ESTIMATION OF SURVIVAL FUNCTIONS FOR INTERVAL-CENSORED SEXUALLY TRANSMITTED DISEASE DATA

by

Jamtsho

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Applied Mathematics
May 2016

Committee Members:

Qiang Zhao, Chair

Alex White

Shuying Sun

FAIR USE AND AUTHOR’S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my supervisor Dr. Qiang Zhao for his unwavering support throughout this process. During this process, I have come to know him as a person who have deep understanding of his pupil's struggle in *R* programming and survival analysis. In addition, he possess great recognition skills that teaching new ideas does take time. During this process, there was no attrition of interest in his part to help me reach my goal. So, thank you from bottom of heart for everything you have taught me, patiently.

I am also grateful to my thesis committee members, Dr. Alex White and Dr. Shuying Sun for accepting to be on my thesis committee. It's a great privilege and honor to be in your presence. Last but not least, I want to thank you both for your helpful suggestions and comments on my thesis.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

LIST OF ILLUSTRATIONS

Illustration **Page**

ABSTRACT

Attention often pivots on the distribution of the sexually transmitted disease (STD) true infection time based on interval-censored event time data. It occurs in biological and medical studies such as medical follow-up studies and clinical trials. The underlying survival function for the interval censored data can be estimated by imputing the unknown infection time from a list of sexual encounter times. Harezlak and Tu (2006) proposed an imputation based method for the estimation of the survival function of the infection time using auxiliary behavioral information provided by daily diaries. In this study, we propose a method by considering a similar situation but using additional information, whether a condom is used or not by the subjects during their coital episodes. We incorporated the STD data introduced in Harezlak and Tu (HT) study into three methods: HT, Turnbull (Turnbull, 1976), and our proposed method and then assessed the estimates of each method. Our proposed method survival estimates behaved close to Turnbull method and even closer to HT method. The lack of true survival estimates between the three methods led us to perform simulation in order to make comparison. Our simulation results of mean integrated squared error (MISE) estimates reveal that the proposed method perform slightly better against HT method when settings have four scheduled visits and close when there are eight and sixteen number of scheduled visits and significantly better in all other scheduled visit times against Turnbull. We also compared biases in terms of sample size ($n = 100$) and level of right censoring (20%, 35%, 50%) in the sample at various time points (0 – 260 days). The biases for the proposed method are smaller when compared against HT and Turnbull method.

I. INTRODUCTION

Survival analysis is widely employed in many fields such as biology, medicine, public health, epidemiology, and economics. Usually, survival analysis is a collection of statistical methods for data analysis for which the outcome variable of interest is *time until an event occurs* (Kleinbaum, 1996).

By time, we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs; alternatively, time can refer to the age of an individual when an event occurs and denote as T_i. In survival analysis, we usually refer to the time variable as survival time, because it gives the time that an individual has "survived" over some follow-up period.

By event, we mean death, disease incidence, relapse from remission, recovery (e.g., return to work) or any designated experience of interest that may happen to an individual. We also typically refer to the event as a failure, because the kind of event of interest usually is death, disease incidence, or some other negative individual experience. However, survival time may be "time to return to work after an elective surgical procedure," in which case failure is a positive event. Now, this leads us to the focus of the study – interval-censored data.

In interval-censored data, the survival time of interest is known only to be placed in an interval, instead of being observed exactly (Sun 1996). By survival time, we mean the time to some event such as death or a disease. In other words, we can think of interval-censored observation as a union of non-overlapping intervals. Examples of interval censored data in human respiratory symptoms, animal carcinogenicity, and epidemiology studies can be found in Kongerud and Samuelsen (1991), Hoel and Walburg (1972), Finkelstein (1986), and Self and Grossman (1986). Interval-censored data also arise in AIDS studies, see for example, Jewell, Malani, and Vittinghoff (1994). In many instances, interval-censored survival data occurs in medical or health studies that require periodic follow-up. Several clinical trials and longitudinal studies have been studied in this category (Finkelstein, 1986). Among medical or health studies, sexually transmitted disease (STD) arises when the subjects are involved in the coital events. In this case, the event (STD positive) was known to fall only in the interval between visits (interval censored) or after the last time the subject was seen (right censored). Thus only an interval given by the last STD negative test and the first STD positive test is known for the STD infection time. In order to perform the analysis of interval-censored STD data, it is essential to understand some common infectious diseases and their consequences.

Sexually transmitted infections (STI) are among the most common infectious diseases worldwide. Among these curable STI, *Chlamydia trachomatis (CT)*, *Neisseria gonorrhoeae (NG),* and *Trichomonas vaginalis (TV)* are the common ones. Detection of the incubation and infection time for these infectious diseases are made possible by the polymerase chain reaction (PCR) (Garrow, Smith, & Harnett, 2002). Potential consequence of these STI in females include pelvic inflammatory disease, ectopic pregnancy, tubal factor infertility, adverse pregnancy outcomes, and potentially increased risk of both transmission and acquisition of human immunodeficiency virus (HIV). Additionally, investigators have shown epidemiologic associations with chlamydia or trichomonas infection and subsequent cervical neoplasia and carcinoma (Pol, Kraft, & Williams, 2006).

Immense progress has been achieved in the development of clinical trials during the past years. Methods have been developed, employed, and advanced that enable the reliable, efficient, and ethical evaluation of the benefits and risks of interventions that target the treatment and prevention of STI diseases. One of the most important components of this development has been the designing of censored data survival analysis technique. For instance, Turnbull (1976) introduced a self-consistency algorithm to estimate the survival function of survival time based the interval censored data. The estimate can be determined iteratively. In conjunction with the development of survival analysis methods, there has been recent development in the application of survival analysis techniques. One such techniques is the result of the availability of R software packages which are now able to run the difficult and computationally intensive algorithms used in these types of analyses relatively quickly and efficiently. Our study rely heavily on R to perform data imputation, iterations, and graphical plots. For reference to the R codes for this study, readers can refer to Appendix R code.

To set the stage for the survival analysis for which the study is being developed, we integrated three data sets: $std2.ic, std2.condom$ and $std2.sextime$ of previous studies by Harezlak and Tu (2006). Harezlak and Tu collected the data for their study: *Estimation of survival functions in interval and right censored data using STD behavioral diaries*. For the purpose of this study, we named their method of estimation as HT method and we briefly discuss how the data were collected. The study included female subjects between the ages of 14 and 17 who had been infected with one of the aforementioned organisms were approached at a county STD clinic. All subjects received appropriate treatment of their infections and were scheduled for return visits at 1, 3, 5, and 7 months. At enrollment, subjects were given pocket size diaries and received instructions to record the occurrence of sexual intercourse, condom use, condom failures, and an array of behavioral factors. At each follow-up visit, the subjects were interviewed and tested for the presence of NG, CT, and TV. Subjects who tested positive were promptly treated. They also received new diaries upon returning the completed ones. An important endpoint of the investigation was the timing of re-infections. In other words, the investigators were interested in estimating the survival function of the recurrent STIs.

The objective of this study is to examine the efficiency of proposed method and to estimate survival functions in the presence of interval censored data using STD behavioral diaries. In addition, the goal of this thesis is to establish a framework of survival analysis procedures for interval censored data with auxiliary information in the software *R* (Appendix A and B). Within this framework, we have set three goals to be achieved by the end of the study.

Goal 1. To estimate and interpret survival function using our assumption information.

Goal 2. To compare survivor functions estimate of proposed method against those of the HT method and the Turnbull self-consistency algorithm.

Goal 3. To assess mean integrated squared error (MISE) and biases against schedule visit time periods through computer simulations.

In the subsequent sections, subsections 1.0 – 1.3 report the theoretical foundations in the non-parametric estimation approach and its application. Subsections 2.1 – 2.4. detail the applicable methods involving Kaplan-Meier, the Turnbull algorithm, HT method, and the proposed method. Subsections 3.1 – 3.3 illustrate a simulation study to conduct and evaluate the operating features of the proposed method and consider a real-world example from an STD investigation. Subsections 4.1 report application results. Subsections 4.2

report simulation results. And finally, Section 5 suggests possible areas for further research and summarizes the main findings from this analysis respectively.

In order to understand the subsequent theoretical foundations in the nonparametric estimation, we begin with basic concepts and notations.

1.0 Survival Function

Let T denote a nonnegative random variable representing the failure (infection) time of individuals in some population ("Nonnegative" means $T \geq 0$.) and T is continuous. Let $F(t)$ denote the cumulative distribution function (c.d.f.) of T with corresponding probability density function (p.d.f.), $f(t)$.

Note $f(t) = 0$ for $t < 0$. Then

$$F(t) = P(T \leq t) = \int_0^t f(x)dx$$

The probability that an individual survives beyond time t is given by the survival function:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx.$$

1.2 Censoring

Censored survival data is fundamentally different from other types of data coming upon in statistical problems in the sense that the response of interest, the time until some specified event, cannot always be fully observed. Instead, causes for censoring exist which can result in the disconnection of observation before the event occurs. When this happens the recorded data do not provide direct information about the time until the event, and so models must be used to relate what one observes to what he wants (Lagakos, 1979). As a simple example of censoring, consider leukemia subjects followed until they go out of remission. If for a given subject, the study ends while the subject is still in remission (i.e., doesn't get the event), then that subject's survival time is considered censored. We know that, for this person, the survival time is at least as long as the period that the person has been followed, but if the person goes out of remission after the study ends, we do not know the complete survival time.

There are generally, three reasons why censoring may occur (Kleinbaum, 1996):

1. a person does not experience the event before the study ends;
2. a person is lost to follow-up during the study period;
3. a person withdraws from the study because of death (if death is not the event of interest) or some other reason (e.g., adverse drug reaction). (Pol, Kraft, & Williams, 2006).

Different types of censoring arise in practice. In a standard survival analysis application, individuals are followed over time for the occurrence of a specific event.

Let T_i be the true failure (infection) time for the ith subject, where $i = 1,,2, \ldots, n$.

1.2 (a) Interval-Censored Data

Interval-censored data are often obtained in longitudinal studies in which subjects are assessed periodically, and the time when the event of interest occurs is not directly observed but is known to have taken place within some time interval $\left((L_i, R_i]\right)$ (He, Kong, & Su, 2013). For instance, in a clinical trial with progression free survival as the outcome of interest subjects may visit a clinic for disease assessment at pre-determined times, and for subjects with disease progression it is known only to have occurred at some time between visits with the exact time of progression being unknown. Now, we will illustrate how the interval censoring occurs.

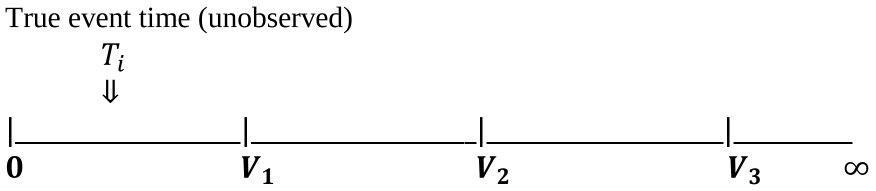


Figure 1. An illustrative example for interval-censored data.

Interval censoring is very common when we have discrete follow-up time (Xiao, Hu, Yu, & Xie, 2014). In a time to event investigation, let the true event time, T_i where $L_i \leq T_i \leq R_i$ of each subject be a non-negative random variable. We follow-up the state (e.g., infected or not infected) of all the subjects according to follow up time table- $(V_0, V_1, \dots, V_m, V_{m+1})$, where $V_0 = 0$ and $V_{m+1} = \infty$. Let T_i be the survival time of the subject i, where $i = 1, 2, \dots, n$. For interval-censored data, exact T_i is unobservable. We could only observe that the event time of one subject is between two adjacent follow-up time points, provided no skipped visits. The observation which we obtain is $(L_i, R_i]$ intervals, all given in the form of $(V_{i-1}, V_i]$. Figure 1 is shown for illustration.

Note that exactly observed, right-censored and left-censored are special cases of $(L_i, R_i]$ interval-censored observations, with $L_i = R_i$ for exactly observed, observations $R_i = \infty$ for right-censored, and $L_i = 0$ for left-censored observations. For interval-censored data, Turnbull proposed an iterative procedure to estimate the survival function *S(t)* corresponding to the interval-censored data.

1.2 (b) Right Censored Data

Let T_i denote the time to the outcome of interest for ith subject under study and C_i the corresponding potential right censoring time for the ith subject. Let δ_i denote the event indicator.

$$\delta_i = \begin{cases} 1, & \text{if the event was observed } (T_i \leq C_i) \\ 0, & \text{if the response was censored } (T_i > C_i) \end{cases}$$

The observable random variables are $X_i = \min(T_i, C_i)$ (Lagakos, 1979).

For example

$$\begin{pmatrix} T_i & C_i & X_i & \delta_i \\ 80 & 100 & 80 & 1 \\ 40 & 80 & 40 & 1 \\ 74+ & 74 & 74 & 0 \\ 85+ & 85 & 85 & 0 \\ 40 & 95 & 40 & 1 \end{pmatrix}$$

When no event times are censored, a non-parametric estimator of $S(t)$ is $1 - F_n(t)$, where $F_n(t)$ is the empirical cumulative distribution function.

When some observations are censored, we can estimate $S(t)$ using the Kaplan-Meier product-limit estimator.

1.2 (c) Left-Censored Data

Left censoring occurs when individuals have experienced the event of interest prior to the start of the period of observation. For instance, if a subject was recruited to a trial and the event of interest had already occurred, their data would be left-censored.

An observed failure time, X_i associated with a subject i in a study is considered to be left censored if it is less than a censoring time C_i. The data observed on the subject i can be recorded as $\{X_i, \delta_i, i = 1, \ldots, n\}$ where

$$X_i = \max(T_i, C_i),\ \delta_i = \begin{cases} 1, & if\ X_i = T_i, \\ 0, & if\ X_i = C_i \end{cases}$$

Example

$$\begin{pmatrix} T_i & C_i & X_i & \delta_i \\ 100- & 100 & 100 & 0 \\ 80- & 80 & 80 & 0 \\ 74 & 74 & 74 & 1 \\ 85 & 85 & 85 & 1 \\ 95- & 95 & 95 & 0 \end{pmatrix}$$

1.3 Application

In this thesis, we propose a multiple imputation procedure for interval-censored data by incorporating coital times and condom use information. The different application areas in which the proposed method can be applied are other clinical or epidemiological investigations where interval censoring arises.

II. METHODS

Nonparametric estimation of the survival function will be discussed in this section. The Kaplan-Meier estimator is widely employed as the nonparametric estimator for right-censored data. Similarly, Turnbull estimator is used as the nonparametric estimation of the survival function for interval-censored data.

2.1 Kaplan–Meier Estimate

We consider the data presented in *Table 1* for the remission times (in weeks) for two groups of leukemia patients: one group of 21 persons has received a certain treatment; the other group of 21 persons has received a placebo. The data came from Freireich et al., Blood, 1963 (Kleinbaum, 1996).

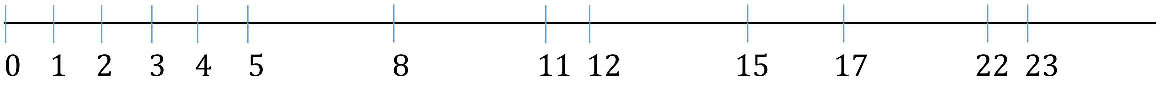
Table 1. Remission times (weeks) data for two groups of leukemia patients. A + indicates a censored observations.

Group	Length of complete remission (in weeks)
Group 1	6, 6,6,7,10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+.
Group 2	1, 1, 2, 2, 3, 4, 4,5,5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.

We first analyze this data for group 2 that has no censored observations. Let T be the random variable for a person's survival time for group 1 and group 2. Since T denotes time, its possible values include all nonnegative numbers. We define the empirical survivor function (esf), $S_n(t)$ as

$$S_n(t) = \frac{\text{\# of observations} > t}{n}$$

The $S_n(t)$ is the proportion of subjects still in remission after t weeks.



The values of the esf for the group 2 are:

t	0	1	2	3	4	5	8	11	12	15	17	22	23
$S_n(t)$	$\frac{21}{21}$	$\frac{19}{21}$	$\frac{17}{21}$	$\frac{16}{21}$	$\frac{14}{21}$	$\frac{12}{21}$	$\frac{8}{21}$	$\frac{6}{21}$	$\frac{4}{21}$	$\frac{3}{21}$	$\frac{2}{21}$	$\frac{1}{21}$	$\frac{0}{21}$

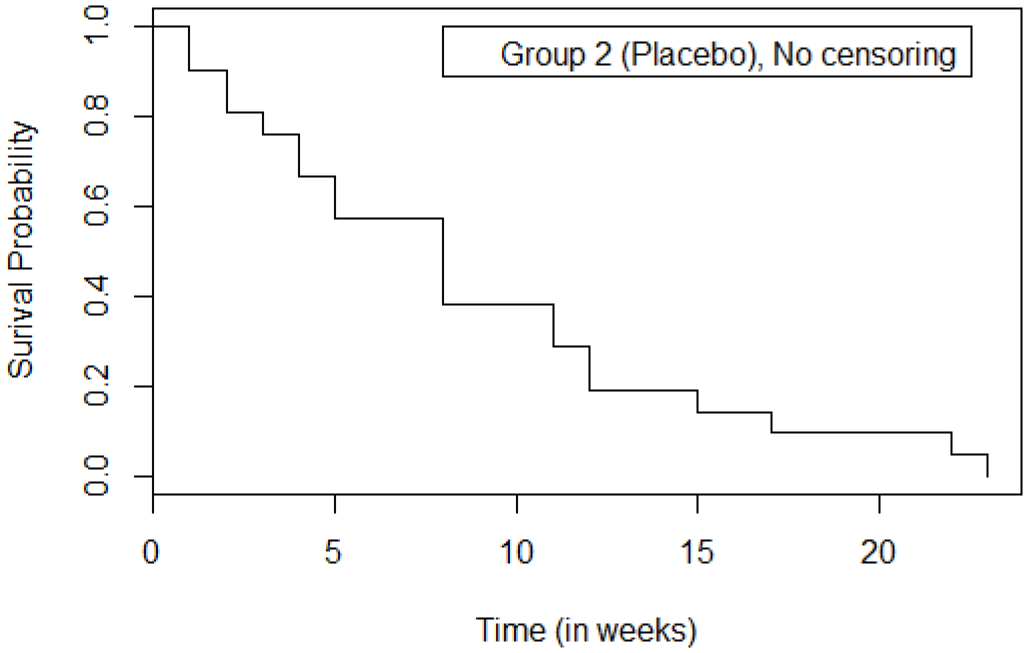


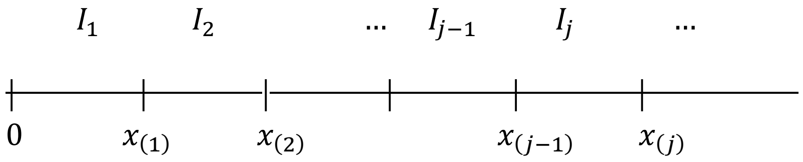
Figure 2. Survival function vs. time for Group 2 (with no censored subjects).

When some observations are right-censored, we can estimate $S(t)$ using the Kaplan-Meier product-limit estimator. Kaplan and Meier proposed the standard estimator of the survival function called Product-Limit estimator (Klein & Moeschberger, 2003).

Assume C_i to be fixed censoring time, then instead of observing the T_i we observe X_i, for i^{th} subject. Thus, for each of the n individuals we observe the pair (X_i, δ_i) where

$$X_i \ = \ min(T_i, C_i) \text{ and } \delta_i \ = \begin{cases} 1 \text{ if } T_i \leq C_i \\ 0 \ \text{ if } C_i < T_i. \end{cases}$$

On a time line we have



where $x_{(j)}$ denotes the jth distinct ordered censored or uncensored observation and is the right endpoint of the interval $I_j, j = 1, 2, \ldots, k$ for some k, and j is for time (Tableman & Sung, 2004).

Define:

n_j = Number of alive (and not censored) just before $x_{(j)}$

d_j = Number of patients died in I_j

$p_j = P$ (Surviving through I_j | Alive at beginning I_j)

$= P(T > x| T > x_{(j-1)})$

$q_j = 1 - p_i = P$ (Die in I_j | Alive at beginning I_j).

Recall the general multiplication rule for joint events A_1 and A_2:

$$P(A_1 \cap A_2) = P(A_2|A_1)P(A_1).$$

From repeated application of this product rule the survivor function can be expressed as

$$S(t) = P(T > t) = \Pi_{x_{(j)} \leq t} p_j.$$

The estimates of p_i and q_i are

$\widehat{q_J} = \frac{d_j}{n_j}$ and $\widehat{p_J} = 1 - \widehat{q_J} = 1 - \frac{d_j}{n_j} = \left(\frac{n_j - d_j}{n_j}\right)$.

The Kaplan-Meier estimator of the survivor function is

$$S(t) = P(T > t) = \Pi_{x_{(j)} \leq t} \widehat{p_J} = \Pi_{x_{(j)} \leq t} \left(\frac{n_j - d_j}{n_j}\right)$$

Let’s consider the Remission times data from *Table 1* on a time line where a “ $+$ ” denotes a right censored observations. The censored time 6+ we place to the right of the observed relapse time 6 since the censored patient at 6 weeks was still in remission. Hence, his relapse time (if it occurs) is greater than 6 weeks.

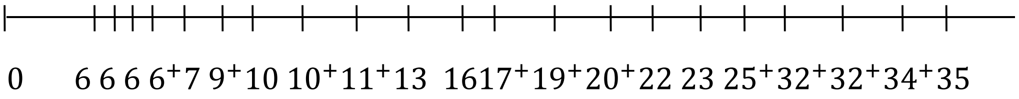

$0 \quad 6\ 6\ 6\ 6^{+}7\ 9^{+}10\ 10^{+}11^{+}13\ \ 1617^{+}19^{+}20^{+}22\ 23\ 25^{+}32^{+}32^{+}34^{+}35$

Table 2. The estimated survival probabilities obtained using the Kaplan-Meier formula.

Time, t	Number of risk	Number of event	Number of censored	Survival estimate, $\hat{S}(t)$.
0	21	0	0	1
6	21	3	1	1 x 18/21 = 0.8571
7	17	1	1	0.8571 x 16/17 = 0.8076
10	15	1	2	0.8076 x 14/15 = 0.7529
13	12	1	0	0.7529 x 11/12 = 0.6902
16	11	1	3	0.6902 x 10/11 = 0.6275
22	7	1	0	0.6275 x 6/7 = 0.5378
23	6	1	5	0.5378 x 5/6 = 0.4482

The Kaplan-Meier curve is a right continuous step function which steps down only at an uncensored observation. A plot of this together with the esf curve is displayed in *Figure 3*. Note the difference in the two curves.

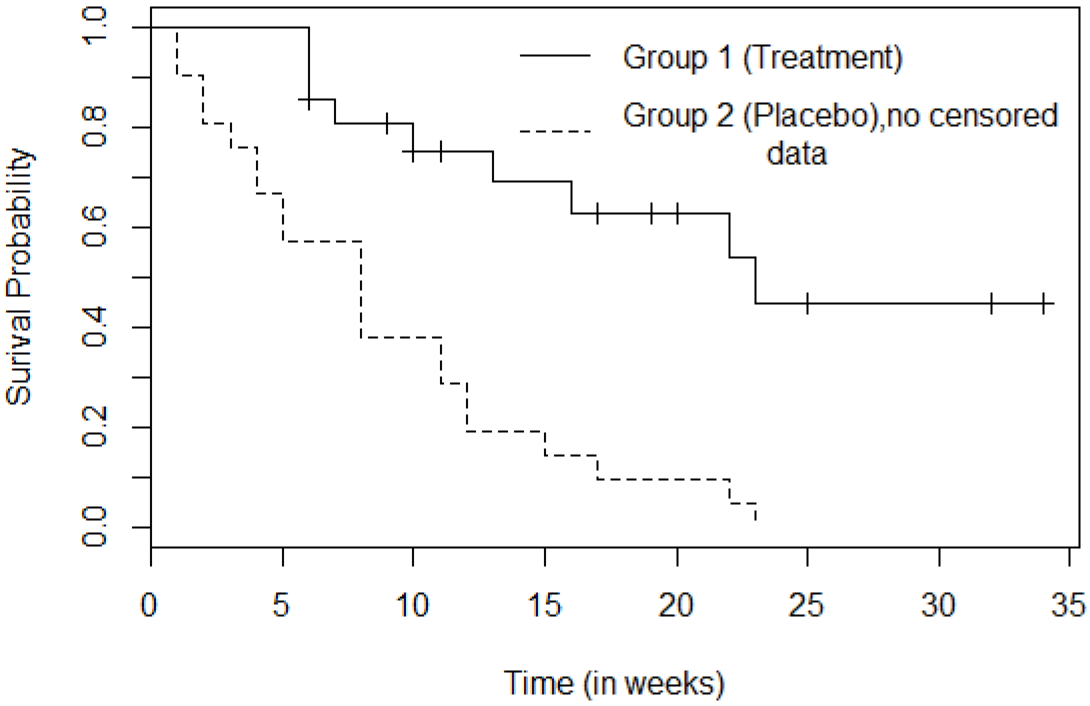


Figure 3. Kaplan-Meier plots of Remission Data: Group 1 and Group 2.

Kaplan-Meier is always greater than or equal to **esf**. When there are no censored observations, the Kaplan-Meier estimate reduces to the **esf**. Note the Kaplan-Meier curve does not jump down to zero as the largest survival time (32+) is censored. *Table 3* displays Kaplan-Meier survival estimates.

Table 3. Kaplan Meier survival estimates.

Time, t (in weeks)	$\hat{S}(t)$
$0 \leq t < 6$	1.000
$6 \leq t < 7$	0.857
$7 \leq t < 10$	0.807
$10 \leq t < 13$	0.753
$13 \leq t < 16$	0.690
$16 \leq t < 22$	0.628
$22 \leq t < 23$	0.538
$23 \leq t < 35$	0.448

2.2 Turnbull's Algorithm

This algorithm is based on an iterative procedure to estimate the survival function *S(t)* corresponding to the interval-censored data (Peto, 1973). To obtain Turnbull's estimate, the end points of the observed intervals are ordered in the same manner as in the Kaplan - Meier estimation. Let $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_m$ be the ordered distinct time points including all left $L_i$and right $R_i$time points in all intervals of $(L_i,\ R_i], i = 1, 2, \dots, n$ from *n* subjects.

Then, for the *i*th subject, define an indicator α_{ij} to keep track of whether the interval (τ_{j-1}, τ_j) is completely within the observed interval $(L_i, R_i]$ as

$$\alpha_{ij} = \{\begin{matrix} 1: & \text{If } (\tau_{j-1}, \tau_j] \in (L_i, R_i] \\ 0: & \text{Othewise} \end{matrix}$$

where α_{ij} also indicates whether the event that occurred in $(L_i, R_i]$ could have occurred at τ_j. Based on this indicator, we can obtain Turnbull's estimator iteratively as follows:

1. Make an initial guess, $\widehat{p_j}^{(0)}$ at $S(\tau_j)$.

$$p_j = S(\tau_{j-1}) - S(\tau_j) \qquad j = 1, 2, \dots, m$$

where p_j is the probability mass over $(\tau_{j-1}, \tau_j]$

2. Update correct estimate of p_j through

$$p_j^{(l)} = \frac{1}{n}\sum_{i}^{n}\left(\frac{p_j^{(l-1)}\alpha_{ij}}{\sum_{k=1}^{m} p_k^{(l-1)}\alpha_{ij}}\right), \quad j = 1, 2, \dots, m$$

where $\frac{p_j^{(l-1)}\alpha_{ij}}{\sum_{k=1}^{m} p_k^{(l-1)}\alpha_{ij}}$ corresponds to $P(T_i \in (\tau_{j-1}, \tau_j] \mid T_i \in (L_i, R_i])$.

$p_j^{(l-1)}$ and $p_j^{(l)}$ denote the current and the updated estimate respectively. $p_j^{(l)}$ can be calculated iteratively having obtained the $\widehat{p_j}^{(0)}$'s. The iteration continues until $\sum_{j=1}^{m}\left|\hat{p}_j^{(l)} - \hat{p}_j^{(l-1)}\right| \leq \epsilon$ is reached. The final estimate is $\hat{\boldsymbol{p}} = (\hat{p}_1, \dots, \hat{p}_m)'$. Finally, the survival function estimate is computed as:

$$\hat{S}(t) = \sum_{\tau_j > t} \hat{p}_j = 1 - \sum_{\tau_j \leq t} \hat{p}_j$$

We now illustrate Turnbull's self-consistency algorithm, we produce a hypothetical data as shown below in *Table 4*.

Table 4. Hypothetical interval-censored data to illustrate Turnbull self-consistent algorithm.

Subjects, ni	Left	Right
1	2	3
2	3	6
3	5	8
4	4	9
5	8	10

Let $\tau_0 = 0, \tau_1, \dots, \tau_8$ denote the ordered distinct time points of

$$\begin{matrix} \{2, & 3, & 3, & 4, & 5, & 6, & 8, & 8, & 9, & 10\} \\ L & L & R & L & L & R & L & R & R & R \end{matrix}$$

at which the esimated survival function can have jumps as shown *Figure 4* in the intervals

$$[3,3], [5,6], [8,8]$$

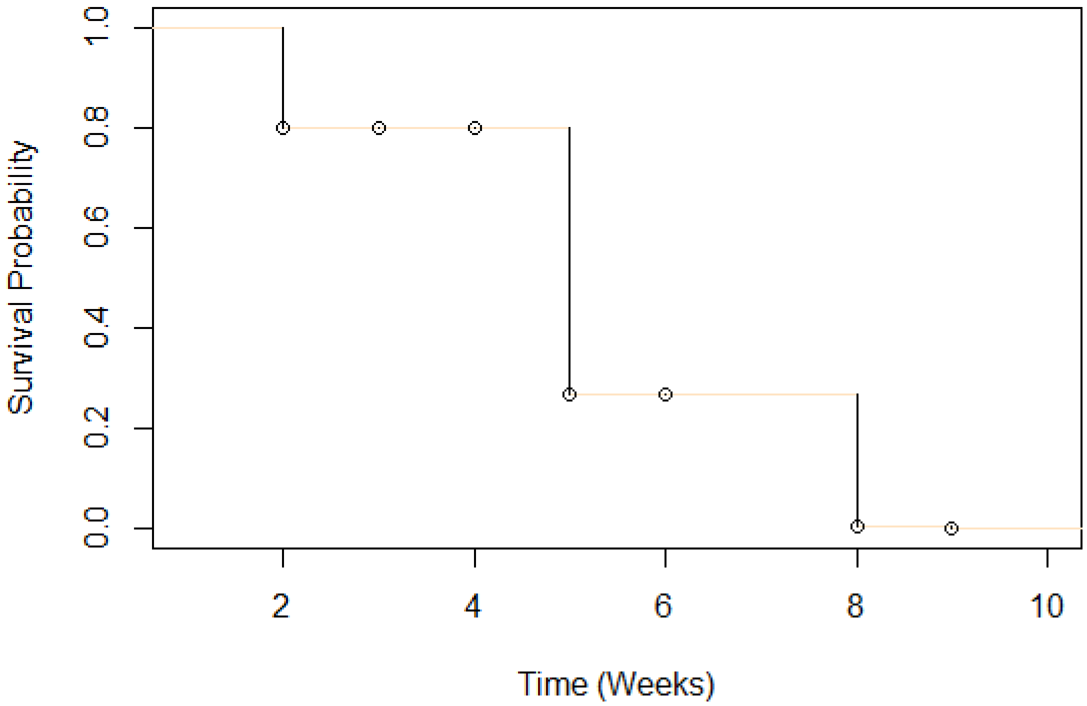


Figure 4. Turnbull survival function estimate for interval-censored data.

Next, we define an indicator function as folllows:

$$\alpha_{ij} = \begin{cases} 1, \ L_i < \tau_j \leq R_i \\ 0, Otherwise \end{cases}$$

and *Table 5* display distinct ordered time points.

Table 5. Ordered distinct time points of $\{0, L_i, R_i, i = 1, \ldots, 5\}$ and $\alpha_{ij} = I(\tau_j \in (L_i, R_i], i = 1, \ldots, 5, j = 1, \ldots, 8$.

Subject, ni	$[L_i, R_i]$	α_{i1}	α_{i2}	α_{i3}	α_{i4}	α_{i5}	α_{i6}	α_{i7}	α_{i8}
		2	3	4	5	6	8	9	10
1	(2,3]	0	1	0	0	0	0	0	0
2	(3,6]	0	0	1	1	1	0	0	0
3	(5,8]	0	0	0	0	1	1	0	0
4	(4,9]	0	0	0	0	1	1	1	0
5	(8,10]	0	0	0	0	0	0	1	1

The initial guess to start the iteration, let $\hat{p}_j^{(0)} = \frac{1}{m}$, where $j = 1, \ldots, 8$ and $m = 8$ in this hypothetical example. Since, we have our initial point, we can now start iteration in the following method.

$$\hat{p}_j^{(1)} = \frac{1}{5} \sum_{i=1}^{5} \frac{\alpha_{ij}\hat{p}_j^{(0)}}{\alpha_{i1}\hat{p}_1^{(0)} + \alpha_{i2}\hat{p}_2^{(0)} + \cdots \alpha_{i8}\hat{p}^{(0)}},$$

$$= \frac{1}{5}\left[\frac{\alpha_{11}\hat{p}_1^{(0)}}{\alpha_{11}\hat{p}_1^{(0)} + \cdots + \alpha_{18}\hat{p}_8^{(0)}} + \cdots + \frac{\alpha_{5j}\hat{p}_j^{(0)}}{\alpha_{51}\hat{p}_1{}^{(0)} + \cdots \alpha_{58}\hat{p}_8^{(0)}}\right]$$

Stopping criteria:

If $\sum_{j=1}^{m} \left|\hat{p}_j^{(1)} - \hat{p}_j^{(0)}\right| > \epsilon$, we let $\hat{p}_j^{(0)} = \hat{p}_j^{(1)}$ and continue the iteration or otherwise stop.

After one iteration, we check if $\sum_{j=1}^{m} \left|\hat{p}_j^{(1)} - \hat{p}_j^{(0)}\right| \leq \epsilon$. If yes, then we stop iteration. Then the $\hat{p}_j^{(1)}$ is the final estimate.

$\hat{S}(t) = \sum_{\tau_j > t} \hat{p}_j = 1 - \sum_{\tau_j \leq t} \hat{p}_j$ is the final survival function estimate from the iterations.

2.3 HT Method

We now discuss the HT method briefly in this section.

Let $G_i = \{E_{ij}: \delta_i = 1 \text{ and } L_i < E_{ij} \leq R_i\}$, where E_{ij}: jth coital episode time in the observaed interval, $j = 1, \dots, n_i$ and $G_i = \phi$ if $\delta_i = 0$, if a subject is not infected (Harezlak & Tu, 2006).

For operational convenience, let $G_i = \{C_i: \delta_i = 0\}$.

Algorithm:

1. For the bth resampled data set, we impute uniformly one infection (or censoring) time for the ith subject from the set G_i and denote it as $X_i^{(b)}$.

   - $X_i^{(b)} = E_{ij}$, if a subject i is infected for some j coital events in the infection interval.
   - $X_i^{(b)} = C_i$, if a subject i is right-censored.

   When this process is completed for all n subjects, we have a complete right censored data set:

   $\{X_i^{(b)}, \delta_i$; for $i = 1, \dots, n$.

2. For $\{X_i^{(b)}, \delta_i$; for $i = 1, \dots, n\}$ obtained in (1), compute the Kaplan-Meier estimate, $\hat{S}^{(b)}(t)$:

   $t_1^{(b)}, t_2^{(b)}, \dots, t_q^{(b)}$ = the resampled distinct infection times.

   $N_r^{(b)}$ = the # of infections at time $t_r^{(b)}$.

   $R(t_r^{(b)})$ = the # of subjects at risk at time $t_r^{(b)}$.

   $$\hat{S}^{(b)}(t) = \prod_{r=1}^{q} \left\{ \frac{1 - N_r^{(b)}}{R\left(t_r^{(b)}\right)} \right\}.$$

3. Repeat steps 1-2 B times and combine the Kaplan-Meier estimates to obtain the HT's survival function estimate: for each data set:

$$\hat{S}^{*}(t) = \frac{1}{B}\sum_{b=1}^{B}\hat{S}^{(b)}(t)$$

2.4 Proposed Method

We now focus our attention to our proposed method of study.

Let T_i be the true failure (infection) time for the ith subject, where $i = 1, \ldots, n$ and C_i be the right-censoring time (dropout time for subject without infection) and $\delta_i = I(R_i \leq C_i)$ be the positive test indicator. We note that $T_i < R_i \leq C_i$ is for an infected subject; and $C_i < R_i$ is for a right-censored subject.

Once a subject is tested positive at a follow-up visit, we can utilize her diaries to find the infection time to a set of coital episodes at which the infection could have been developed. We denote such a set of likely infection times in infected subject i as

$\mathbb{G}_i = \{u_{ij} : \ \delta_i = 1 \text{ and } L_i < u_{ij} \leq R_i\}$,

where u_{ij} is the jth coital episode time in $(L_i, R_i], j = 1, \ldots, n_i$, and n_i is the number of coital events recorded.

- u_{i1}, first coital episode time

  .
- .

  .
- u_{in_i}, last episode time

However, if a subject i is not infected before droping out or at the end of the study,

$\mathbb{G}_i = \emptyset$ if $\delta_i = 0$. Nonetheless for ease, we let $\mathbb{G}_i = \{C_i : \delta_i = 0\}$. The elements of $\mathbb{G}_i$ form the support of the predictive distribution for the ith subject in our imputation scheme.

For each subject, we propose a multiple imputation procedure that resamples (with replacement) the possible true infection times from the times of coital episodes and taking into account whether condom was used at each time. The infection time of a study subject is considered right-censored if subject remained uninfected at her last clinical visit. Since the clinical visit times are pre-scheduled at $1, 3, 5,$ and 7 months according to the study design, the right-censoring time, C_i, would be independent of the true infection, T_i, for all i (Harezlak & Tu, 2006). On the other hand, it is interval-censored if she was infected between her visits. Assume T_i is independent of $(L_i, R_i]$. We use the following procedure to get an estimate of the survival function of the infection time, $\hat{S}(t)$.

The proposed method consists of three steps.

Step 1: Impute the interval-censored data to obtain a right-censored data.

We use B as the total number of imputation data sets and superscript (b) to represent the bth imputed data set. For each data set, denote the imputed right-censored failure time data by $\{X_i^{*(b)}\ , \delta_i^{(b)}, i = 1, \dots, n\}$ based on the n subjects in the original data, where $X_i^{*(b)}$ denotes the imputed failure time from $\mathbb{G}_i$, and $\delta_i^{(b)}$is the event indicator.

For the bth data set, we select one time point from the set $\mathbb{G}_i$ of each subject i randomly where the selection is affected by whether the condom was used or not. Specifically, let

$$C_{ij} = \begin{cases} 1, & if\ condom\ was\ used, \\ 0, & otherwise \end{cases}$$

Given the number of coital episodes recorded, n_i, odds of being infected in favor of not using condom (m: 1), the probability that u_{ij} is selected is

$\pi_{ij} = \frac{C_{ij}+m(1-C_{ij})}{Q}$, where $Q = \sum_{j=1}^{n_i} C_{ij} + m\left(n_i - \sum_{j=1}^{n_i} C_{ij}\right)$, $\sum_{j=1}^{n_i} \mathrm{C_{ij}}$ is total number of coital episodes of subject i where condom was used, and $(n_i - \sum_{j=1}^{n_i} C_{ij})$ is total number of coital episodes where condom was not used. Note that $\sum_{j=1}^{n_i} \pi_{ij} = 1$, and

$$\pi_{ij} = \begin{cases} \frac{1}{Q}, & C_{ij} = 1 \\ \frac{m}{Q}, & C_{ij} = 0 \end{cases}$$

If T_i is not right-censored, $X_i^{*(b)} = u_{\{ij\}}$, randomly generated based on the probabilities above and let $\delta_i^{(b)} = 1$. If T_i is right-censored, $X_i^{*(b)} = C_i$ and $\delta_i^{(b)} = 0$. Once this process is completed for all n subjects, we have a complete right censored data set $\{(X_i^{*(b)}, \delta_i^{(b)})$, $i = 1, \dots, n\}$ for the bth iteration.

Step 2. Using $\{(X_i^{*(b)}, \delta_i^{(b)}), i = 1, \dots, n\}$ generated from step 1, we compute the Kaplan-Meier survival curve estimate $\hat{S}^{(b)}(t)$. Let $t_1^{(b)}, t_2^{(b)}, \dots, t_q^{(b)}, q = 1, \dots, q_{(b)}$, where $q_{(b)}$ is distinct time points for exact observations for bth iteration, d_j's and n_j's are the number of subjects being infected at time $t_j^{(b)}$; and the number of subjects at risk right before time $t_j^{(b)}$, respectively. Then, we have

$$\hat{S}^{(b)}(t) = \Pi_{j:t_j^{(b)} \le t} \left(1 - \frac{d_j}{n_j}\right)$$

Step 3. We repeat Steps 1 and 2 B times and compute a survival function estimate by combining the Kaplan-Meier estimates obtained for each data set:

$$\hat{S}(t) = \frac{1}{B}\sum_{b=1}^{B}\hat{S}^{(b)}(t)$$

III. SIMULATION

Our goal is to compare three methods for right and interval censored data in a specific study setting. For the best comparison we should relate the estimates obtained by analysis of censored datasets to the true survival function which is only known for simulated datasets. Therefore we have to simulate datasets and cannot use observed data for our comparison. To perform a data simulation, one needs to specify many parameters. The name and nature of these parameters is described in a later section. Dependent on the parameter values different percentages of right-censoring of datasets can be simulated.

3.1 Data Generation Procedures

In order to test the performance of the proposed method against HT's and the Turnbull's methods, we conduct a simulation study using R program. Below, we describe the data generating steps. The steps include subjects' visits, infection time, censorings, coital event times, and condom usage.

1. Generate nv follow-up visit times for the ith subject as $V_{i1}, \dots, V_{i,nv}$ from a uniform distribution $V_{iv} \sim Uniform\ (60 \times v - 5, 60 \times v + 5)$ for $i = 1, \dots, n$ and $v = 1, \dots, nv$. V_{iv} is based on official times of visits with a uniform perturbation, nv (number of scheduled visit times) = 4, 8 or 16. For instance, *Illustration 1* shows a uniform distribution $V_{iv} \sim Uniform\ (60 \times v - 5, 60 \times v + 5)$ for four follow-up visits.

| − − − − − − −| − − − − − − −| − − − − − − −| − − − − − − −| → Time (days).
0 V_{i1} V_{i2} V_{i3} V_{i4}
60 ± 5 120 ± 5 180 ± 5 240 ± 5

Illustration 1. Follow up visit times.

2. Generate right-censoring time: $C_i \sim$ Discrete Uniform $\{V_{i1}, \dots, V_{i4}\}$ as shown in *Illustration 2*. We assume every follow-up time can be the last visit of the subject. For instance, if C_i for $i = 1, \dots, n$ positioned at V_{i2} (as shown by the "↓" arrow) in *Illustration 2*, this corresponds to the subjects' last visits.

↓ C_i

| − − − − − − −| − − − − − − −| − − − − − − −| − − − − − − −| → Time (days).
0 V_{i1} V_{i2} V_{i3} V_{i4}

Illustration 2. Right censoring time.

3. Generate infection time $T_i \sim$ Weibull (a, b), where a is shape parameter and b is the scale parameter of the Weibull to control right censoring. We choose $(1.2, 0.8, 1) \in a$ corresponding to increasing, decreasing, and constant hazard function, respectively. b is chosen to achieve a certain percentage of right-censored observations. We choose 20%, 35%, and 50% right censoring representing light, moderate and heavy right-censoring, respectively. We have right censoring (RC) if $T_i > C_i$ and interval censoring (IC) if $T_i \le C_i$.

4. Determine interval observation $I_i = (L_i, R_i]$. If $T_i > C_i$, the observation is right-censored and $I_i = (C_i, \infty)$, and $\delta_i = 0$.

Otherwise, let $V_{i0} = 0$. We define d_{ij} for $j = 1, 2$; number of visits skipped by the subject i; d_{i1} is number of visits skipped before T_i and d_{i2} is the number of visits skipped after T_i. If d_{i1} and d_{i2} are zero then the subject i does not skip any visit.

If $V_{ij-1} < T_i \leq V_{ij}$ for some j, generate d_{i1} and d_{i2} independently from $Discrete\ Uniform\ \{0, 1, \ldots, D\}$ and let $l =$ max $\{0, j-1-d_1\}$ and $r =$ min $\{v, j+d_2\}$. Then, $I_i = (V_{il}, V_{ir}]$. Note that D is an integer chosen to control overall width of the observed intervals. If $D = 0$, the subjects do not skip a visit as in HT study.

5. To generate the coital event times in an interval for an infected subject i, let J_i be the number of scheduled follow-up periods in $(L_i, R_i]$. We first generate the number of coital event times n_i from the binomial distribution ($n_i \sim Bin(64J_i, \frac{1}{8})$) when $j_i = 1$ as in HT method between two consecutive visits with the mean equal to 8 and standard deviation equal to 2.646 (Variance equal to 7). The Binomial parameters are chosen to mirror the STD diary data. Let one of the coital event times be the true infection time, while the rest of $(n_i - 1)$ coital episodes are generated from $Uniform\ (L_i, R_i]$.

   Example in *Illustration 3*. $J_i = 2$ (covers two consecutive visits $(V_{i1}, V_{i2}]$ and $(V_{i2}, V_{i3}]$, $n_i = 6$ (represented by "×") and $n_i \sim Bin(64J_i, \frac{1}{8})$, $u_{ij} \sim Uniform(L_i, R_i]$.

   $\downarrow T_i$

   $| - - - (- \times - \times - \times -| -\times - \times - \times - \times -] - - - - - - - | \rightarrow$ Time (days).

   0 V_{i1} V_{i2} V_{i3} V_{i4}

   Illustration 3. Coital events in the interval $(\mathrm{L_i}, \mathrm{R_i}]$.

6. Determine condom use: We assume coital events that are at or close to T_i have higher chance of condom being not used. The closeness is measured by $|u_{ij} - T_i|$.

To illustrate the condom use in our simulation, we use the following procedure to estimate the probabilities for condom use at each of the $n_i$coital events.

a. Consider a proportion $(prop)$ of sexual event times that are "close" to infection time, T_i.

b. Consider ceiling $(N_i \times prop)$: number of the coital event times, u_{ij} that are close.

c. Determine whether condom is used: for u_{ij}'s close to T_i: generate C_{ij}~Ber$(p_1 = P(CondomUse))$ and u_{ij}'s further from T_i: C_{ij}~Ber$(\ p_2 = P(CondomUse))$.

3.2 Estimation of Mean Integrated Squared Error (MISE) for Survival Function

As stated earlier, one of the goals of this thesis is to examine estimates of the survival function, in this case whether the subjects have used condoms or not during the coital episode. Using this information, we intend to compare our proposed methods against the HT and Turnbull method. The comparisons are based on the MISE. Since the true survival function is known in the simulations, we report MISE $(E(\int_t \left(\hat{S}(t) - S(t)\right)^2 dt))$ estimate for each method based on

$$MISE = \frac{\sum_{r=1}^{M} \sum_{j=1}^{m^{(r)}} \left[\hat{S}^{(r)}\left(\tau_j^{(r)}\right) - S\left(\tau_j^{(r)}\right)\right]^2)}{M},$$

where $S(\cdot)$ is the true survival function and $\hat{S}^{(r)}(\cdot)$ is an estimated survival function based on the rth set of interval-censored data generated, and $\tau_j^{(r)}$'s are distinct time points provided by Turnbull's method which are based on the observed intervals of the rth dataset. Moreover, based on the M estimated survival functions an average survival probability was computed at each time point over the union of M sets of distinct time points.

Under this data simulation scheme, we generate M = 500 data sets with $n = 100$ subjects. With 20, 35, 50 percentage of right-censored data sets which means we can observe 80, 65, and 50 percentage of interval-censored data sets.

In the simulations, we generate the infection/survival time from Weibull (a, b). Each simulation run corresponds to a unique combination of the values shown in *Table 6* and variables defined subsequently. Our next step is to run the integers corresponding (input or parameter) to the variables from *Table 6* and *Table 7* into *R* function.

Table 6. Weibull distribution and percentage of right censoring.

a	Right-censoring (%)
0.8	20
1.0	35
1.2	50
0.8	20
1.0	35
1.2	50
0.8	20
1.0	35
1.2	50

Table 7. Parameters used for simulation.

Description	Parameter Value
No. of scheduled visits after being enrolled in study	
Distribution of Vij ~Uniform	nv = 4, 8, 16
BinN: Ni ~ Bin (BinN, 1/8)to generate number of coital event times Ni	BinN = 64, 32, 16
Vector of probabilities of being censored at $V_{\{iv\}}$'s	p = (1/nv, ...,1/nv4)
Controls width of $(L_i, R_i]$ for an interval-censored data	D = 0
Probability of condom use at events close to T_i	p1 = 0.1
Probability of condom use at events close to T_i	p2 = 0.9
Proportion of events that are defined as close to T_i	prop = 1/10
Odds in favor of no condom use is m: 1	m = 10
max $\{L_i, R_i, i = 1, \dots, n\}$, where $R_i = \infty$; $maxday$	maxday = 260
Tolerance	tol = 1e-7

3.3 Estimation of Bias for Survival Function

By looking at the bias one can find out more about the general direction of the error and about its share of the MISE. From the interval censored data, we computed $Bias(\hat{S}(t))$ at some given t by

1) Finding the mean of the column corresponding to t, $\frac{1}{M}\sum_{r=1}^{M}\hat{S}(t)^{(r)}$,
2) Then, estimating the $Bias$ of $\hat{S}$ at t:

$$\widehat{Bias}\left(\hat{S}(t)\right) = \frac{1}{M}\sum_{r=1}^{M}\hat{S}(t)^{(r)} - S(t),\, t \in \{1, 2, \dots, 260\}$$

where infection times $T{\sim}Weibull$(a = shape, b = scale), $S(t)$ =exp$\left[-\left(\frac{1}{b}\cdot t\right)^{a}\right]$. Note, we will use the same parameters from *Table 6* and *Table 7* for biases estimation as well.

IV. RESULTS

4.1 Application Results

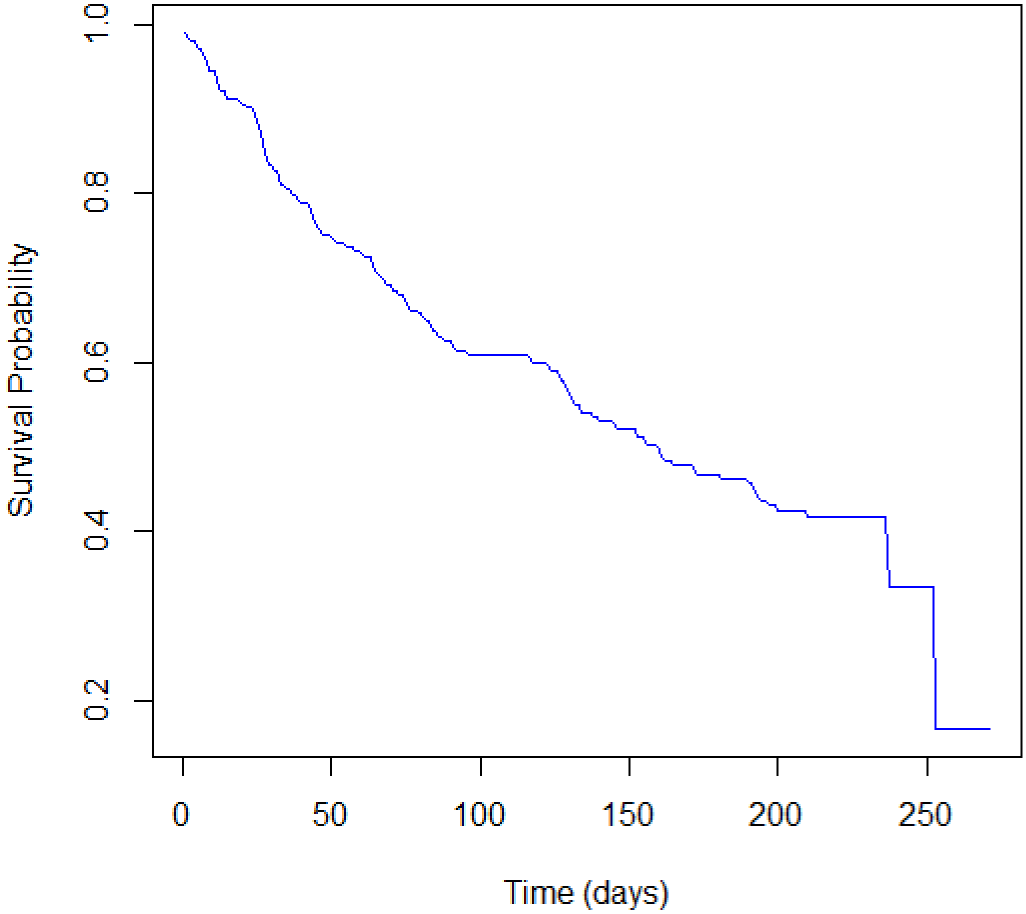


Figure 5. Estimated survival function using the proposed method.

The survival curve for the proposed method is shown in *Figure 5*. The y-axis depicts the probability of survival estimate and x-axis depicts time in days. At the start of the study all subjects are not yet infected. Each time a subject was infected the line takes a tick downward to indicate that the number of subjects are still uninfected has decreased (Lindsey & Ryan, 1998). As we move out over time, we can see that there is less chance subjects are still uninfected. The percentage of patient infected beyond day 100 is roughly

60 whereas the percentage of patient infected within 100 days is 40. At 50 percent corresponding with 150 days more than half subjects were infected with STD. By the end of the study period we see that we have subjects who used condom or not have a probability of less than 20 percent still uninfected.

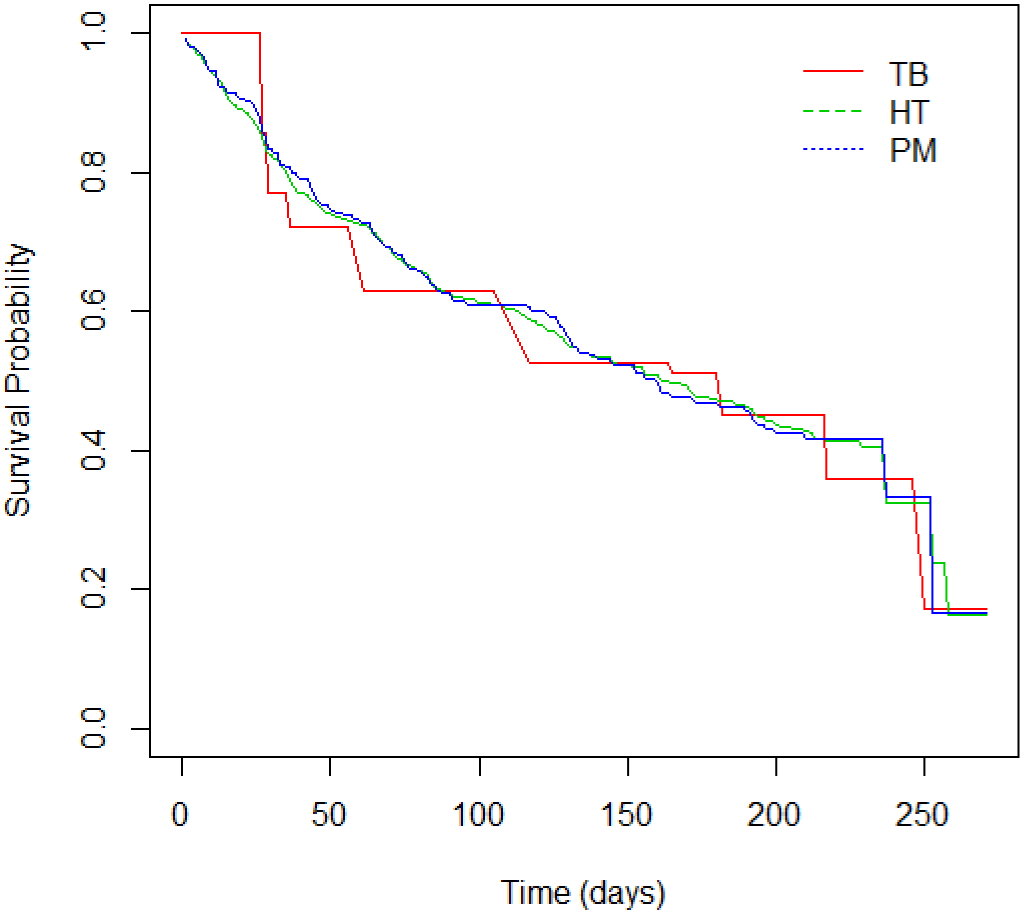


Figure 6. Depicts survival curves: Turnbull (top left), Harezlak and Tu (top right), Proposed (bottom left), and overlayed together (bottom right): *PM (Survival Proposed Method)*, HT (Survival Harezlak and Tu Method), and TB (Turnbull).

.

The survival curves for the three methods are shown in *Figure 6*. This plot indicate that subjects with Proposed and HT methods have better prognosis up until 150 days than Turnbull method. Moreover, it appears that the difference between the two methods is smaller than Turnbull throughout. Overall, there does not appear to be a dramatic difference between all three.

4.2 Simulation Results

4.2 (a). MISE Results

We now report the results of a simulation study comparing the performances of the three methods, namely (1) the Turnbull method proposed by B. W. Turnbull (Turnbull, 1976), (2) the HT method proposed by Tu and Harezlak (Harezlak & Tu, 2006), and (3) our proposed method.

4.2 (a) (i). MISE values for sixteen number of scheduled visits (nv =16):

Table *8* shows the computed MISE values for sixteen number of scheduled visits with uniform distribution (vdbn = 1).

Based on the result, we can see that the bigger the percentage of right censoring is bigger in general for the MISE estimates. To compare the methods against each other, MISE for the Turnbull method (MISE-T) shows higher MISE than the other two methods. There is not a significant difference in MISE between the HT (MISE-HT) and Proposed method (MISE-P) in this settings. Therefore, we concluded that the HT and Proposed method are significantly better than Turnbull method in our simulation based on sixteen number of scheduled visits after enrolled in the study.

Table 8. MISE values of three methods using simulated data: n = 100, M = 500, and R = 20; Weibull shape parameters 0.8, 1.0, 1.2 and 20, 35, 50 percent right-censoring (RC).

% RC	MISE-T	MISE-HT	MISE-P
20	0.8990644	0.4346771	0.4381402
20	0.6912520	0.3184233	0.3374687
20	0.6306498	0.2935330	0.3161075
35	1.0873459	0.7278090	0.7537602
35	1.0040464	0.6734604	0.6967028
35	0.8803182	0.5615514	0.5968780
50	1.1593806	0.8769619	0.9072753
50	1.1160584	0.8125789	0.8475597
50	1.1179644	0.8109986	0.8485383

4.2 (a) (ii). MISE values for four number of scheduled visits (nv = 4):

Table 9 shows the computed MISE values for four number of scheduled visits with uniform distribution (vdbn = 1). We observer from *Table 9* that the MISE values for Proposed method are continuously increasing as the percentage of right-censoring is increasing. But, in the case of other two methods, MISE is decreasing as the percentage of right-censoring is increasing. This phenomenon cannot be explained easily. That means the accuracy of the estimates is continuously decreasing. A potential explanation for this increase in MISE with increase in right-censoring is the decrease in the number of patients during the course of the study. Furthermore, the presence of C_{ij} in $(L_i, R_i]$ becomes more important in wider intervals. Overall, with a fewer scheduled visits, our method performs significantly better than the other two methods.

Table 9. MISE values of three methods using simulated data: n = 100, M = 500, and R = 20; Weibull shape parameters 0.8, 1.0, 1.2 and 20, 35, 50 percent right-censoring (RC).

% RC	MISE-T	MISE-HT	MISE-P
20	9.347350	1.612659	0.591491
20	7.2165327	1.0258137	0.4577241
20	6.1247636	0.7865365	0.3886110
35	3.2984721	0.7488157	0.6520763
35	4.4665377	0.7291003	0.6219940
35	3.9284993	0.5786846	0.5495322
50	3.3405317	0.7660412	0.6644610
50	2.8938508	0.6408469	0.6389000
50	2.7860596	0.6457157	0.7027665

4.2 (a) (iii). MISE values for eight number of scheduled visits (nv = 8):

The computed MISE values for eight number of scheduled visits with uniform distribution (vdbn = 1) are displayed in *Table 10*.

Table 8 depicts similar to *Table 10*.

Table 10. MISE values of three methods using simulated data: n = 100, M = 500, and R = 20; Weibull shape parameters 0.8, 1.0, 1.2 and 20, 35, 50 percent right-censoring (RC).

% RC	MISE-T	MISE-HT	MISE-P
20	2.5510057	0.5729613	0.4442325
20	1.9232118	0.4038000	0.3805245
20	1.5336815	0.3104111	0.3105633
35	1.9109433	0.6624638	0.6646909
35	1.6205296	0.5761161	0.6208983
35	1.4731332	0.5206552	0.5705658
50	1.5562110	0.7555878	0.7874246
50	1.5540335	0.7850342	0.8352650
50	1.5633537	0.7653063	0.8317692

4.2 (a) (iv). MISE values for sixteen number of scheduled visits (nv = 16):

The computed MISE values shown in *Table 11* for sixteen number of scheduled visits with number of visit times to be constant distribution (vdbn = 3).

Table 11. MISE values of three methods using simulated data: n = 100, M = 500, and R = 20; Weibull shape parameters 0.8, 1.0, 1.2 and 20, 35, 50 percent right-censoring (RC).

% RC	MISE-T	MISE-HT	MISE-P
20	1.3880587	0.4024294	0.4091177
20	1.1528719	0.3484485	0.3703734
20	1.0123281	0.2855357	0.3040276
35	1.2178128	0.6859460	0.7125337
35	1.0706326	0.6011174	0.6326794
35	1.0847910	0.5867803	0.6218295
50	1.1369364	0.8480740	0.8748477
50	1.0870349	0.7838913	0.8134785
50	1.1480899	0.8255319	0.8644642

4.2 (a) (v). MISE values for four number of scheduled visits ($nv = 4$):

The computed MISE values for four number of scheduled visits with number of visit times to be constant distribution (vdbn = 3) are displayed in *Table 12*. In this scenario, our method seems to perform better against the Turnbull and HT method with 20 and 35 percent right-censoring. Although, our method performs significantly better against Turnbull method in all percent right-censoring, but it does not perform better against HT method with 50 percent right-censoring.

Table 12, MISE values of three methods using simulated data: n = 100, M = 500, and R = 20; Weibull shape parameters 0.8, 1.0, 1.2 and 20, 35, 50 percent right-censoring (RC).

% RC	MISE-T	MISE-HT	MISE-P
20	11.978181	1.570904	0.577512
20	10.2951322	1.0326246	0.4586705
20	9.4465559	0.7362482	0.3595773
35	4.3100187	0.8005320	0.7048722
35	6.2314232	0.6444899	0.5486675
35	6.1388079	0.5721504	0.5204539
50	4.2273473	0.7820437	0.7002792
50	3.9911653	0.6353892	0.6496869
50	3.9629056	0.5937034	0.6525611

4.2 (a) (vi). MISE values for eight number of scheduled visits (nv = 8):

The computed MISE values for eight number of scheduled visits with number of visit times to be constant distribution (vdbn = 3) are displayed in *Table 13*. We conclude under eight number of scheduled visits, our method perform significantly better against Turnbull method but perform satisfactorily against the HT method only for 20 percent right-censoring.

Table 13. MISE values of three methods using simulated data: n = 100, M = 500, and R = 20; Weibull shape parameters 0.8, 1.0, 1.2 and 20, 35, 50 percent right-censoring (RC).

% RC	MISE-T	MISE-HT	MISE-P
20	3.8879142	0.5904986	0.4479362
20	3.2639250	0.4270929	0.3950388
20	2.9508277	0.3481985	0.3439841
35	2.4669358	0.6344398	0.6329138
35	2.3156698	0.6174921	0.6576491
35	2.1377473	0.4818234	0.5451871
50	1.8202397	0.7274015	0.7464793
50	1.7712613	0.7704176	0.8302618
50	1.6701244	0.6329738	0.6942970

4.2 (b). Bias Results

Bias occurs when there is a systematic difference between the results from a study and the true state of affairs (Petrie & Sabin, 2005). By looking at the bias one can find out more about the general direction of the error.

4.2 (b) (i). Bias values for four number of scheduled visits (nv = 4):

Result from section 4.1 (b) showed the least MISE, we now compare the bias of three methods: Turnbull bias (Bias-T), HT method bias (Bias-HT), and Propose bias (Bias-P) at each time points (60, 120, 180, 240) with four non-perturbed scheduled visits.

The size of bias values is displayed in *Table 14* for this scenario. In *Table 9* the

MISE values with respect to this scenario for right-censored estimates are reflected in increased and decreased bias values as *Table 14*.

The biases obtained from four scheduled visits are much smaller for proposed method compared to other two methods at all percentage of right-censoring considered. The reason can be explained by having a wider $(L_i, R_i]$ which means less informative. It is interesting to note that there exists negative biases and positive biases in all methods. The bias for proposed method is almost zero for 20 percent right-censoring time points at 120, 180, and 240. A similar situation is seen for 35 and 50 percent right-censoring same time points. However, biases values are much more different in other two methods.

Table 14. Bias: n = 100, M = 500, R = 20, Weibull shape: 0.8, 1.0,1.2; and 20, 35, 50 percent right-censoring.

Time points	% RC	Bias-T	Bias-HT	Bias-P
60	20	0.04430400	0.109621468	0.0313317055
120	20	-0.01494402	0.012951529	0.0046611261
180	20	0.00472166	-0.002338721	-0.0009732671
240	20	-0.01032978	-0.013774516	-0.0047920508
60	20	0.051355200	0.091826309	0.027394609
120	20	-0.014378792	0.025366729	0.010448417
180	20	0.003684749	0.005119891	0.002301323
240	20	-0.007592578	-0.008673577	-0.002104848
60	20	0.058068314	0.079214330	0.024563448
120	20	-0.013026156	0.034938984	0.014483880
180	20	0.002344709	0.011502692	0.006128206
240	20	-0.003771359	-0.003395231	0.001339812
60	35	0.031457393	0.044925319	0.011673007
120	35	-0.012835622	0.006901286	0.003492425
180	35	0.006130413	-0.004319034	-0.001331350
240	35	-0.015800980	-0.024257837	-0.010184884
60	35	0.041508570	0.052156176	0.013562391
120	35	-0.014654658	0.017691572	0.007145820
180	35	0.010658967	0.004983518	0.004025459
240	35	-0.009875546	-0.017031667	-0.005980096
60	35	0.047957923	0.037610708	0.010339887
120	35	-0.015605400	0.020795686	0.005634600
180	35	0.003816715	0.006155429	0.002358016
240	35	-0.010157952	-0.016326204	-0.007148097
60	50	0.032681702	0.046110617	0.012254537
120	50	-0.012948062	0.005417002	0.002075243
180	50	0.003625843	-0.004646005	-0.000942266
240	50	-0.005599059	-0.020475542	-0.007052641
60	50	0.036500574	0.033850642	0.011494744
120	50	-0.003178681	0.014097122	0.007114672
180	50	0.008108355	0.003344834	0.001282178
240	50	-0.013887848	-0.020855639	-0.007776449
60	50	0.034860431	0.020154238	0.005750489
120	50	-0.005178909	0.014472285	0.004463457
180	50	0.003509867	0.004565061	0.001407979
240	50	-0.013828951	-0.021135544	-0.010854704

We observe from above table and we conclude proposed method perform significantly better against the other two methods.

4.2 (b) (ii). Bias Figures

In order to observe a clear differences, we construct bias figures. *Figure 7* - *Figure 9* shows the bias values at four time points with uniform distribution (vdbn =1). *Figure 7* - *Figure 9* compare and contrast the biases of three methods. In these figures, the bias among different time points (~60 – 100) is very high in Turnbull and HT method, and significantly lower in Proposed method. In addition, there is a small difference in variability in bias among different time points between HT and Proposed method which is depicted by *Table 14*.

For the most part in proposed method, especially, in early stages the right-censoring bias is positive, in late stage it is somewhat negative (see *Figure 7* - *Figure 9*). One can see in early stage (time points ~0 day to 150 days) that much bigger bias exists..

The bias is close to zero in mid time points for the proposed method. In addition, in late stage (time points 150 days and >250 days) the situation is more balanced i.e. close to zero or zero lines. The reason behind balanced biases can explained by more subjects are being censored especially as the study get closer to the end.

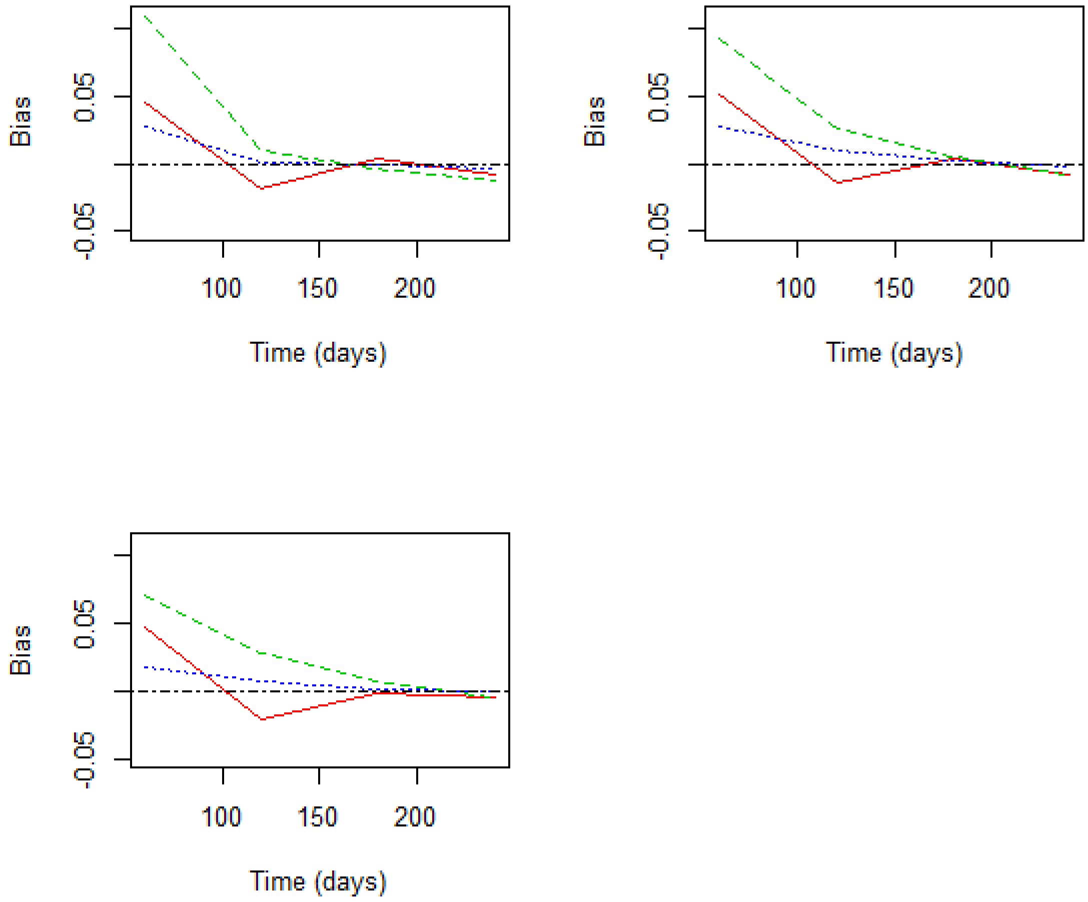


Figure 7. Bias: n= 100, M= 500, R= 20, nv = 4, vdbn = 1; Weibull shape and scale parameter: 0.8 (top left), 1.0 (top right), 1.2 (bottom left); 20 percent right-censoring rate. Legend: straight line: Turnbull method; long dotted line: HT method; dotted line: Proposed method.

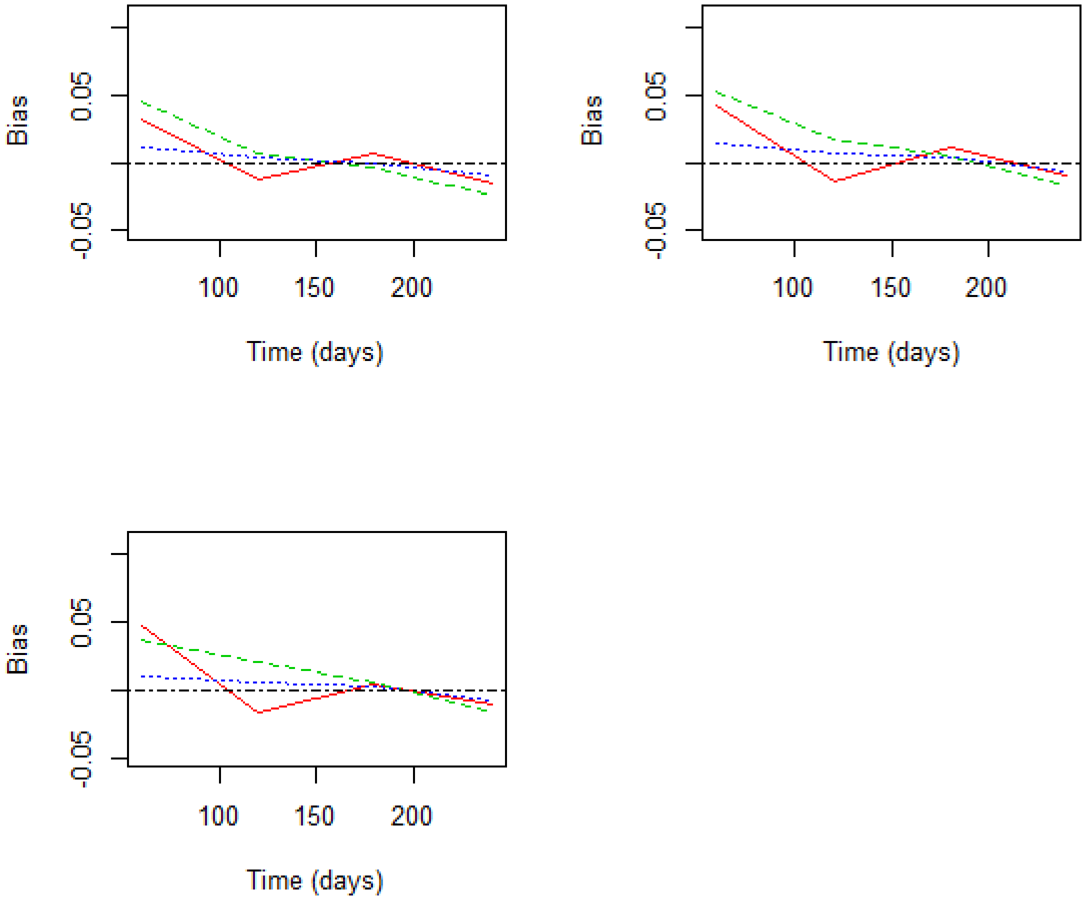


Figure 8. Bias: n= 100, M= 500, R= 20, nv = 4, vdbn = 1; Weibull shape and scale parameter: 0.8 (top left), 1.0 (top right), 1.2 (bottom left); 35 percent right-censoring rate. Legend: straight line: Turnbull method; long dotted line: HT method; dotted line: Proposed method.

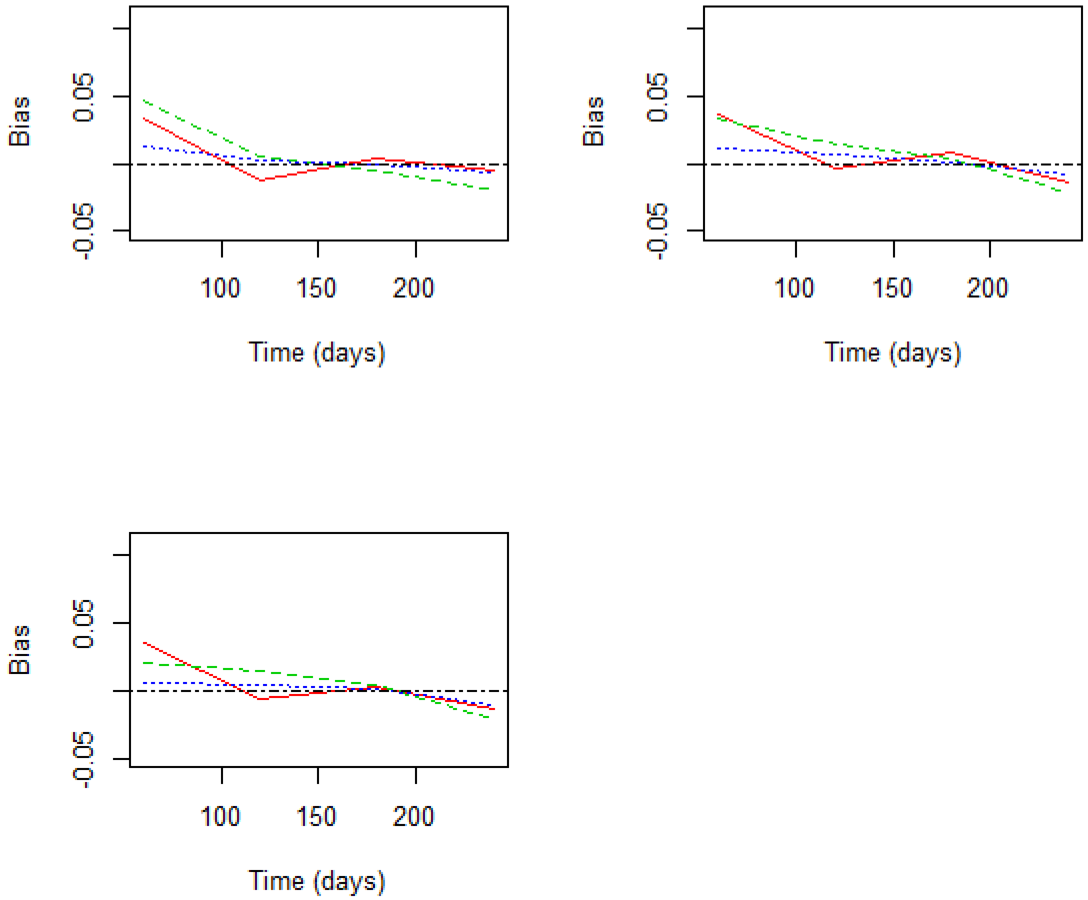


Figure 9. Bias: n = 100, M= 500, R= 20, nv = 4, vdbn = 1; Weibull shape and scale parameter: 0.8 (top left), 1.0 (top right), 1.2 (bottom left); 50 percent right-censoring rate. Legend: straight line: Turnbull method; long dotted line: HT method; dotted line: Proposed method.

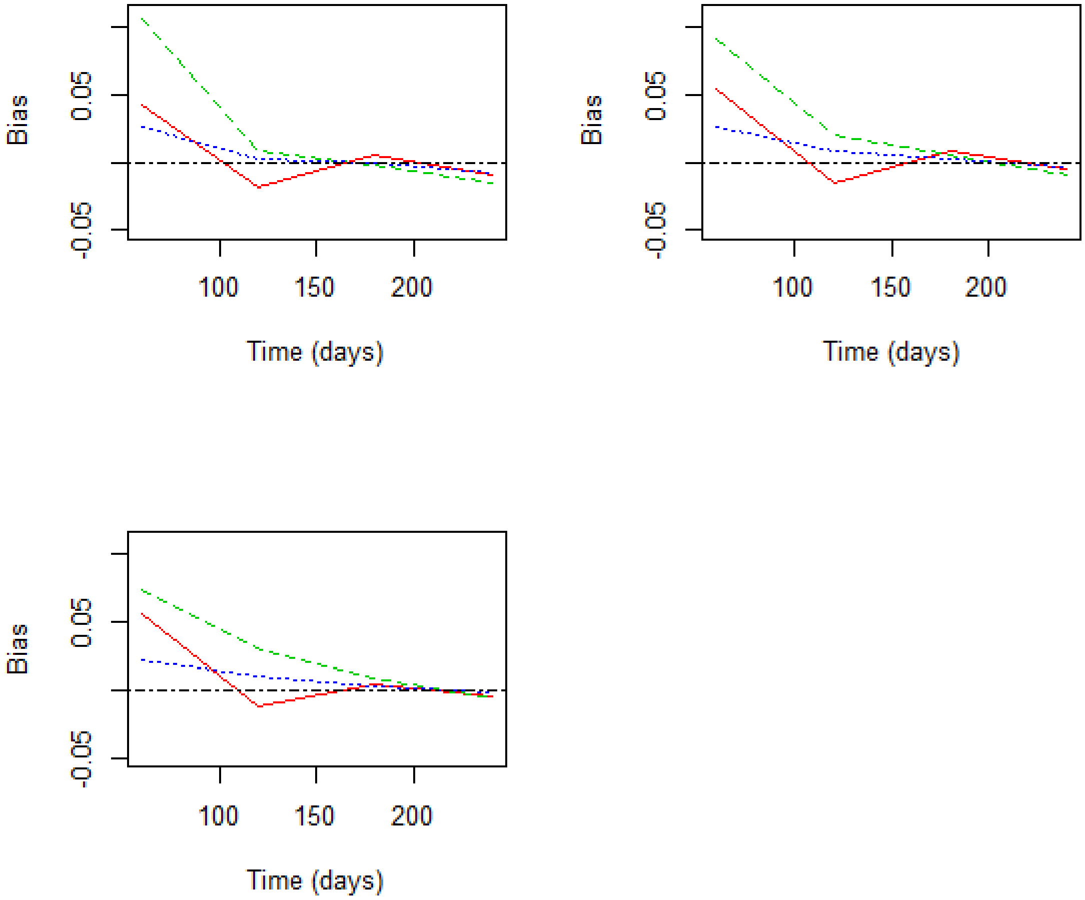


Figure 10. Bias: n = 100, M= 500, R= 20, nv = 4, vdbn = 3; Weibull shape and scale parameter: 0.8 (top left), 1.0 (top right), 1.2 (bottom left); 20 percent right-censoring rate. Legend: straight line: Turnbull method; long dotted line: HT method; dotted line: Proposed Method.

Figure 10 shows a bias that is based on 100 sample and with exact visit times (vdbn =3). From the above figure, we observe that proposed method bias is continuously decreasing as the time point increases. At time point ~175, the bias is almost zero. Beyond time point 220, the bias falls below zero (negative bias).

4.2 (b) (iii). More Bias Figures

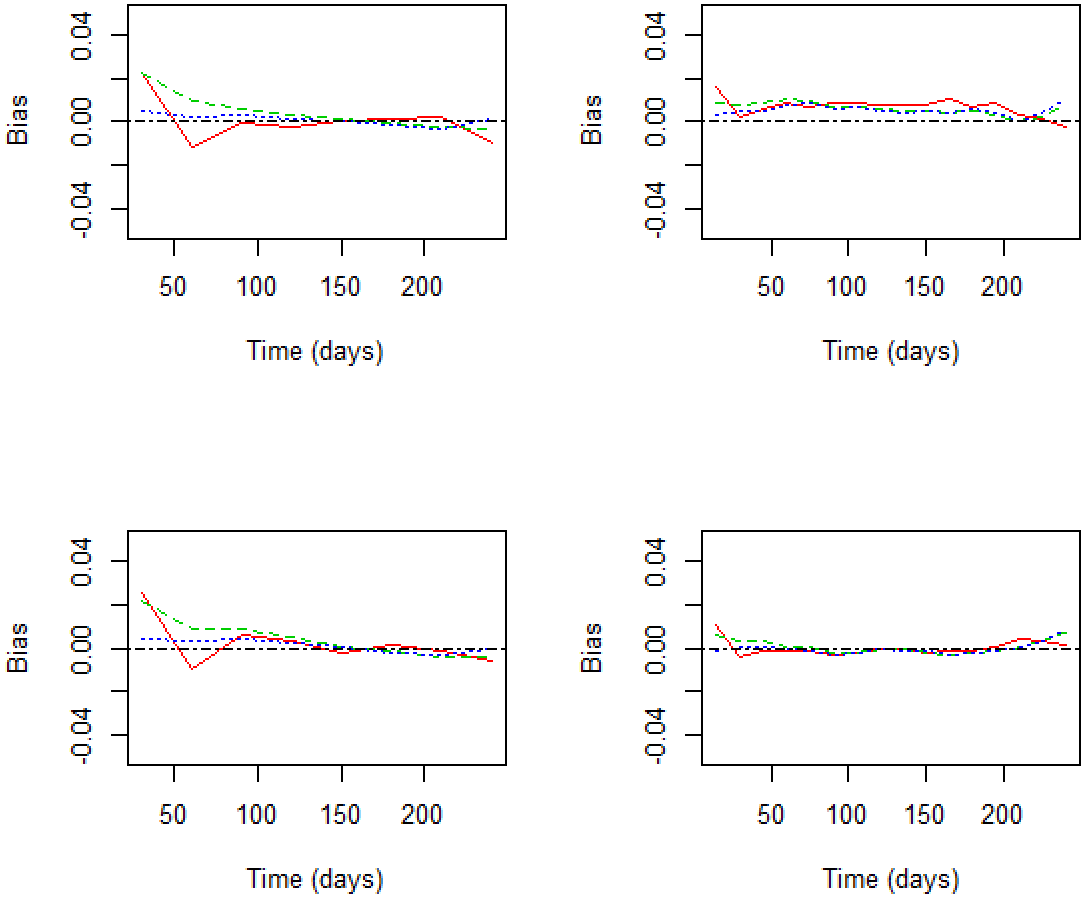


Figure 11. Bias: n = 100, M = 500, R = 20. Top (left): nv = 8, vdbn = 1; Top (right): nv = 16, vdbn = 1; Bottom (left): nv = 8, vdbn = 3; Bottom (right): nv = 16, vdbn = 3. Weibull shape (a = 1.0), 35 percent right-censoring rate. Legend: straight line: Turnbull method; long dotted line: HT method; dotted line: Proposed method.

Now, we can sum up by stating that it appears the estimation gets more accurate when there is 35 percent right-censorings due to dropouts. This is true particularly at later stages of the study (see *Figure 11*). This implies a common sense, because in interval-censored study we might have to wait prolong time until we can observe the target number of events. It also makes sense because more patients drop out from the study and more events at later time points. As a result we will have precise estimation (almost no bias) at later time points.

V. DISCUSSION

In this study, we propose a procedure that makes use of the auxiliary information provided by behavioral diaries. The proposed method uses an imputation-based approach to estimate the survival function of the infection time based on interval-censored data and is able to incorporate the important behavioral information provided by STD diaries. Our simulation study provides an overview of the behavior of interval-censored methods in the context of condom use of STD behavioral diaries. In a study like this one, it is certainly feasible to make it usable on a regular basis in clinical trials. Our simulation study suggests that by utilizing the diary information such as coital events and condom use, the MISE result reveal that proposed method performs better if not comparable to HT method and performs significantly better when compared against the Turnbull method in the settings mentioned in subsection 4.2 (a)(i) – 4.2 (a)(vi). Our proposed method is able to produce better bias results when the number of scheduled visits is four. Under this setting, the proposed method biases perform significantly better against Turnbull and HT method.

While analyzing the results it becomes clear that many combination of parameters could be used in this R function to improve knowledge about the behavior of the estimation methods for interval-censored with auxiliary information. Also, the way the data is generated in simulation seems “backwards” comparing with what happens in reality. It seems more natural that we generate the sexual times first, then condom use, censoring time, infection time based on condom use, and observed interval. However, it is hard to generate the infection time, especially based on a continuous parametric distribution, which can be used to compute the MISE and biases easily. Assuming that a subject’s sexual behaviors do not change significantly over time. We believe that the result may not differ much compared with determining the observed interval first and then coital event times and condom use, as don in our simulation.

VI. CONCLUSION AND FURTHER RESEARCH

In conclusion, the proposed method could not determine the comparisons among the methods as a result of nonexistence of true survival estimates. This led us to perform simulation study where we are able to determine the true survival estimates. Consequently, we are able to make the following remarks based on our simulation study:

1. In general, HT's and our method are comparable against each other and both significantly perform better against Turnbull's method.
2. Our method performs significantly better against the HT and the Turnbull method when the number of scheduled visits is four.
3. All settings have positive and negative biases in all percentages of right-censoring.

This study illustrates the importance of the decisions the subjects have to make involving in coital episode in order to prevent from the STDs. Condom use reduces the heterosexual transmission of sexually transmitted disease (STD). Hence, valid measures of condom use are critical for evaluating interventions to increase condom use and assessing whether condom use confounds the association between STDs and other risk factors.

Although, we are able to make some remarks, we could not make definite conclusion about our results based on limited number of settings we have performed. Therefore, we suggest the following investigation in the future to get better comparison results:

1. More settings should be examined to the idea alongside the MISE. For example, in our simulation, we didn't allow skipping visits. Allowing skip visits may result in a lower MISE because, as the observed intervals are wider and thus less informative, any information such as coital event times or condom use becomes

more helpful.

2. Investigate other potential influences on the accuracy of self-reported STD history such as method, promptness, and setting of STD treatment. Because this study was limited to a specific sample of females, further study is needed to determine the extent of underreported STD incidence and treatment among other populations (e.g., males, adults, and other ethnic groups).
3. Valid measures of condom use are needed to assess infection status; future research is needed to develop more accurate measures.

APPENDIX SECTION

APPENDIX A: R FUNCTIONS FOR IMPLEMENTING FOR METHODS

5.1. GenerateDiscreteDBN

Generate a random number given a p.m.f. where the random variable takes m values x with probabilities p returns a random number in ascending sorted x.

```
> GenerateDiscreteDBN <- function(x, p, tol)
+ {
+   #if(sum(p) == 1.0)
+   if(1.0 - tol <= sum(p) && sum(p) <= 1.0 + tol)
+   {
+     cp = cumsum(p)
+     u = runif(1, 0, 1)
+     j= 1
+     while(u > cp[j])
+     {
+       j = j + 1
+     }
+     #  out = list(x[j], j)
+     # names(out) = c("x", "j")
+     # print(c(u,",", cp,",", out))
+     # out
+     #   x[j]
+   }
+   else
+   {
+     print("sum(p) != 1 in GenerateDiscreteDBN().")
+     print(sum(p))
+     print(p)
+     break
+   }
+   x[j]
+ }
```

5.2. Discrete Uniform

Randomly draw a non-negaive integer between [a and b].

```
> DiscreteUniform = function(a, b)
+ {
+
+   u = runif(1, a, b+1)
+   if(u == b + 1)
+     u = b + 1 - 0.000001
+   floor(u);
+
+ }
```

5.3. Alpha Matrix

This function creates (and returns a matrix with entries 0 or 1.

```
> Alphamatrix <- function(left, right, times)
+ {
+   n <- length(left)
+   m <- length(times)
+   aa <- matrix(0, n, m)
+   for(k in 1:n)
+   {
+     aa[k,  ][left[k] < times & right[k] >= times] <- 1
+   }
+   return(aa)
+ }
```

5.4. Turnbull Algorithm

This function calculates estimator of p.d.f. of survival time for interval-censored data with diary information using the Turnbull's self-consistent algorithm. Note that intervals are considered as $(L_i, R_i]$ so p(0) = 0. Input includes $(L_i, R_i]$. Ouput has two columns: column 1 is time and column 2 has the corresponding density value.

```
> Turnbull=function(data, criterion)
+ {
+   n <- nrow(data)                          # sample size
+   times = sort(unique(c(data[, 1], data[, 2])))
+   m <- length(times)                                            # No of distinct values
+   alpha <- Alphamatrix(data[, 1], data[, 2], times)
+   epsilon <- 1
+   p1 <- rep(1/m, m)
+   while(epsilon > criterion) {
+     p2 <- alpha %*% matrix(p1, m, 1)        # n-vector of denominator of (2) of manuscript
+     p2 <- matrix(1, 1, n) %*% (alpha * matrix(1/p2[, 1], n, m))
+     p2 <- (p2[1,  ] * p1)/n
+     epsilon <- sum(abs(p2 - p1))
+     p1 <- p2
+   }
+   result <- cbind(times, p1)
+   return(result)
+ }
```

5.5. CompleteSurvMatrix

This function completes survival probability at time points (1: maxday) for each subject.

Does not use linear interpolation: example, if S(3) = a and S(6) = b with a > b, then S(4) = S(5) = a becuause S is a right-continuous step function.

```
> CompleteSurvMatrix <- function(SS, B, maxday)
+ {
+   for(i in 1:B)
+   {
+     prob = 1
+     for(j in 1:maxday)
+     {
+       if(is.na(SS[i, j]))
+         SS[i,j] = prob
+       else
+         prob = SS[i,j]
+     }
+   }
```

5.6. Imputation

Change input from ImputingData (), but two functions are the same. Input: icdata $(L_i, R_i]$ = number of coital events recorded, sextimes = times of coital events, output: imputed right-censored data.

```
> ImputingData2 <-function(icdata, ni, sextimes, n, zero)
```

```
+ {
+   delta = rep(0, n)
+   t = rep(NA, n)
+   for(i in 1:n)
+   {
+     if(icdata[i,2]==Inf) # RC
+     {
+       t[i] = icdata[i,1] # ti = Ri since ti is in (Li, Ri]
+     }
+     else #IC
+     {
+       if(ni[i] == 0) # no coital events
+       {
+         t[i] = icdata[i,2] # ti = Ri since ti is in (Li, Ri]
+       }
+       else # ni > 0 coital events
+       {
+         prob = rep(1/ni[i], ni[i]) # equal prob for each coital event
+         t[i] = GenerateDiscreteDBN(sextimes[i, 1:ni[i]], prob, zero)
+       }
+       delta[i] = 1
+     }
+   }
+   cbind(t, delta)
+ }
```

5.7. HT Method

Input: icdata $(L_i, R_i]$, $n_i =$ number of coital events recorded, sextimes = times of coital events, R = number of imputed datasets, totaldays = max $\{L_i$ and $R_i <$ infty$\}$.

Output: time points 1: maxday and R estimated survival functions.

maxday: max$\{L_i, R_i, i = 1, \dots, n\}$, where R_i ! = infty. Maxday = 271 or STD data; maxday = 260 can be used for simulated data.

Change input from HTMethod(), but two functions are the same.

```
> HTMethod2 <- function(icdata, ni, sextimes, R, totaldays, zero)
+ {
+   n = length(ni)
+   SS = matrix(NA, nrow = R, ncol = totaldays)
+   for(r in 1:R)
+   {
```

```
+     datar = ImputingData2(icdata, ni, sextimes, n, zero)
+     #print(datar)
+     km = survfit(Surv(as.numeric(datar[,1]), as.numeric(datar[,2])) ~ 1)
+     times = km$time
+     S = km$surv
+     SS[r, times] = S
+   }
+   SS = CompleteSurvMatrix(SS, R, totaldays)
+   # print(SS[1:5, 1:30])
+   S = apply(SS, 2, mean)
+   t = 1:totaldays
+   cbind(t, S)
+ }
```

5.8. ProposeMethod

Input: icdata $(L_i, R_i]$, n_i = number of coital events recorded, sextimes = times of coital events, R = number of imputed datasets, totaldays=max{L_i and R_i < infty}. Odds: odds of being infected for not using condom (1: odds). Condom use = type of condom use at each coital event (structure: $n \times$ total data). Output: time points 1:maxday and R estimated Survival functions, maxday: max$\{L_i, R_i, i = 1, \ldots, n\}$, where R_i ! = infty. maxday = 271 for STD data; maxday = 260 can be used for simulated data.

```
> ProposedMethod <- function(icdata, ni, sextimes, R, condomuse, odds, totaldays, zero)
+ {
+   n = length(ni)
+   SS = matrix(NA, nrow = R, ncol = totaldays)
+   for(r in 1:R)
+   {
+     datar = NewImputingData(icdata, condomuse, odds, ni, sextimes, n, zero)
+     #print(datar)
+     km = survfit(Surv(as.numeric(datar[,1]), as.numeric(datar[,2])) ~ 1)
+     times = km$time
+     S = km$surv
+     SS[r, times] = S
+   }
+   SS = CompleteSurvMatrix(SS, R, totaldays)
+   # print(SS[1:5, 1:30])
+   S = apply(SS, 2, mean)
+   t = 1:totaldays
+   cbind(t, S)
+ }
```

5.9. NewImputingData

Change input from ImputingData(), but two functions are the same. Input: icdata $(L_i, R_i]$, condom use indicators, n_i = number of coital events recorded, sextimes = times of coital events. Output: imputed right-censored data.

```
> NewImputingData <-function(icdata, condomuse, odds, ni, sextimes, n, zero)
+ {
+   delta = rep(0, n)
+   t = rep(NA, n)
+   for(i in 1:n)
+   {
+     if(icdata[i,2]==Inf) # RC
+     {
+       t[i] = icdata[i,1] # ti = Ri since ti is in (Li, Ri]
+     }
+     else #IC
+     {
+       if(ni[i] == 0) # no coital events
+       {
+         t[i] = icdata[i,2] # ti = Ri since ti is in (Li, Ri]
+       }
+       else # ni > 0 coital events
+       {
+         count = sum(condomuse[i, 1:ni[i]])
+         Li = count + odds*(ni[i] - count)
+         p = rep(1/Li, ni[i])
+         prob = p +  p * (odds-1) * (rep(1, ni[i]) - condomuse[i, 1:ni[i]])
+         # prob = rep(1/ni[i], ni[i]) # equal prob for each coital event
+         t[i] = GenerateDiscreteDBN(sextimes[i, 1:ni[i]], prob, zero)
+       }
+       delta[i] = 1
+     }
+   }
+   cbind(t, delta)
+ }
```

5.10. ComputeCij

```
> ComputeCij=function(ni, condomuse)
+ {
+   n = length(ni)
+   cij = matrix(NA, nrow=n, ncol=dim(condomuse)[2])
+   for(i in 1:n)
+   {
+     if (ni[i] > 0)
+     {
+       for (j in 1:ni[i])
+       {
+         if(condomuse[i, j] == 3)
+           cij[i, j] = 1
+         else
+           cij[i, j] = 0
+       }
+     }
+   }
+   cij
+ }
```

5.11. CondomUseProb

The function returns probabilities for condom use at each of the Ni coital events. Input: N_i = number of coital events > 0, T_i = true infection time, $u_{\{ij\}}$ = vector of coital event times and $T_{i,j} = 1,\ldots,N_{i-1}$. Output: Prob of using condom at $u_{\{ij\}}$'s. Method: the first ceiling(N_i × prop) event times that are close to T_i gets p1 and the rest gets p2 as P(condom use). Problem: probs are too small for large Ni.

```
> CondomUseProb = function(Ni, Ti, uij, p1, p2, prop)
+ {
+     if(Ni > 0)
+     {
+         prob = rep(p2, Ni)  # p2 assigned to events for from Ti
+           prob[sort(abs(uij-Ti), index=T)$ix[1:ceiling(Ni*prop)]] = p1     #ceiling function
returns a value >= 1.
+         prob
+     }
+     else
+         warning("No coital events. So quit!")
+ }
```

APPENDIX B: R FUNCTIONS USED IN SIMULATION

6.1. GeneratingDataNew

This function generate STD data with coital event times and also, you get condom use information. n = sample size, nv = No. of scheduled visits after being enrolled in study, $vdbn$ = distribution of V_{ij}. $vdbn$ = 1 (Uniform); $vdbn$ = 2 (Triangle); $vdbn$ = 3 (constant): number of visit time to be constant, $BinN$: N_i ~ Bin(BinN, 1/8): number to generate number of sex times (N_i), p = vector of probabilities of being censored at $V_{\{i\}}$'s. Use discrete uniform (has to satisfy % of right-censoring): number of chance of being right censored at 4 visit times is 0.25, $B =$ controls width of $(L_i, R_i]$ for an interval-censored data, a & b: shape and scale parameter of a Weibull distribution, $p1$: prob of condom use at events close to T_i, $p2$: prob of condom use at events not close to T_i, $prop$: proportion of events that are defined as close to T_i, $maxday$ is an estimated upper limit for number of coital events, $output$: a data frame with fields icdata= (L_i, R_i, N_i), coitaltimes, and condomuse.

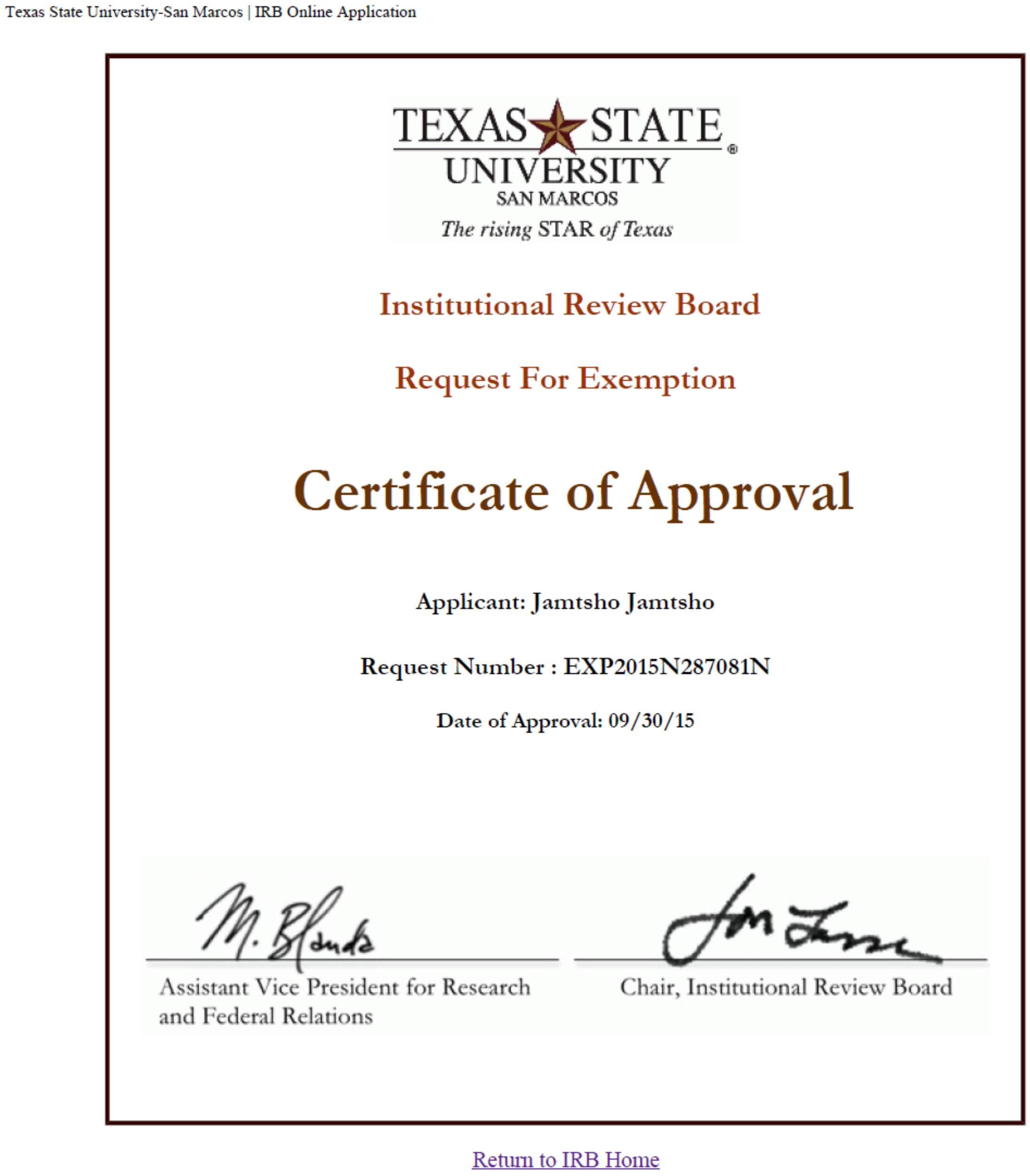
```
> GeneratingDataNew = function(n, nv, vdbn, BinN, p, B, a, b, p1, p2, prop, maxday, zero)
+ {
+   icdata = matrix(NA, nrow = n, ncol = 2) # for storing Li, Ri
+   Ni = rep(NA, n)
+   coitaltimes = matrix(NA, nrow=n, ncol=maxday)
+   condomuse = matrix(NA, nrow=n, ncol=maxday)
+   Vi = rep(NA, nv)
+   maxevents = 0 # counting max # of coital events
+   cnt = 0 # counting No. of right-censored obs
+   for(i in 1:n)
+   {
+     for(j in 1:nv)
+     {
+       if(vdbn == 1)
+         Vi[j] = round(runif(1, 60*j - 5, 60*j + 5)) # uniform perturbed visit time rounded to
an integer
+       else if(vdbn == 2)
+         Vi[j] = round(rtriangle(1, 60*j - 20, 60*j + 20))
+      else
+        Vi[j] = 60*j
+    }
+    Ci = GenerateDiscreteDBN(Vi, p, zero) # last visit time or right-censoring time
+    Ti = ceiling(rweibull(1, a, b))  # to make sure Ti >= 1
+    if(Ti > Ci) # right-censored; no coital events observed
+    {
+      icdata[i, 1] = Ci
+      icdata[i, 2] = Inf
+      Ni[i] = 0
+      cnt = cnt + 1
+    }
+    else  # interval-censored
+    {
+      index = 0  # for the position of Ti
+      for(j in 1:nv)
+      {
+        if(Ti > Vi[j])
+          index = index + 1
+        else
+          break
+      } # Ti is between Vi[index] and Vi[index+1]
+      d1 = DiscreteUniform(0, B)
+      d2 = DiscreteUniform(0, B)
+      LIndex = max(0, index - d1)
+      RIndex = min(nv, (index + 1) + d2)
+      VVi = c(0, Vi)
+      icdata[i, 1] = VVi[LIndex + 1] # +1 because LIndex is for vector Vi
+      icdata[i, 2] = VVi[RIndex + 1] # +1 because RIndex is for vector Vi
+       Ni[i] = rbinom(1, BinN*(RIndex - LIndex), 1/8) # No. of coital events in interval
(V_{i,v-1}, V_{i, v}] ~ Bin(64, 1/8)
+      maxevents = max(Ni[i], maxevents)
+      if(Ni[i] == 0)
+      {
+        Ni[i] = 1 # must have a coital event since true infection
          #time Ti is between (Li, Ri]
+        coitaltimes[i,1] = Ti
+        condomuse[i,1] = 0  # condom not used at Ti
+      }
+      else if(Ni[i] == 1)
+      {
+        coitaltimes[i,1] = Ti
+        condomuse[i,1] = 0  # condom not used at Ti
+      }
+      else  # Ni > 1: generating (Ni - 1) coital event times
+      {
```

```
+        coitaltimes[i,1:Ni[i]] = sort( c(Ti, round(runif(Ni[i]-1, icdata[i, 1]+1, icdata[i, 2]) )
) ) # +1: to match (Li, Ri]
+        prob.condom = CondomUseProb(Ni[i], Ti, coitaltimes[i,1:Ni[i]], p1, p2, prop)
+        u = runif(Ni[i], 0, 1)
+        condomuse[i,1:Ni[i]] = as.numeric(u < prob.condom)
+      }
+    }
+   }
+   if(maxevents <= maxday)
+   {
+     coitaltimes = coitaltimes[,1:maxevents] # remove columns with no coital times
+     condomuse = condomuse[,1:maxevents]
+     #print(coitaltimes)
+   }
+   else
+   {
+     print("# of Coital Events Exceeds Maxday, So Quit!")
+     break
+   }
+  #data = cbind(icdata, Ni, coitaltimes)
+  # print(c("% of RT-censoring:", cnt/n))
+   data = data.frame()
+   class(data) = "STD Data"
+   data$icdata = cbind(icdata, Ni)
+   data$coitaltimes = coitaltimes
+   data$condomuse = condomuse
+   data
+ }
```

6.2. ComparingMethodNew

```
> ComparingMethodsNew <- function(n, nv, btwvisits, vdbn, BinN, p, B, a, b, pclose, pfar,
propclose, odds, tol, M, R, maxday, zero)
+ {
+
+    print(c("vdbn, n, nv, btwvisits, a, b, M, R, B, odds"))
+    print(c(vdbn, n, nv, btwvisits, a, b, M, R, B, odds))
+    miseT = 0
+    miseHT = 0
+    miseP = 0
+
+    S.T = matrix(NA, nrow=M, ncol=maxday+1)  # columns are for times: 0 - maxday
+    S.HT = matrix(NA, nrow=M, ncol=maxday+1)
+    S.P = matrix(NA, nrow=M, ncol=maxday+1)
+    for(r in 1:M)
+   {
+     if((r+99)%% 100 == 0)
+       print(c(">", r))
+      data.r = GeneratingDataNew(n, nv, btwvisits, vdbn, BinN, p, B, a, b, pclose, pfar,
propclose, maxday, zero)
+
+     resultT = Turnbull(data.r$icdata[,1:2], tol)
+     times = resultT[,1]
+     m = length(times)
+     ST = 1 - cumsum(resultT[,2])
+
+      resultHT = HTMethod2(data.r$icdata[,1:2], data.r$icdata[,3], data.r$coitaltimes, R,
maxday, zero)
+
+      resultP = ProposedMethod(data.r$icdata[,1:2], data.r$icdata[,3], data.r$coitaltimes,
R, data.r$condomuse, odds, maxday, zero)
+
+     ################
+     # Compute MISE #
+     ################
+       # SHT and SP will be used to compute MISE based on times, which are ordered
distinct times of Turnbull
+     SHT = resultHT[times[2:(m-1)],2] # exclude at times that possibly be 0 or Inf b/c S(t)
is 1 or 0 anyway
+     SP = resultP[times[2:(m-1)],2]   # exclude at times that possibly be 0 or Inf b/c S(t) is
1 or 0 anyway
+     if(times[1] == 0)  # t1 = 0
+     {
+       SHT = c(1, SHT)  # attach S(0) = 1 in front
+       SP = c(1, SP)    # attach S(0) = 1 in front
+     }
+     else  # t1 != 0
+     {
+       SHT = c(resultHT[times[1],2], SHT)  # attach S(t1) in front
+       SP = c(resultP[times[1],2], SP)     # attach S(t1) in front
+     }
+     if(times[m] == Inf)  # tm = Inf
+     {
+       SHT = c(SHT, 0)  # attach S(Inf) = 0 at the end
+       SP = c(SP, 0)    # attach S(Inf) = 0 at the end
+     }
+     else  # tm != Inf
+     {
+       SHT = c(SHT, resultHT[times[m],2])  # attach S(tm) at the end
+       SP = c(SP, resultP[times[m],2])     # attach S(tm) at the end
+     }
+
+      SWeibull = 1 - pweibull(times, a, b)
+      miseT = miseT + sum((ST-SWeibull)^2) / M
+      miseHT = miseHT + sum((SHT-SWeibull)^2) / M
+      miseP = miseP + sum((SP-SWeibull)^2) / M
+
+      ########################
+      # Store Survival Probs #
+      ########################
+      if(times[m] == Inf)  # unable to use Inf as a column number
+      {
+         S.T[r, times[-m] + 1] = ST[-m] # + 1 because times[1] = 0; or column 1 of S.T is S(0) = 1
+      }
+      else
+      {
+        S.T[r,times + 1] = ST # + 1 because times[1] = 0
+      }
+       S.HT[r, 1:maxday + 1] = resultHT[, 2]  # +1 b/c HT method gives survival probs at 1:maxday so that S.HT[,1] = NA (for t = 0)
+         S.P[r, 1:maxday + 1] = resultP[, 2]  # +1 b/c HT method gives survival probs at 1:maxday so that S.P[,1] = NA (for t = 0)
+      S.HT[r, 1] = 1 # at time 0
+      S.P[r, 1] = 1 # at time 0
+    } # end of r loop
+    mise1 = c(miseT, miseHT, miseP)
+    print(mise1)
+
+
+ #    times = (0:maxday)[!apply(is.na(S.T), 2, all)]  # pick days on which >= 1 survival prob among M iterations
+ #   d = length(times)
+ #    Sout.T = S.T[,!apply(is.na(S.T), 2, all)] # pick days on which >= 1 survival prob among M iterations
+ #   Sout.HT = S.HT[,!apply(is.na(S.T), 2, all)] # use those from Turnbull
+ #   Sout.P = S.P[,!apply(is.na(S.T), 2, all)] # use those from Turnbull
+ #   Sout.T = CompleteSurvMatrix(Sout.T, M, d)
+ #   Sout.HT = CompleteSurvMatrix(Sout.HT, M, d)
+ #   Sout.P = CompleteSurvMatrix(Sout.P, M, d)
+
+    Sout.T = CompleteSurvMatrix(S.T, M, maxday + 1)
+    Sout.HT = S.HT #CompleteSurvMatrix(S.HT, M, maxday + 1) # no need to complete matrix b/c S.HT has no NA
+    Sout.P = S.P #CompleteSurvMatrix(S.P, M, maxday + 1)  # no need to complete matrix b/c S.P has no NA
+
+    S.mean.T = apply(Sout.T, 2, mean, na.rm=T)  # na.rm = T ?
+    S.mean.HT = apply(Sout.HT, 2, mean, na.rm=T)
+    S.mean.P = apply(Sout.P, 2, mean, na.rm=T)
+    #list(S.mean.T, S.mean.HT, S.mean.P, times)
+
+ # SWeibull2 = 1 - pweibull(times,a,b)
+         SWeibull2 = 1 - pweibull(0:maxday,a,b)
+   BiasT = S.mean.T -SWeibull2
+         BiasHT = S.mean.HT -SWeibull2
+         BiasP = S.mean.P -SWeibull2
+         #list(BiasT, BiasHT, BiasP, 1:maxday)
+ #       bias = cbind(times, BiasT, BiasHT, BiasP)
+         bias = cbind(0:maxday, BiasT, BiasHT, BiasP)
+
+         # MISE based on all times 0 - maxday #
+         miseT = 0
+         miseHT = 0
+         miseP = 0
+         for(r in 1:M)
+         {
+           miseT = miseT + sum((Sout.T[r,] - SWeibull2)^2)/M
+           miseHT = miseHT + sum((Sout.HT[r,] - SWeibull2)^2)/M
+           miseP = miseP + sum((Sout.P[r,] - SWeibull2)^2)/M
+         }
+         mise2 = c(miseT, miseHT, miseP)
+         print(mise2)
+
+         print(bias[(1:nv)*btwvisits+1,]) # only print out biases at time = btwvisits, 2 btwvisits, ..., nv btwvisits.
+         bias
+  }
```

APENDIX C: IRB



Texas State University-San Marcos | IRB Online Application

TEXAS STATE UNIVERSITY
SAN MARCOS
The rising STAR *of Texas*

Institutional Review Board

Request For Exemption

Certificate of Approval

Applicant: Jamtsho Jamtsho

Request Number : EXP2015N287081N

Date of Approval: 09/30/15

Assistant Vice President for Research and Federal Relations

Chair, Institutional Review Board

Return to IRB Home

REFERENCES

Chen, D.-G., & Peace, K. E. (2011). *Clinical Trial Data Analysis Using R.* Boca Raton: CRC Press.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure. *Biometrics,* 42, 845-854.

Garrow, S., Smith, D., & Harnett, G. (2002). The diagnosis of chlamydia, gonorrhoea, and trichomonas infections by self obtained low vaginal swabs, in remote northern Australian clinical practice. *Sexually Transmitted Infections,* 78(4): 278-281.

Harezlak, J., & Tu, W. (2006). Estimation of survival functions in interval and right censored. *Stat Med*, 4053–4064.

He, P., Kong, G., & Su, Z. (2013). Estimating the survival functions for right-censored and. *Contemporary Clinical Trials, 35*, 122-127.

Hoel, D. G., & Walburg, H. E. (1972). Statistical analysis of survival experiments. *Journal of the National Cancer Institute., 49*, 361-372.

Jewell, N. P., Malani, H. M., & Vittinghoff, E. (1994). Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS. *Journal of American Statistical Association, 89,* 7-18.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival Analysis.* New York: Springer Science+Business Media, Inc.

Kleinbaum, D. G. (1996). *Survival Analysis: A self-Learning Text.* New York: Springer-Verlag.

Kongerud, J., & Samuelsen, S. O. (1991). A longitudinal study of respiratory symptoms in aluminum potroom workers. *American Review of Respiratory Diseases, 144*, 10-16.

Lagakos, S. W. (1979, March). General Right Censoring and Its Impact on the Analysis of Survival Data. *Biometrics, 35*(1), 139-156.

Lindsey, J. C., & Ryan, L. M. (1998). Tutorial in biostatistics: methods for interval-censored data. *Statistics in Medicine*, 12:219–238.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics* , 22:86–91.

Petrie, A., & Sabin, C. (2005). *Medical Statistics at a Glance.* Malden, Mass: Blackwell.

Pol, B. V., Kraft, C. S., & Williams, J. A. (2006). Use of an Adaptation of a Commercially Available PCR Assay Aimed at Diagnosis of Chlamydia and Gonorrhea To Detect Trichomonas vaginalis in Urogenital Specimens. *Journal of Clinical Microbiology*, 44(2): 366–373.

Self, S. G., & Grossman, E. A. (1986). Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics, 42*, 521-530.

Sun, J. (1996). A NON-PARAMETRIC TEST FOR INTERVAL-CENSORED FAILURE TIME DATA WITH APPLICATION TO AIDS STUDIES. *STATISTICS IN MEDICINE, 15*, 1387-1395.

Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data.* New York: Springer Science+Business Media, Inc.

Tableman, M., & Sung, K. J. (2004). *Survival Analysis Using S: Analysis of Time-to-Event Data.* Boca Raton: Chapman & Hall/CRC.

Turnbull, B. W. (1976). The Empirical Distrubution Function with Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society, Vol. 38, Series B*, 290-295.

Xiao, X., Hu, Q., Yu, D., & Xie, M. (2014). Study of an imputation algorithm for the analysis of interval-censored data. *Journal of Statistical Computation and Simulation, 84*(3), 477-490.