Comparing survival functions with interval-censored data in the presence of an intermediate clinical event

Sohee Kim

The Graduate School
Yonsei University
Department of Biostatistics and Computing

Comparing survival functions with interval-censored data in the presence of an intermediate clinical event

A Dissertation
Submitted to the Department of Biostatistics and Computing
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Sohee Kim

December 2015

This certifies that the dissertation of Sohee Kim is approved.

Thesis Supervisor: Dr. Chung Mo Nam

Dr. Jinheum Kim

Dr. Byung-Ho Nam

Dr. Inkyung Jung

Dr. Sohee Park

The Graduate School
Yonsei University
December 2015

Contents

List of Figures

List of Tables

Abstract

The primary endpoint is considered a time metric such as the time taken to reach a well-defined endpoint in the presence of an intermediate clinical event (IE). A subject may experience an IE including an intervention during the period of follow-up. When an IE occurs, it may change the survival distribution. When deriving the statistics of Nam and Zelen (2001), the data was divided into two parts according to an IE; one part was right-censored data that does not reach the time of experiencing an IE and the other part was left-truncated and right-censored data, which is having an IE and truncation at that time. Considering the primary endpoint was interval-censored, we extended the approach of Finkelstein (1986) as accommodating left-truncated data. After that, for convenience, we applied multiple imputation techniques to left truncated and interval-censored data. Firstly, we proposed a uniform method to impute data of uniform weight in a characterized set. The second method applies the non-parametric maximum likelihood estimator (NPMLE) from the original data as weight. We applied two forms of variance, which are formed by adding and subtracting within- and between variance for all proposed models. Both variances worked efficiently, but the first one was slightly over- while the second one was marginally underestimated. Through simulation, the stratified log-rank test is unsatisfactory when the proportion of an IE is different between two groups. When the survival distribution is changed after experiencing an IE in addition to the difference proportion of an IE for two groups, the log-rank test is not appropriate. The proposed methods satisfied a nominal level of 0.05 and had superior power to compare the proposed methods with the log-rank and stratified log-rank tests for all

scenarios.

KEY WORDS: Interval-censored, Intermediate clinical event, comparing survival functions

1 Introduction

1.1 Background

We consider the primary endpoint is a time metric such as the time taken to reach a well-defined endpoint in the presence of an intermediate clinical event (IE). A subject may experience an IE including an intervention during the period of follow-up. When an IE occurs, it may change the survival distribution.

For example, we can consider the objective response rate (ORR) in a randomized cancer clinical trial with advanced cancers, which is often used as a primary or secondary endpoint. Anderson et al. (1983) studied the landmark method of survival by ORR, which is a measure of tumor shrinkage for the subject with a measurable lesion. ORR is estimated by the proportion of responders who had a complete response (disappearance of tumor) or partial response (30% or more reduction in size from the baseline estimated using RECIST 1.1 (Eisenhauer et al., 2009)). When a subject becomes a responder, it can be regarded as the subject experienced an IE. To be a responder, subjects should survive at least the respond time. Shortly, an IE needs guaranteed time to occur. Therefore, the length of survival itself will influence the chance of an IE. This is called a length-biased problem.

An example of a length-biased problem is the heart transplantation study (Mantel and Byar, 1974). It is necessary to know whether a heart transplant would be beneficial. The waiting time of subjects who eventually have heart transplant must be long enough to receive treatment, whereas there is no requirement for not having a heart transplant.

It can be considered sequential therapies in the last example. For randomized sequential therapies, the total progression-free survival (PFS) as the primary end-

point is usually defined as the interval between the randomization (the start date of first-line therapy) to the disease progression or death during second-line therapy. After first-line therapy, a subject can be dropped for various reasons and does not receive second-line therapy. In this case, first-line events were used and subjects without tumor progression or death during second-line therapy were censored. It is difficult to know the pure effect of sequential therapies using standard approaches because of its length-biased characteristics. Survivors have more chance of being treated with second-line therapy.

For example, in SWITCH research (Eichelberg et al., 2015), the efficacy of Sorafenib followed by Sunitinib (So-Su, n = 182) versus the reverse sequence (Su-So, n = 183) was determined for treating metastatic renal cell cancer. The proportion that was administered in second-line therapy was higher in So-Su (57% vs 42%, P value < 0.01). The total PFS and PFS of first-line therapy did not show the statistically significant difference (12.5 mo vs. 14.9 mo (P value = 0.5) and, 5.9 mo vs. 8.5 mo (P value = 0.8), respectively), whereas the PFS of second-line therapy showed a shorter duration in Su-So (5.4 mo vs. 2.8 mo, P value < 0.001). Therefore, in the current paper, we will assess subjects receiving second-line therapy as experiencing an IE. If we consider the subject receiving second-line therapy as experiencing an IE, we can compare the difference of survival functions through the analysis of the proportion of subjects having second-line therapy and the duration of first-/second-line therapy with different hazards assumptions. As shown in the simulation, especially when the proportion of subjects experiencing an IE is different, the standard log rank test is not appropriate.

1.2 Literature review

To resolve length-biased problems, Mantel and Byar (1974) studied time-dependent cox regression. Anderson et al. (1983) researched the landmark study that selects a fixed time after the initiation of therapy as a landmark. This study determined the response of the subject at a landmark time and only survivors at the landmark time were analyzed. Lefkopoulou and Zelen (1995), Nam and Zelen (2001) derived statistical tests based on the score test to verify whether an IE induced a change in the survival distribution.

In addition, when the primary outcome is interval-censored, the situation is more complicated. Interval-censored data is data for which all that is known is that the event occurred in some interval $(L_i, R_i]$ with $L_i < T_i \leq R_i$, but the exact time is not known; where T_i is survival time, L_i is the last visit that ith subject does not have an event, R_i is the closest visit of observing the event for the ith subject. If the event occurs exactly at the moment of a visit, then we have an exact survival time with $T_i = L_i = R_i$. However, the event of interest did not occur until the last visit, $R_i = \infty$. PFS is considered interval-censored data. When subjects are not in progression at the last assessment visit, but are in progression at the next visit, it will only be known that the event occurs in the known interval. Law and Brookmeyer (1992) studied a naïve simple midpoint imputation approach for interval-censored data. It was shown that when the two groups had different censoring mechanisms and intervals were wide and varied, the naïve imputation method was invalid. Therefore, a specific method to process interval-censored data is needed.

There is much literature to estimate and test the survival function of interval-

censored data. Peto (1973) proposed the non-parametric maximum likelihood estimator (NPMLE) using Newton-Rapshon algorithm. Turnbull (1976) characterized the NPMLE in the presence of arbitrarily censoring and truncation. The self-consistent algorithm was used to estimate the NPMLE, which is a special case of the Expectation-maximization (EM) algorithm (Dempster et al., 1976). It was shown that the convergence of the EM algorithm does not guarantee convergence to the global maximum if it does not meet Kuhn-Tucker conditions (Kuhn and Tucker, 1951, Gentleman and Geyer, 1994). Finkelstein (1986) studied the weighted log-rank test on interval-censored data under the proportional hazard model. Sun (1996) studied a log-rank type test under the logistic model by applying Turnbull's algorithm to estimate the pseudo-risk and failure sets. Zhao and Sun (2004) improved the previous study of Sun (1996) that could reduce to the original log-rank test in the case of right-censored data by using a multiple imputation technique to estimate the covariance matrix. Kim et al. (2006) studied a log-rank type test that did not use an iterative algorithm. A uniform weights algorithm was proposed where a subject contributed uniformly to each mass point s_k; the set s_k consisted of all the distinct $(L_i, R_i]$ observed interval. Huang et al. (2008) proposed a log-rank type test similar to Zhao and Sun (2004) but used different covariance matrix estimator. Fay and Shaw (2010) released the **interval** R package to perform weighted log-rank tests for interval-censored data. In the package, the score test with several options (Fay (1996), Finkelstein (1986) and Sun (1996)), based on a general score test for interval-censored data (Fay, 1999), permutation methods (Heinze et al., 2003) and a weighted log-rank type test (Huang et al., 2008) was developed.

We also reviewed studies regarding left-truncated and interval-censored (LTIC) data. When deriving the statistics of Nam and Zelen (2001), the data was divided

into two parts according to an IE; one part was right-censored data where the subject did not reach the time of experiencing an IE (a subject might experience an IE eventually, but did not experience an IE at that time) and the other part was left-truncated and right-censored data where subjects had an IE and truncation at that time. We considered the interval-censored data like PFS, and therefore will have interval-censored data for not having an IE and LTIC data for having an IE. For LTIC data, there is a limited literature. Frydman (1994) corrected Turnbull's characterization as accommodating both truncation and interval-censoring time points. Alioum and Commenges (1996) extended it to the regression model under the proportional assumption. Pan and Chappell (1999) noted that NPMLE was inconsistent for the early time with LTIC data while conditional NPMLE was shown to be consistent. Pan and Chappell (2002) considered the estimation of the parameters in the Cox model with LTIC data. Shen (2014a) studied a rank-based test of survival function in LTIC. However, the length-biased problem was not considered in those methods.

The data was separable according to the IE in proposed method. This concept has similarities with the multi-state model. Frydman (1995) and Joly et al. (2002) studied for interval-censored intermediate event and exactly observed outcome (absorbing state). Joly et al. (2012) and Frydman et al. (2013) studied for all transition times are interval censored for multi-state model. They did not use MI while Yu et al. (2010) studied MI on illness-death model applied with PMDA and ANDA methods (Wei and Tanner, 1991, Pan, 2000a). There was no studies an IE as exact data and interval-censored end point.

Most of the proposed methods for interval-censored data used intensively iterative computation. To avoid this, an imputation method was considered. For the

survival analysis, the event times for interval-censored observations are unknown, so they can be regarded as missing by assuming independent censoring mechanisms. We can obtain complete or (left-truncated and) right-censored data after imputation of the (left-truncated and) interval-censored data. Then we can analyze the imputed set using standard statistical methods. For missing data, Rubin (1987) suggested multiple imputation (MI). MI accounts for true variability by incorporating the between-imputation variability (Rubin, 1987). For right-censored data, Wei and Tanner (1991) proposed two semiparametric algorithm motivated by the data augmentation algorithm of Tanner and Wong (1987). Two implementations were considered for this, poor man's data augmentation (PMDA) and asymptotic normal data augmentation (ANDA). Pan (2000b) studied a two-sample test with interval-censored data via MI based on the approximate Bayesian bootstrap. Pan (2000a) also proposed a MI using cox regression for interval-censored data by adapting Wei and Tanner (1991)'s method. Hsu et al. (2007) studied the MI for interval-censored data with auxiliary variables. Zhao and Sun (2004) and Kim et al. (2006) used MI techniques for computing the variance of statistics. Huang et al. (2008) proposed log-rank tests via multiple imputations. After estimating the NPMLE using Turnbull's algorithm, they imputed the exact time for all data points including right-censoring data, from the conditional probability of NPMLE. Yu et al. (2010), Shen (2014b) extended the approaches of MI using cox regression (Pan, 2000a) to accommodate left-truncation.

1.3 Outlines

The purpose of the current paper is to investigate distribution-free methods with interval-censored endpoints in the presence of an IE when comparing the survival for two groups. In Section 2.1, we reviewed previous methods. In Section 2.2, we derived the statistics based on score test for interval-censored data. After constructing the likelihood, Finkelstein (1986)'s reparametrization was used with accommodating left-truncation. Then we proposed several multiple imputation methods based on those of Kim et al. (2006) and Huang et al. (2008) in Section 2.3. After imputation, the statistics are based on those of Nam and Zelen (2001). The simulation study provided a log-rank test, stratified log-rank test, and the proposed method. The simulations to assess the performance of the proposed methods are presented in Section 3. The real data application is in Section 4, followed by the the discussion.

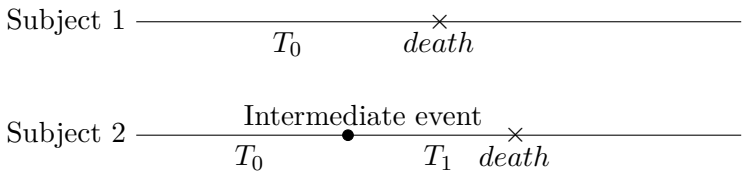

Figure 1: Representation of two subjects with or without an intermediate clinical event (IE).

2 Methods

2.1 Previous methods

2.1.1 Analysis for length biased data

Nam and Zelen (2001) studied a length biased problem with right-censored data in the presence of an IE. The random variables W and T were defined as the waiting time for experiencing an IE and time to death. Z is the indicator of experiencing an IE($Z = 1$) or not($Z = 0$). $Z = 1$ implies waiting time W is less than the event time T ($Z = I\{W \leq T\}$). The random variables T_0 and T_1 are defined as the times to death conditional on $Z = 0$ or 1 respectively. For example, a subject who experienced an IE at W_i and the failure time is T_i, T_{0i} implies the duration from initiation to W_i, T_{1i} implies the duration from W_i to failure time T_i; i.e. $T = (1-Z)T_0 + ZT_1$ (Figure 1).

The probability density function and tail probabilities of W, T_0, T_1 are denoted by $g(w), q_0(t), q_1(t)$, and $G(w), Q_0(t), Q_1(t)$, respectively. A subject who does not experience an IE means that the waiting time for an IE has been right-censored; i.e. $f(t, z = 0) = q_0(t)G(t)$. When a subject experienced an IE, $Z = 1$. Namely, a subject experience an IE at W_i, the survival distribution is changed at w and an event occurs at t; i.e. $f(t, w, z = 1) = Q_0(w)g(w)\frac{q_1(t)}{Q_1(w)}$. Next, the hypothesis

$H_0 : q_{0a}(t) = q_{0b}(t), q_{1a}(t) = q_{1b}(t)$ versus the general alternative which is the complement of H_0 could be considered. Note that the hypotheses are independent of the waiting time distribution. Using this information, we can construct the likelihood as below. Let $\delta_i = 0$ or 1 depending on whether the ith subject is right-censored or not.

$$\begin{cases} \{Q_0(t)G(t)\}^{(1-\delta)(1-z)} & \text{for } \delta = 0, z = 0 \\ \{q_0(t)G(t)\}^{\delta(1-z)} & \text{for } \delta = 1, z = 0 \\ \left\{Q_0(w)g(w)\frac{Q_1(t)}{Q_1(w)}\right\}^{(1-\delta)z} & \text{for } \delta = 0, z = 1 \\ \left\{Q_0(w)g(w)\frac{q_1(t)}{Q_1(w)}\right\}^{\delta z} & \text{for } \delta = 1, z = 1 \end{cases} \tag{1}$$

For $\delta = 0$ and $z = 0$, it implies a subject does not experience an IE($G(t)$) and event($Q_0(t)$) until time t. If an event occurs at time $t(= 1)$ without experiencing an IE ($z = 0$), an IE is right censored at time $t(G(t))$. If a subject experienced an IE at $w(z = 1, g(w))$, it means that the survival without an IE exceed the observed waiting time ($Q_0(w)$). The survival distribution is changed to Q_1 as truncated at $w(\frac{1}{Q_1(w)})$. After experiencing an IE, the subject may have an event at time $t(q_1(t))$ or not have an event until time t ($Q_1(t)$). The score test was derived using a proportional hazards model. Define $Q_{ia}(t) = Q_{ib}(t)^{\beta_i}$ for $i = 0, 1$, where $\beta_i = \exp(\gamma_i)$. Define

$$x = \begin{cases} 1 & \text{if observation is from } A \\ 0 & \text{otherwise,} \end{cases} \quad \delta = \begin{cases} 1 & \text{if observation is non-censored} \\ 0 & \text{otherwise.} \end{cases}$$

The loglikelihood for a single observation is

$$
\begin{aligned}
&l(\beta_0, \beta_1 | x, z, w, t, \delta) \\
&= x\Big\{(1-z)[\delta log\beta_0 + \beta_0 logQ_{0b}(t)\}] + z[\delta log\beta_1 + \beta_1 log\frac{Q_{1b}(t)}{Q_{1b}(w)} + \beta_0 logQ_{0b}(w)]\Big\} \\
&\qquad + \text{terms not involving } (\beta_0, \beta_1).
\end{aligned}
$$

After evaluating at $\beta_0 = \beta_1 = 0$, the score test is

$$
\begin{aligned}
S_0 &= \left(\frac{\partial l_N(\beta_0, \beta_1)}{\partial \beta_0}\right)_{\beta_0=1} = \sum_{k=1}^{N} x_k \Big\{(1-z_k)[\delta_k + logQ_{0b}(t_k)] + z_k logQ_{0b}(w_k)\Big\} \\
S_1 &= \left(\frac{\partial l_N(\beta_0, \beta_1)}{\partial \beta_1}\right)_{\beta_1=1} = \sum_{k=1}^{N} x_k \Big\{z_k\delta_k + log\frac{Q_{1b}(t_k)}{Q_{1b}(w_k)}\Big\}
\end{aligned}
$$

The results can be rewritten using counting process notation. For this purpose, define $N(t) = I(T \leq t, \delta = 1), Z(t) = I(W \leq t)$ and $R(t) = I(T \geq t)$. Let $s_k = x_k z_k(t_k) dN_k(t_k), n_k = \sum_{j=1}^{N} x_j R_j(t_k) z_j(t_k)$, and $N_k = \sum_{j=1}^{N} R_j(t_k) z_j(t_k)$. The statistics $\hat{S}_1$ can be written as

$$
\hat{S}_1 = \sum_{k=1}^{N} x_k z_k(t_k) dN_k(t_k) - \sum_{k=1}^{N} p_k dN_k(t_k), \quad p_k = n_k/N_k
$$

and under the null hypothesis has mean zero and variance $V(\hat{S}_1) = \sum_{k=1}^{N} p_k(1-p_k) dN_k(t_k)$. The statistics $\hat{S}_0$ can be written as

$$
\hat{S}_0 = \sum_{k=1}^{N} x_k(1 - z_k(t_k)) dN_k(t_k) - \sum_{k=1}^{N} \pi_k dN_k(t_k), \quad \pi_k = m_k/M_k,
$$

where $r_k = x_k(1 - z_k(t_k)) dN_k(t_k), m_k = \sum_{j=1}^{N} x_j R_j(t_k)(1 - z_j(t_k))$, and $M_k =$

$\sum_{j=1}^{N} R_j(t_k)(1 - z_j(t_k))$. The variance is $V(\hat{S}_0) = \sum_{k=1}^{N} \pi_k(1 - \pi_k)dN_k(t_k)$. Hence, an appropriate chi-square statistic with 2 degress of freedom for testing H_0 is $\chi_2^2 = \hat{S}_1^2/V(\hat{S}_1) + \hat{S}_0^2/V(\hat{S}_0)$.

2.1.2 Analysis for interval-censored data

Finkelstein (1986) developed a method for fitting the proportional hazards regression model for interval-censored response time data. The data can be represented as $(L_i, R_i]$ and X_i. X_i is the r-dimensional vector of covariates. For the right-censored data, $R_i = \infty$, while for exact observations, $L_i = R_i$. The independence of censoring mechanism is assumed. The likelihood is proportional to

$$L = \prod_{i=1}^{N} [G(L_i|\mathbf{x}_i) - G(R_i|\mathbf{x}_i)], \tag{2}$$

where $G(s|\mathbf{x}) = Pr(S > s|\mathbf{X} = \mathbf{x})$. It can be characterized the set of times $0 = s_0 < s_1 < ... < s_m = \infty$ from each of L_i and R_i. The contribution of the ith observation to the likelihood (2) can be expressed as

$$\sum_{j=1}^{m} \alpha_{ij}[G(s_{j-1}|\mathbf{x}_i) - G(s_j|\mathbf{x}_i),$$

$$\text{where } \alpha_{ij} = \begin{cases} 1 & \text{if } (s_{j-1}, s_j] \subseteq A_i \\ 0 & \text{otherwise} \end{cases}$$

Under the assumption of proportional hazards, the probability of surviving beyond time s_j for an subject with covariates $\mathbf{x}_i$ is $G(s_j|\mathbf{x}_i) = [G(s_j)]^{\exp \mathbf{x}_i \beta}$. With

reparametrization $\gamma_j = \log[-\log G(s_j)]$, it can be expressed

$$L = \sum_{i=1}^{N} \log \sum_{j=1}^{m} \alpha_{ij} \{\exp[-\exp(\mathbf{x}_i\beta + \gamma_{j-1}) - \exp[-\exp(\mathbf{x}_i\beta + \gamma_j)]\}. \tag{3}$$

Note that since $G_0 = 1(\gamma_0 = \infty)$ and $G_m = 0(\gamma_m = \infty)$, L is a function of γ_j and β for $j = 1, ..., m$. After calculating the first and second derivatives for the likelihood (3), we can estimate γ_j and β using a Newton-Raphson algorithm. When comparing the survival curves, a test for $\beta = 0$ is of interest. The score statistic for testing $\beta = 0$ is

$$U = \frac{\partial \log L}{\partial \beta}\bigg|_{\beta=0} = \sum_{j=1}^{m} \Big[\sum_{i \subseteq F'_j} \mathbf{x}_i w_{rij} \log \hat{p}_j - \sum_{i \subseteq D'_j} \mathbf{x}_i w_{dij} \log \hat{p}_j/(1-\hat{p}_i) \Big], \tag{4}$$

$$\text{where } w_{rij} = \sum_{k=j}^{m} \alpha_{ij}\hat{g}_k / \sum_{n} \alpha_{in}\hat{g}_n = \Pr(S > s_j | \mathbf{A}),$$

$$w_{dij} = \alpha_{ij}\hat{g}_j / \sum_{n} \alpha_{in}\hat{g}_n = \Pr(s_{j-1} < S \leq s_j | \mathbf{A}),$$

$$\hat{g}_j = \hat{G}_{j-1} - \hat{G}_j.$$

F'_j is pseudo risk set of all subjects who have a nonzero probability of being at risk at interval $(s_{j-1}, s_j]$. D'_j is a pseudo response set of all subjects who have a nonzero probability of having an event at interval $(s_{j-1}, s_j]$.

Kim et al. (2006) studied a log-rank type test that did not use an iterative algorithm. They proposed uniform weights algorithm, where a subject contributed uniformly to each mass point. s_j for $j = 1, ..., m$ consists of left and right endpoints which may have masses. Let $T_i \subset A_i$, where $A_i = (L_i, R_i]$, $\delta = 0$ or 1, $0 = s_0 < s_1 < ... < s_m$, $\alpha_{ij} = 1$ if $L_i < s_j \leq R_i$ or 0 otherwise. The pseudo-risk set at each point is

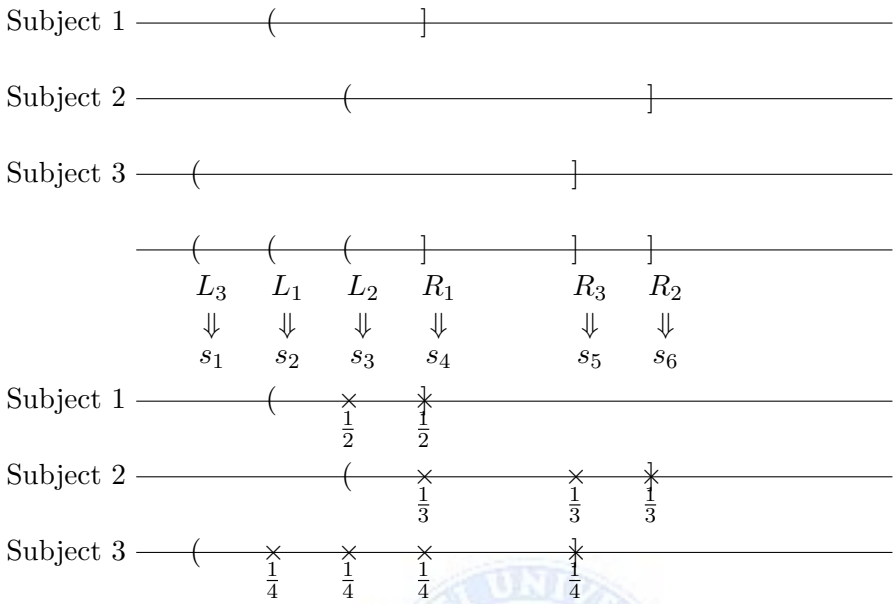

Figure 2: Example of uniform weights of three subjects.

$\frac{\sum_{k=j}^{m} \alpha_{ik}}{\sum_{h=1}^{m} \alpha_{ih}}$ while the pseudo-failure set at each point is $\frac{\delta_i \alpha_{ij}}{\sum_{h=1}^{m} \alpha_{ih}}$ for $\delta = 1$ if an event is occurred, or $\delta = 0$ if not. Figure 2 shows the example of uniform weight for 3 subjects. For example, there are 2 time points that may have a positive mass for subject 1. Therefore, the weight is $\frac{1}{2}$ for s_2 and s_3, respectively. The pseudo risk set for (s_2, s_3) of subject 1 are $(1, \frac{1}{2})$, respectively. The pseudo failure set for same time points of subject 1 is $(\frac{1}{2}, \frac{1}{2})$. With the pseudo-risk and pseudo-failure set, a log-rank type test statistic can be derived. Through simulation, the satisfactory type I error and power was shown.

2.1.3 Analysis for left truncated and interval-censored data

Let F be the cumulative distribution function of a real-valued random variable X. Turnbull (1976) characterized a set C of disjoint intervals $A_i = [L_i, R_i]$ in which

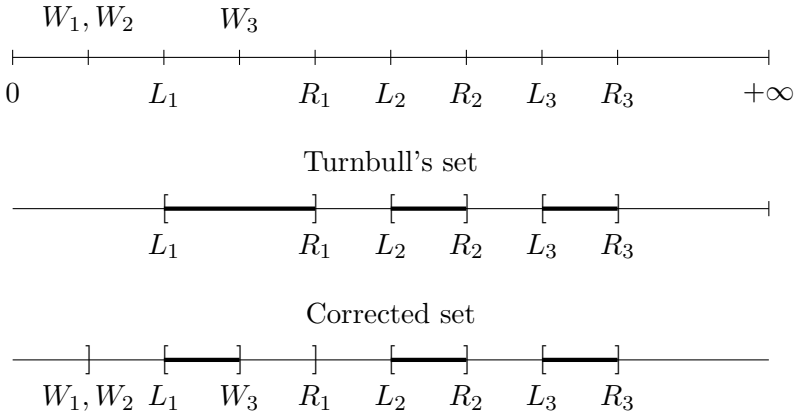

Figure 3: Example of characterizing sets of three subjects.

the NPMLE of F may have positive masses. The NPMLE can be obtained using the self-consistency algorithm. However, as noted by Frydman (1994), when there exist both truncation and censoring, the set C is not applicable. For example, we consider a small disjoint data set with three observations: $\{(A_i, B_i), 1 \leq i \leq 3\}$. $A_i = [L_i, R_i]$ with $L_1 < L_2 < L_3$. $B_i = (W_i, \infty)$ with $W_1 = W_2 = 0 < L_1 < W_3 < R_2$. The set C from Turnbull is simply $C = \cup_{i=3}^{3} A_i$. Please refer to Figure 3. The likelihood is $L(F) = \prod_{i=1}^{3} \{F(R_i+) - F(L_i-)\}/\{1 - F(W_3+)\}$. Note that $F(W_1) = F(W_2) = 0$. For fixed values of $F(R_i+)$ and $F(L_i-)$ the likelihood depends on the value of F at W_3. This violate Turnbull (1976)'s lemma 2 . The correct characterization of $\hat{F}$ the set C should be redefined as $C' = [L_1, W_3] \cup [L_2, R_2] \cup [L_3, R_3]$.

2.1.4 Multiple imputation techniques

In the survival analysis, the event times for interval-censored observations is unknown, so it can be regarded as missing by assuming independent censoring mechanisms. After imputation, the interval-censored data is reduced to right-censored

data. True variability may be underestimated when using simple imputation. In multiple imputations, the variance is adjusted to within-imputation covariance and between-imputation variance. Pan (2000a) studied multiple imputation method to estimate the coefficient of the cox regression with interval-censored data. They imputed m sets from the initial estimate of the cox proportional hazard model and baseline survival. Right-censored data was maintained without manipulation. With imputed exact time, a Cox model was fitted to obtain new estimates and the baseline survival. The iteration was repeated until estimates converged. Huang et al. (2008) studied a log-rank test via multiple imputation for interval censored data. After estimating the NPMLE by using Turnbull's algorithm, they imputed the exact time for all data points including right-censored data from the conditional probability of NPMLE. Unlike Pan (2000a), Huang et al. (2008) used a large imputation number (M=100). The covariance matrix estimator was formed of subtracting within-imputation covariance and between-imputation variance (Follmann et al., 2003). In rth imputation for $r = 1, ..., M$, U^r and V^r was obtained by a log-rank test from the imputed-data set (i.e. data is exact or right-censored).

$$\hat{V} = \frac{\sum_{r=1}^{M} V^r}{M} - \frac{\sum_{r=1}^{M} [U^r - \bar{U}][U^r - \bar{U}]^T}{M-1}, \text{where } \bar{U} = \frac{\sum_{r=1}^{M} U^r}{M}.$$

They showed through intensive simulations (100,000 replications) that the subtracting of within-imputation covariance and between-imputation variance is more suitable than adding that which was used by Zhao and Sun (2004). In a simulation result, Zhao and Sun (2004)'s variance was marginally overestimated while Huang et al. (2008)'s variance was slightly underestimated. In the current paper, we applied both forms of variance.

2.2 Notation and framework

For simplicity, we assumed that the IE is binary and only two treatment group exists. Let W and T be a positive real-valued random variable as the waiting time for an IE to occur and time to an event, respectively. Assume the binary random variable Z to be $Z = I\{X \leq T\}$. The random variables T_0 and T_1 are defined as the times to event conditional on $Z = 0$ or 1 respectively, i.e. $T = (1 - Z)T_0 + ZT_1$. T_0 means failure time was observed without an IE and T_1 means failure time was observed after an IE occurred. Survival time with an intermediate event implies that the survival time should exceed the waiting time for an IE. This reflects length bias phenomenon; i.e. an individual has to live long enough to have experienced the IE.

The density functions of W, T_0, T_1 are $g(x)$, $q_0(t)$, and $q_1(t)$ respectively. Also the survival distribution functions are $G(w) = Pr[W > w], Q_0(t) = Pr[T_0 > t]$ and $Q_1(t) = Pr(T_1 > t)$.

We further assumed that the failure time T is interval-censored. Therefore, for the ith subject, we did not observe T_i exactly but observed $T_i \in A_i$, where $A_i = (L_i, R_i]$ was the interval in which the failure occurred. Note that for right-censored observations, $R_i = \infty$, while for the exact observations, $L_i = R_i$. Let $\delta_i = 0$ or 1 depending on whether the ith subject is right-censored or not. For the model with $Z = 1$, it implies that the waiting time was observed before the failure time. Therefore T_1 is left truncated at the waiting time W and interval-censored. Let $\{B_i, 1 \leq i \leq N\}$ is the truncation sets, i.e. $B_i = (W_i, \infty)$.

We now characterized the set with all observed points including left-truncated data as mentioned by Frydman (1994). We considered the set of N independent

paired $\{A_i, B_i\}$. We assumed $A_i \subseteq B_i$. We characterized the following $\tilde{L}^k$ and $\tilde{R}^k$, where $k = 0, 1$. For the survival distribution of T_0, it implied that subjects observed an event before having an IE. In the case of having an IE, it was regarded as right censoring at W_i; i.e. $(W_i, \infty]$.

$\tilde{L^0} = \{L_i; 1 \leq i \leq N\} \cup \{W_i; 1 \leq i \leq N\}$

$\tilde{R^0} = \{R_i; 1 \leq i \leq N\} \cup \{\infty\}$

For survival distribution of T_1, a time point W_i, L_i and R_i of a subject who experienced an IE was included.

$\tilde{L^1} = \{L_i; 1 \leq i \leq N\}$

$\tilde{R^1} = \{R_i; 1 \leq i \leq N\} \cup \{W_i; 1 \leq i \leq N\} \cup \{\infty\}$

When an IE occurs, it implies that the survival distribution may change to T_1. Note that the initial survival distribution is $Q_0(t) = Pr(T_0 > t)$. When an IE occurs, the waiting time is a change point of distribution for survival. That is, for the distribution of T_0, the information of the event exceeding the waiting time can no longer be observed. Therefore, the waiting time is treated as righted censored for T_0. The event time exceeding the waiting time is not included in set $\tilde{L^0}$ and $\tilde{R^0}$. For the survival distribution having an IE, the waiting time W is a left-truncated time. The subject who does not experience an IE is not included in set $\tilde{L^1}$ and $\tilde{R^1}$. Therefore, the waiting time for an IE is included in $\tilde{L}^0$ for T_0 as right-censored at W, whereas it is included in $\tilde{R}^1$ for T_1.

The distinct endponits were set as C^k in which all the timepoints $\tilde{L^k}$ and $\tilde{R^k}$ were ordered and labeled $0 = s_0^k < s_1^k < ... < s_m^k = \infty$ for $i = 1, ..., n, j = 1, ..., m, k = 0, 1$. We assumed that the censoring and truncating mechanisms were independent of the failure times. Please see the Figure 4- 6.

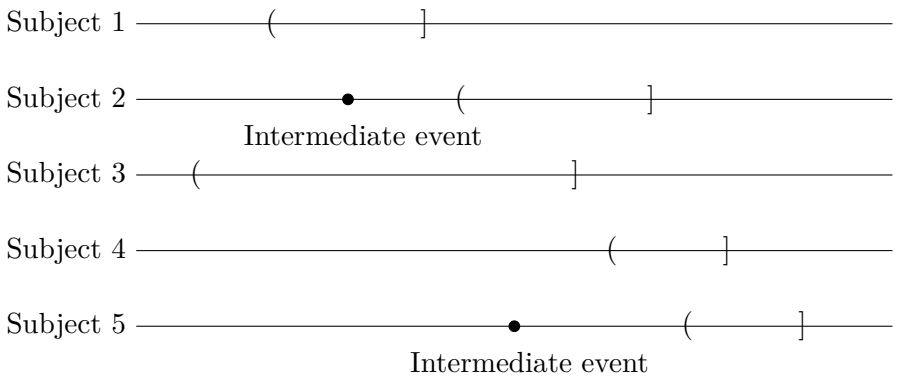

Figure 4: An example of 5 subjects with or without an intermediate clinical event.

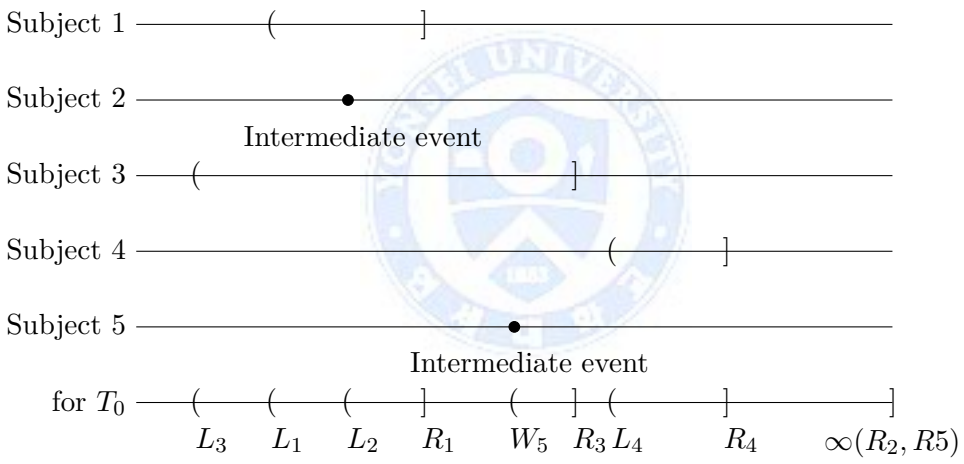

Figure 5: An example of constructing sets of 5 subjects for T0.

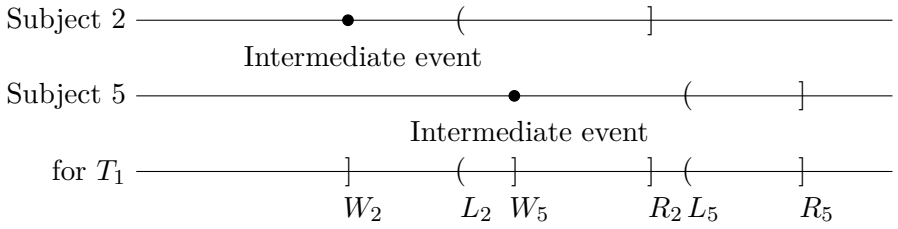

Figure 6: An example of constructing sets of 5 subjects for T1.

2.3 Proposed method

2.3.1 Likelihood with interval-censored data

The likelihood with an interval-censored outcome in the presence of an IE consisted of two parts according to Z. The likelihood of the observation is proportional to $L = \prod L_{1i} \prod L_{2i}$, where

$$\begin{cases} L_{1i} = \{(Q_0(L_i) - Q_0(R_i))G(L_i)^{1-\delta}G(R_i)^{\delta}\}^{(1-z)} & \text{for } z = 0 \\ L_{2i} = \left\{Q_0(W_i)g(W_i)\frac{Q_1(L_i)-Q_1(R_i)}{Q_1(W_i)}\right\}^{z} & \text{for } z = 1 \end{cases} \tag{5}$$

As mentioned before, $R_i = \infty$ for right-censored data. In the case of a subject who did not experience an IE ($z = 0$) but an event occurred ($\delta = 1$), we only know that an IE did not occur until R_i. If a subject who did not experience an IE and an event ($z = 0, \delta = 0$), the last observed time point was L_i where $R_i = \infty$. So the likelihood has a term of $G(L_i)^{1-\delta}$ and $G(R_i)^{\delta}$. When a subject experienced an IE at W_i, the subject survived without an IE until at that time. After experiencing an IE, a survival distribution of the subject was changed to Q_1.

The $Q_0(L_i) - Q_0(R_i)$ contribution of the ith observation to the likelihood can be expressed as $\sum_j \alpha_{ij}^k \{Q(s_{j-1}^k) - Q(s_j^k)\}$ (Finkelstein, 1986) and for left truncation, it can be expressed as $\sum_j \nu_{ij}\{Q(s_{j-1}^1) - Q(s_j^1)\}$ that we accommodate truncate time to the characterized set, where

$$\alpha_{ij}^k = \begin{cases} 1 & \text{if } L_i < s_j^k \leq R_i \\ 0 & \text{otherwise} \end{cases} \qquad \nu_{ij} = \begin{cases} 1 & \text{if } W_i < s_j^1 \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{for } i = 1, ..., n, j = 1, ..., m, k = 0, 1.$$

Thus, the log of likelihood (5) is expressed as

$$
\begin{aligned}
L_{1i} &= \left[G(L_i)^{1-\delta} G(R_i)^{\delta} \sum_j \alpha_{ij}^0 \{Q_0(s_{j-1}^0) - Q_0(s_j^0)\}\right]^{(1-z)} \\
l_1 &= \log \text{ of } \prod_i L_{1i} = \sum_i \log L_{1i} \\
&= (1-z) \sum_i \Big[\log \sum_j \alpha_{ij}^0 \{Q_0(s_{j-1}^0) - Q_0(s_j^0)\} \\
&\qquad + (1-\delta)\log G(L_i) + \delta \log G(R_i)\Big].
\end{aligned}
$$

Under proportional assumption, $Q_k(s_j^k) = Q_k^0(s_j^k)^{\exp(x_i\beta_k)}$ where Q_k^0 is baseline survival function at time s_j^k for k = 0, 1. For conveinence, Finkelstein reparametrization is used as $Q_k^0(s_j^k) = \exp(-\exp(\gamma_{kj}))$. Therefore, $\log Q_k(s_j^k) = \exp(x_i\beta_k)\log Q_k^0(s_j^k) = \exp(x_i\beta_k)(-\exp(\gamma_{kj})) = -\exp(x_i\beta_k + \gamma_{kj})$. The log of likelihood is rewritten as

$$
\begin{aligned}
l_1 &= \sum \log L_{1i} \\
&= (1-z) \sum_i \log \sum_j \alpha_{ij}^0 \{\exp(-\exp(x_i\beta_0 + \gamma_{0j-1})) - \exp(-\exp(x_i\beta_0 + \gamma_{0j}))\} \\
&\qquad + \text{terms not involve } \beta \text{ or } \gamma. \qquad (6)
\end{aligned}
$$

We now calculate the first and second derivatives for the likelihood function (6) to estimate observed Fisher information matrix. It closely follows the paper by Finkelstein (1986). To calculate derivatives of L_{1i}, we need the information below. We omit the k for convenience in calculating L_{1i} part because all time points are in

s_j^0 for $j = 1, ..., m$ at L_{1i} .

$$
\begin{aligned}
\text{let } u_{0j} &= -\exp(x_i\beta_0 + \gamma_{0j}) = \log Q_0(s_j) \\
\frac{\partial u_{0j}}{\partial \beta_0} &= -x_i \exp(x_i\beta_0 + \gamma_{0j}) = x_i \log Q_0(s_j) \\
\frac{\partial u_{0j}}{\partial \gamma_{0j}} &= -\exp(x_i\beta_0 + \gamma_{0j}) = \log Q_0(s_j) \\
\frac{\partial \exp(u_{0j})}{\partial \beta_0} &= \frac{\partial u_{0j}}{\partial \beta_0}\exp(u_{0j}) = \exp(u_{0j}) x_i \log Q_0(s_j) = x_i Q_0(s_j)\log Q_0(s_j) = x_i h_{ij} \\
\frac{\partial \exp(u_{0j})}{\partial \gamma_{0j}} &= Q_0(s_j)\log Q_0(s_j) = h_{ij} \\
\frac{\partial \sum \alpha_{ij} g_{ij}}{\partial \beta_0} &= \sum \alpha_{ij}(x_i Q_0(s_{j-1})\log Q_0(s_{j-1}) - x_i Q_0(s_j)\log Q_0(s_j)) \\
&= \alpha_{ij}[h_{ij-1} - h_j]x_i \\
\frac{\partial \sum \alpha_{ij} g_{ij}}{\partial \gamma_{0j}} &= -\alpha_{ij} Q_0(s_j)\log Q_0(s_j) + \alpha_{ij+1} Q_0(s_j)\log Q_0(s_j) \\
&= (\alpha_{ij+1} - \alpha_{ij})h_{ij} \\
\text{where } g_{ij} &= Q_0(s_{j-1}) - Q_0(s_j), h_{ij} = Q_0(s_j)\log Q_0(s_j) (= u_{0j}\exp u_{0j})
\end{aligned}
$$

First derivatives can be calculated by differentiation β_0 and γ_{0j}.

$$
\begin{aligned}
\frac{\partial l_1}{\partial \beta_0} &= (1-z)\sum_i \frac{1}{\sum_j \alpha_{ij} g_{ij}} \sum_j \alpha_{ij}[h_{ij-1} - h_{ij}]x_i \\
\frac{\partial l_1}{\partial \gamma_{0j}} &= (1-z)\sum_i \frac{(\alpha_{ij+1} - \alpha_{ij})h_{ij}}{\sum_j \alpha_{ij} g_{ij}} = (1-z)\sum_i \mu_{ij} h_{ij} \\
\text{where } \mu_{ij} &= \frac{\alpha_{ij+1} - \alpha_{ij}}{\sum_j \alpha_{ij} g_{ij}}.
\end{aligned}
$$

To calculate the second derivatives

$$\begin{aligned}
\frac{\partial h_{ij}}{\partial \beta_0} &= \frac{\partial (u_{0j}\exp(u_{0j}))}{\partial \beta_0} = \frac{\partial u_{0j}}{\partial \beta_0}\exp(u_{0j}) + u_{0j}\frac{\partial u_{0j}}{\partial \beta_0}\exp(u_{0j}) \\
&= x_i Q_0(s_j)\log Q_0(s_j) + x_i(\log Q_0(s_j))^2 Q_0(s_j) \\
&= x_i b_{ij} \\
\frac{\partial h_{ij}}{\partial \gamma_{0j}} &= \frac{\partial u_{0j}}{\partial \gamma_{0j}}\exp u_{0j} + u_{0j}\frac{\partial u_{0j}}{\partial \gamma_{0j}}\exp u_{0j} \\
&= Q_0(s_j)\log Q_0(s_j) + (\log Q_0(s_j))^2 Q_0(s_j) = b_{ij} \\
\frac{\partial \mu_{ij}}{\partial \beta_0} &= -\frac{\alpha_{ij+1}-\alpha_{ij}}{(\sum_j \alpha_{ij} g_{ij})^2}\frac{\partial \sum \alpha_{ij} g_{ij}}{\partial \beta_0} \\
&= -\frac{\alpha_{ij+1}-\alpha_{ij}}{(\sum_j \alpha_{ij} g_{ij})^2}\sum_j \alpha_{ij}(h_{ij-1}-h_{ij})x_i \\
\frac{\partial \mu_{ij}}{\partial \gamma_{0j}} &= \frac{\alpha_{ij+1}-\alpha_{ij}}{(\sum_j \alpha_{ij} g_{ij})^2}\frac{\partial \sum \alpha_{ij} g_{ij}}{\partial \gamma_{0j}} \\
&= -\frac{\alpha_{ij+1}-\alpha_{ij}}{(\sum_j \alpha_{ij} g_{ij})^2}(\alpha_{ij+1}-\alpha_{ij})h_{ij} = \frac{(\alpha_{ij+1}-\alpha_{ij})^2}{(\sum_j \alpha_{ij} g_{ij})^2}h_{ij} = -\mu_{ij}^2 h_{ij}
\end{aligned}$$

Second derivatives can be written as

$$\begin{aligned}
\frac{\partial^2 l_1}{\partial \gamma_{0j}^2} &= (1-z)\sum_i \big(-\mu_{ij}^2 h_{ij}^2 + \mu_{ij} b_{ij}\big) \\
\frac{\partial^2 l_1}{\partial \beta_0 \partial \gamma_{0j}} &= (1-z)\sum_i \left\{ \big(-\frac{\alpha_{ij+1}-\alpha_{ij}}{(\sum_j \alpha_{ij} g_{ij})^2}\sum_j \alpha_{ij}(h_{ij-1}-h_{ij})x_i\big)h_{ij} + \mu_{ij} x_i b_{ij} \right\} \\
&= (1-z)\sum_i x_i \left\{ \mu_{ij} b_{ij} - \mu_{ij} h_{ij}\frac{\sum_j \alpha_{ij}(h_{ij-1}-h_{ij})}{\sum_j \alpha_{ij} g_{ij}} \right\} \\
\frac{\partial^2 l_1}{\partial \beta_0^2} &= (1-z)\sum_i x_i^2 \left\{ \ -\frac{\sum_j \alpha_{ij}(b_{ij-1}-b_{ij})}{\sum_j \alpha_{ij} g_{ij}} + \left[\frac{\sum_j \alpha_{ij}(h_{ij-1}-hij)}{\sum_j \alpha_{ij} g_{ij}}\right]^2 \right\}
\end{aligned}$$

Then we can establish the observed Fisher information matrix $A_k = \begin{bmatrix} A_{k,11} & A_{k,12} \\ A_{k,21} & A_{k,22} \end{bmatrix}$, where $A_{k,11} = -\partial^2 L/\partial\gamma^2, A_{k,12} = A_{k,21} = -\partial^2 L/\partial\gamma\partial\beta, A_{k,22} = -\partial^2 L/\partial\beta^2$.

For L_{2i}, the log of likelihood is expressed as

$$\begin{aligned} l_2 &= \sum_i \log L_{2i} \\ &= z \sum_i \big[\log \sum_j \nu_{ij} \{ \exp(-\exp(x_i\beta_0 + \gamma_{0j-1})) - \exp(-\exp(x_i\beta_0 + \gamma_{0j})) \\ &\quad - \log \sum_j \nu_{ij} \{ \exp(-\exp(x_i\beta_1 + \gamma_{1j-1})) - \exp(-\exp(x_i\beta_1 + \gamma_{1j})) \\ &\quad + \log \sum_j \alpha_{ij} \{ \exp(-\exp(x_i\beta_1 + \gamma_{1j-1})) - \exp(-\exp(x_i\beta_1 + \gamma_{1j})) \} \\ &\quad + \text{terms not involve } \beta \text{ or } \gamma. \end{aligned} \tag{7}$$

The first derivatives of L_{2i} are below.

$$\begin{aligned}
\frac{\partial l_2}{\partial \beta_0} &= z \sum_i \frac{1}{\sum_j \nu_{ij} g^0_{ij}} \sum_j \nu_{ij}[h^0_{ij-1} - h^0_j] x_i \\
\frac{\partial l_2}{\partial \gamma_{0j}} &= z \sum_i \frac{(\nu_{ij+1} - \nu_{ij}) h^0_{ij}}{\sum_j \nu_{ij} g^0_{ij}} = z \sum_i \mu^0_{ij}(\nu) h^0_{ij} \\
\frac{\partial l_2}{\partial \beta_1} &= z \sum_i \left[\frac{1}{\sum_j \alpha^1_{ij} g^1_{ij}} \sum_j \alpha^1_{ij}[h^1_{ij-1} - h^1_j] x_i \right. \\
&\quad \left. - \frac{1}{\sum_j \nu_{ij} g^1_{ij}} \sum_j \nu_{ij}[h^1_{ij-1} - h^1_{ij}] x_i \right] \\
\frac{\partial l_2}{\partial \gamma_{1j}} &= z \sum_i \{\mu^1_{ij}(\alpha) h^1_{ij} - \mu^1_{ij}(\nu) h^1_{ij})\} \\
\text{where } h^k_{ij} &= Q_k(s^k_j) \log Q_k(s^k_j), g^k_{ij} = Q_k(s^k_{j-1}) - Q_k(s^k_j) \\
\mu^k_{ij}(parameter) &= \frac{parameter_{ij+1} - parameter_{ij}}{\sum_j parameter_{ij} g^k_{ij}}, parameter = \alpha \text{ or } \nu.
\end{aligned}$$

The second derivatives of L_{2i} are added $\log Q_0(W_i)$ or $-\log Q_1(W_i)$ term from l_1. The score statistics are,

$$\begin{aligned}
S_0 = \left.\frac{\partial l}{\partial \beta_0}\right|_{\beta_0=0} &= (1-z) \sum_{i=1}^{N} \frac{1}{\sum_j \alpha^0_{ij} g^0_{ij}} \sum_j \alpha^0_{ij}[h^0_{ij-1} - h^0_{ij}] x_i \\
&\quad + z \sum_{i=1}^{N} \frac{1}{\sum_j \nu_{ij} g^0_{ij}} \sum_j \nu_{ij}[h^0_{ij-1} - h^0_j] x_i \\
S_1 = \left.\frac{\partial l}{\partial \beta_1}\right|_{\beta_1=0} &= z \sum_i \left[\frac{1}{\sum_j \alpha^1_{ij} g^1_{ij}} \sum_j \alpha^1_{ij}[h^1_{ij-1} - h^1_j] x_i \right. \\
&\quad \left. - \frac{1}{\sum_j \nu_{ij} g^1_{ij}} \sum_j \nu_{ij}[h^1_{ij-1} - h^1_{ij}] x_i \right] \qquad (8)
\end{aligned}$$

For $Z = 0$, we characterize the set C^0 that the waiting time W_i is regarded as right-censored at that time. So we can rewrite ν_{ij} as α^0_{ij}. Then we can reduce S_0

as $\sum_{i=1}^{N} \frac{\sum_j \alpha_{ij}[h^0_{ij-1} - h^0_{ij}]x_i}{\sum_j \alpha_{ij} g^0_{ij}}$. For $Z = 1$, we have both interval-censored and left truncated terms. We developed the likelihood and score test based on the methods of Finkelstein (1986). However, it is not easy to estimate the variance matrix in some conditions. Finkelstein (1986) used a discrete baseline survival, and the estimation was based on a full likelihood using the proportional hazards model. The number of parameters could increase with the number of event times, rendering numerically unstable optimization. Kim et al. (2006), Huang et al. (2008) were studied to compare several methods for interval-censored data through simulations, the results showed that the Finkelstein method did not satisfy the nominal level for some conditions. Kim et al. (2006) showed that Finkelstein's method overestimated the nominal level 0.05 for most configurations (They had 54 configurations varying shape parameter of a Weibull distribution, censoring fraction, and depth for $n = 50$ or $n = 100$). Similar result was observed in Huang et al. (2008). They showed that estimated size of the Finkelstein test was biased over 0.05 for various censoring fraction conditions with 100,000 replications ($n = 50$ or 100). The current study focuses on comparing two survival curves in the presence of an IE. Here, we did not use Finkelstein's form directly, and move to another simple way.

2.3.2 Multiple imputation

MI converts interval-censored data to right-censored data so standard methods can be applied. This method can simplify the complicated situation and some non-iterative methods have been suggested (Pan, 2000b, Kim et al., 2006, Huang et al., 2008). We considered uniform weight method and weighted weight method. The uniform method closely followed that of Kim et al. (2006) who proposed a log-rank type test using uniform assumptions with controlling type I error, and having

acceptable power. The proposed method did not need intensive computation and was easy to implement. The weighted method closely followed Huang et al. (2008) with accomodating left truncation. After imputation, the proposed score statistics by Nam and Zelen (2001) was used.

2.3.3 Uniform weight method

Kim et al. (2006) assumed the true failure time of a subject may be uniformly distributed over $\{s_j, L_i < s_j \leq R_i, \text{for } j = 1, ..., m\}$ for each subject. They calculated a pseudo-risk and failure set based on uniform weights. They used MI techniques to estimate a variance matrix. In the current paper, we used MI techniques for the whole process includiing imputing a true failure time under the same assumption. We use a moderate imputation number (M=10) as Pan (2000a) recommended.

Step0: Set $r = 1$, where r denotes an imputation number.

Step1. Charaterize the set C^k with $\tilde{L^k}, \tilde{R^k}$ for each of T_k for $k = 0, 1$.

$\tilde{L^0} = \{L_i; 1 \leq i \leq N\} \cup \{W_i; 1 \leq i \leq N\}$, $\tilde{R^0} = \{R_i; 1 \leq i \leq N\} \cup \{\infty\}$

$\tilde{L^1} = \{L_i; 1 \leq i \leq N\}$, $\tilde{R^1} = \{R_i; 1 \leq i \leq N\} \cup \{W_i; 1 \leq i \leq N\} \cup \{\infty\}$.

The distinct endponits set C^k in which all the timepoints $\tilde{L^k}$ and $\tilde{R^k}$ are ordered and labeled $0 = s_0^k < s_1^k < ... < s_m^k = \infty$ for $i = 1, ..., n, j = 1, ..., m, k = 0, 1$.

Step2: If the ith observation is interval-censored, generate a value randomly sampled from a set $C^k = \{s_j^k, L_i < s_j^k \leq R_i, \text{ for } j = 1, ..., m\}$. Note that after imputed exact time, $T_0^{(r)}$ is right-censored data while $T_1^{(r)}$ is left-truncated and right-censored data. For making $T_0^{(r)}$, we censored the data at W_i if $Z_i = 1$. For making $T_1^{(r)}$, we only

use the data with $Z_i = 1$.

$$T_0^{(r)} = \begin{cases} T_{0i}^{(r)} = L_i & \text{if } \delta_i = 0, Z_i = 0 \\ T_{0i}^{(r)} = W_i, \delta_i = 0 & \text{if } Z_i = 1 \\ T_{0i}^{(r)} = \text{random sample from the set} & \\ \qquad \{s_j^0, L_i < s_j^0 \leq R_i, \text{ for } j = 1, ..., m\} & \text{if } \delta_i = 1, Z_i = 0 \end{cases}$$

$$T_1^{(r)} = \begin{cases} T_{1i}^{(r)} = L_i & \text{if } \delta_i = 0, Z_i = 1 \\ T_{1i}^{(r)} = \text{random sample from the set} & \\ \qquad \{s_j^1, L_i < s_j^1 \leq R_i, \text{ for } j = 1, ..., m\} & \text{if } \delta_i = 1, Z_i = 1 \end{cases}$$

Step3. Based on the rth imputed (left-truncated) right-censored data, compute the Nam and Zelen (2001) statistics and its covariance $S_k^{(r)}, V(\hat{S}_k)^{(r)}$ for $r = 1, ..., M, k = 0, 1$.

Step4. Repeate Step2 to Step3 M(> 0) times and obtain M paired of $(S_k^{(r)}, V(\hat{S}_k)^{(r)})$, where $r = 1, ..., M, k = 0, 1$.

Step5: Compute the sum of the average within-imputation covariance associated with S_k and the between-imputation variance of S_k.

$$\bar{S_k} = \frac{1}{M} \sum_{r=1}^{M} S_k^{(r)},$$

$$V(\hat{S}_k)_{mi,a} = \frac{1}{M} \sum_{r=1}^{M} \hat{V}_{S_k}^{(r)} + (1 + \frac{1}{M}) \frac{1}{M-1} \sum_{r=1}^{M} (S_k^{(r)} - \bar{S_k})^2$$

Here, we applied two types of variances. One added within- and between variance and the other subtracted two variances as in Huang et al. (2008). The first term

can be seen above, whereas the second term is formed as

$$V(\hat{S}_k)_{mi,b} = \frac{1}{M}\sum_{r=1}^{M}\hat{V}_{S_k}^{(r)} - \frac{1}{M-1}\sum_{r=1}^{M}(S_k^{(r)} - \bar{S}_k)^2.$$

Thus, we can test H_0 based on

$$\chi_2^2 = \bar{S_0}^2/V(\hat{S}_0)_{mi,l} + \bar{S_1}^2/V(\hat{S}_1)_{mi,l} \quad \text{for } l = a, b.$$

2.3.4 Weighted weight method based on NPMLE

We proposed another weighted method based on NPMLE. We estimated the NPMLE from the original data set by Turnbull's algorithm and used the NPMLE as weights for the imputation. The data was LTIC when having an IE, therefore we characterized the set that may have a positive mass including truncated points same as above method.

Step 1. Estimate the NPMLE from the original data set.

Step 2. Using the NPMLE as weight, impute the data conditional on $\{L_i < T_i^{(r)} \leq R_i\}$.

$$\begin{cases} T_{0i}^{(r)} = L_i & \text{if } \delta_i = 0, z_i = 0 \\ T_{0i}^{(r)} = W_i, \delta_i = 0 & \text{if } z_i = 1 \\ T_{0i}^{(r)} = \text{random sampling from the distribution NPMLE} \end{cases}$$

$$\begin{cases} T_{1i}^{(r)} = L_i & \text{if } \delta_i = 0, z_i = 1 \\ T_{1i}^{(r)} = \text{random sampling from the distribution NPMLE} \end{cases}$$

Step 3. Based on the rth imputed (left-truncated) right-censored data, we can calculate the average Nam and Zelen (2001) statistics and variance as uniform weight method.

3 Simulation studies

3.1 Data generation

The IE such as a heart transplant or the start of second-therapy can be observed exactly. Hence, we can assume that the waiting time W for an IE is exactly observed while the failure time T such as recurrence or progression is interval-censored. At first, we generate the true failure time T_0 and waiting time W from a survival distribution below;

$G_l(w) = e^{-\mu_l w}, Q_{0l}(t_0) = e^{-\lambda_{0l} t}$ for $l = A, B$.

Note that the probability of experiencing an IE is $\theta_l = \frac{\mu_l}{\mu_l + \lambda_{0l}}$. If $W > T_0$, then $T = T_0$. If $W \leq T_0$, generate a random variable T_1 from the truncated probability distribution function $q_1(t)/Q_1(w)$ with $W \leq T_1$, where $Q_{1t}(t) = e^{-\lambda_{1l} t}$ for $l = A, B$. Therfore, T_1 should be larger than W, so we can generate $Q_1(t) \sim U(0, Q_1(W))$. The value of λ_{1l} was chosen from the mean time to failure, $m_1 = \{1, 1.25, 1.5, 2\}$. In our simulations, $\theta_A = 0.5, \theta_B = \{0.3, 0.4, 0.5\}, \lambda_{0A} = \lambda_{0B} = 1, m_{1A} = 1, m_{1B} = \{1, 2\}$.

The first examination time E was $Uniform(0, \psi)$. For a subject having an IE, the first examination time E was equal to or greater than the waiting time $W (E \sim Uniform(W, W + \psi))$. The length of the time interval between two follow-up examinations was assumed as a constant, len = 0.5. We fixed ψ as the same as len. If we have p examinations, survival time T_i is accordingly observed in one of intervals $(0, E_i], (E_i, E_i + len), ..., (E_i + p * len, \infty)$. Here, we did not restrict the number of follow-up visits, because a subject having an IE should survival at the waiting time and has more chance to follow up for longer. We assumed that every subject visited at the first examination time, E. After that, there is a probability that a subject might not comply with the follow-up visits. For a visit that may be

Table 1: Design parameters and their values included in the simulations.

Parameters	Values	Description
θ_B	0.3, 0.4, 0.5	The probability of experiencing an IE
μ_l	1, 3/7, 2/3	For the waiting time W
λ_0	1	For the true failure time T_0
m_1, l	1,1.25, 1.5, 2	The mean time to failure of T_1
len	0.5	The visit interval
d	(0, 0), (0.1, 0.2)	Follow-up missing rate at (first year, therafter)
c_p	0, 0.3	Censoring fraction
n	50, 100, 200	Sample size for each group

missing we considered two settings. One was each subject would not miss any of the follow-up visits. The other was a subject might miss any of the follow-up visits and was more likely to miss later visits (i.e. 0.1 for first years, and then 0.2 thereafter).

Supposing that a censoring indicator δ having 0 or 1 is generated from the Bernoulli distribution with a success probability c_p. c_p is set as 0 or 0.3. If $\delta = 0$, the observation on T_i is right-censored and $\delta = 1$, the observation on T_i is observed on $(L_i, R_i]$. For right-censored data ($\delta = 0$), we set L as it is but R to be infinity. For $\delta = 1$, we maintained the endpoints as they were. The sample sizes were chosen as 50, 100 and 200 for each group. The results reported were based on 1000 replications for each scenario. All design parameters and their values included in simulation are listed in Table 1.

For comparison, we included the log-rank test and stratified log-rank test (the stratum is having an IE or not) along with our proposed tests. For the log-rank and stratified log-rank test, the true failure times were used rather than the interval-censored ones. We used two variance forms, which are formed by (1) adding and (2) subtracting of within- and between variance.

3.2 Results

The results of the simulations are summarized from Tables 2 to 9. Table 2 to 5 shows the estimate of the upper 5% of each of the five tests under the null hypothesis, whereas Tables 6 to 9 shows the power under the alternative hypothesis for each scenario. The proposed methods showed the appropriate 5% significant level under all scenarios. For the variance with adding form (1), the methods marginally overestimated the variance so the effect sizes were less than 0.05 for most of scenarios. For the variance with subtracting form (2), the methods slightly underestimated the variance.

When the proportion of an IE is different between two groups (i.e. θ_A is not equal to θ_B.), the stratified log-rank test is unsatisfactory. Even though the survival distribution is similar between two groups and only the proportion of an IE is different, the stratified log-rank test determined incorrectly that the survival distribution were different in many cases. The log-rank test satisfies the nominal significance level when the survival functions are not changed after experiencing an IE regardless of the proportion. When the survival distribution is changed after experiencing an IE (i.e., λ_{0A} is not equal to λ_{1A}.) in addition to the difference proportion of an IE, however, the log-rank test is not appropriate. The comparison of the uniform and weighted weight multiple imputation methods, did not show considerable differences.

When $\theta_A = \theta_B = 0.5$, the simulations in Table 2 confirmed that all tests gave the correct 5% significance level. Hence, the power calculations were restricted to this case. The value of the other parameters was: $\lambda_{0A} = \lambda_{0B} = 1, m_{1A} = 2$. Only the mean time to failure was changed for m_{2B}. When the sample size is increased, the

value of censoring fraction c_p is decreased, or the difference of mean time to failure is increased, it is expected that the power of the tests can be improved. In all cases, the proposed methods have superior power by taking advantage of the knowledge of the intermediate clinical event.

In the appendices, the power of proposed methods was shown, which had acceptable nominal levels for $\theta_A \neq \theta_B$(Table 10-13). We also simulated larger imputation numbers (M=100) with same scenarios and it showed similar results with moderate imputation numbers (M=10) (Tables 14-25).

Table 2: Empirical 5% level tests by varying θ_B, m_{1A}, and m_{1B} with $\theta_A = 0.5$ when all events are observed in some intervals and there are no missed visits.

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.056	0.049	0.053	0.058	0.055	0.058
(0.5, 0.5)	(1, 1)	(1, 1)	0.061	0.047	0.044	0.051	0.047	0.051
(0.5, 0.4)	(1, 1)	(2, 2)	0.055	0.093	0.044	0.047	0.046	0.050
(0.5, 0.4)	(1, 1)	(1, 1)	0.066	0.111	0.048	0.060	0.051	0.058
(0.5, 0.3)	(1, 1)	(2, 2)	0.099	0.213	0.041	0.052	0.042	0.048
(0.5, 0.3)	(1, 1)	(1, 1)	0.045	0.245	0.044	0.054	0.045	0.054
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.057	0.058	0.052	0.058	0.052	0.056
(0.5, 0.5)	(1, 1)	(1, 1)	0.058	0.056	0.045	0.051	0.044	0.055
(0.5, 0.4)	(1, 1)	(2, 2)	0.076	0.133	0.048	0.054	0.049	0.054
(0.5, 0.4)	(1, 1)	(1, 1)	0.043	0.143	0.045	0.050	0.045	0.054
(0.5, 0.3)	(1, 1)	(2, 2)	0.139	0.365	0.049	0.052	0.047	0.054
(0.5, 0.3)	(1, 1)	(1, 1)	0.048	0.460	0.044	0.051	0.040	0.055
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.055	0.048	0.049	0.055	0.048	0.053
(0.5, 0.5)	(1, 1)	(1, 1)	0.043	0.052	0.049	0.058	0.050	0.057
(0.5, 0.4)	(1, 1)	(2, 2)	0.092	0.238	0.044	0.049	0.043	0.050
(0.5, 0.4)	(1, 1)	(1, 1)	0.053	0.293	0.044	0.050	0.044	0.055
(0.5, 0.3)	(1, 1)	(2, 2)	0.210	0.646	0.047	0.051	0.048	0.056
(0.5, 0.3)	(1, 1)	(1, 1)	0.056	0.726	0.048	0.050	0.043	0.048

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 3: Empirical 5% level tests by varying θ_B, m_{1A}, and m_{1B} with $\theta_A = 0.5$ when all events are observed in some intervals and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter.

(θ_A, θ_B)	(λ_A, λ_B)	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.054	0.058	0.048	0.052	0.044	0.056
(0.5, 0.5)	(1, 1)	(1, 1)	0.055	0.050	0.042	0.052	0.044	0.053
(0.5, 0.4)	(1, 1)	(2, 2)	0.073	0.105	0.045	0.051	0.045	0.056
(0.5, 0.4)	(1, 1)	(1, 1)	0.060	0.124	0.042	0.058	0.042	0.060
(0.5, 0.3)	(1, 1)	(2, 2)	0.098	0.212	0.048	0.059	0.044	0.057
(0.5, 0.3)	(1, 1)	(1, 1)	0.057	0.236	0.046	0.057	0.047	0.055
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.051	0.048	0.051	0.058	0.052	0.058
(0.5, 0.5)	(1, 1)	(1, 1)	0.053	0.067	0.040	0.046	0.041	0.046
(0.5, 0.4)	(1, 1)	(2, 2)	0.069	0.148	0.044	0.049	0.046	0.049
(0.5, 0.4)	(1, 1)	(1, 1)	0.047	0.173	0.040	0.045	0.040	0.050
(0.5, 0.3)	(1, 1)	(2, 2)	0.137	0.372	0.049	0.056	0.050	0.060
(0.5, 0.3)	(1, 1)	(1, 1)	0.049	0.462	0.042	0.060	0.046	0.062
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.059	0.057	0.054	0.060	0.056	0.057
(0.5, 0.5)	(1, 1)	(1, 1)	0.055	0.042	0.042	0.049	0.043	0.056
(0.5, 0.4)	(1, 1)	(2, 2)	0.096	0.221	0.054	0.058	0.054	0.062
(0.5, 0.4)	(1, 1)	(1, 1)	0.061	0.282	0.045	0.053	0.044	0.052
(0.5, 0.3)	(1, 1)	(2, 2)	0.232	0.621	0.051	0.056	0.050	0.056
(0.5, 0.3)	(1, 1)	(1, 1)	0.053	0.747	0.045	0.051	0.043	0.052

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 4: Empirical 5% level tests by varying θ_B, m_{1A}, and m_{1B} with $\theta_A = 0.5$ when censoring fraction is 0.3, but there are no missed visits.

(θ_A, θ_B)	(λ_A, λ_B)	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.055	0.054	0.050	0.056	0.050	0.054
(0.5, 0.5)	(1, 1)	(1, 1)	0.057	0.064	0.043	0.045	0.042	0.044
(0.5, 0.4)	(1, 1)	(2, 2)	0.057	0.085	0.053	0.055	0.051	0.054
(0.5, 0.4)	(1, 1)	(1, 1)	0.047	0.091	0.052	0.056	0.050	0.056
(0.5, 0.3)	(1, 1)	(2, 2)	0.085	0.181	0.048	0.055	0.052	0.058
(0.5, 0.3)	(1, 1)	(1, 1)	0.047	0.200	0.043	0.048	0.044	0.049
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.048	0.045	0.048	0.054	0.049	0.055
(0.5, 0.5)	(1, 1)	(1, 1)	0.058	0.056	0.047	0.055	0.048	0.055
(0.5, 0.4)	(1, 1)	(2, 2)	0.069	0.121	0.050	0.052	0.048	0.051
(0.5, 0.4)	(1, 1)	(1, 1)	0.049	0.134	0.044	0.049	0.044	0.050
(0.5, 0.3)	(1, 1)	(2, 2)	0.103	0.283	0.055	0.060	0.056	0.058
(0.5, 0.3)	(1, 1)	(1, 1)	0.040	0.341	0.043	0.050	0.044	0.047
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.046	0.051	0.047	0.050	0.048	0.049
(0.5, 0.5)	(1, 1)	(1, 1)	0.047	0.051	0.048	0.052	0.048	0.055
(0.5, 0.4)	(1, 1)	(2, 2)	0.081	0.193	0.041	0.045	0.045	0.047
(0.5, 0.4)	(1, 1)	(1, 1)	0.055	0.229	0.049	0.057	0.053	0.057
(0.5, 0.3)	(1, 1)	(2, 2)	0.166	0.494	0.042	0.048	0.048	0.050
(0.5, 0.3)	(1, 1)	(1, 1)	0.051	0.607	0.040	0.046	0.042	0.046

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 5: Empirical 5% level tests by varying θ_B, m_{1A}, and m_{1B} with $\theta_A = 0.5$ when censoring fraction is 0.3 and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter.

(θ_A, θ_B)	(λ_A, λ_B)	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.050	0.056	0.049	0.055	0.045	0.055
(0.5, 0.5)	(1, 1)	(1, 1)	0.065	0.060	0.044	0.058	0.043	0.055
(0.5, 0.4)	(1, 1)	(2, 2)	0.058	0.100	0.051	0.060	0.049	0.062
(0.5, 0.4)	(1, 1)	(1, 1)	0.052	0.090	0.042	0.053	0.048	0.053
(0.5, 0.3)	(1, 1)	(2, 2)	0.079	0.162	0.049	0.054	0.052	0.055
(0.5, 0.3)	(1, 1)	(1, 1)	0.047	0.200	0.048	0.058	0.043	0.054
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.052	0.055	0.045	0.049	0.048	0.051
(0.5, 0.5)	(1, 1)	(1, 1)	0.044	0.052	0.044	0.054	0.044	0.054
(0.5, 0.4)	(1, 1)	(2, 2)	0.075	0.105	0.052	0.056	0.053	0.057
(0.5, 0.4)	(1, 1)	(1, 1)	0.052	0.133	0.045	0.060	0.049	0.060
(0.5, 0.3)	(1, 1)	(2, 2)	0.110	0.258	0.046	0.058	0.046	0.054
(0.5, 0.3)	(1, 1)	(1, 1)	0.052	0.336	0.041	0.052	0.042	0.051
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.059	0.059	0.042	0.047	0.045	0.048
(0.5, 0.5)	(1, 1)	(1, 1)	0.050	0.054	0.052	0.059	0.050	0.056
(0.5, 0.4)	(1, 1)	(2, 2)	0.078	0.180	0.048	0.054	0.050	0.053
(0.5, 0.4)	(1, 1)	(1, 1)	0.057	0.219	0.044	0.050	0.043	0.051
(0.5, 0.3)	(1, 1)	(2, 2)	0.168	0.485	0.047	0.051	0.050	0.052
(0.5, 0.3)	(1, 1)	(1, 1)	0.060	0.582	0.040	0.049	0.043	0.050

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 6: Empirical power of tests by varying m_{1B} when all events are observed in some intervals and there are no missed visits.

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.106	0.101	0.113	0.117	0.113	0.124
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.196	0.183	0.246	0.261	0.247	0.264
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.367	0.298	0.514	0.526	0.514	0.525
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.169	0.156	0.197	0.206	0.203	0.209
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.363	0.315	0.497	0.515	0.498	0.509
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.645	0.557	0.838	0.851	0.837	0.850
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.286	0.254	0.386	0.398	0.385	0.400
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.662	0.562	0.803	0.809	0.801	0.809
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.932	0.858	0.990	0.990	0.990	0.991

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 7: Empirical power of tests by varying m_{1B} when all events are observed in some intervals and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter.

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.120	0.108	0.111	0.136	0.110	0.128
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.222	0.181	0.250	0.283	0.245	0.281
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.386	0.320	0.480	0.513	0.484	0.509
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.181	0.146	0.201	0.214	0.204	0.216
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.373	0.315	0.471	0.501	0.474	0.505
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.647	0.564	0.824	0.841	0.826	0.841
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.310	0.289	0.364	0.387	0.360	0.384
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.652	0.575	0.808	0.821	0.812	0.821
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.925	0.860	0.991	0.991	0.990	0.991

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 8: Empirical power of tests by varying m_{1B} when censoring fraction is 0.3, but there are no missed visits.

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.077	0.099	0.120	0.126	0.118	0.124
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.157	0.158	0.214	0.226	0.225	0.223
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.268	0.249	0.405	0.425	0.409	0.425
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.127	0.117	0.165	0.172	0.165	0.174
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.271	0.231	0.391	0.405	0.393	0.405
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.493	0.421	0.703	0.713	0.705	0.711
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.243	0.194	0.311	0.322	0.311	0.320
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.500	0.420	0.707	0.713	0.709	0.713
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.797	0.677	0.950	0.955	0.951	0.954

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 9: Empirical power of tests by varying m_{1B} when censoring fraction is 0.3 and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter.

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.101	0.099	0.110	0.120	0.110	0.119
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.161	0.147	0.204	0.220	0.200	0.218
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.266	0.229	0.388	0.417	0.391	0.414
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.113	0.114	0.145	0.160	0.143	0.155
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.258	0.218	0.380	0.407	0.376	0.402
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.474	0.400	0.707	0.724	0.704	0.723
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.248	0.202	0.297	0.312	0.301	0.310
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.507	0.432	0.695	0.711	0.695	0.706
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.802	0.720	0.957	0.960	0.956	0.959

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

4 Real example

In this section we illustrated the proposed method for the sequential therapy of randomized clinical trial comparing Sorafenib (So) followed by Sunitinib (Su) (So-Su, n = 182) versus the reverse sequence (Su-So, n = 183) for metastatic renal cell cancer (SWITCH trial). The primary endpoint was total PFS which was defined as the interval between the randomization (the start date of first-line therapy) to the disease progression or death during second-line therapy. For subjects who did not switch to per-protocol second-line therapy, first-line events were used. Subjects without tumor progression or death during second-line therapy were censored. Secondary endpoints included first-line PFS (time from randomization to progression or death during first-line therapy); second-line PFS (time from first day of second-line therapy to progression or death during second-line therapy). Details of the study have been published in Eichelberg et al. (2015).

We have chosen this study to illustrate our methods because it presented interesting aspects of intermediate clinical events. The proportion that has been administered a second-line therapy was higher in So-Su (57% vs 42%, P value $<$ 0.01). The total PFS and PFS of first-line did not show a statistically significant difference (12.5 mo vs. 14.9 mo (P value = 0.5), 5.9 mo vs. 8.5 mo (P value = 0.9), respectively), whereas the PFS of second-line therapy showed a shorter duration in Su-So (5.4 mo vs. 2.8 mo, P value $<$ 0.001). If we consider receiving second-line therapy as experiencing an IE, we can compare the difference of survival functions by taking advantage of the knowledge of the information of the proportion of having second-line therapy and the duration of first-line/second-line therapy with different hazards assumption.

Unfortunately, it is difficult to obtain the raw data of this study. Therefore, we regenerated the data using of the survival rate that was extracted from the Kaplan–Meier (KM) graph by using program (http://arohatgi.info/WebPlotDigitizer/app/) and number of patients at risk set for each time. The number events during the interval could be estimated, which assumes that censoring is uniform over the intervals defined by the numbers at risk (Williamson et al., 2002).

$$s^*_{j,i} = s^*_{j,i-1}\left[1 - \frac{d^*_{j,i}}{n_{j,i-1} - (c^*_{j,i}/2)}\right]$$

$$n_{j,i} = n_{j,i-1} - d^*_{j,i} - c^*_{j,i}$$

where $d^*_{j,i}$ =number of events in $[t_{i-1}, t_i)$ and $c^*_{j,i}$ =number censored in $[t_{i-1}, t_i)$. Rearranging above equation gives

$$d^*_{j,i} = \frac{(n_{j,i-1} + n_{j,i})(s^*_{j,i-1} + s^*_{j,i})}{(s^*_{j,i-1} + s^*_{j,i})}$$

$$c^*_{j,i} = \frac{2(n_{j,i-1}s^*_{j,i} + n_{j,i}s^*_{j,i-1})}{(s^*_{j,i-1} + s^*_{j,i})}.$$

The KM graphs on the total, first-line, and second-line PFS with risk tables were provided in Eichelberg et al. (2015). The KM graphs from regenerated data have confirmed the similar results. The interval of radiological assessment follow-up was 12 weeks. As in simulation, we assume that the loss rate of radiological assessment was 0.1 for first year and then 0.2 thereafter.

As the result, proposed methods showed the significant difference of two arms (P value < 0.01), whereas log rank test and stratified log rank test did not (P value > 0.5). We also applied the method based on cox model (Yu et al., 2010, Shen,

2014a) and the result was similar. The hypothesis on (β_0, β_1) are separable as noted by Nam and Zelen (2001). Therefore we can test whether distribution is different for each parameter; i.e., $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. One degree of freedom is used in a chi-square test $\chi_1^2 = \hat{S}_1^2 / V(\hat{S}_1)$ of this hypothesis. In this case, we did not reject the null hypothesis of β_0 (P value = 0.6) but reject the null hypothesis of β_1 (P value < 0.001).

5 Discussion

We proposed a general method of comparing two interval-censored samples in the presence of an intermediate clinical event. When an IE occurs, it may change the survival distribution. One may want to know that whether an IE affects the survival as longer lives have more opportunity of receiving intervention. Therefore, we needed a specific method to resolve length-biased problems. For example, it can be considered a second-line therapy as an IE. The focus of the current study was to compare two survival functions incorporating the information of an IE. The outcome of prior studies that assessed the length-biased problems was exact data. In a cancer trial, the PFS is the popular outcome which is well known as interval censored data. Therefore, we considered interval-censored outcomes. When the statistics of Nam and Zelen (2001) was derived, the data was divided into two parts depending on an IE; one part was right-censored data and the other part was left truncated and right-censored data. For interval censored outcome, we have interval censored and LTIC data. We extended the score test of Finkelstein (1986) accommodating with left truncated.

We proposed non-iterative methods to impute LTIC data; uniform weight and weighted weight method based on NPMLE, respectively. In the uniform weight method, we assumed the true failure time of a subject might be uniformly distributed over $\{s_j, L_i < s_j \leq R_i, \text{ for } j = 1, ..., m\}$ for each subject like Kim et al. (2006). We used a MI technique for the whole process including imputing a true failure time while Kim et al. (2006) used a MI technique to estimate variance matrix. Uniform weight assumption in the characterized set is convenient to implement in practice. It is simple and fast. It can be extended to other complicated problem.

Second, we proposed a weighted weight method based on NPMLE. After characterized the set that may have a positive mass including truncated points as pointed by Frydman (1994), Turnbull' s algorithm was used to estimate the NPMLE. The performance of imputation procedures depends highly on the performance of the NPMLE. In the case of left-truncated and interval-censored data, NPMLE is not consistent, whereas conditional NPMLE is still consistent (Pan and Chappell, 1999). However, the problem is limited to the early time point. Here we did not use any special correction because our purpose was not to obtain NPMLE. As shown in the simulation, it was not considerably different from other proposed methods.

Pan (2000a) imputed the exact failure time from the coefficient and the baseline survival after fitting the cox model for interval censored data. They repeated the algorithm until the coefficient β^h converged, where h denotes number of iteration. Yu et al. (2010), Shen (2014a) extended Pan's method to accomodate left truncation. We applied the method based on cox model to the real example and the result was similar with proposed methods.

We appled two forms of variance that were formed by being added and subtracted. Both variance methods function efficiently, but the first one was marginally overestimated and the second one is slightly underestimated. This phenomenon is same as Huang et al. (2008).

We proposed to impute interval-censored observations and keep right-censored data as it is unlikely Huang et al. (2008). As Pan (2000b) illustrated, we can apply standard software after imputation and there is insufficient information to impute from right-censored observations. In clinical trials, many subjects may be right-censored at the last follow-up time or study closing time, and we have no information to impute exact survival times thereafter.

In the simulation, we used a moderate (M=10) and large (M=100) imputation numbers. The results were not considerably difference, so we recommended using a moderate imputation number.

We assumed that the intermediate clinical event was exactly observed; i.e., we can determine the exact date of transplant or the start date of second-line therapy. Further research is needed when an IE is considered as interval-censored.

Appendices

Table 10: Empirical power of tests by varying m_{1B} when all events are observed in some intervals and there are no missed visits (M=10).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.114	0.129	0.119	0.131
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.216	0.236	0.222	0.242
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.445	0.467	0.443	0.469
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.096	0.102	0.098	0.103
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.186	0.203	0.188	0.206
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.379	0.400	0.377	0.404
$n = 100$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.164	0.171	0.162	0.168
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.466	0.479	0.469	0.484
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.783	0.796	0.785	0.804
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.172	0.182	0.174	0.184
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.361	0.376	0.361	0.381
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.682	0.703	0.687	0.706
$n = 200$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.379	0.387	0.379	0.399
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.756	0.771	0.766	0.779
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.985	0.987	0.985	0.988
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.280	0.296	0.286	0.302
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.655	0.662	0.657	0.664
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.966	0.970	0.969	0.970

III = Uniform weight method, IV = Weighted weight method.

(1) added within- and between variance, (2) substracted within- and between variance

Table 11: Empirical power of tests by varying m_{1B} when all events are observed in some intervals and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter (M=10).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.095	0.100	0.094	0.103
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.197	0.215	0.200	0.213
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.346	0.358	0.345	0.358
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.096	0.103	0.099	0.106
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.132	0.139	0.130	0.138
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.295	0.317	0.303	0.320
$n = 100$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.133	0.136	0.132	0.138
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.340	0.353	0.346	0.356
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.650	0.666	0.658	0.665
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.136	0.141	0.139	0.144
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.302	0.309	0.302	0.314
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.559	0.566	0.560	0.572
$n = 200$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.287	0.296	0.289	0.298
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.639	0.653	0.639	0.647
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.950	0.954	0.953	0.956
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.219	0.225	0.220	0.229
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.539	0.554	0.541	0.553
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.896	0.897	0.895	0.897

III = Uniform weight method, IV = Weighted weight method.

(1) added within- and between variance, (2) substracted within- and between variance

Table 12: Empirical power of tests by varying m_{1B} when censoring fraction is 0.3, but there are no missed visits (M=10).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.095	0.113	0.093	0.109
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.191	0.215	0.189	0.212
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.414	0.446	0.421	0.446
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.094	0.108	0.093	0.107
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.182	0.209	0.185	0.207
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.367	0.400	0.372	0.402
$n = 100$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.150	0.165	0.156	0.162
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.442	0.466	0.440	0.473
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.749	0.774	0.758	0.777
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.159	0.178	0.159	0.174
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.369	0.400	0.376	0.405
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.667	0.697	0.679	0.711
$n = 200$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.338	0.364	0.345	0.374
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.741	0.757	0.743	0.759
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.981	0.983	0.980	0.983
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.257	0.275	0.258	0.277
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.638	0.661	0.652	0.673
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.946	0.953	0.952	0.958

III = Uniform weight method, IV = Weighted weight method.

(1) added within- and between variance, (2) substracted within- and between variance

Table 13: Empirical power of tests by varying m_{1B} when censoring fraction is 0.3 and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter (M=10).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.095	0.108	0.096	0.102
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.190	0.201	0.193	0.206
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.358	0.377	0.359	0.378
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.081	0.094	0.083	0.096
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.150	0.165	0.153	0.163
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.259	0.290	0.264	0.290
$n = 100$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.144	0.158	0.150	0.158
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.339	0.356	0.336	0.354
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.670	0.688	0.672	0.687
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.126	0.143	0.131	0.148
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.277	0.294	0.285	0.301
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.542	0.560	0.544	0.566
$n = 200$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.293	0.308	0.302	0.311
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.655	0.671	0.659	0.673
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.940	0.945	0.938	0.943
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.196	0.212	0.193	0.211
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.547	0.564	0.559	0.573
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.879	0.884	0.878	0.884

III = Uniform weight method, IV = Weighted weight method.

(1) added within- and between variance, (2) substracted within- and between variance

Table 14: Empirical 5% level tests by varying θ_B, m_{1A}, and m_{1B} with $\theta_A = 0.5$ when all events are observed in some intervals and there are no missed visits (M=100).

(θ_A, θ_B)	(λ_A, λ_B)	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.040	0.050	0.048	0.051	0.044	0.050
(0.5, 0.5)	(1, 1)	(1, 1)	0.061	0.047	0.046	0.049	0.046	0.049
(0.5, 0.4)	(1, 1)	(2, 2)	0.055	0.093	0.044	0.048	0.044	0.050
(0.5, 0.4)	(1, 1)	(1, 1)	0.066	0.111	0.051	0.057	0.047	0.056
(0.5, 0.3)	(1, 1)	(2, 2)	0.099	0.213	0.043	0.049	0.044	0.047
(0.5, 0.3)	(1, 1)	(1, 1)	0.045	0.245	0.044	0.051	0.044	0.056
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.057	0.058	0.050	0.055	0.051	0.056
(0.5, 0.5)	(1, 1)	(1, 1)	0.058	0.056	0.043	0.054	0.046	0.057
(0.5, 0.4)	(1, 1)	(2, 2)	0.076	0.133	0.046	0.055	0.049	0.055
(0.5, 0.4)	(1, 1)	(1, 1)	0.043	0.143	0.045	0.052	0.045	0.053
(0.5, 0.3)	(1, 1)	(2, 2)	0.139	0.365	0.046	0.053	0.047	0.054
(0.5, 0.3)	(1, 1)	(1, 1)	0.048	0.460	0.040	0.049	0.040	0.053
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.055	0.048	0.050	0.052	0.049	0.055
(0.5, 0.5)	(1, 1)	(1, 1)	0.043	0.052	0.049	0.058	0.050	0.058
(0.5, 0.4)	(1, 1)	(2, 2)	0.090	0.221	0.052	0.058	0.053	0.057
(0.5, 0.4)	(1, 1)	(1, 1)	0.053	0.293	0.046	0.053	0.046	0.053
(0.5, 0.3)	(1, 1)	(2, 2)	0.210	0.646	0.046	0.054	0.046	0.054
(0.5, 0.3)	(1, 1)	(1, 1)	0.054	0.732	0.044	0.047	0.044	0.051

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 15: Empirical 5% level tests by varying θ_B, m_{1A}, and m_{1B} with $\theta_A = 0.5$ when all events are observed in some intervals and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter (M=100).

(θ_A, θ_B)	(λ_A, λ_B)	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.054	0.058	0.045	0.054	0.044	0.053
(0.5, 0.5)	(1, 1)	(1, 1)	0.061	0.069	0.043	0.060	0.042	0.058
(0.5, 0.4)	(1, 1)	(2, 2)	0.073	0.105	0.045	0.056	0.047	0.054
(0.5, 0.4)	(1, 1)	(1, 1)	0.060	0.124	0.042	0.056	0.041	0.057
(0.5, 0.3)	(1, 1)	(2, 2)	0.098	0.212	0.045	0.055	0.046	0.055
(0.5, 0.3)	(1, 1)	(1, 1)	0.057	0.236	0.042	0.058	0.041	0.057
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.051	0.048	0.046	0.055	0.049	0.053
(0.5, 0.5)	(1, 1)	(1, 1)	0.053	0.067	0.039	0.048	0.042	0.049
(0.5, 0.4)	(1, 1)	(2, 2)	0.073	0.145	0.051	0.061	0.053	0.062
(0.5, 0.4)	(1, 1)	(1, 1)	0.047	0.173	0.038	0.048	0.040	0.048
(0.5, 0.3)	(1, 1)	(2, 2)	0.137	0.372	0.049	0.060	0.049	0.058
(0.5, 0.3)	(1, 1)	(1, 1)	0.049	0.462	0.041	0.056	0.045	0.056
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.059	0.057	0.054	0.060	0.053	0.058
(0.5, 0.5)	(1, 1)	(1, 1)	0.055	0.042	0.040	0.053	0.042	0.051
(0.5, 0.4)	(1, 1)	(2, 2)	0.096	0.221	0.051	0.057	0.050	0.058
(0.5, 0.4)	(1, 1)	(1, 1)	0.061	0.282	0.041	0.054	0.044	0.051
(0.5, 0.3)	(1, 1)	(2, 2)	0.232	0.621	0.051	0.053	0.047	0.055
(0.5, 0.3)	(1, 1)	(1, 1)	0.067	0.732	0.044	0.059	0.045	0.056

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 16: Empirical 5% level tests by varying θ_B, m_{1A}, and m_{1B} with $\theta_A = 0.5$ when censoring fraction is 0.3, but there are no missed visits (M=100).

(θ_A, θ_B)	(λ_A, λ_B)	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.055	0.054	0.053	0.057	0.052	0.058
(0.5, 0.5)	(1, 1)	(1, 1)	0.057	0.064	0.043	0.046	0.041	0.044
(0.5, 0.4)	(1, 1)	(2, 2)	0.060	0.091	0.052	0.064	0.055	0.059
(0.5, 0.4)	(1, 1)	(1, 1)	0.047	0.091	0.048	0.053	0.046	0.056
(0.5, 0.3)	(1, 1)	(2, 2)	0.085	0.181	0.050	0.054	0.050	0.056
(0.5, 0.3)	(1, 1)	(1, 1)	0.047	0.200	0.044	0.047	0.043	0.048
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.048	0.045	0.049	0.055	0.048	0.054
(0.5, 0.5)	(1, 1)	(1, 1)	0.058	0.056	0.051	0.055	0.051	0.055
(0.5, 0.4)	(1, 1)	(2, 2)	0.069	0.121	0.050	0.051	0.049	0.052
(0.5, 0.4)	(1, 1)	(1, 1)	0.049	0.134	0.044	0.047	0.042	0.047
(0.5, 0.3)	(1, 1)	(2, 2)	0.103	0.283	0.058	0.060	0.056	0.061
(0.5, 0.3)	(1, 1)	(1, 1)	0.040	0.341	0.044	0.051	0.044	0.048
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.046	0.051	0.046	0.050	0.048	0.050
(0.5, 0.5)	(1, 1)	(1, 1)	0.047	0.051	0.047	0.058	0.050	0.056
(0.5, 0.4)	(1, 1)	(2, 2)	0.081	0.193	0.042	0.047	0.044	0.046
(0.5, 0.4)	(1, 1)	(1, 1)	0.055	0.229	0.051	0.058	0.051	0.056
(0.5, 0.3)	(1, 1)	(2, 2)	0.166	0.494	0.045	0.051	0.045	0.050
(0.5, 0.3)	(1, 1)	(1, 1)	0.051	0.607	0.041	0.045	0.043	0.047

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 17: Empirical 5% level tests by varying θ_B, m_{1A}, and m_{1B} with $\theta_A = 0.5$ when censoring fraction is 0.3 and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter (M=100).

(θ_A, θ_B)	(λ_A, λ_B)	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.050	0.056	0.045	0.051	0.046	0.051
(0.5, 0.5)	(1, 1)	(1, 1)	0.065	0.060	0.044	0.055	0.045	0.056
(0.5, 0.4)	(1, 1)	(2, 2)	0.058	0.100	0.050	0.057	0.049	0.056
(0.5, 0.4)	(1, 1)	(1, 1)	0.052	0.090	0.045	0.053	0.046	0.052
(0.5, 0.3)	(1, 1)	(2, 2)	0.079	0.162	0.047	0.055	0.049	0.054
(0.5, 0.3)	(1, 1)	(1, 1)	0.047	0.200	0.044	0.058	0.041	0.057
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.052	0.055	0.045	0.049	0.043	0.050
(0.5, 0.5)	(1, 1)	(1, 1)	0.044	0.052	0.043	0.054	0.043	0.053
(0.5, 0.4)	(1, 1)	(2, 2)	0.075	0.105	0.050	0.056	0.051	0.059
(0.5, 0.4)	(1, 1)	(1, 1)	0.052	0.133	0.045	0.056	0.048	0.057
(0.5, 0.3)	(1, 1)	(2, 2)	0.110	0.258	0.047	0.055	0.049	0.052
(0.5, 0.3)	(1, 1)	(1, 1)	0.052	0.336	0.043	0.050	0.044	0.051
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 2)	0.059	0.059	0.043	0.047	0.045	0.049
(0.5, 0.5)	(1, 1)	(1, 1)	0.054	0.062	0.051	0.058	0.048	0.060
(0.5, 0.4)	(1, 1)	(2, 2)	0.078	0.180	0.045	0.051	0.046	0.048
(0.5, 0.4)	(1, 1)	(1, 1)	0.057	0.219	0.045	0.048	0.044	0.051
(0.5, 0.3)	(1, 1)	(2, 2)	0.168	0.485	0.050	0.052	0.048	0.053
(0.5, 0.3)	(1, 1)	(1, 1)	0.060	0.582	0.042	0.053	0.044	0.049

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 18: Empirical power of tests by varying m_{1B} when all events are observed in some intervals and there are no missed visits (M=100).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.106	0.101	0.112	0.121	0.116	0.125
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.196	0.183	0.250	0.263	0.253	0.265
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.367	0.298	0.506	0.528	0.508	0.526
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.169	0.156	0.198	0.207	0.199	0.210
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.363	0.315	0.490	0.510	0.491	0.506
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.645	0.557	0.840	0.848	0.839	0.850
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.286	0.254	0.388	0.398	0.391	0.399
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.662	0.562	0.805	0.812	0.804	0.810
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.932	0.858	0.988	0.991	0.990	0.991

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 19: Empirical power of tests by varying m_{1B} when all events are observed in some intervals and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter (M=100).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.120	0.108	0.114	0.131	0.110	0.132
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.222	0.181	0.249	0.279	0.249	0.284
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.386	0.320	0.484	0.512	0.483	0.511
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.181	0.146	0.198	0.214	0.199	0.214
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.373	0.315	0.471	0.499	0.475	0.498
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.647	0.564	0.824	0.840	0.825	0.841
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.310	0.289	0.356	0.380	0.362	0.382
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.652	0.575	0.814	0.827	0.812	0.825
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.925	0.860	0.988	0.991	0.989	0.991

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 20: Empirical power of tests by varying m_{1B} when censoring fraction is 0.3, but there are no missed visits (M=100).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.077	0.099	0.122	0.125	0.122	0.126
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.157	0.158	0.211	0.225	0.217	0.224
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.268	0.249	0.410	0.426	0.408	0.421
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.127	0.117	0.166	0.177	0.169	0.174
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.271	0.231	0.395	0.408	0.393	0.406
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.493	0.421	0.704	0.710	0.705	0.716
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.243	0.194	0.314	0.320	0.313	0.323
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.500	0.420	0.709	0.717	0.711	0.716
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.797	0.677	0.949	0.951	0.948	0.953

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 21: Empirical power of tests by varying m_{1B} when censoring fraction is 0.3 and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter (M=100).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	I	II	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.101	0.099	0.111	0.120	0.111	0.120
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.161	0.147	0.207	0.218	0.205	0.218
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.266	0.229	0.395	0.423	0.395	0.422
$n = 100$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.113	0.114	0.142	0.155	0.145	0.153
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.258	0.218	0.387	0.402	0.383	0.396
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.474	0.400	0.699	0.721	0.699	0.723
$n = 200$								
(0.5, 0.5)	(1, 1)	(2, 1.5)	0.248	0.202	0.300	0.314	0.300	0.319
(0.5, 0.5)	(1, 1)	(2, 1.25)	0.507	0.432	0.699	0.712	0.695	0.710
(0.5, 0.5)	(1, 1)	(2, 1.0)	0.802	0.720	0.958	0.964	0.961	0.963

I = log-rank, II = Stratified log-rank, III = Uniform weight method, IV = Weighted weight method.
(1) added within- and between variance, (2) substracted within- and between variance

Table 22: Empirical power of tests by varying m_{1B} when all events are observed in some intervals and there are no missed visits (M=100).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.116	0.129	0.116	0.132
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.220	0.235	0.216	0.236
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.444	0.468	0.444	0.471
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.096	0.104	0.099	0.105
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.178	0.203	0.185	0.202
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.377	0.399	0.382	0.405
$n = 100$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.161	0.165	0.158	0.164
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.466	0.480	0.461	0.485
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.784	0.799	0.785	0.800
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.172	0.180	0.175	0.185
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.362	0.378	0.370	0.385
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.680	0.696	0.685	0.701
$n = 200$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.380	0.390	0.382	0.392
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.762	0.773	0.766	0.778
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.980	0.983	0.982	0.983
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.295	0.310	0.300	0.314
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.654	0.667	0.658	0.669
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.947	0.953	0.949	0.953

III = Uniform weight method, IV = Weighted weight method.

(1) added within- and between variance, (2) substracted within- and between variance

Table 23: Empirical power of tests by varying m_{1B} when all events are observed in some intervals and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter (M=100).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.092	0.112	0.094	0.113
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.187	0.212	0.190	0.218
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.418	0.449	0.419	0.452
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.093	0.109	0.096	0.114
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.188	0.208	0.190	0.208
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.372	0.401	0.376	0.402
$n = 100$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.151	0.162	0.156	0.167
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.447	0.469	0.444	0.472
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.750	0.771	0.758	0.776
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.154	0.174	0.157	0.178
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.367	0.395	0.376	0.406
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.672	0.700	0.678	0.705
$n = 200$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.338	0.372	0.349	0.372
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.743	0.754	0.745	0.761
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.980	0.982	0.981	0.981
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.287	0.308	0.296	0.320
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.660	0.684	0.668	0.690
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.946	0.952	0.946	0.954

III = Uniform weight method, IV = Weighted weight method.

(1) added within- and between variance, (2) substracted within- and between variance

Table 24: Empirical power of tests by varying m_{1B} when censoring fraction is 0.3, but there are no missed visits (M=100).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.092	0.100	0.090	0.098
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.198	0.215	0.201	0.215
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.350	0.362	0.349	0.363
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.097	0.101	0.099	0.101
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.127	0.137	0.124	0.138
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.299	0.314	0.302	0.323
$n = 100$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.132	0.136	0.133	0.139
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.346	0.356	0.344	0.356
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.654	0.668	0.656	0.667
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.135	0.141	0.138	0.144
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.307	0.311	0.308	0.318
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.553	0.566	0.560	0.568
$n = 200$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.280	0.286	0.281	0.286
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.639	0.648	0.642	0.647
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.934	0.936	0.930	0.936
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.222	0.229	0.224	0.233
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.561	0.570	0.558	0.572
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.896	0.902	0.896	0.904

III = Uniform weight method, IV = Weighted weight method.

(1) added within- and between variance, (2) substracted within- and between variance

Table 25: Empirical power of tests by varying m_{1B} when censoring fraction is 0.3 and there are some missed visits with a probability 0.1 for the first year and then 0.2 thereafter (M=100).

(θ_A, θ_B)	$(\lambda_{0A}, \lambda_{0B})$	(m_{1A}, m_{1B})	III-(1)	III-(2)	IV-(1)	IV-(2)
$n = 50$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.096	0.108	0.099	0.111
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.193	0.204	0.190	0.201
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.360	0.377	0.358	0.377
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.077	0.094	0.082	0.093
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.149	0.163	0.149	0.167
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.259	0.286	0.260	0.285
$n = 100$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.146	0.154	0.146	0.154
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.336	0.355	0.337	0.351
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.669	0.687	0.673	0.691
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.127	0.145	0.129	0.142
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.279	0.291	0.280	0.294
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.544	0.563	0.547	0.563
$n = 200$						
(0.5, 0.4)	(1, 1)	(2, 1.5)	0.271	0.282	0.274	0.286
(0.5, 0.4)	(1, 1)	(2, 1.25)	0.643	0.660	0.637	0.661
(0.5, 0.4)	(1, 1)	(2, 1.0)	0.931	0.935	0.931	0.936
(0.5, 0.3)	(1, 1)	(2, 1.5)	0.227	0.244	0.227	0.244
(0.5, 0.3)	(1, 1)	(2, 1.25)	0.548	0.562	0.546	0.566
(0.5, 0.3)	(1, 1)	(2, 1.0)	0.886	0.894	0.890	0.896

III = Uniform weight method, IV = Weighted weight method.

(1) added within- and between variance, (2) substracted within- and between variance

Bibliography

C. M. Nam and M. Zelen, "Comparing the Survival of Two Groups with an Intermediate Clinical Event," *Lifetime Data Analysis*, vol. 7, no. 1, pp. 5–19, 2001.

D. M. Finkelstein, "A proportional hazards model for interval-censored failure time data." *Biometrics*, vol. 42, no. 4, pp. 845–854, 1986.

J. R. Anderson, K. C. Cain, and R. D. Gelber, "Analysis of survival by tumor response." *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 1, no. 11, pp. 710–719, 1983.

E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *European Journal of Cancer*, vol. 45, no. 2, pp. 228–247, 2009.

N. Mantel and D. Byar, "Evaluation of response-time data involving transient states: an illustration using heart-transplant data," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 81–86, 1974.

C. Eichelberg, W. L. Vervenne, M. De Santis, L. Fischer von Weikersthal, P. J. Goebell, C. Lerchenmüller, U. Zimmermann, M. M. Bos, W. Freier, S. Schirrmacher-Memmel, M. Staehler, S. Pahernik, M. Los, M. Schenck, A. Flörcken, C. van Arkel, K. Hauswald, M. Indorf, D. Gottstein, and M. S. Michel, "SWITCH: A Randomised, Sequential, Open-label Study to Evaluate the Efficacy and Safety of Sorafenib-sunitinib Versus Sunitinib-sorafenib in the Treatment of Metastatic Renal Cell Cancer," *European Urology*, vol. 68, no. 5, pp. 837–847, 2015.

M. Lefkopoulou and M. Zelen, "Intermediate clinical events, surrogate markers and survival," *Lifetime Data Analysis*, vol. 1, no. 1, pp. 73–85, 1995.

C. G. Law and R. Brookmeyer, "Effects of mid-point imputation on the analysis of doubly censored data," *Statistics in Medicine*, pp. 1905–1910, 1992.

R. Peto, "Experimental Survival Curves for Interval-Censored Data," *Applied Statistics*, vol. 22, no. 1, pp. 86–91, 1973.

B. W. Turnbull, "The Empirical Distribution Function with Arbitrarily Grouped , Censored and Truncated Data point in time is to be incorporated," *Journal of the Royal Statistical Society*, vol. 38, no. 3, pp. 290–295, 1976.

A. P. Dempster, N. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B*, pp. 290–295, 1976.

H. W. Kuhn and A. W. Tucker, "Nonlinear Programming," *Econometrica*, vol. 19, p. 50, 1951.

R. Gentleman and C. J. Geyer, "Maximum likelihood for interval censored data : and computation Consistency," *Biometrika*, vol. 81, no. 3, pp. 618–623, 1994.

J. Sun, "A Non-parametric test for interval censored failure time data with application to aids studies," *Statistics in Medicine*, vol. 15, no. July 1995, pp. 1387–1395, 1996.

Q. Zhao and J. Sun, "Generalized log-rank test for mixed interval-censored failure time data," *Statistics in Medicine*, vol. 23, no. 10, pp. 1621–1629, 2004.

J. Kim, D. R. Kang, and C. M. Nam, "Logrank-type tests for comparing survival curves with interval-censored data," *Computational Statistics and Data Analysis*, vol. 50, no. 11, pp. 3165–3178, 2006.

J. Huang, C. Lee, and Q. Yu, "A generalized log-rank test for interval-censored failure time data via multiple imputation," *Statistics in Medicine*, vol. 27, no. 17, pp. 3217–3226, 2008.

M. P. Fay and P. a. Shaw, "Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The interval R Package," *Journal Of Statistical Software*, vol. 36, no. 2, pp. 1–34, 2010.

M. P. Fay, "Rank invariant tests for interval censored data under the grouped continuous model." *Biometrics*, vol. 52, no. 3, pp. 811–822, 1996.

——, "Comparing several score tests for interval censored data," *Statistics in Medicine*, vol. 18, no. 3, pp. 273–285, 1999.

G. Heinze, M. Gnant, and M. Schemper, "Exact Log-Rank Tests for Unequal Follow-Up," *Biometrics*, vol. 59, no. 4, pp. 1151–1157, 2003.

H. Frydman, "A note on nonparametric estimation of the distribution function from interval-censored and truncated observations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 56, no. 1, pp. 71–74, 1994.

A. Alioum and D. Commenges, "A proportional hazards model for arbitrarily censored and truncated data." *Biometrics*, vol. 52, no. 2, pp. 512–524, 1996.

W. Pan and R. Chappell, "A Note on Inconsistency of NPMLE of the Distribution

Function from Left Truncated and Case I Interval Censored Data," *Lifetime Data Analysis*, vol. 5, no. 3, pp. 281–291, 1999.

——, "Estimation in the cox proportional hazards model with left-truncated and interval-censored data." *Biometrics*, vol. 58, no. 1, pp. 64–70, 2002.

P.-S. Shen, "Nonparametric tests for left-truncated and interval-censored data," *Journal of Statistical Computation and Simulation*, vol. 85, no. 8, pp. 1544–1553, 2014.

H. Frydman, "Nonparametric estimation of a Markov 'illness-death' process from interval-censored observations, with application to diabetes survival data," *Biometrika*, vol. 82, no. 4, pp. 773–789, 1995.

P. Joly, D. Commenges, C. Helmer, and L. Letenneur, "A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia." *Biostatistics*, vol. 54, no. 1, pp. 433–443, 2002.

P. Joly, T. A. Gerds, V. Qvist, D. Commenges, and N. Keiding, "Estimating survival of dental fillings on the basis of interval-censored data and multi-state models," *Statistics in Medicine*, vol. 31, no. 11-12, pp. 1139–1149, 2012.

H. Frydman, T. Gerds, R. Grøn, and N. Keiding, "Nonparametric estimation in an "illness-death" model when all transition times are interval censored," *Biometrical Journal*, vol. 55, no. 6, pp. 823–843, 2013.

B. Yu, J. S. Saczynski, and L. Launer, "Multiple imputation for estimating the risk of developing dementia and its impact on survival," *Biometrical Journal*, vol. 52, no. 5, pp. 616–627, 2010.

G. C. Wei and M. a. Tanner, "Applications of multiple imputation to the analysis of censored regression data." *Biometrics*, vol. 47, no. 4, pp. 1297–1309, 1991.

W. Pan, "A Multiple Imputation Approach to Cox Regression with Interval-Censored Data," *Biometrics*, no. March, pp. 199–203, 2000.

D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, 1987.

M. A. Tanner and W. H. Wong, "The Calculation of Posterior Distributions by Data Augmentation: Rejoinder," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 548–550, 1987.

W. Pan, "A Two-Sample Test with Interval Censored Data via Multiple Imputation," 2000.

C. H. Hsu, J. M. G. Taylor, S. Murray, and D. Commenges, "Multiple imputation for interval censored data with auxiliary variables," *Statistics in Medicine*, vol. 26, no. 4, pp. 769–781, 2007.

P. S. Shen, "Proportional hazards regression with interval-censored and left-truncated data," *Journal of Statistical Computation and Simulation*, vol. 84, no. 2, pp. 264–272, feb 2014.

D. Follmann, M. A. Proschan, and E. Leifer, "Multiple outputation: Inference for complex clustered data by averaging analyses from independent data," *Biometrics*, vol. 59, no. 2, pp. 420–429, 2003.

P. R. Williamson, C. T. Smith, J. L. Hutton, and A. G. Marson, "Aggregate data meta-analysis with time-to-event outcomes," *Stat Med*, vol. 21, no. 22, pp. 3337–3351, 2002.

국문 요약

구간 중도 절단된 자료에서 임상적으로 의미 있는 중도 사건 발생 여부에 따른 생존 분포 비교

본 연구에서는 임상적 중간 사건이 존재하면서 일차 목적이 시간에 대한 변수인 경우를 고려하였다. 추적조사 기간 동안 시험대상자는 특정한 임상적중간 사건을 경험할 수 있으며 이런 임상적 중간 사건이 발생할 경우 생존분포가 변할 수 있다. 임상적 중간 사건이 존재할 때는 기간 차이 치우침 (length biased) 문제를 함께 고려해야 한다. 여기서는 임상적 중간 사건에 대한 영향을 직접적으로 추정하기 보다는 임상적 중간 사건이 존재할 때 두 군의 생존 함수를 비교하는 방법에 대해 연구하고자 하였다.

남정모와 Zelen(2001) 의 연구에서 임상적 중간 사건을 고려해서 식을 전개해보면 통계량이 두 부분으로 나뉘어 진다; 첫번째는 임상적 중간 사건을 경험하기 전에 대한 것으로 우측 중도 절단 (right censoring) 자료 형태를 가지며 두번째는 임상적 중간 사건을 경험한 후에 대한 것으로 좌측 절단 및 우측 중도 절단 (left truncation and right censoring) 자료 형태를 가진다. 기간 차이 치우침을 가진 자료에서의 이전 연구는 일차 목적에 대한 변수가 정확하게 관측될 경우에 대해 연구하였는데 본 논문에서는 구간중도절단자료 형태를 고려하고자 한다. 그 경우 다뤄야 하는 자료 형태는는 구간 중도 절단 혹은 좌측 절단 및 구간 중도 절단 형태가 된다. 구간 중도 절단 자료 형태를 고려하지 않고 일반 방법을 쓰면 치우침이 발생할 수 있다는 것은 잘 알려져 있다.

(좌측 절단 및) 구간 중도 절단 자료 형태를 고려하여 먼저 Finkelstein 방법을 확장하여 우도 함수를 얻고 통계량을 계산하였다. 일반적으로 구간 중도 절단 자료에서는 각 구간의 값을 추정하기 위해서 매우 반복이 많이 발생한다. 여기서는 실제적으로 구현이 쉬운 다중 대체 방법을 사용하여 (좌측 절단 및) 우측 중도 절단 형태로 자료를 대체하여 남정모와 Zelen(2001) 통계량을 적용하였다. 균등 가중치 방법, 비모수적

최대우도추정치를 기반으로 한 가중치 방법을 제안하였다. 다양한 시나리오로 모의실험을 진행하였고, 층화 로그 순위 방법의 경우 임상적 중간 사건 경험 후 생존함수가 변하지 않더라도 임상적 중간 사건의 발생 비율이 다르면 귀무 가설을 잘 못 기각하는 경우가 많았다. 두 군의 생존 함수는 동일하지만, 임상적 중간 사건의 발생 비율이 다르고 임상적 중간 사건 경험 후 생존 함수가 달라지는 경우에는, 로그 순위 방법도 귀무 가설을 잘 못 기각하는 경향을 보였다. 제안된 방법은 모든 시나리오에서 명목수준 0.05를 잘 만족하였고 로그순위 및 층화 로그순위 방법보다 더 높은 검정력을 보였다. 그룹 내 분산과 그룹 간 분산을 합하여 사용할 경우 분산이 과대추정되는 경향이 있어 Huang 등(2008)이 제안한 바와 같이 그룹 내 분산과 그룹 간 분산의 차이로 구한 추정 분산으로도 결과를 요약하였다.

핵심이 되는 말: 구간 중도 절단, 임상적 중간 사건, 생존함수 비교