SciencePG
Science Publishing Group

Non-parametric Analysis of Interval-Censored Survival Data with Application to a Phase III Metastatic Colorectal Cancer Clinical Trial

Yeqian Liu, Junyu Chen

Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA

Email address:
yeqian.liu@mtsu.edu (Y. Liu)

To cite this article:
Yeqian Liu, Junyu Chen. Non-parametric Analysis of Interval-Censored Survival Data with Application to a Phase III Metastatic Colorectal Cancer Clinical Trial. *Biomedical Statistics and Informatics*. Vol. 6, No. 1, 2021, pp. 14-22. doi: 10.11648/j.bsi.20210601.13

Abstract: In oncology clinical trials, the exact time of event occurrence such as tumor progression is usually unknown but the time interval within which the event occurs is known. The determination of such survival time can be subject to measurement error and influenced by the timing of scheduled assessment. Ignoring interval-censored survival time could lead to serious estimation bias. In addition, a crucial characteristic of interval-censored data is how frequently the measurement interval is taken, which directly determine the efficiency of statistical inference. Therefore, it is highly desirable to find statistical methods that are robust to different assessment frequencies. We compare conventional imputation-based approach with non-parametric approaches to handle interval-censored survival data. We apply these approaches to both hypothesis test and the estimations of hazard and survival functions. Empirical performance of these methods are assessed through extensive simulation studies with various sample sizes. A phase III randomized clinical trial on metastatic colorectal cancer is analyzed by using conventional approaches and non-parametric interval-censored analysis approaches. Out findings suggest that the phase III colorectal cancer clinical trial failed to show a clinical benefit of adding bevacizumab (B) to standard chemotherapy (CT), and the proposed non-parametric interval-censored analysis approaches outperforms the conventional approach for routine applications to oncology clinical trials to analyze interval-censored survival data.

Keywords: Interval-censoring, Finkelstein's Score Test, Generalized Log-rank Test, Non-parametric Maximum Likelihood Estimation, EM Algorithm

1. Introduction

Interval-censored time-to-event data occur naturally and frequently in randomized clinical trials, where the exact time of event occurrence is unknown but the time interval within which the event occurs is known. The left-point of the time interval in the interval-censored data represents the last time the individual is known to be event-free, and the right-point of the interval represents the earliest time that the individual is recorded with an event. There are two important special cases of interval-censored data. The first case is current status data, where only the observation time and whether or not the event has occurred at the time are known. The second case is grouped time-to-event data, where the interval-censored time for each subject is a member of a collection of non-overlapping intervals, and multinomial distribution can be used on the number of subjects in the given intervals. This paper focuses on case II interval-censored data.

In oncology clinical trials, progression-free survival is the time from randomization date to the time of disease progression or death. Due to the latency of disease progression, the exact time of disease progression is never known. Most progression-free survival are interval-censored time-to-event data, since determination of such survival time is always subject to measurement error and influenced by the timing of scheduled assessment. Several researchers have studied the impacts of bias due to ignoring interval-censored survival time. For example, Panageas et al. discussed that ignoring the interval censored data structure leads to overestimation of median progression-free survival [1]; Hess et al. discussed that unscheduled assessments may falsely conclude the significance of treatment effect [2]; Penson et al. considered

that different measurement intervals between treatment arms could lead to estimation bias. [3]. These researchers recommended to use consistent and symmetric interval assessment across treatment arms whenever possible, and use interval censoring analysis methodology for progression-survival data to minimize potential bias. Following these recommendations, this paper aims to provide recommendations of interval-censoring analysis for progression-free survival data.

Intuitively, when comparing with right-censored time-to-event data, interval-censored data is subject to loss of information. As a result, a crucial characteristic of interval-censored data analysis is how frequently the measurement interval is taken, which directly determine the efficiency of statistical inference. Meanwhile, the assessment schedule is often predetermined by various external factors such as evaluation cost, patient convenience and clinical practices. Therefore, it is highly desirable that statistical methods are robust to different assessment frequencies and schedules for progression-free survival data.

Statistical methods for right-censored data are widely used in pharmaceutical industry. For example, we have Kaplan-Meier estimator for non-parametric estimator of survival function; log-rank test for non-parametric test of treatment effect; semi-parametric Cox proportional hazards model is used for treatment effect estimation. The corresponding statistical methods for interval-censored data have also been developed in the past three decades, for example, nonparametric estimation of survival function by Turnbull [4], Gentle- man and Geyer [5], and Titman [6]; comparison of survival functions by Zhao and Sun [7], and Sun et. al. [8]; nonparametric proportional hazards model by Finkelstein [9] and Withana [10]. Complete references of statistical methods for interval-censored data can be found in the review paper by Zhang and Sun [11]. However, very few of these new methods have been directly compared with right-point and mid-point imputations which are widely used as the conventional approaches for interval-censored data.

Recently, Sun and Chen [12] compared the conventional methods with Finkestein's method of proportional hazard model [9] when analyzing interval-censored time-to-event data based on Monte Carlo simulation studies, and argued that Finkelstein's method for interval-censored data is superior to conventional approaches for interval-censored data. Their conclusions are based on limited scenarios and may not be valid all the time under different scenarios. In addition, the statistical methods for hypothesis testing for interval-censored data are also of our interests, which was not extensively studied by Sun and Chen. We conduct extensive Monte Carlo simulations under various scenarios which may occur in clinical trials. We compare the conventional imputation-based approaches with Finkelstein's method in terms of estimation, as well as Finkelstein's score test, generalized log-rank tests [7, 8] in terms of hypothesis testing.

The rest of paper is organized as follows, in section 2 we first introduce some notation and then review the idea behind non-parametric interval-censored data analysis approaches. In section 3, we present some results obtained from an extensive simulation study where the pros and cons of different statistical methods are discussed. In section 4, a Phase III randomized clinical trial on metastatic colorectal cancer is analyzed by the methods mentioned in section 2. We compare the performances of previously mentioned methods. Section 5 contains some discussion and concluding remarks.

2. Inference Procedure

2.1. Conventional Approach

The goal of right-point and mid-point imputations is to transform the interval-censored data into right-censored data. Right-point imputation uses the right-point of the time interval as the true event time, while the mid-point imputation uses the average of left-point and right-point of the time interval as the true event time [13]. After either right-point or mid-point imputation, one can use standard statistical methods for right-censored data, such as Kaplan-Meier estimator, log-rank test, and Cox proportional hazard model for estimation, inference, and hypothesis testing. When the assessment intervals are symmetric between treatment groups, both imputations have the same ranks for the time. Therefore, rank-based methods such as log-rank test and Cox proportional hazard model, gives similar results for right-point and mid-point imputations. Moreover, the assessment intervals usually have heavy ties in most clinical trials, the methods of handling such ties should be very carefully chosen. We recommend Efron's method [14] to deal with tied event times. Among methods of handling ties, which include Breslow's method, Efron's method, and exact method [15], Efron's method yield estimate that is fairly close to the one given by exact method and it is more computationally efficient.

2.2. Non-parametric Methods for Proportional Hazard Model

Let (L_i, R_i) be the observed event intervals with $i = 1, \cdots, n$ and $0 \leq L_i \leq R_i \leq \infty$. A subject is right-censored when $R_i = \infty$. Create m nonoverlapping intervals (s_j, s_{j+1}], where $0 = s_0 < s_1 < \cdots < s_m = \infty$, define increasing ordered intervals of $\{0, \{L_i\}_{i=1}^n, \{R_i\}_{i=1}^n, \infty\}$. For each treatment arm, the hazard function $\lambda(t \mid Z)$ and the survival function $S(t \mid Z)$ are set to be constant. Then the full log-likelihood function can be written as

$$L = \sum \log\big(S(L_i \mid Z_i) - S(R_i \mid Z_i)\big) \qquad (1)$$

and the likelihood contribution of the i-th patient is

$$S(L_i \mid Z_i) - S(R_i \mid Z_i) = \sum_{j=1}^m \alpha_{ij} \left(S\big(s_{j-1} \mid Z_i\big) - S\big(s_j \mid Z_i\big)\right) \quad (2)$$

where $\alpha_{ij} = 1$ if $L_i < s_j \leq R_i$ and $\alpha_{ij} = 0$ otherwise. When the problem reduces to one-sample, the problem of finding the non-parametric maximum likelihood estimator (NPMLE) of S becomes that of maximizing L under the

constraint that $\sum_{j=1} \left(S(s_{j-1}) - S(s_j) \right) = \sum_{j=1}^m p_j = 1$ and $p_j \geq 0$. Different methods to maximize the likelihood function for one-sample problem have been proposed, for example, EM self-consistency algorithm by Turnbull [4], as well as Gentleman and Geyer [5], and iterative convex minorant (ICM) algorithm by Groeneboom and Wellner [6].

Consider Cox proportional hazard model, $S(t \mid Z) = [S_0(t)]^{\exp(\beta^T Z)}$, where $S_0(t)$ is the baseline survival function. Let $\gamma_j = \log\left(\log\left(S_0(s_j)\right)\right)$, then Equation 2 can be re-parameterized as $S(s_j \mid Z_i) = \exp\left(-\exp(\beta^T Z_i + \gamma_j)\right)$ where γ_j strictly increases in j with $\gamma_0 = -\infty$ and $\gamma_m = \infty$. Maximum likelihood estimates (MLE) of β and γ_j are obtained by maximizing the re-parameterized log-likelihood function together using Newton-Raphson algorithm under the constraint $\gamma_1 < \cdots < \gamma_m$. Overserved Fisher's information matrix is used to obtain the standard error of the estimator. The approach actually simplifies the situation to a finite-dimensional parametric estimation problem. As a result, Finkelstein's maximum likelihood estimation becomes more computationally intensive as the number of intervals gets larger.

2.3. Nonparametric Comparisons of Survival Functions

Suppose there are K treatment arms in a clinical study and let $S^{(k)}(t)$ denote the survival function of the kth arm with k=1, $\cdots$, K. The null hypothesis to test is

$$H_0: S^{(1)}(t) = S^{(2)}(t) = \cdots = S^{(K)}(t) \text{ for all } t$$

When assuming each of the *n* subjects receives one of the *K* treatments, the data for the *K* samples can be represented as $\{((L_i, R_i), Z_i), 1 \leq i \leq n\}$, where Z_i is the $K \times 1$ vector of treatment indicators that are associated with subject *i* with interval-censored time $(L_i, R_i]$ whose *k* element is 1 if it is from the *k*-th population, and 0 otherwise.

2.3.1. Finkelstein's Score Test

For right-censored data, the log-rank test can be obtained as a score test on the proportional hazards regression model. One way to compare survival functions for interval-censored data is to perform a score test on a regression model for interval-censored data. The survival functions are compared by performing the score test for β=0 based on Finkelstein's method for proportional hazard model, where β is the vector of regression coefficients for Z_i. The score statistic for testing β=0 is

$$U = \left. \frac{\partial \log(L(\beta, \gamma_1, \cdots, \gamma_m))}{\partial \beta} \right|_{\beta=0} \quad (3)$$

2.3.2. Generalized Log-rank Test I

Zhao and Sun [7] proposed a rank-based approach that is a direct generalization of the log-rank test for right-censored data. For each pair of (i,j), define α_{ij} to be the indicator of the event $s_j \in (L_i, R_i]$, $1 \leq i \leq n, 1 \leq j \leq m$. For subject *i*, define $\delta_i = 0$ if the observation on T_i is right-censored and 1 otherwise; $\rho_{ij} = I\left(\delta_i = 0, L_i \geq t_j\right)$, which is equal to 1 if T_i

is right-censored and subject *i* is still at risk at $t_j -$. The log-rank statistic $U = (U_1, \cdots, U_K)^T$ is thus defined as

$$U_k = \sum_{j=1}^m \left(d_{jk} - \frac{n_{jk} d_j}{n_j} \right) \quad (4)$$

for k=1, $\cdots$, K, where

$$d_{jk} = \sum_{i=1}^n \delta_i \frac{\alpha_{ij} [\hat{s}_0(t_j-) - \hat{s}_0(t_j)]}{\sum_{u=1}^{m+1} \alpha_{iu} [\hat{s}_0(t_u-) - \hat{s}_0(t_u)]} Z_{ik} \quad (5)$$

$$n_{jk} = \sum_{r=j}^{m+1} \sum_{i=1}^n \delta_i \frac{\alpha_{ij} [\hat{s}_0(t_j-) - \hat{s}_0(t_j)]}{\sum_{u=1}^{m+1} \alpha_{iu} [\hat{s}_0(t_u-) - \hat{s}_0(t_u)]} Z_{ik} + \sum_{i=1}^n \rho_{ij} Z_{ik} \quad (6)$$

and $d_j = \sum_{k=1}^K d_{jk}$ and $n_j = \sum_{k=1}^K n_{jk}$, which can be regarded as the estimates of the total observed failure and risk numbers, respectively, at time t_j under H_0. It can be easily shown that if right-censored data are available, the statistic U reduce to the log-rank test statistic. Zhao and Sun [7] also proposed a multiple imputation approach to estimate the covariance matrix Σ of U. The null hypothesis of the homogeneity of the K populations can be tested by comparing the test statistic $T = U^T \Sigma^- U$ to a χ^2 distribution with K-1 degrees of freedom, where Σ^- is a generalized inverse of Σ.

2.3.3. Generalized Log-rank Test II

Sun, Zhao and Zhao [8] proposed a new class of K-sample test for interval-censored data which includes Finkelstein's score test statistic [9] as a special case. The K-sample test statistic is defined as

$$U_n = \sum_{i=1}^n K_i(L_i, R_i) \quad (7)$$

Where

$$K_i(L_i, R_i) = \sum_{i=1}^n Z_i \frac{\xi[\hat{s}_0(L_j) - \hat{s}_0(R_j)]}{\hat{s}_0(L_j) - \hat{s}_0(R_j)} \quad (8)$$

where ξ is a known function over (0, 1). When $\xi(u) = u \log(u)$ this test statistic reduces to Finkelstein's score test statistic. Denote the first components of U_n as U_n^*, Σ as the covariance matrix of U_n, Σ^* is derived by deleting the last row and column of Σ, whose expression is provided by the authors. The null hypothesis of the homogeneity of the K populations can be tested by comparing the statistic $\frac{1}{n} U_n^{*T} \Sigma^{*-1} U_n^*$ to a χ^2 distribution with K-1 degrees of freedom.

3. Simulation Studies

3.1. Data Generation

We generate data to simulate a hypothetical oncology Phase III clinical trial with two arms based on allocation ratio of 1:1. The sample size is set to be 200, 400 or 600 and the number of replications is 1000. The survival time is set to follow an exponential distribution with median equals to 8 weeks, 12 weeks, or 24 weeks for control arm C. The hazard ratio between Treatment arm (T) and Control arm © is assumed to be either 0.5 or 0.78. In simulations, each exact failure time is censored by a pre-specified time invterval to simulate the non-informative censoring. We also report the results from

Cox-regression and log-rank test based on exact event times in order to control the random error from simulations. The cut-off date is set to be equal for all patients. The overall duration is chose to have an approximately 80% or 60% event rate, respectively.

Table 1 provides overall study duration and the maximum

number of assessments under different assessment schedules. The maximum number of assessments ranges from 3 to 38 for treatment comparison, which cover a wide range of scenarios in practice. To keep the event rates remaining at desired proportions, we add an additional assessment at the end of the study.

Table 1. *Overall duration and maximum number of assessments.*

Median Survival	Hazard Ratio	Event Proportion	Overall Duration	Maximum number of assessment		
				per 6 weeks	per 8 weeks	per 12 weeks
8 weeks in control	1	80%	60	10	8	5
	1	60%	28	5	4	3
	0.5	80%	76	13	10	7
	0.5	60%	32	6	4	3
	0.67	80%	76	13	10	7
	0.67	60%	32	6	4	3
	0.78	80%	64	11	8	6
	0.78	60%	32	6	4	3
12 weeks in control	1	80%	88	15	11	8
	1	60%	40	7	5	4
	0.5	80%	116	20	15	10
	0.5	60%	48	8	6	4
	0.67	80%	108	18	14	9
	0.67	60%	48	8	6	4
	0.78	80%	96	16	12	8
	0.78	60%	48	8	6	4
24 weeks in control	1	80%	172	29	22	15
	1	60%	80	14	10	7
	0.5	80%	226	38	29	19
	0.5	60%	98	17	13	9
	0.67	80%	206	35	26	18
	0.67	60%	92	16	12	8
	0.78	80%	192	32	24	16
	0.78	60%	86	15	11	8

3.2. Simulation Results

Tables 2, 3 summarize point estimates under equal assessment schedules, with different median survival times in control arm (8 weeks, 12 weeks, or 24 weeks), and hazard ratios between treatment and control arms (HR=0.5 or 0.78), based on total sample size 200. Results from Cox regression of exact failure times are used as benchmarks. We use relative bias in order to make fair comparison for simulation results between different true parameter values. As we can see from the results, point estimates based on Finkelstein's method is almost unbiased under different scenarios, while point estimates based on conventional method are always negatively bias (away from null) and over-estimate treatment effects. The biases for conventional method become worse as assessment frequency decreases (assessment interval 8 weeks), right-censoring proportion increases (> 20%), as well as treatment effect between treatment and control arms attenuates. The estimates based on Finkelstein's method are very robust to the assessment frequency and censoring in general. Based on our extensive simulations (some results are not shown here), when maximum number of assessment is fewer than 5 times, conventional method would have about at least 10% negative bias at log hazard ratio scale, while for Finkelstein's method, the bias is at most 5%. Conventional methods and Cox model with exact failure times yield similar

standard deviations. However, Finkelstein's method overestimates the standard deviations. The 95% coverage probability of both conventional method and Finkelstein's method are fairly close to the Cox model with exact failure time.

Tables 4, 5 summarize the empirical Type I error rates at α=5% (two-sided) based on sample size of 200 and 400, respectively. Consistent with the findings on point estimation results in Tables 2, 3, the score test based on conventional method tends to be conservative when assessment frequency decreases and censoring proportion increases. Finkelstein's Type I errors increases as number of assessments increases. On the other hand, Finkelstein's Wald test Type I error rate are very well controled. Log-rank test of mid- point imputations also performs well under most scenarios. Generalized log-rank tests seem to perform consistently well among all tests evaluated in our simulation, when comparing the log-rank tests of exact time. In addition, type I errors based on any of these interval-censored methods tend to be slightly inflated when event rate is low.

Tables 6, 7 summarize the empirical power at α=5% (two-sided) based on sample size of 200 and 400. The findings are consistent with each other as well. Finkelstein's score test and Wald test have comparable power as compared to the log-rank test of conventional method. When comparing with the log-rank tests of exact failure time, it is clear to see that as

assessment frequency decreases and censoring proportion increases, the interval-censored data based tests become less powerful, as the information contained in the data decreases.

In conclusion, we find that the conventional approaches over-estimate treatment effect in point-estimation.

In particular, when the assessment frequency is low, conventional methods may give severely biased estimation.

For hypothesis testing, when assessments are balanced between treatment arms, conventional approaches and interval-censoring methods are comparable. In particular, conventional approach score test is more conservative in terms of type I error, while conventional approach score test is less powerful than log-rank tests.

Table 2. *Summary of Point Estimation and Inference for β=-0.693 (HR=0.5), n=200.*

Scenario	Method	Assess Interval	Bias	Std. Dev.	M. C. Std. Err.	95% Cov. Prob.
Median 8 Weeks 80% Event	Exact Time	-	0.6%	0.167	0.167	95.1%
	Conventional	6 weeks	-1.5%	0.165	0.167	95.2%
		8 weeks	-2.6%	0.166	0.167	94.9%
		12 weeks	-7.0%	0.162	0.167	94.1%
	Finkelstein	6 weeks	1.1%	0.170	0.171	94.7%
		8 weeks	1.2%	0.172	0.171	95.0%
		12 weeks	0.1%	0.175	0.173	94.1%
Median 8 Weeks 60% Event	Exact Time	-	1.7%	0.193	0.195	95.8%
	Conventional	6 weeks	-1.8%	0.193	0.196	95.5%
		8 weeks	-5.5%	0.192	0.196	95.6%
		12 weeks	-12.8%	0.187	0.196	92.6%
	Finkelstein	6 weeks	1.2%	0.199	0.198	95.7%
		8 weeks	-0.4%	0.202	0.200	95.6%
		12 weeks	-3.3%	0.208	0.205	94.7%
Median 12 Weeks 80% Event	Exact Time	-	-0.7%	0.164	0.167	96.7%
	Conventional	6 weeks	0.3%	0.160	0.167	96.7%
		8 weeks	-0.4%	0.160	0.167	96.5%
		12 weeks	-2.0%	0.158	0.167	96.7%
	Finkelstein	6 weeks	2.1%	0.163	0.172	96.6%
		8 weeks	1.9%	0.164	0.171	96.7%
		12 weeks	1.8%	0.164	0.171	96.4%
Median 12 Weeks 60% Event	Exact Time	-	-0.2%	0.205	0.195	93.6%
	Conventional	6 weeks	-1.7%	0.202	0.196	93.5%
		8 weeks	-2.8%	0.202	0.196	92.7%
		12 weeks	-6.9%	0.197	0.196	93.2%
	Finkelstein	6 weeks	0.1%	0.206	0.197	93.6%
		8 weeks	-0.3%	0.207	0.198	93.2%
		12 weeks	-1.8%	0.208	0.200	93.9%
Median 24 Weeks 80% Event	Exact Time	-	-0.2%	0.168	0.167	95.3%
	Conventional	6 weeks	-0.3%	0.168	0.167	95.8%
		8 weeks	-0.6%	0.168	0.167	95.9%
		12 weeks	-1.0%	0.168	0.167	95.9%
	Finkelstein	6 weeks	1.0%	0.170	0.177	96.6%
		8 weeks	0.9%	0.170	0.174	96.3%
		12 weeks	0.9%	0.170	0.172	96.0%
Median 24 Weeks 60% Event	Exact Time	-	-0.2%	0.205	0.195	93.6%
	Conventional	6 weeks	0.0%	0.195	0.195	95.8%
		8 weeks	-0.7%	0.195	0.194	95.3%
		12 weeks	-1.5%	0.196	0.195	95.7%
	Finkelstein	6 weeks	1.2%	0.197	0.199	96.2%
		8 weeks	0.7%	0.197	0.198	95.7%
		12 weeks	0.3%	0.200	0.197	95.6%

Table 3. *Summary of Point Estimation and Inference for β=-0.248 (HR=0.78), n=200.*

Scenario	Method	Assess Interval	Bias	Std. Dev.	M. C. Std. Err.	95% Cov. Prob.
Median 8 Weeks	Exact Time	-	1.7%	0.159	0.161	96.1%
	Conventional	6 weeks	-0.9%	0.157	0.161	96.0%
		8 weeks	-2.8%	0.157	0.161	96.0%
		12 weeks	-8.0%	0.153	0.161	95.5%

Scenario	Method	Assess Interval	Bias	Std. Dev.	M. C. Std. Err.	95% Cov. Prob.
80%		6 weeks	2.4%	0.162	0.165	96.0%
Event	Finkelstein	8 weeks	1.9%	0.164	0.165	95.8%
		12 weeks	1.0%	0.168	0.168	95.2%
	Exact Time	-	0.3%	0.188	0.185	95.2%
Median		6 weeks	-4.0%	0.186	0.185	94.3%
8 Weeks	Conventional	8 weeks	-7.6%	0.183	0.185	95.7%
		12 weeks	-15.6%	0.178	0.185	95.0%
60%		6 weeks	-0.5%	0.193	0.188	93.6%
Event	Finkelstein	8 weeks	-1.8%	0.194	0.190	95.2%
		12 weeks	-4.3%	0.203	0.196	94.2%
	Exact Time	-	-0.6%	0.163	0.161	95.7%
Median		6 weeks	-1.7%	0.161	0.161	95.7%
12 Weeks	Conventional	8 weeks	-2.5%	0.161	0.161	95.7%
		12 weeks	-5.4%	0.157	0.161	95.4%
80%		6 weeks	0.4%	0.165	0.165	95.6%
Event	Finkelstein	8 weeks	0.3%	0.165	0.165	95.7%
		12 weeks	-0.8%	0.165	0.165	95.7%
	Exact Time	-	1.3%	0.187	0.185	95.5%
Median		6 weeks	-0.7%	0.185	0.185	95.3%
12 Weeks	Conventional	8 weeks	-1.6%	0.184	0.186	96.2%
		12 weeks	-6.4%	0.180	0.186	95.7%
60%		6 weeks	1.3%	0.189	0.187	94.9%
Event	Finkelstein	8 weeks	1.4%	0.190	0.188	95.5%
		12 weeks	-0.6%	0.191	0.190	95.2%
	Exact Time	-	-0.1%	0.164	0.162	95.5%
Median		6 weeks	-0.4%	0.164	0.162	95.5%
24 Weeks	Conventional	8 weeks	-0.7%	0.164	0.162	95.0%
		12 weeks	-1.3%	0.163	0.162	95.4%
80%		6 weeks	1.0%	0.166	0.169	95.9%
Event	Finkelstein	8 weeks	0.9%	0.167	0.167	95.6%
		12 weeks	0.8%	0.166	0.166	95.3%
	Exact Time	-	-0.1%	0.187	0.187	95.2%
Median		6 weeks	-0.5%	0.187	0.187	95.6%
24 Weeks	Conventional	8 weeks	-0.9%	0.188	0.187	95.5%
		12 weeks	-2.0%	0.186	0.187	95.9%
60%		6 weeks	0.8%	0.189	0.190	95.7%
Event	Finkelstein	8 weeks	0.5%	0.190	0.189	95.5%

Table 4. *Summary of Type I error rate (%) at α=5% (two-sided), n=200 Median Event Assess Conventional Finkelstein Gen. Log-rank Exact.*

Survival	Prop.	Interval	Score	Log-rank	Score	Wald	Test I	Test II	Log-rank
		6 weeks	5.7	6.7	7.1	6.6	6.7	6.2	
	80%	8 weeks	5.9	6.6	6.7	6.3	6.6	6.0	6.5
8		12 weeks	5.6	7.0	7.3	7.0	7.0	6.9	
weeks		6 weeks	3.8	4.1	4.5	4.4	4.1	4.3	
	60%	8 weeks	4.1	4.5	5.0	4.9	4.5	4.8	4.3
		12 weeks	3.3	5.2	6.2	5.8	5.2	5.3	
		6 weeks	5.8	5.7	6.3	5.6	5.7	5.6	
	80%	8 weeks	4.7	5.0	5.5	5.3	5.0	5.0	5.1
12		12 weeks	5.3	5.9	6.0	5.7	5.9	5.6	
weeks		6 weeks	5.2	5.4	5.6	5.6	5.4	5.5	
	60%	8 weeks	4.6	5.1	5.4	5.4	5.1	5.3	5.1
		12 weeks	3.9	4.3	5.0	4.7	4.3	4.4	
		6 weeks	5.3	5.2	6.1	5.0	5.2	5.0	
	80%	8 weeks	5.2	5.3	5.9	4.8	5.3	5.0	5.1
24		12 weeks	4.9	5.0	5.5	4.9	5.0	4.8	
weeks		6 weeks	4.3	4.1	4.5	3.9	4.1	4.2	
	60%	8 weeks	4.4	4.6	4.9	4.4	4.6	4.2	4.3
		12 weeks	4.6	4.5	4.7	4.7	4.5	4.6	

Table 5. *Summary of Type I error rate (%) at α=5% (two-sided), n=400 Median Event Assess Conventional Finkelstein Gen. Log-rank Exact.*

Survival	Prop.	Interval	Score	Log-rank	Score	Wald	Test I	Test II	Log-rank
		6 weeks	5.0	5.1	5.4	5.4	5.1	5.3	
	80%	8 weeks	4.5	4.8	5.0	5.0	4.8	4.9	4.9
8		12 weeks	3.9	5.8	5.8	5.7	5.8	5.6	
weeks		6 weeks	5.3	5.6	5.7	5.7	5.6	5.7	
	60%	8 weeks	4.8	5.9	5.9	5.6	5.9	5.4	4.6
		12 weeks	4.2	5.3	5.8	5.6	5.3	5.3	
		6 weeks	4.5	4.7	4.7	4.5	4.7	4.5	
	80%	8 weeks	5.1	5.3	5.5	5.1	5.3	5.2	5.2
12		12 weeks	4.3	4.8	5.1	5.1	4.8	4.9	
weeks		6 weeks	5.3	5.4	5.6	5.6	5.4	5.5	
	60%	8 weeks	5.1	5.4	5.2	5.2	5.4	5.2	5.2
		12 weeks	4.5	5.6	5.6	5.5	5.6	5.3	
		6 weeks	4.9	4.9	5.3	4.6	4.9	4.6	
	80%	8 weeks	4.7	4.8	5.3	4.7	4.8	4.6	4.8
24		12 weeks	4.2	4.3	4.5	4.3	4.3	4.3	
weeks		6 weeks	5.2	5.5	5.7	5.3	5.5	5.4	
	60%	8 weeks	5.5	5.7	5.7	5.4	5.7	5.4	5.3
		12 weeks	5.4	5.5	5.7	5.7	5.5	5.6	

Table 6. *Summary of power (%) at α=5% (two-sided), HR=0.5, n=200 Median Event Assess Conventional Finkelstein Gen. Log-rank Exact.*

Survival	Prop.	Interval	Score	Log-rank	Score	Wald	Test I	Test II	Log-rank
		6 weeks	97.6	97.7	97.9	97.8	97.7	97.8	
	80%	8 weeks	97.4	97.7	97.7	97.7	97.7	97.6	98.3
8		12 weeks	96.0	97.3	97.4	97.3	97.3	97.0	
weeks		6 weeks	92.9	93.5	93.4	93.4	93.5	93.4	
	60%	8 weeks	89.1	90.7	90.8	90.5	90.7	90.5	95.7
		12 weeks	72.5	80.5	80.4	80.3	80.5	80.4	
		6 weeks	99.0	99.0	99.3	99.2	99.0	99.0	
	80%	8 weeks	98.8	98.9	98.9	98.9	98.8	98.8	99.1
12		12 weeks	98.6	98.9	98.9	98.8	98.9	98.9	
weeks		6 weeks	94.9	95.2	95.1	95.0	95.1	95.1	
	60%	8 weeks	93.1	93.6	93.8	93.6	93.6	93.6	96.3
		12 weeks	89.0	90.8	90.7	90.7	90.6	90.6	
		6 weeks	98.7	98.7	99.2	98.6	98.7	98.8	
	80%	8 weeks	98.7	98.7	99.2	98.7	98.7	98.7	98.7
24		12 weeks	98.7	98.7	98.9	98.6	98.7	98.7	
weeks		6 weeks	95.0	95.5	95.9	95.3	95.5	95.6	
	60%	8 weeks	95.2	95.2	95.5	95.4	95.2	95.4	95.6
		12 weeks	95.0	94.9	95.1	95.1	94.9	94.9	

Table 7. *Summary of power (%) at α=5% (two-sided), HR=0.78, n=400 Median Event Assess Conventional Finkelstein Gen. Log-rank Exact.*

| Survival	Prop.	Interval	Score	Log-rank	Score	Wald	Test I	Test II	Log-rank
		6 weeks	77.5	77.8	78.3	78.1	77.8	78.0	
	80%	8 weeks	74.4	75.5	75.8	75.4	75.5	75.3	77.9
8		12 weeks	70.1	73.0	73.5	73.2	73.0	72.9	
weeks		6 weeks	65.2	66.5	66.7	66.6	66.5	66.4	
	60%	8 weeks	59.5	61.7	62.0	61.9	61.7	61.9	68.1
		12 weeks	51.1	56.4	56.9	56.8	56.4	56.6	
		6 weeks	77.4	77.8	78.0	77.6	77.8	77.5	
	80%	8 weeks	77.0	77.3	77.9	77.5	77.3	77.3	78.0
12		12 weeks	74.4	75.7	76.0	75.5	75.7	75.4	
weeks		6 weeks	66.1	66.3	66.5	66.4	66.3	66.4	
	60%	8 weeks	65.2	66.5	66.6	66.6	66.5	66.2	68.2
		12 weeks	59.7	61.5	61.8	61.7	61.5	61.7	
		6 weeks	78.0	78.6	79.6	77.8	78.6	78.2	
	80%	8 weeks	78.3	78.6	79.1	78.3	78.6	78.5	78.0
24		12 weeks	77.5	78.0	78.2	77.7	78.0	77.6	
weeks		6 weeks	64.1	64.0	64.5	64.1	64.0	64.0	
	60%	8 weeks	64.3	64.5	64.7	64.5	64.5	64.2	65.5
		12 weeks	63.3	63.7	64.1	63.7	63.7	63.5	

4. An Application

Now we apply the methodologies proposed in previous sections to a Phase III colorectal cancer clinical trial (ITACa). The full analysis set includes 376 patients, among which 176 patients in treatment (CT+B) arm A and 194 patients in control (CT alone) arm B. At the time of final analysis, 343 failure events were observed (306 disease progressions and 37 deaths), 163 in arm A and 180 in arm B.

The unstratified log-rank test shows that survival distributions between arm A and B are not significant at level $\alpha=0.05$, with p-value 0.681. The estimated hazard ratio between group B verse A is 1.027, with p-value 0.772. The unstratified log-rank test shows that survival distributions between group A and B are not significant at level $\alpha=0.05$, with p-value 0.63. The estimated hazard ratio between group B verse A is 0.974 with p-value 0.541.

To illustrate the nonparametric interval-censored analysis methods, we compare the results by the conventional method with mid-point imputation, Finkelstein's method, and generalized log-rank tests. When we consider interval-censored data structure, the reported assessment date serves as the right point for events, and the left point for censoring. The assessment date prior to the reported assessment date is used as the left point for the events. If the recorded progression date is the first assessment date after randomization, the left point is set to be 0.

For overall survival, the mid-point imputation log-rank test, generalized log-rank test I and test II are neither significant at $\alpha=0.05$ with p-values 0.87, 0.653 and 0.591, respectively. The hazard ratio based on Fikelstein's method is 1.109. For progression free survival, the mid-point imputation log-rank test, generalized log-rank test I and test II are neither significant at $\alpha=0.05$ with p-values 0.591, 0.577 and 0.565 respectively. The hazard ratio based on Fikelstein's method is 0.991. We also compared the nonparametric estimates of survival functions for treatment and control groups, based on right-point imputation, mid-point imputation, and interval-censoring EM-ICM method. The results, as well as the median of overall survival and progression-free survival between two arms, are shown in Figures 1 and 2.
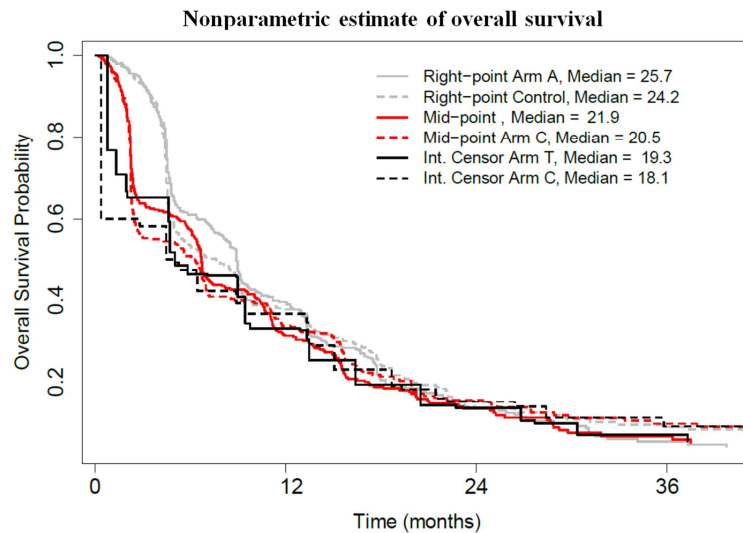
Figure 1. *Nonparametric Survival Function Estimates: Overall Survival.*
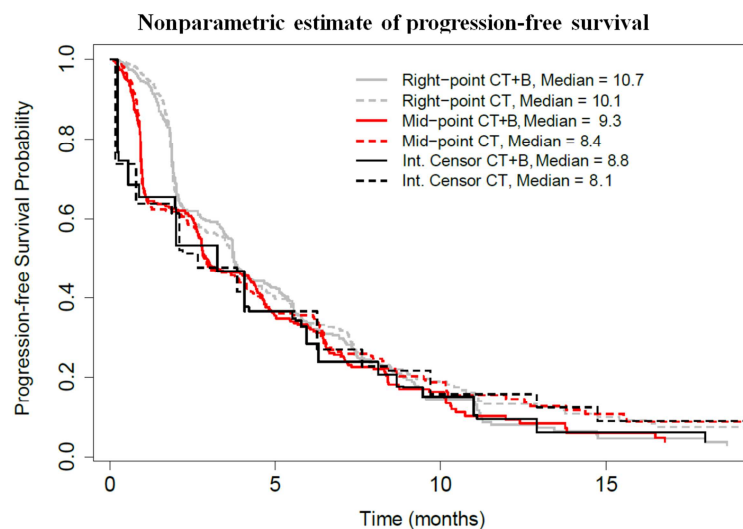
Figure 2. *Nonparametric Survival Function Estimates: Progression-free Survivall.*

In conclusion, the phase III colorectal cancer clinical trial failed to show a clinical benefit of adding bevacizumab (B) to standard chemotherapy (CT). A possible future research direction is to identify the biomarkers that could predict the sensitivity of anti-angiogenic drugs.

5. Discussions and Conclusions

In this paper, we compare various approaches to handle interval-censored survival data, e.g., conventional approaches and non-parametric approaches. The performance of these methods are evaluated through an extensive simulation study.

We discovered that when assessment interval are exactly symmetric across treatment arms, the conventional approach (right-point imputation) performs similarly to mid-point imputation, since the ranks of event times are same. We show that regular Cox regression could be severely biased when assessment is less frequent or event proportion is low. Finkelstein's non-parametric maximum likelihood estimation method, as a natural extension of Cox proportional hazard model for right-censored data, performs uniformly better in various scenarios we examined. It is remarkably robust to different assessment schedules and event proportions. Both Wald test and score test based on Finkelstein's method, as well as generalized log-rank tests, have performed well with generally acceptable Type I error rates and power relative to the convectional approach with log-rank test at given sample sizes.

In conclusion, when analyzing interval-censored survival data, we recommend to always consider and assess the possibility of evaluation-time bias. In practice, we strongly recommend adopting consistent and symmetric interval assessments across treatment arms whenever possible, and use sensitivity analysis to investigate the robustness of analysis results. Based on Monte Carlo simulation we conduct, we conclude that interval-censoring methods, e.g., Finkelstein's method for point estimation, Finkelstein's score test and generalized log-rank tests for hypothesis testing, are preferred when analyzing such data if possible. However, interval-censoring methods may be less efficient when sample size is small or moderate, and the corresponding computation may be too intensive when too many events occur.

Acknowledgements

References

[1] Panageas, K., Ben-Porat, L., Dickler, M. N., Chapman, P. B. & Schrag, D. (2007). When you look matters: the effect of assessment schedule on progression-free survival. Journal of the National Cancer Institute. 99 (6): 428.

[2] Hess, L., Brnabic, A., Mason, O., Lee, P. & Barker, S. (2019). Relationship between Progression-free Survival and Overall Survival in Randomized Clinical Trials of Targeted and Biologic Agents in Oncology. Journal of Cancer. 10, 3717-3727.

[3] Penson, D., Armstrong, A., Concepcion, R., Wu, K., Wang, F., Krivoshik, A., Phung, D. & Higano, C. (2016). Sensitivity analyses for progression-free survival (PFS) and radiographic PFS (rPFS) from the phase II STRIVE trial comparing enzalutamide (ENZA) with bicalutamide (BIC) in men with castration-resistant prostate cancer (CRPC). Journal of Clinical Oncology. 34, 169-169.

[4] Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society. Series B. 38 (3): 290–295.

[5] Gentleman, R. & Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. Biometrika. 81 (3): 618.

[6] Titman, A. (2017). Non-parametric maximum likelihood estimation of interval-censored failure time data subject to misclassification. Statistics and Computing. 27. 10.1007/s11222-016-9705-7.

[7] Zhao, Q. & Sun, J. (2004). Generalized log-rank test for mixed interval-censored failure time data. Statistics in medicine. 23 (10): 1621–1629.

[8] Sun, J., Zhao, Q. & Zhao., X. (2005). Generalized log-rank tests for interval-censored failure time data. Scandinavian Journal of Statistics. 32 (1): 49–57.

[9] Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. Biometrics. 42 (4): 845–854.

[10] Withana, G. P., Chaudari, M., Mcmahan, C. & Kosorok, M. (2020). A proportional hazards model for interval-censored subject to instantaneous failures. Lifetime Data Analysis. 26. 10.1007/s10985-019-09467-z.

[11] Zhang, Z. & Sun, J. (2009). Interval censoring. Statistical methods in medical research. 19, 53-70.

[12] Sun, X. & Chen, C. (2010). Comparison of Finkelstein's method with the conventional approach for interval-censored data analysis. Statistics in Biopharmaceutical Research. 2 (1): 97–108.

[13] Chen, D., Sun, J. & Peace, K. E. (2002). Interval-Censored Time-to-Event Data: Methods and Applications. 10.13140/2.1.3493.2169.

[14] Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association, 72 (359): 557–565.

[15] Kalbfleisch, J. D. & Prentice, R. L. (2002). The statistical analysis of failure time data. Wiley-Interscience.