Vesole, D., Williams, M., Abonour, R., Siegel, D., Katz, M., Greipp, P., and Eastern Cooperative Oncology Group (2010). Lenalidomide plus high-dose dexamethasone versus lenalidomide plus low-dose dexamethasone as initial therapy for newly diagnosed multiple myeloma: an open-label randomised controlled trial. *Lancet Oncology*, 11:29–37.

Royston, P. and Parmar, M. (2011). The use of restricted mean survival time to estimate the treatment effect in randomzed clincial trials when the proportional hazards assumption is in doubt. *Journal of American Statistical Association*, 30:2409–2421.

Ruberg, S. J. (1986). Efficiencies of some two–sample location tests for a broad class of distributions. *Communications in Statistics-Theory and Methods*, 15(10):2991–3004.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*.

Schoenfeld, D. and Richter, J. (1982). Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics*, 38:163–170.

Slud, E. (1991). Relative efficiency of the logrank test within a multiplicative intensity model. *Biometrika*, 18:1172–1187.

Tian, L., Zhao, L., and We, LJ. (2014). Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, 15:222–233.

Trinquart, L., Jacot, J., Conner, S. C., and Porcher, R. (2016). Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*, page JCO642488.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M., and Wei, LJ. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32:2380–2385.

Uno, H., Tian, L., Claggett, B., and Wei, L. (2015). A versatile test for equality of two survival functions based on weighted differences of kaplan–meier curves. *Statistics in medicine*, 34(28):3680–3695.

Wei, L. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*.

Zhang, D. and Quan, H. (2009). Power and sample size calculation for log-rank test with a time lag in treatment effect. *Statistics in Medicine*, 28:864–879.

Zhao, L., Tian, L., Uno, H., Solomon, S., Pfeffer, M., Schindler, J., and Wei, LJ. (2012). Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical Trials*, 9:570–577.

Zucker, D. (1998). Restricted mean life with covariates: Modification and extension of a useful survival analysis method. *Journal of American Statistical Association*, 93:702–709.

Figure 1: (a) The KM curves in two arms of the ECOG-ACRIN study; (b) The survival curves based on the Weibull model in two arms of the ECOG-ACRIN study without assuming the PH assumption; (c) The survival curves based on the Weibull model in the high dose arm of the ECOG-ACRIN study assuming the PH assumption with a HR of 0.7; solid: high dose arm; dotted: low dose arm.

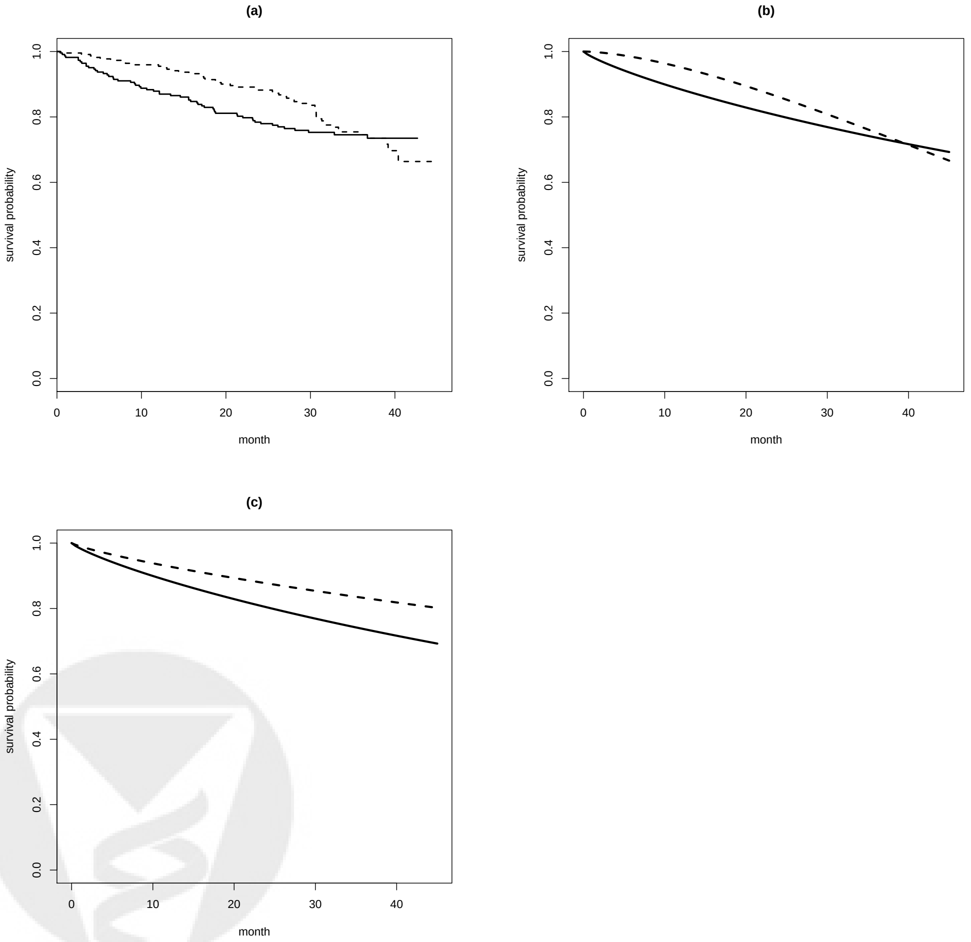
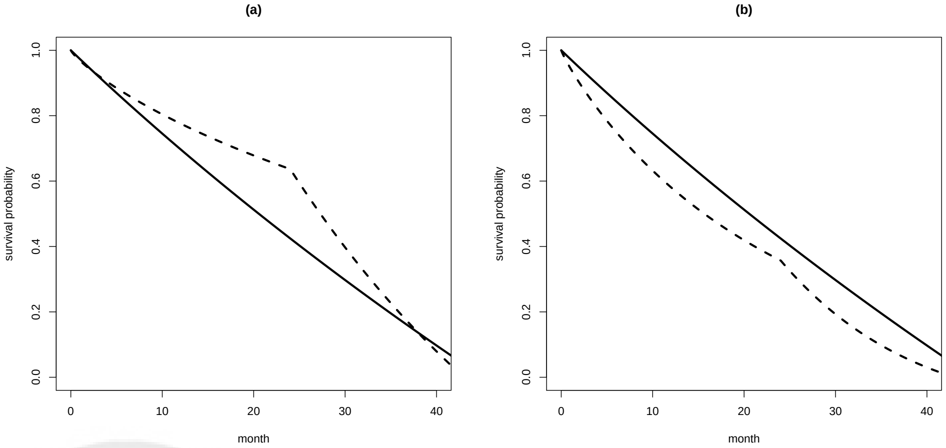


Figure 2: The weight functions for t-MST and HR based tests; solid: weight function for t-MST based test; dotted: weight function for HR based test; (a) light drop-out rate with $\lambda_c = 0.06\%$ and (b) high drop-out rate with $\lambda_c = 2.5\%$


(a)
(b)
survival probability
month