Harvard University

Harvard University Biostatistics Working Paper Series

Year 2013 *Paper* 159

A versatile test for equality of two survival functions based on weighted differences of Kaplan-Meier curves

Hajime Uno[*] Lu Tian[†]

Brian Claggett[‡] L. J. Wei[**]

[*]Dana Farber Cancer Institute and Harvard University

[†]Stanford University School of Medicine

[‡]Harvard University

[**]Harvard University, wei@hsph.harvard.edu

A versatile test for equality of two survival functions based on weighted differences of Kaplan-Meier curves

Hajime Uno, Lu Tian, Brian Claggett, and L. J. Wei

Abstract

With censored event time observations, the logrank test is the most popular tool for testing the equality of two underlying survival distributions. Although this test is asymptotically distribution-free, it may not be powerful when the proportional hazards assumption is violated. Various other novel testing procedures have been proposed, which generally are derived by assuming a class of specific alternative hypotheses with respect to the hazard functions. The test considered by Pepe and Fleming (1989) is based on a linear combination of weighted differences of two Kaplan-Meier curves over time and is a natural tool to assess the difference of two survival functions directly. In this article, we take a similar approach, but choose weights which are proportional to the observed standardized difference of the estimated survival curves at each time point. The new proposal automatically makes weighting adjustments empirically. The new test statistic is aimed at a one-sided general alternative hypothesis, and is distributed with a short right tail under the null hypothesis, but with a heavy tail under the alternative. The results from extensive numerical studies demonstrate that the new procedure performs well under various general alternatives. The survival data from a recent cancer comparative study are utilized for illustrating the implementation of the process.

A versatile test for equality of two survival functions based on weighted differences of Kaplan-Meier curves

Hajime Uno[*], Lu Tian[†], Brian Claggett[‡], L. J. Wei[§]

May 24, 2013

Abstract

With censored event time observations, the logrank test is the most popular tool for testing the equality of two underlying survival distributions. Although this test is asymptotically distribution-free, it may not be powerful when the proportional hazards assumption is violated. Various other novel testing procedures have been proposed, which generally are derived by assuming a class of specific alternative hypotheses with respect to the hazard functions. The test considered by Pepe and Fleming (1989) is based on a linear combination of weighted differences of two Kaplan-Meier curves over time and is a natural tool to assess the difference of two survival functions directly. In this article, we take a similar approach, but choose weights which are proportional to the observed standardized difference of the estimated survival curves at each time point. The new proposal automatically makes weighting adjustments empirically. The new test statistic is aimed at a one-sided general alternative hypothesis, and is distributed with a short right tail under the null hypothesis, but with a heavy tail under the alternative. The results from extensive numerical studies demonstrate that the new procedure performs well under various general alternatives. The survival data from a recent cancer comparative study are utilized for illustrating the implementation of the process.

Key words: Logrank test, Perturbation resampling method, Proportional hazards, Robust tests.

[*] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, U.S.A.

[†] Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California 94305, U.S.A.

[‡] Department of Biostatistics, Harvard University, Boston, MA 02115, U.S.A.

[§] Department of Biostatistics, Harvard University, Boston, MA 02115, U.S.A.

1 Introduction

In summarizing the comparisons of two survival distributions with censored event time observations, it is conventional to provide a plot of two Kaplan-Meier (KM) curves along with a p-value from the two-sample logrank test. Note that the logrank test statistic reflects the difference of two underlying hazard functions, not of the KM curves directly (Pepe and Fleming, 1989). Asymptotically, the logrank test is valid nonparametrically, but may perform rather poorly when the proportional hazards (PH) assumption does not hold (Tarone and Ware, 1977; Lagakos and Schoenfeld, 1984). As an example, in a randomized clinical trial (E4A03) recently conducted by the Eastern Cooperative Oncology Group, low-dose and high-dose dexamethasone for treating newly diagnosed multiple myeloma patients were compared with respect to the patient's overall survival (Rajkumar et al., 2010). Of a total of 445 enrolled patients, 222 were assigned to the low-dose and 223 to the high-dose group. This trial was terminated at the first interim analysis conducted in March 2007 as a result of the superior performance of the low-dose group with respect to overall survival. Twenty eight patients who were still receiving high-dose treatment were then switched to low-dose treatment, and the trial continued to further study the patients' long term survival profiles. Figure 1 presents the KM curves of overall survival, based on the data as of November 2009, for the two dose groups. The two curves are markedly separated before 30 months of follow-up, but then appear to cross near the end of the study. Since there were relatively few patients in the original high-dose group switching to the low dose after the interim analysis, the differential patterns of the KM curves is intriguing. Visually, it appears that the low-dose does have a short-term survival advantage over the high dose. However, the two-sided p-value from the logrank test is 0.46, and the p-value from the Peto-Prentice-Wilcoxon test is 0.28. Neither test gives strong evidence that the low-dose group is better than the high-dose group under the intent-to-treat principle. In this example, the PH assumption is likely violated, and the logrank test may not be powerful for detecting the difference of the survival functions.

To avoid a "fishing expedition" in which tests of equality of the survival curves are selected ad hoc, a pre-specified test needs to be well described in the study protocol or the statistical analysis plan before unblinding the data. Unfortunately, in general, little or no information is available regarding the profile of the potential difference between two survival curves at the design stage. To this end, various flexible test procedures have been proposed. For example, the $G^{\rho,\gamma}$ tests, which are constructed assuming a class of survival distributions indexed by ρ and γ under the alternative hypothesis (Fleming and Harrington, 1991), a linear combination or the maximum of several test statistics (Tarone, 1981; Fleming and Harrington, 1984, 1991; Gastwirth, 1985; Zucker and Lakatos, 1990; Self, 1991; Lee, 1996; Kosorok and Lin, 1999), and other novel robust procedures (Lai and Ying, 1991; Pecková and Fleming, 2003) have been extensively studied. The pros and cons of those procedures were discussed in a recent paper by Yang and Prentice (2010). Note that the above novel alternatives to the logrank test are more or less built with respect to a family of pre-specified survival functions. The test recently proposed by Yang and Prentice (2010) is a weighted logrank test whose weights are obtained by fitting the data to the model proposed by Yang and Prentice (2005), which includes the proportional hazards and the proportional odds models as special cases.

Note that a class of novel tests based on the weighted Kaplan-Meier (WKM) statistics proposed by Pepe and Fleming (1989; 1991) did not get much attention in practice. Specifically, let $\hat{S}_1(\cdot)$ and $\hat{S}_2(\cdot)$ be the KM estimators for the two groups to be compared. A WKM test statistic is

$$WKM = \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} \int_0^\tau \hat{W}(t)\hat{D}(t)dt, \tag{1.1}$$

where $\hat{D}(t) = \hat{S}_2(t) - \hat{S}_1(t)$, $\tau = \sup\left[t : \min\left\{\hat{K}_1(t), \hat{K}_2(t)\right\} > 0\right]$, $\hat{K}_i(\cdot)$ denotes the left-continuous version of the KM estimator for the censoring survival function for the ith group, n_i is the sample size in group i, $(i = 1, 2)$, and $\hat{W}(\cdot)$ is the data-dependent weight function. Since we conventionally present the KM curves as in Figure 1 to show the temporal profile of the group difference, it seems natural to provide a test which directly compares two survival functions, rather than their hazard functions. For the class of test statistics in (1.1), Pepe and Fleming (1989) considered two weighting schemes:

$$\frac{\hat{K}_1(t)\hat{K}_2(t)}{\hat{q}_1\hat{K}_1(t) + \hat{q}_2\hat{K}_2(t)}, \tag{1.2}$$

and

$$\left\{ \frac{\hat{K}_1(t)\hat{K}_2(t)}{\hat{q}_1\hat{K}_1(t) + \hat{q}_2\hat{K}_2(t)} \right\}^{1/2}, \tag{1.3}$$

where $\hat{q}_i$ is the proportion of subjects assigned to group i. Note that their weighting schemes only depend on the censoring distributions.

In this article, we also consider tests similar to (1.1) for testing against a one-sided alternative hypothesis, i.e., one survival curve is greater than the other for a time interval of the follow-up. However, in choosing $\hat{W}(\cdot)$, instead of considering only the observed censoring distribution, we propose to put more weight at the time points t such that the difference $\hat{D}(t)$ is "large". One possible choice is to let $\hat{W}(t) = \hat{D}(t)$; the resulting test statistic, however, is like a "chi-square" statistic and tends to have a rather long right tail under the null hypothesis. It follows that this omnibus test may not be powerful for certain alternatives. In the next section, we propose a simple weighting scheme whose weight at time t depends on $\hat{D}(t)$. Under the null, the distribution of the test statistic has a relatively short tail, but under a general one-sided alternative, the observed test statistic tends to be large and likely to reject the null hypothesis. The new test statistic is not derived from any pre-specified class of distributions like most existing test procedures in the literature. Instead, it automatically chooses the weights adaptively based on the size of $\hat{D}(\cdot)$ or a function thereof, to effectively differentiate the null and alternative hypotheses. For the above cancer study survival data, the resulting one-sided p-value of the proposed test is 0.005. The details of implementing the new test are given in Section 2. We also conducted extensive numerical studies to assess the performance of the proposed procedure.

2 Combining weighted differences of Kaplan-Meier curves

Let $S_1(\cdot)$ and $S_2(\cdot)$ be two survival functions for failure times T_1 and T_2, respectively. Let C_1 and C_2 be the corresponding censoring times. Also, let $\{(T_{1j}, C_{1j}),\ j = 1, \ldots, n_1\}$ and $\{(T_{2j}, C_{2j}),\ j = 1, \ldots, n_2\}$ be independent random copies from (T_1, C_1) and (T_2, C_2), respectively. Due to censoring, one can only observe $\{(X_{ij}, \Delta_{ij});\ i = 1, 2,\ j = 1, \ldots, n_i\}$, where $X_{ij} = \min(T_{ij}, C_{ij})$ and Δ_{ij} is equal to 1 if $T_{ij} \leq C_{ij}$ and 0 otherwise. Let $D(\cdot) = S_2(\cdot) - S_1(\cdot)$. Let $[0, \zeta]$ be a given time interval, and we assume that $\Pr(X_i > \zeta) > 0,\ i = 1, 2$. We are interested in testing the null hypothesis that $D(t) = 0$, for $t \in [0, \zeta]$, against a general one-sided alternative, that is, $D(\cdot) \geq 0$ with at least one $t \in [0, \zeta]$, such that $D(t) > 0$. Now, let $\hat{D}(\cdot)$ be equal to $\hat{S}_2(\cdot) - \hat{S}_1(\cdot)$, $\hat{\sigma}(\cdot)$ be its standard error estimate, and $Z(\cdot) = \hat{D}(\cdot)/\hat{\sigma}(\cdot)$, which is distributed approximately $N(0, 1)$ under the null hypothesis.

Instead of utilizing $Z(t)$ as a test statistic at a specific time point t for testing the null hypothesis, we consider a test statistic which is a weighted integration of standardized differences between two survival curves over $[0, \zeta]$. For example, one potential class of test statistics is

$$V = \int_0^\zeta \hat{W}(t) Z(t) dt, \tag{2.1}$$

where $\hat{W}(\cdot)$ is a data-dependent weight function. Note that (2.1) is slightly different from (1.1). We replace $\hat{D}(\cdot)$ by $Z(\cdot)$ due to the fact that we are primarily interested in hypothesis testing. Note that we define $Z(t) = 0$ for $\hat{D}(t) = \hat{\sigma}(t) = 0$, that is, $\hat{S}_1(t) = \hat{S}_2(t) = 1$ where no events have been observed by t.

Heuristically, a test based on V would perform well if $\hat{W}(t)$ is proportional to $E(Z(t))$ under alternatives. That is, $\hat{W}(t)$ is large for a large observed $Z(t)$. A natural choice is to let $\hat{W}(t) = Z(t)$. However, as discussed in the Introduction, the distribution of this "chi-square-like" statistic may have a rather long right tail under the null hypothesis and the test may not perform well for specific alternatives. On the other hand, when $\hat{W}(\cdot)$ is constant over time, the distribution of such a "normal-like" statistic is centered around zero, and has a short tail, under the null, but this test may only be powerful when $Z(t)$ is approximately constant over $[0, \zeta]$. Therefore, the question is how to choose the weight function such that the distribution of the resulting statistic has a short right tail under the null, but the observed V is large under the alternative. A possible solution is to expand on the idea proposed by Xu et al. (2003) for combining a small number of dependent test statistics for linkage or association across multiple phenotypic traits. Specifically, let $c \in [0, \eta]$, where η is a constant, say 4. Let $\hat{W}_c(t) = \max\{Z(t), c\}$ and

$$V_1(c) = \int_0^\zeta \hat{W}_c(t) Z(t) dt. \tag{2.2}$$

For a fixed c, say 1.65, under the null hypothesis, since $Z(t)$ is approximately $N(0, 1)$, $\hat{W}_c(t) \sim 1.65$ for most $t \in [0, \zeta]$. It follows that the distribution of $V_1(c)$, which is similar to a linear combination of dependent standard normal random variables, would not have a long right tail. On the other hand, under an alternative hypothesis, for a large observed $Z(t)$, (≥ 1.65), $\hat{W}(t) = Z(t)$ and the resulting observed $V_1(c)$ would be large.

On the other hand, the choice of $c = 1.65$ may not work well for cases in which $D(\cdot)$ is positive for a large portion of time points but the observed $Z(\cdot)'$s are less than 1.65 due to, for example, the low observed mortality rates. Therefore, it is not obvious how to select such a threshold value c a priori.

Here, we propose a simple, automatic way to choose c adaptively to construct a test statistic based on $\{V_1(c), 0 \leq c \leq \eta\}$. First, suppose that we can generate a good approximation to the null distribution of the process $V_1(c)$ indexed by $c \in [0, \eta]$. Let $v_1(c)$ be the observed value of $V_1(c)$ and its p-value $p(c)$ can be obtained via the approximation to the null distribution of $V_1(c)$. Let $p_b = \min\{p(c) : c \in [0, \eta]\}$, the most significant $p(c)$ in $c \in [0, \eta]$. A small p_b would support the alternative hypothesis. The question is how to choose the threshold value for claiming a "statistical significance" based on p_b. That is, one needs to obtain the null distribution of $P_b = \min\{P(c) : c \in [0, \eta]\}$, the random counterpart of p_b, where $P(c) = S_{V_1(c)}(V_1(c))$, and $S_{V_1(c)}(v)$ is the survival function of $V_1(c)$. Using the standard martingale theory, it can be shown that $Z(\cdot)$ converges weakly to a limiting Gaussian process $G(\cdot)$ (Gill, 1983). In Appendix A, we show that $V_1(c)$ and $P(c)$, as processes in c, converge weakly to $\psi(c) = \int_0^\zeta \max\{G(t), c\}G(t)dt$ and $U(c) = S_{\psi(c)}(\psi(c))$, respectively, where $S_{\psi(c)}(v)$ is the survival function of $\psi(c)$.

To empirically approximate the limiting distribution of this process under the null, one may utilize a perturbation-resampling method, which has been applied successfully to various problems in survival analysis (Lin et al., 1994; Parzen et al., 1997). Specifically, the distribution of the process $\left\{\hat{S}_i(t) - S_i(t)\right\}$, $i = 1, 2$ can be approximated by that of

$$Q_i(t) = -\hat{S}_i(t) \sum_{j=1}^{n_i} \left[\left\{ \sum_{k=1}^{n_i} I(x_{ik} \geq x_{ij}) \right\}^{-1} \delta_{ij} I(x_{ij} \leq t) \xi_{ij} \right],$$

where (x_{ij}, δ_{ij}) is the observed value of (X_{ij}, Δ_{ij}), $I(\cdot)$ is the indicator function, and $\{\xi_{ij},\ i = 1, 2,\ j = 1, \ldots,\ n_i\}$ is a random sample from a distribution with mean 0 and variance 1, for example, the standard normal distribution. In practice, the null distribution of $V_1(c)$ can be approximated by generating M sets of $\{\xi_{ij}\}$. For each realized set $\{\xi_{ij}\}$, we compute

$$V_1^*(c) = \int_0^\zeta W_c^*(t) Z^*(t) dt, \tag{2.3}$$

where $Z^*(\cdot) = \{Q_2(\cdot) - Q_1(\cdot)\} / \hat{\sigma}(\cdot)$ and $W_c^*(\cdot) = \max\{Z^*(\cdot),\ c\}$. The set $\mathcal{D}$ of M realizations $\{V_1^*(c),\ c \in [0, \eta]\}$ serves as a reference set for the proposed test. For each of the M sets, we compute (2.3), and obtain the corresponding $P(c)$, using the reference set $\mathcal{D}$, denoted by $P^*(c)$. The null distribution of P_b can be estimated using the M realizations of $P_b^* = \min\{P^*(c) : 0 \leq c \leq \eta\}$ based on $\{V_1^*(c),\ c \in [0, \eta]\}$. The bona fide p-value of the proposed test is then given by $\Pr(P_b^* < p_b)$. In Appendix B, we show that conditional on the observed data, under the null, the limiting distributions of $V_1^*(c)$ and $P^*(c)$ are $\psi(c)$ and $U(c)$, respectively, and thus the null distribution of P_b can be approximated well by $\min\{P^*(c) : 0 \leq c \leq \eta\}$.

Another potential class of test statistics is

$$V_2(c) = \int_0^\zeta \hat{W}_c(t) Z(t) d\bar{N}(t), \tag{2.4}$$

where $\bar{N}(t) = (n_1 + n_2)^{-1} \sum_{i=1}^2 \sum_{j=1}^{n_i} I(X_{ij} \leq t)\Delta_{ij}$. Note that the null distribution of P_b can also be approximated via the aforementioned procedure. The role of $\bar{N}(\cdot)$ as the integrator serves as another weighting function for $Z(\cdot)$, that is, the weight is heavy for the time intervals with large numbers of observed events. The parameter η in the proposed test can be any positive constant. Empirically we find that the choice $\eta = 4$ works well since it is unlikely $Z(\cdot)$ would be beyond 4 under the null (see Section 3 for details).

3 The E4A03 example and numerical comparison studies

First, we apply the proposed tests to the survival data from E4A03 for comparing the low- and high-dose groups discussed in the Introduction section. We calculate $V_1(c)$ and $V_2(c)$ over the range [0, 40] months at each value of $c = 0,\ 0.1,\ 0.2, \ldots, 4$, where we use Greenwood's formula for estimating variances of $\hat{S}_1(\cdot)$ and $\hat{S}_2(\cdot)$ to calculate the standard error of $\hat{D}(\cdot)$. The observed p_b values are 0.0044 and 0.0018 based on $V_1(c)$ and $V_2(c)$, respectively, both of which are obtained at $c = 0$ in this example. To construct the null reference sets for these two tests, we generate $M = 5000$ realized samples $\{\xi_{ij}\}$ from $N(0, 1)$ to obtain the null distributions of P_b. The resulting one-sided bona-fide p-values, $\Pr(P_b < p_b)$, are 0.0048 and 0.0020 respectively. For comparison, we also analyze the data with the two WKM tests, (1.2) and (1.3), proposed by Pepe and Fleming (1989); the corresponding p-values are 0.007 and 0.014, respectively. The test proposed by Yang and Prentice (2010) gives p=0.138. Recall that the Peto-Prentice-Wilcoxon test and logrank test yield p=0.142 and p=0.233, respectively.

We conduct an extensive simulation study to examine the performance of the new tests. First, we assess the size and power of the tests under a similar setting to the E4A03 trial. The pattern illustrated in Figure 2(a), with no difference between the two survival functions, is considered for evaluating the empirical Type I error rate. The curve in Figure 2(a) is the survival function of a Weibull distribution derived as the best approximation of the low-dose group data from E4A03 using the maximum likelihood method. Figure 2(b) shows survival functions of Weibull distributions that approximate the low- (solid line) and high-dose (dashed line) groups data from the E4A03 trial, respectively. For the underlying censoring time distributions, we consider four scenarios: (i) no censoring, (ii) light censoring, (iii) heavy censoring and (iv) the observed censoring patterns from the E4A03 trial, obtained by fitting the data with a Weibull distribution model with the low- and high-dose groups combined. For all censoring configurations, we consider an administrative censoring at 40 months. We graphically present those censoring patterns in Figure 3.

We generate 2000 independent samples, with sample size $n = 300$ (per group), from the distributions described above (Figures 2 and 3). Note that, for each subject, we generate a survival time and an underlying censoring time and compute the observable time (i.e., the minimum of the survival time and the censoring time) and the censoring indicator. Similar to the analysis for E4A03 presented above, we let $[0, \zeta] = [0, 40]$ (months), $M = 5000$, and $\{\xi_{ij}\}$ is from the standard normal distribution, and c is evaluated in increments of 0.1 up to a maximum of $\eta = 4$. For comparators, we include the logrank, Peto-Prentice-Wilcoxon, Yang-prentice, and Pepe-Fleming tests. Because the results of Pepe-Fleming tests with the weights (1.2) and (1.3) are similar, we present the results with the weight (1.2) only.

Table 1 shows the results of this numerical study. The empirical Type I error rates are nearly identical to their nominal level of 0.05. The new tests appear to be consistently more powerful than their comparators. For this cancer study data, the logrank test performs rather poorly with respect to power, as expected.

We also examined other scenarios to evaluate the performance of the new proposals. In one of the numerical studies, we consider the following patterns of differences between two survival functions illustrated in Figure 4. Specifically, five different patterns are examined: 4(a) no difference, 4(b) PH difference (with true hazard ratio of 0.8), 4(c) early difference, 4(d) middle difference, and 4(e) late difference. For this numerical study, the sample size is $n = 200$ per arm. Other configurations, such as the underlying censoring distributions (Figure 3), the number of iterations, etc. are the same as described above.

Table 2 presents the results of the study. The results from pattern 4(a) show that the empirical significance levels of all tests are close to their nominal value of 0.05. Under the PH alternative (pattern 4(b)), the test V_2 is comparable with other tests which are not designed for this specific alternative. However, the test V_1 performs as well as the logrank test in this scenario. For the early difference alternative (pattern 4(c)), the Peto-Prentice-Wilcoxon test gives a higher power than the logrank as we expect, but the new tests are even more powerful than the Wilcoxon. For the pattern 4(d), our tests are more powerful than the other tests. For the late difference (pattern 4(e)), the Peto-Prentice-Wilcoxon test is worst, and V_1 is the best among all the test procedures considered. Note that V_2 is not as powerful under this setting due to the fact that most of the failures are observed prior to the time at which the survival curves separate.

We also observe how the distribution of the selected value of c corresponding to p_b, the smallest $p-$values, varies across simulation scenarios. For each simulation scenario, we obtain the selected optimal c for each of the generated 2000 independent samples, and draw the histogram. Figure 5 shows the results from simulation scenarios corresponding to Figure 4(a-e) with (i) no censoring, for the test based on V_1. The selected values of c rarely reach 4 in our simulations, which suggests that $[0, 4]$ is a reasonable search range for c in practice.

In summary, while the test based on V_2 performs particularly well when the survival curves separate early, the test based on V_1 appears to be more generally useful, demonstrating power equal to that of the logrank test under the PH alternative and *exceeding* the power of all comparator tests under all other scenarios, including those proposed by Pepe and Fleming(1989). Such robust performance characteristics are likely indicative of the fact that the proposed tests are not derived to detect a specific departure from the null hypothesis. Unlike other procedures, the new proposals are not derived under the assumption of a specific type of alternative hypotheses.

4 Remarks

Unlike the estimates of the hazard function or the cumulative version thereof, the Kaplan-Meier plots are informative and easily interpretable in describing the temporal profile between-group differences. It seems natural to have a companion two-sample test for a global statistical assessment of such differences. The logrank test statistic is sensitive for testing the equality of two hazard functions, especially under the PH model, but not necessarily for comparing two survival functions directly. The new proposal, on the other hand, is shown to be a useful tool to serve this purpose.

Since the weighted logrank test is constructed by combining 2x2 tables at the observed event times, in principle, one may utilize the same idea to choose the weight associated with each table. However, the operating characteristics of the resulting test procedure are not clear. Further research is needed along this line. The new tests automatically and empirically adjust the weighting functions, which are not required to be pre-specified in the protocol or statistical analysis plan, and are not restricted to be powerful only for specific alternatives.

Although hypothesis testing provides statistical evidence of treatment difference, the estimation of the treatment difference is also important for assessing the magnitude of the difference. The standard companion quantification procedure to the logrank test is the hazard ratio estimation. However, when the PH assumption is violated, the resulting estimate is difficult to interpret, as it is not simply an average of the true hazard ratio over time (Struthers and Kalbfleisch, 1986; Xu and O'Quigley, 2000). An alternative, model-free approach is to use the restricted mean event time as the parameter of interest, which can be estimated by the area under the KM curve. Inferences about the difference or ratio of two restricted mean survival times can then be made (Royston and Parmar, 2011; Zhao et al., 2012; Tian et al., 2013), which are both based directly on the KM curves. Further research is warranted to connect the proposed testing procedure with these closely related estimators.

Appendix

A Weak convergence of $V_1(c)$ and $P(c)$

Note that, under the null hypothesis, $Z(t)$ converges weakly to $G(t)$ as a process. Assuming that $c_2 - c_1 = o(1)$, we have

$$|V_1(c_2) - V_1(c_1)| \leq |c_2 - c_1| O\left(\int_0^\zeta |Z(t)|dt\right) = o_p(1). \tag{A.1}$$

Coupled with the continuous mapping theorem, (A.1) implies that $V_1(c)$ converges weakly to $\psi(c) = \int_0^\zeta \max\{G(t), c\}G(t)dt$.

Next we show that under H_0, $P(c) \to U(c), c \in [0, \eta]$ in distribution as $n \to \infty$, where the limiting process $U(c) = S_{\psi(c)}\{\psi(c)\}$, and $S_{\psi(c)}(v) = \text{pr}\{\psi(c) \geq v\}$, the survival function of $\psi(c)$. Here we consider the survival function of $V_1(c)$, $S_{V_1(c)}(v) = \text{pr}\{V_1(c) \geq v\}$ as well. To simplify the notation in the following argument, we use $S_c(v)$ and $\hat{S}_c(v)$ for $S_{\psi(c)}(v)$ and $S_{V_1(c)}(v)$, respectively. Because of the weak convergence of $V_1(c)$, for any $c_1, c_2, \cdots, c_K \in [0, \eta]$, we have

$$\sup_{1\leq k\leq K} \sup_{v\in(-\infty,\infty)} |\hat{S}_{c_k}(v) - S_{c_k}(v)| = o(1)$$

as $n \to \infty$. Thus,

$$\sup_{1\leq k\leq K} |P(c_k) - S_{c_k}\{V_1(c_k)\}| = \sup_{1\leq k\leq K} |\hat{S}_{c_k}\{V_1(c_k)\} - S_{c_k}\{V_1(c_k)\}| = o(1), \quad \text{a.s}$$

which implies that

$$\begin{aligned}
&\Big| \text{pr}\left\{P(c_1) \geq p_1, \cdots, P(c_K) \geq p_K\right\} - \text{pr}\left\{U(c_1) \geq p_1, \cdots, U(c_K) \geq p_K\right\} \Big| \\
\leq &\Big| \text{pr}\left\{P(c_1) \geq p_1, \cdots, P(c_K) \geq p_K\right\} - \text{pr}\left[S_{c_1}\{V_1(c_1)\} \geq p_1, \cdots, S_{c_K}\{V_1(c_K)\} \geq p_K\right] \Big| \\
+ &\Big| \text{pr}\left\{V_1(c_1) \leq S_{c_1}^{-1}(p_1), \cdots, V_1(c_K) \leq S_{c_K}^{-1}(p_K)\right\} - \text{pr}\left\{\psi(c_1) \leq S_{c_1}^{-1}(p_1), \cdots, \psi(c_K) \leq S_{c_K}^{-1}(p_K)\right\} \Big| \\
= &o(1) \quad \text{for} \quad 0 \leq p_1, \cdots, p_K \leq 1.
\end{aligned}$$

Thus, for any $c_1, c_2, \cdots, c_K \in [0, \eta]$, the joint distribution of $\{P(c_1), \cdots, P(c_K)\}$ converges to that of $\{U(c_1), \cdots, U(c_K)\}$. In addition, (A.1) implies that, for $c_2 - c_1 = o_p(1)$,

$$
\begin{aligned}
|P(c_2) - P(c_1)| &= |S_{c_2}\{V_1(c_2)\} - S_{c_1}\{V_1(c_1)\}| + o_p(1) \\
&\le |S_{c_2}\{V_1(c_2)\} - S_{c_1}\{V_1(c_2)\}| + |S_{c_1}\{V_1(c_2)\} - S_{c_1}\{V_1(c_1)\}| + o_p(1) \\
&= o_p(1).
\end{aligned}
$$

Thus, under the null, $P(c)$ converges weakly to the process $U(c)$ indexed by c.

B Approximation of the null distribution of P_b by P_b^*

Let $\mathcal{O}$ be the observed data. Let $\hat{S}_c^*(v) = \mathrm{pr}\{V_1^*(c) \ge v | \mathcal{O}\}$. Note that under the null, $Z^*(t)|\mathcal{O}$ converges weakly to $G(t)$ as a process almost surely. Coupled with the continuous mapping theorem, $V_1^*(c)|\mathcal{O}$ converges weakly to $\psi(c)$ almost surely. Thus, using the same argument in the Appendix A, we can show that, for any $c_1, c_2, \cdots, c_K \in [0, \eta]$,

$$
\sup_{1 \le k \le K} \sup_{v \in (-\infty, \infty)} |\hat{S}_{c_k}^*(v) - S_{c_k}(v)| = o(1)
$$

almost surely as $n \to \infty$, which implies $\{P^*(c_1), \cdots, P^*(c_K)\} \mid \mathcal{O} \to \{U(c_1), \cdots, U(c_K)\}$. Also, $|P^*(c_2) - P^*(c_1)| = o_p(1)$, for $|c_1 - c_2| = o(1)$ is derived in the same way. Therefore, $P^*(c) = \hat{S}_c^*\{V_1^*(c)\}$, conditional on the observed data, converges weakly to the process $U(c), c \in [0, \eta]$ almost surely. Therefore, the null distribution of $P_b = \min\{P(c) : 0 \le c \le \eta\}$ can be approximated by $P_b^* = \min\{P^*(c) : 0 \le c \le \eta\}$.

Acknowledgements

The authors thank the Eastern Cooperative Oncology Group for permission to use the data.

References

Fleming, T. R. and D. P. Harrington (1984). *Evaluation of censored survival data test procedures based on single and multiple statistics.*, pp. 97 – 123. Topics in Applied Statistics. New York: Marcel Dekker.

Fleming, T. R. and D. P. Harrington (1991). *Counting processes and survival analysis.* Wiley.

Gastwirth, J. L. (1985). The Use of Maximin Efficiency Robust Tests in Combining Contingency Tables and Survival Analysis. *Journal of the American Statistical Association 80*(390), 380–384.

Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *The Annals of Statistics*, 49–58.

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

Lagakos, S. and D. Schoenfeld (1984). Properties of Proportional-Hazards Score Tests under Misspecified Regression Models. *Biometrics 40*(4), 1037–1048.

Lai, T. L. and Z. Ying (1991). Rank Regression Methods for Left-Truncated and Right-Censored Data. *The Annals of Statistics 19*(2), 531–556.

Lee, J. W. (1996). Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics*, 721–725.

Lin, D. Y., T. R. Fleming, and L. J. Wei (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika 81*(1), 73–81.

Parzen, M. I., L. J. Wei, and Z. Ying (1997). Simultaneous confidence intervals for the difference of two survival functions. *Scandinavian Journal of Statistics 24*(3), 309–314.

Pecková, M. and T. R. Fleming (2003). Adaptive test for testing the difference in survival distributions. *Lifetime data analysis 9*, 223–238.

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

Pepe, M. S. and T. R. Fleming (1991). Weighted Kaplan-Meier Statistics: Large Sample and Optimality Considerations. *Journal of the Royal Statistical Society Series B 53*(2), 341–352.

Rajkumar, S. V., S. Jacobus, N. S. Callander, R. Fonseca, D. H. Vesole, M. E. Williams, R. Abonour, D. S. Siegel, M. Katz, P. R. Greipp, and E. C. O. 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. *The lancet oncology 11*(1), 29–37.

Royston, P. and M. K. B. Parmar (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. *Statistics in Medicine 30*(19), 2409–2421.

Self, S. G. (1991). An adaptive weighted log-rank test with application to cancer prevention and screening trials. *Biometrics 47*(3), 975–986.

Struthers, C. A. and J. D. Kalbfleisch (1986). Misspecified Proportional Hazard Models. *Biometrika 73*(2), 363–369.

Tarone, R. E. (1981). On the distribution of the maximum of the logrank statistic and the modified Wilcoxon statistic. *Biometrics 37*, 79–85.

Tarone, R. E. and J. Ware (1977). On Distribution-Free Tests for Equality of Survival Distributions. *Biometrika 64*(1), 156–160.

Tian, L., L. Zhao, and L. Wei (2013). On the restricted mean event time in survival analysis. *Havard University Working Paper Series.* (156).

Xu, R. and J. O'Quigley (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics 1*(4), 423–439.

Xu, X., L. Tian, and L. J. Wei (2003). Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics 4*(2), 223–229.

Yang, S. and R. Prentice (2005). Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika 92*(1), 1–17.

Yang, S. and R. Prentice (2010). Improved logrank-type tests for survival data using adaptive weights. *Biometrics 66*(1), 30–38.

Zhao, L., L. Tian, H. Uno, S. D. Solomon, M. A. Pfeffer, J. S. Schindler, and L. J. Wei (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*(5), 570–577.

Zucker, D. M. and E. Lakatos (1990). Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika 77*(4), 853–864.

Table 1: Size and power of logrank [LR], Peto-Prentice-Wilcoxon [PP], Pepe-Fleming tests based on (1.2)[PF], Yang-prentice [YP] and the new tests based on (2.1) [V_1] and (2.4) [V_2]. 2(a) no difference and 2(b) difference observed in E4A03 (see Figure 2).

Size of tests

Survival distributions: 2(a)

Test	Censoring			
	(i)	(ii)	(iii)	(iv)
LR	0.051	0.049	0.045	0.051
PP	0.051	0.047	0.044	0.049
YP	0.061	0.060	0.056	0.065
PF	0.053	0.045	0.044	0.044
V_1	0.052	0.042	0.040	0.047
V_2	0.055	0.042	0.041	0.043

Power of tests

Survival distributions: 2(b)

Test	Censoring			
	(i)	(ii)	(iii)	(iv)
LR	0.111	0.123	0.166	0.230
PP	0.176	0.196	0.247	0.307
YP	0.200	0.214	0.278	0.330
PF	0.631	0.625	0.627	0.725
V_1	0.840	0.828	0.827	0.847
V_2	0.830	0.827	0.837	0.868

Table 2: Size and power of logrank [LR], Peto-Prentice-Wilcoxon [PP], Pepe-Fleming tests based on (1.2)[PF], Yang-prentice [YP] and the new tests based on (2.1) [V_1] and (2.4) [V_2], to detect the difference between two survival curves, based on 2000 of iterations, with sample size 200 per arm. 4(a) no difference, 4(b) proportional hazards, 4(c) early difference, 4(d) difference in middle, and 4(e) late difference (see Figure 4).

Censoring	(i) no censoring					(ii) light censoring				
Test	Survival distributions					Survival distributions				
	4(a)	4(b)	4(c)	4(d)	4(e)	4(a)	4(b)	4(c)	4(d)	4(e)
	Size	Power				Size	Power			
LR	0.046	0.703	0.294	0.401	0.205	0.045	0.686	0.310	0.410	0.200
PP	0.051	0.605	0.832	0.387	0.098	0.050	0.599	0.851	0.374	0.089
YP	0.062	0.730	0.651	0.451	0.244	0.054	0.713	0.690	0.455	0.242
PF	0.051	0.693	0.303	0.526	0.268	0.046	0.681	0.320	0.524	0.259
V_1	0.057	0.703	0.951	0.620	0.383	0.051	0.699	0.959	0.614	0.378
V_2	0.051	0.627	1.000	0.558	0.160	0.049	0.600	1.000	0.547	0.148

Censoring	(iii) heavy censoring					(iv) observed in E4A03				
Test	Survival distributions					Survival distributions				
	4(a)	4(b)	4(c)	4(d)	4(e)	4(a)	4(b)	4(c)	4(d)	4(e)
	Size	Power				Size	Power			
LR	0.051	0.669	0.332	0.408	0.187	0.043	0.684	0.304	0.415	0.260
PP	0.048	0.582	0.864	0.355	0.088	0.050	0.603	0.847	0.386	0.093
YP	0.062	0.695	0.719	0.458	0.235	0.055	0.710	0.681	0.469	0.317
PF	0.052	0.657	0.338	0.517	0.233	0.047	0.671	0.327	0.567	0.253
V_1	0.054	0.677	0.960	0.609	0.352	0.063	0.714	0.971	0.645	0.401
V_2	0.049	0.585	1.000	0.513	0.124	0.049	0.605	1.000	0.560	0.136

Figure 1: Overall survival curves for low-dose arm (solid line) and high-dose arm (dashed line) with the E4A03 data

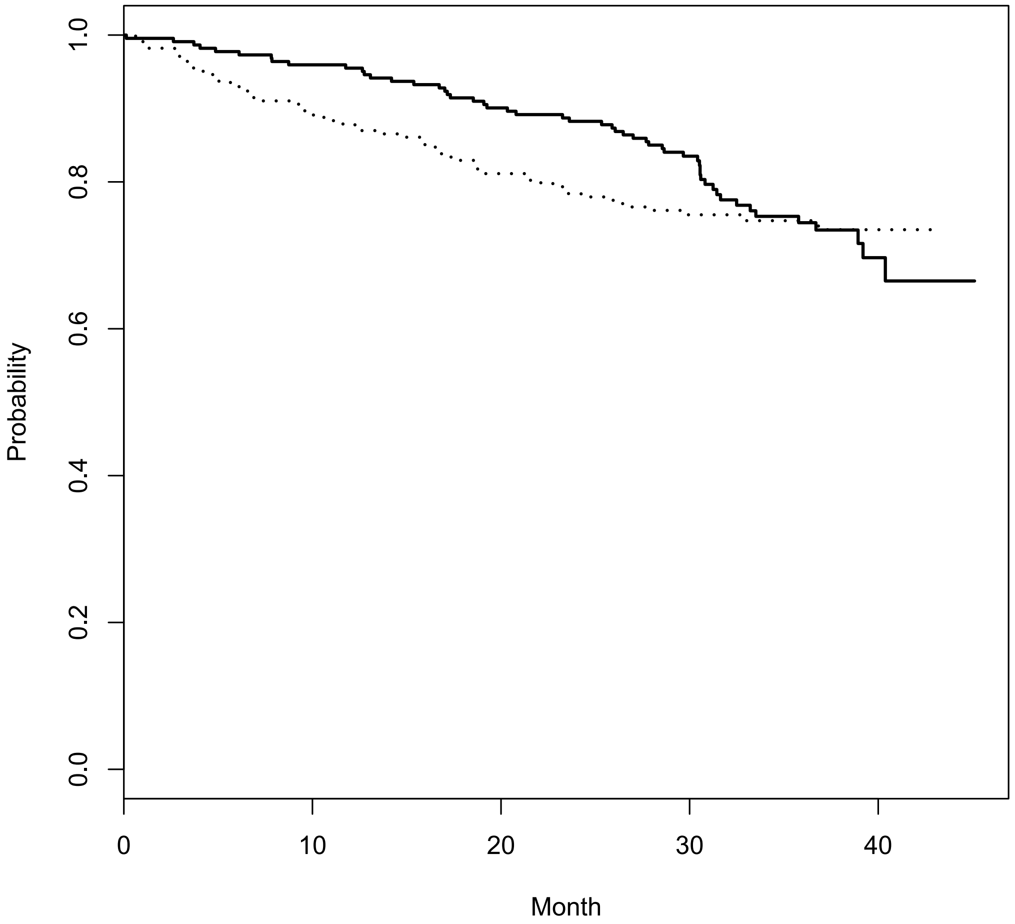


Figure 2: Comparison of survival functions considered in simulation studies

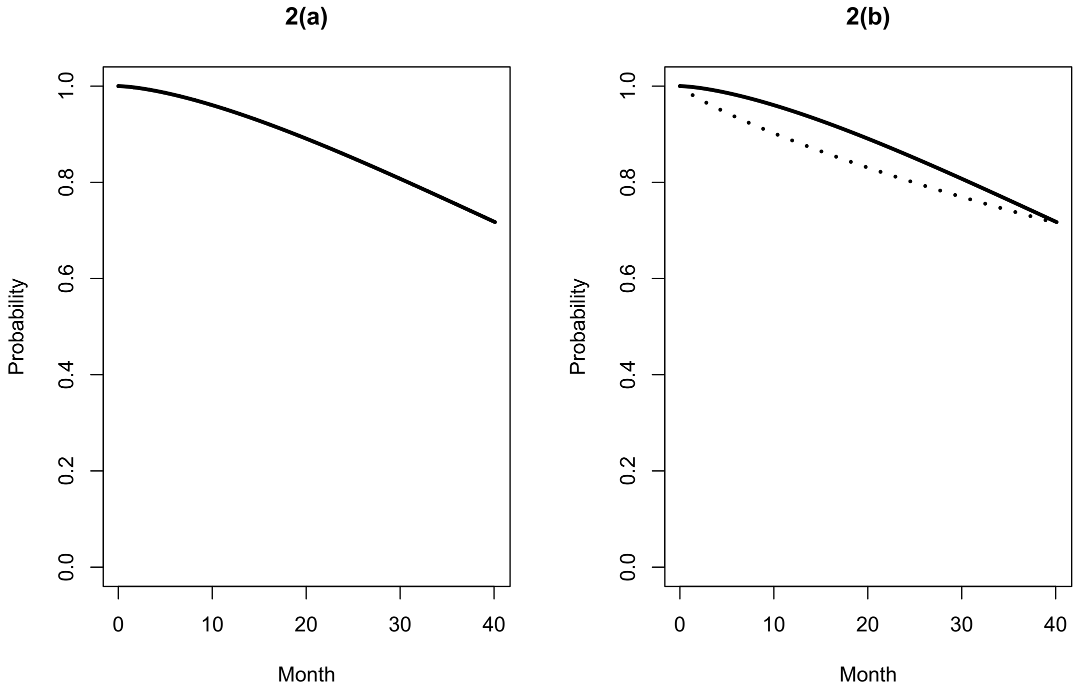


Figure 3: Survival functions of the underlying censoring distributions considered in the simulations.

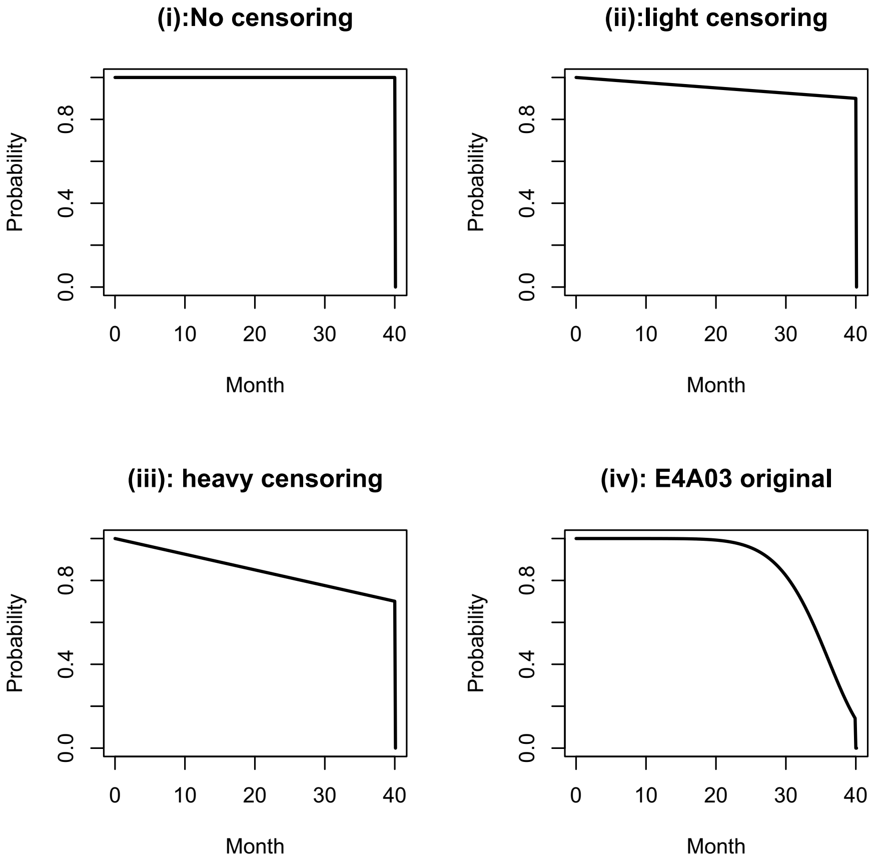


Figure 4: Comparison of survival functions considered in simulation studies

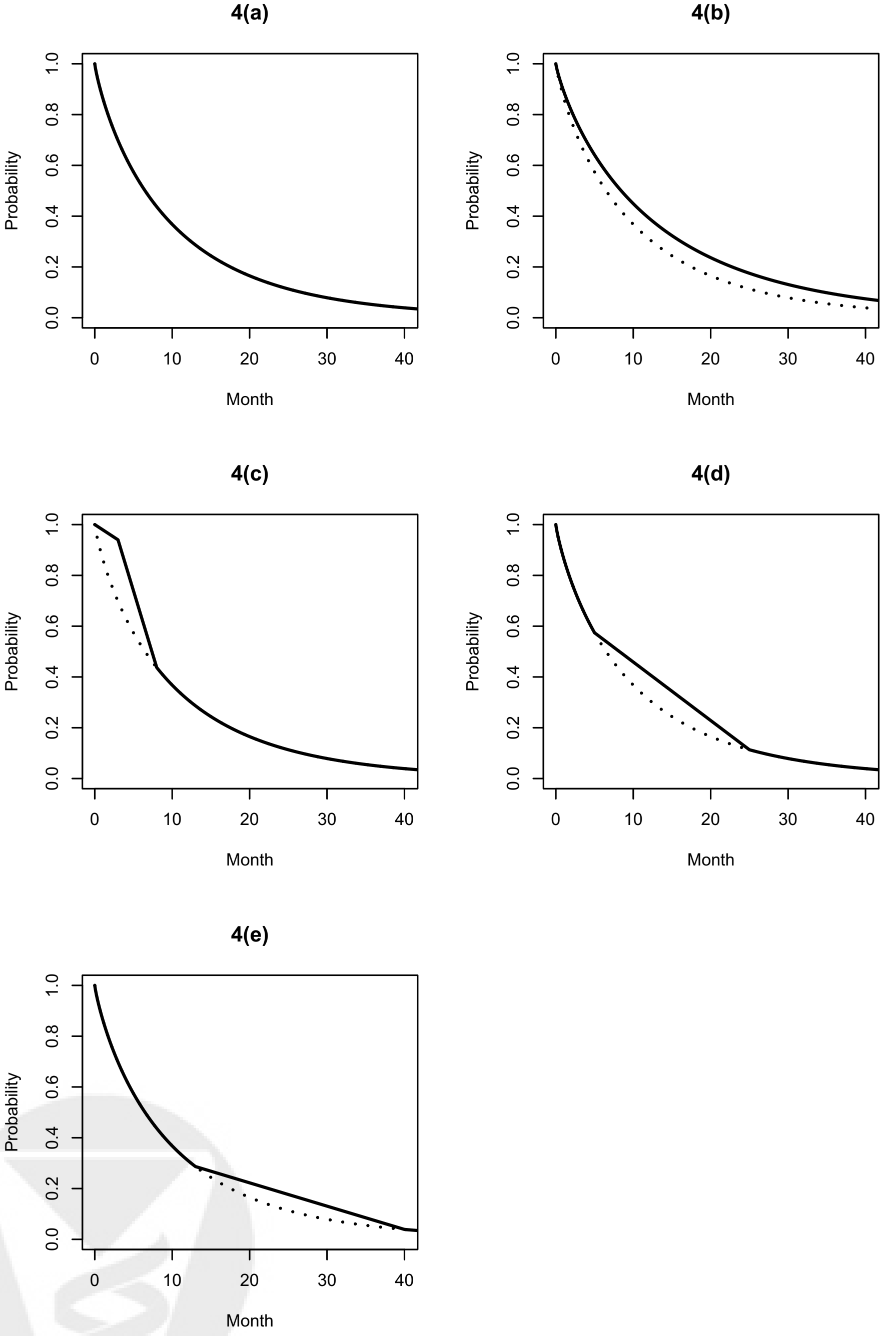
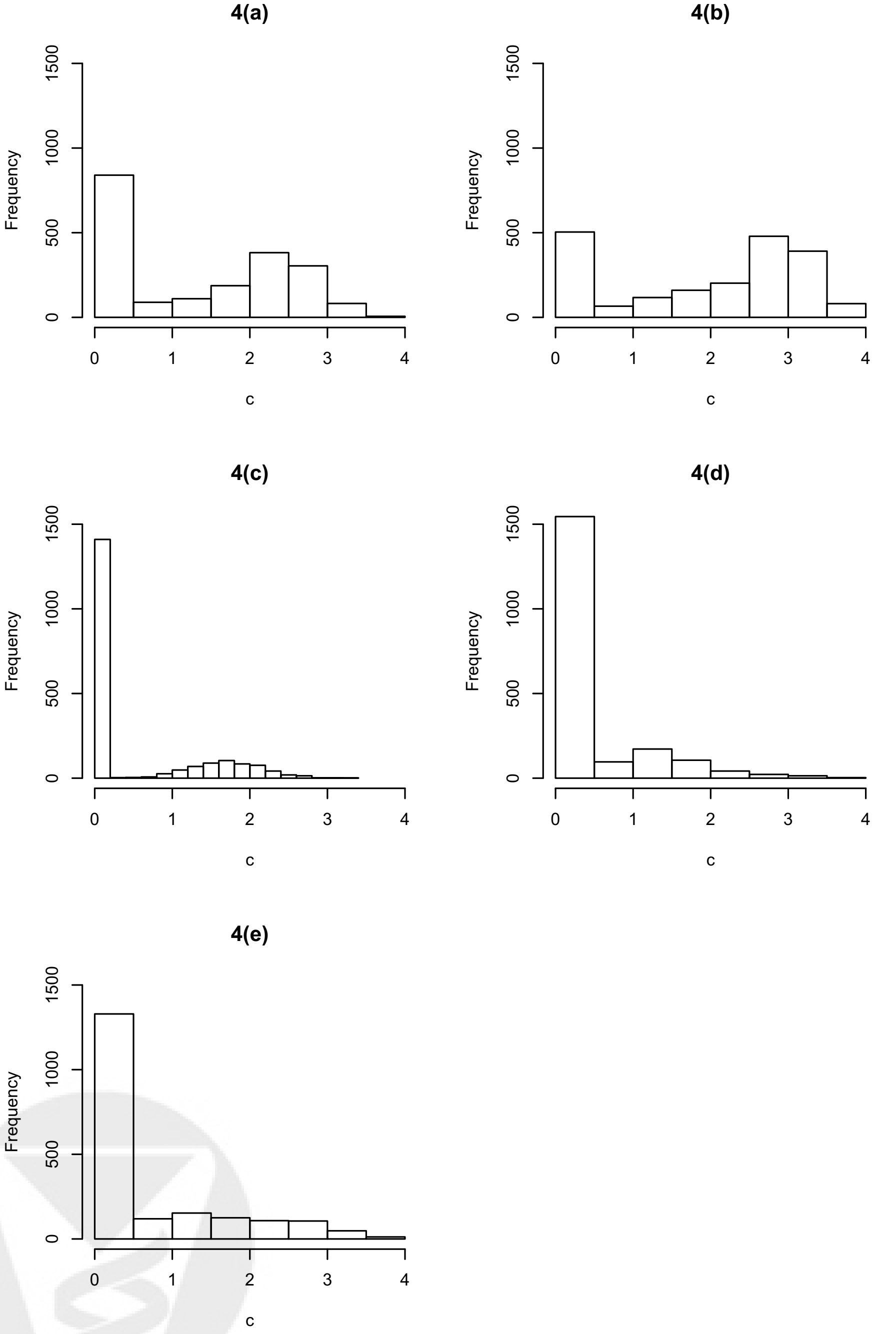


Figure 5: Frequency of the selected value of c for $V_1(c)$ in simulation studies


4(a)
4(b)
4(c)
4(d)
4(e)
Frequency
c