ESTIMATING THE SURVIVAL DISTRIBUTION FOR RIGHT-CENSORED DATA WITH DELAYED ASCERTAINMENT

by

Eun-Jin Kim

B.S. in Biology and Statistics, University of Wisconsin-Madison, 2007

Submitted to the Graduate Faculty of

The Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Eun-Jin Kim

It was defended on

December 7, 2011

and approved by

Stewart Anderson, PhD
Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Jong-Hyeon Jeong, PhD
Associate Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Maria Mori Brooks, PhD
Associate Professor, Department of Epidemiology
Graduate School of Public Health, University of Pittsburgh

Thesis Advisor: Stewart Anderson, PhD
Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Stewart Anderson, PhD

ESTIMATING THE SURVIVAL DISTRIBUTION FOR RIGHT-CENSORED DATA WITH DELAYED ASCERTAINMENT

Eun-Jin Kim, M.S.

University of Pittsburgh, 2011

In many clinical trials, patients are not followed continuously. This means their vital status may not be immediately recorded. In such cases, the results from the Kaplan-Meier estimator or the log rank test, popular methods used for survival analysis, may be biased or inconsistent. Hu and Tsiatis first produced a new estimator to estimate survival distribution for right-censored data with delayed ascertainment, Van der Laan and Hubbard modified their estimator. We investigate each of these proposed estimators and their properties. Using simulations, we compare these new estimators to each other and to the Kaplan-Meier estimator using different sample sizes, different failure rates, and different maximum delay times. The public health importance of this thesis is that we can partially alleviate the problem caused by delayed ascertainment in the analysis of right-censored time to event data by choosing the most accurate and consistent estimator that accounts for the delayed ascertainment. The reduction of bias in analyses of public health data ensures that such studies are reliable so that proper inference can be made and hence, potential public health policy can be based on an accurate decision making process.

Keywords: Ascertainment; Right censored; Survival analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to offer sincere gratitude to my advisor Dr. Stewart Anderson for his continuous encouragement, guidance, time and concern. I would also like to thank my thesis committee members Dr. Jong-Hyeon Jeong and Dr. Maria Brooks for their support and comments.

I would like to take the opportunity to thank my family mom, dad, grandfather, So-Jung, and Yang-Un for their endless love and support. I would also like to thank my friends for their encouragement.

Finally, I would like to thank Dr. Roslyn Stone and Dr. John Wilson for their encouragement and giving me the opportunity to have teaching experiences.

1.0 INTRODUCTION

In some clinical studies, the outcome of interest is the length of time to an event, not just whether the event occurs. However, for a variety of reasons, during the collection of time to event data, some subjects are not followed until they ultimately experience an event. For example, a subject may drop out of a study before the time of analysis or could be lost to follow-up because they moved and cannot be contacted. Another possibility is that the study ends before a subject experiences the event being studied. In these cases, their failure times are said to be right-censored. The term “right” is used since all we know is that the event occurs sometime after the censoring time (Klein and Moeschberger, 2005). Survival data can be described using the survival function, $S(t)$, the probability that the failure time (time to event) is greater than time t. One of the popular estimators for determining the survival function, given the right-censored data, is the Kaplan-Meier estimator. This estimator requires the key assumption that the censoring is independent of the failure times (Miller, 1981).

1.1 SURVIVAL DATA WITH DELAYED ASCERTAINMENT

In clinical trials, patients are not often followed continuously and their vital and disease status might be recorded only after a certain time has passed. This delayed recording is called “delayed ascertainment.” In this case, the key assumption of the Kaplan-Meier estimator, independent censoring, is not met (Hu and Tsiatis, 1996). Three different situations with delayed ascertainment are described in Figure 1. In that figure, *E* represents when patients entered the study, *T* represents when patients experienced an event, *A* is the time when *T* was recorded into dataset, and *C* is the time when the analysis was performed. Case I represents a situation where the event and its ascertainment occurred after time *C*. Case II represents a situation where the event occurred before time *C*, but it was reported after time *C*. Finally, case III represents a situation where the event and its ascertainment occurred before time *C*. At time *C*, only the last case (Case III) is reported as a subject with the event even though the second subject (Case II) also experienced an event prior to the time of analysis.

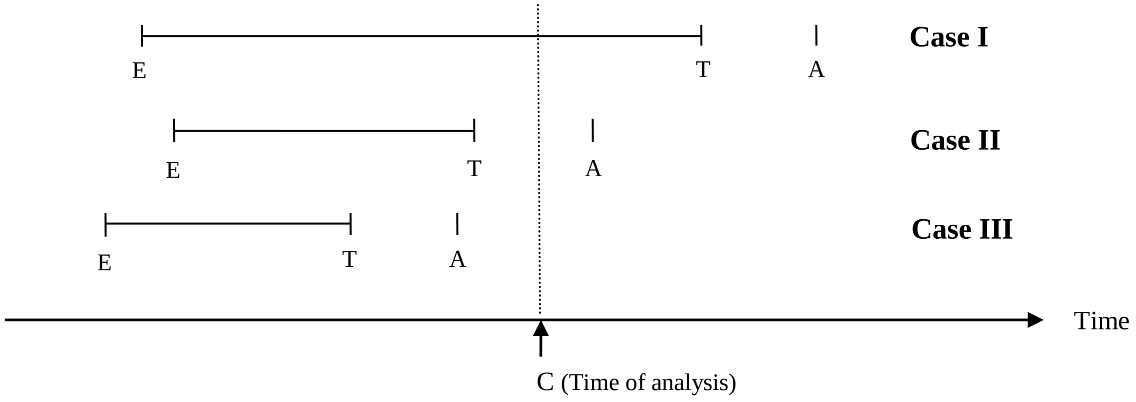

Figure 1. Survival data with delayed ascertainment

The failure times of Cases I and II were right-censored because their recordings indicated that the event had not occurred at time *C*. When a patient's status is not up to date at the time of analysis, the censoring process depends on both *A* and *T*. This causes the Kaplan-Meier estimator to be biased or inconsistent. To alleviate this problem, Hu and Tsiatis (1996) developed a new estimator that accounts for delayed ascertainment. Later, Van der Laan and Hubbard (1998) offered a new estimator, which is a modification the Hu and Tsiatis estimator.

1.2 THE STRUCTURE OF DATA WITH DELAYED ASCERTAINMENT

To describe the structure of survival data with delayed ascertainment, we will focus on the vital status as the event of interest. We denote T as the time from entry into the study until death.

1.2.1 Uncensored data

We will first consider the situation where all patients have experienced the outcome of interest - death. We will follow many of the notational conventions used by Hu and Tsiatis (1996). Assume that a random sample of n subjects participated in the study and regularly visit hospitals. Let U_{ji} (j = 1,…, k-1) be the time to j^{th} visit from time of entry into the study for i^{th} subject and A_{ji} (j = 1,…, k-1) be the time when the status at U_{ji} is recorded. Additionally, let A_{ki} be the time when T_i is recorded. We can describe the status process of each subject using a random vector $\{(U_{1i}, A_{1i}), \ldots , (U_{k-1\ i}, A_{k-1\ i}), (T_i, A_{ki})\}$, where $U_{ji} \leq A_{ji}$ (j = 1,…, k-1). Here, the number of hospital visits, k, is random and depends on i.

The distribution of the above vector is complex since each patient has different number of hospital visits. We can describe this vector using a bivariate process $\{V_i(t), R_i(t); t \geq 0\}$, where $R_i(t)$ is a function indicating whether or not the i^{th} subject died prior to time t and $V_i(t)$ is the first time at which we know the subject's vital status at time t. If a subject is alive at time t (Figure 2, Case I), $V_i(t)$ is $A_{ji,}$ which is the reporting time of U_{j*i} where U_{j*i} is the minimum U_{ji} greater than time t. On the other hand, if a subject dies before time t (Figure 2, Case II), $V_i(t)$ is A_{ki}, the time when T_i is reported. When a subject dies before time t but their failure time is not yet recorded at time t (Figure 2, Case III), $V_i(t)$ is still A_{ki} by the definition.

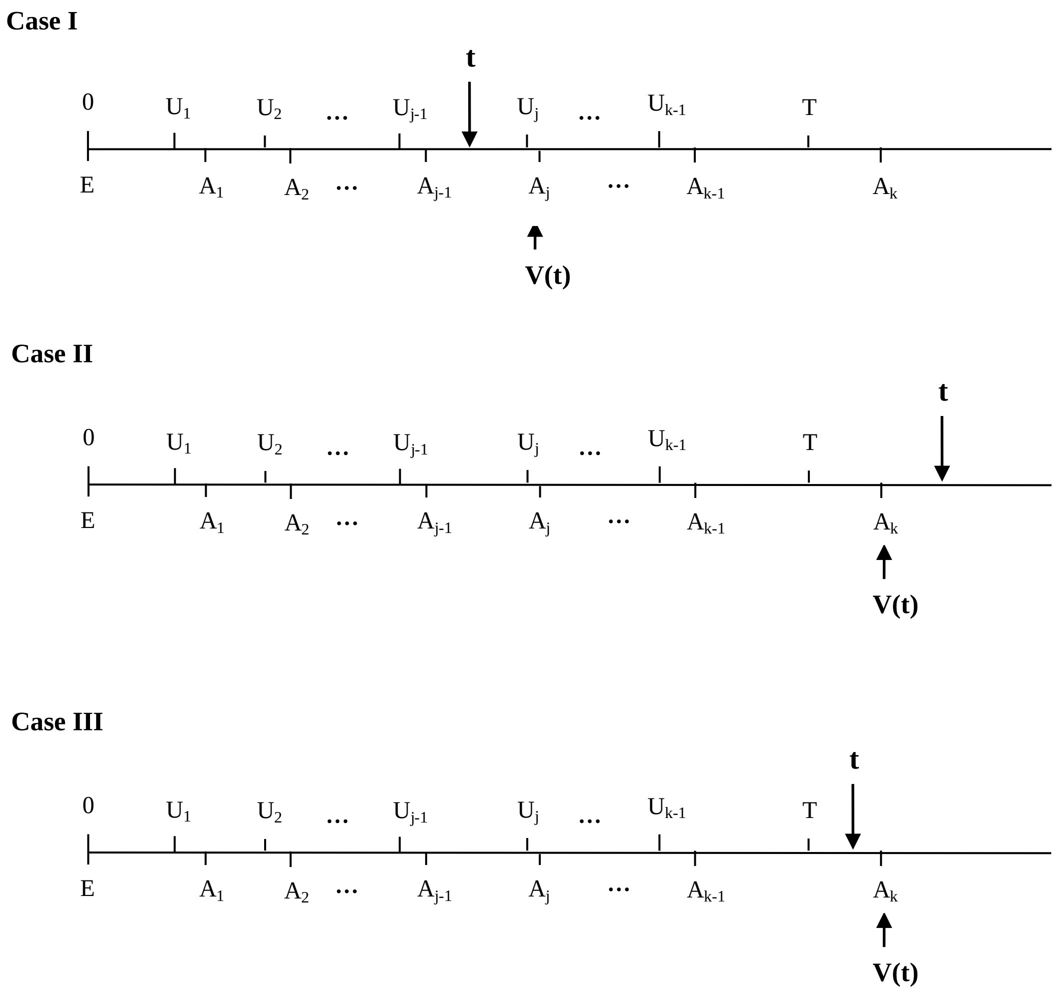

Figure 2. Uncensored data with delayed ascertainment

1.2.2 Right-censored data

As we discussed before, survival data in many clinical trials is right-censored. In this thesis, for simplification purpose, we will focus on the situation when a survival time is right-censored only due to incomplete follow-up. With right-censored data, we define *C* to be the censoring time and *F* to be the follow-up time, the time from a subject's entry to time when the analysis is performed. If a subject is known to be dead before the time of analysis (Figure 3, Case I), the censoring time used for the censoring variable is the time when the analysis is performed. However, if a subject is known to be alive up to the time of analysis (Figure 3, Case II), then we are not certain whether that subject died between the last monitored time and the time of analysis. Van der Laan and Hubbard (1998) state that it is a common practice to use the last monitored time as the censoring time since all we know is that the failure time, T, is greater than the time of the last visit.

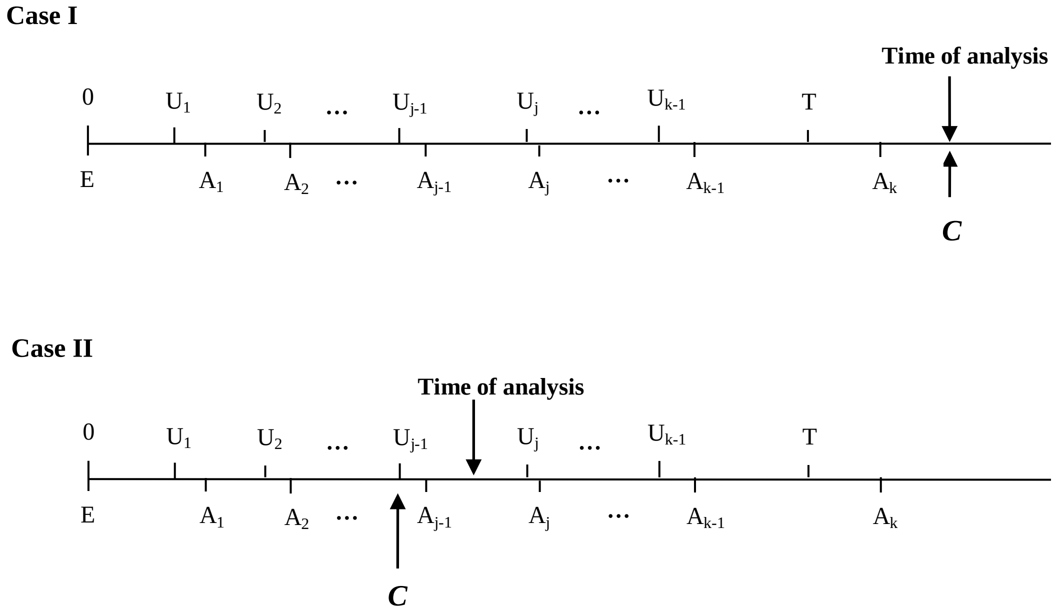

Figure 3. Right-censored data with delayed ascertainment

If $V(t)$ is greater than the censoring time, then we cannot really assess a subject's vital status at time t. Therefore, let $X_i(t)$ be the minimum of $V_i(t)$ and C_i, and $\Delta_i(t)$ be the indicator of whether we can classify a subject's status at time t. We can describe the observed right-censored data using $\{X_i(t), \Delta_i(t), R^*_i(t); t \geq 0\}$, where $R^*(t)$ is the vital status indicator, which can be defined only if a subject's status at time t is known.

2.0 ESTIMATORS

2.1 HU AND TSIATIS ESTIMATOR

The primary interest in survival analysis is to estimate the survival function $S(t)$ = pr($T \geq t$), where T is the time to death. Using the notation of Hu and Tsiatis, this can be expressed as 1 - pr{$R(t)$ =1}. Hu and Tsiatis make two assumptions. First, they assume that one knows a subject died by time t before time t + $C(t)$, where $C(t)$ is a nonrandom value and represents the maximal reporting delay. This assumption can be expressed as pr{$V(t) \leq t + C(t)| R(t) = 1$} = 1. Second, they assume that the follow-up time is independent of both the failure time distribution and the ascertainment process. Under their first assumption, the survival distribution function, $S(t)$, can be expressed as 1 - pr{$V(t) \leq t + C(t)$, $R(t) = 1$} (Appendix A.1). Hu and Tsiatis define cause-specific hazard functions $\lambda_j(t, u)$ and a sub-distribution function $G_1(t, u)$ to describe the survival distribution function using the process {$V(t)$, $R(t)$}:

$$\lambda_j(t,u) := \lim_{h\to 0} h^{-1} \,\mathrm{pr}\{u \leq \mathrm{V}(t) < u + h\,, \mathrm{R}(t) = j|\mathrm{V}(t) \geq u\} \;\; (j = 0{,}1)$$

and

$$G_1(t,v) \coloneqq \mathrm{pr}\{\mathrm{V}(t) \leq v, \mathrm{R}(t) = 1\} = \int_0^v exp - \{\Lambda_1(t,x) + \Lambda_0(t,x)\}\, \lambda_1(t,x)dx\,,$$

where $\Lambda_j(t,x) = \int_0^x \lambda_j(t,u)du$ $(j = 0{,}1)$. Let $\lambda^*_j(t, u)$ be cause-specific hazard functions for $V(t)$ with the observable random variables {$X_i(t)$, $\Delta_i(t)$, $R^*_i(t)$; $t \geq 0$}. Appendix A.2 shows that $\lambda^*_j(t, u)$ is the same as $\lambda_j(t, u)$ under their second assumption. Therefore, the survival distribution function can be expressed as a function of the observable random variables $\{X_i(t), \Delta_i(t), R^*_i(t); t \geq 0\}$:

$$S(t) = 1 - \text{pr}\{R(t) \ = 1\} = 1 - \text{pr}\{V(t) \leq t + C(t), R(t) = 1\} = 1 - G_1\{t, t + C(t)\}$$

$$= 1 - \int_0^{t+C(t)} exp - \{\Lambda_1(t,x) + \Lambda_0(t,x)\}\, \lambda^*{}_1(t,x)dx\ , \qquad (2.1.1)$$

where $\Lambda_j(t,x) = \int_0^x \lambda^*{}_j(t,u)du\ (j = 0,1)$.

To estimate the above survival function, Hu and Tsiatis use the counting process theory (Appendix A.3) and derive the following estimator:

$$\widehat{S}_{HT}(t) = 1 - \sum_{i=1}^{n} \frac{\widehat{E}\{\ X_i(t), \Delta_i(t)\}}{Y\{t, X_i(t)\}}\ I\{X_i(t) \leq t + C(t), R_i(t) = 1, \Delta_i(t) = 1\}\ , \qquad (2.1.2)$$

where $\widehat{E}\{X_i(t), \Delta_i(t)\}$ is the Kaplan-Meier estimator obtained from the data $\{X_i(t), \Delta_i(t)\}$ and $Y\{t, X_i(t)\}$ is the number of subjects with $X_i(t)$ greater than or equal to t.

When vital status at every given time point is recorded immediately, the Hu and Tsiatis estimator and the Kaplan-Meier estimator are the same. In some clinical trials, the maximal reporting delay time $C(t)$ is given. When this value is unknown, Hu and Tsiatis suggest choosing a value large enough.

2.2 VAN DER LAAN AND HUBBARD ESTIMATOR

Van der Laan and Hubbard (1998) extend Hu and Tsiatis' study and modify their estimator. They use the 'inverse probability of censoring weighted' (IPCW) representation from Robins (1993). They assume that the censoring time is independent of both the failure time distribution and the ascertainment process. They define $\bar{G}(V(t))$ as the conditional expectation given the failure time distribution T and ascertainment process $V(t)$:

$$\bar{G}(V(t)) = E[\Delta(t)|T, V(t)] = pr(F \geq V(t)|T, V(t)).$$

Under their assumption, they make a key identity that $F(t)$ = pr($T \leq t$) = $E[\frac{I(T\leq t)\Delta(t)}{\bar{G}(X(t))}]$. The estimator of $F(t)$ is $\frac{1}{n}\sum_{i=1}^{n}\frac{I(T_i\leq t)\Delta_i(t)}{\bar{G}_n(X_i(t))}$, where $\bar{G}_n(X_i(t))$ is an estimator of $\bar{G}(X(t))$. They define $\bar{F}_{V(t)}(x)$ as pr($V(t) \geq x$), and $\bar{F}_{X(t)}(x)$ as pr($X(t) \geq x$). Under their assumption, $\frac{1}{\bar{G}(x)} = \frac{\bar{F}_{V(t)}(x)}{\bar{F}_{X(t)}(x)}$ and the modified estimator is expressed as:

$$\widehat{S}_{VH}(t) = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\bar{F}_{V(t),KM}(X_i(t))}{\bar{F}_{X(t),n}(X_i(t))}\ R_i(t)\Delta_i(t)\ , \qquad (2.2.1)$$

where $F_{V(t),\mathrm{KM}}$ is the Kaplan-Meier estimator obtained from the data $\{X_i(t), \mathit{\Delta}_i(t)\}$ and $F_{X(t),\mathrm{n}}(x)$ is the proportion of subjects with $X_i(t)$ greater than or equal to x.

The Van der Laan and Hubbard estimator is equal to Hu and Tsiatis estimator when the maximal reporting delay, $C(t)$, value is equal to a positive infinite value. So they conclude that the $C(t)$ term in the Hu and Tsiatis estimator is unnecessary.

3.0 SIMULATION RESULTS

3.1 SIMULATION SETUP

A series of simulations were performed to compare the new estimators to each other and to the Kaplan-Meier estimator. For simplicity, it was assumed that the monitoring times of the “alive” status were recorded immediately, i.e., $A_{ji} = U_{ji}$ (j = 1,…, k-1). The time when each subject entered the study was uniformly generated on an interval from zero to three years and the time of analysis was three years. The time to failure was generated from an exponential distribution with means of one year and two years. Furthermore, the failure times were generated independently from the follow-up times, the times from entry to the time of analysis. First, the case where the failures were immediately recorded was simulated. Then, the cases with delayed reporting were simulated. The delays were generated from uniform distributions along three different intervals: (0, 0.5), (0, 1) and (0, 2). The times to periodic hospital visits were simulated using a Poisson process with a mean inter-arrival time of six months. We investigated the magnitude of bias and the consistency of each estimator. We used the last monitoring time as the censoring time when a subject was censored. However, Hu and Tsiatis used the follow-up time as the censoring time. We used both censoring time definitions and compared the estimates.

For all simulations, the Kaplan-Meier estimator, the Hu and Tsiatis estimator and the Van der Laan and Hubbard estimator were calculated at three time points: one, two and three years.

For the Hu and Tsiatis estimator, the $C(t)$ values used were 0 and 1. For each scenario, sample sizes of 60 and 100 were simulated 500 times. These simulations were performed using the R language. A typical simulation of three estimators and two $C(t)$ values for the three landmark times with the sample size of 60 required just under 10 minutes of CPU time. With the sample size of 100, it required about 19 minutes of CPU time.

3.2 RESULTS

Results of our simulations are summarized in Tables 1 - 9. Tables 1 – 4 show the estimated survival distribution values at each time point, compared to the true survival distribution. For each table, t represents the time points (in years) at which estimators were calculated; $S(t)$ represents the true survival distribution; KM, the empirical average of the Kaplan-Meier estimator; $1 - G_1(t,t)$, the empirical average of the Hu and Tsiatis estimator with $C(t)$ value of 0; $1 - G_1(t,t+1)$, the empirical average of the Hu and Tsiatis estimator with $C(t)$ value of 1; and VH represents the empirical average of the Van der Laan and Hubbard estimator. Tables 5 – 8 display empirical standard deviations of the estimators. Table 9 displays empirical averages of the Hu and Tsiatis estimators when the follow-up time was used as the censoring time.

When there was no delay, the Hu and Tsiatis estimator and the Van der Laan and Hubbard estimator gave the same empirical average values as the Kaplan-Meier estimator. This is because two new estimators reduce to the Kaplan-Meier estimator if a subject's status is always recorded immediately. New estimators and the Kaplan-Meier estimator underestimated the true survival distribution at every time point.

The Hu and Tsiatis estimator with the maximum delay value, $C(t)$, of 0 and the Kaplan-Meier estimator gave the same estimate values since the Hu and Tsiatis estimator reduces to the Kaplan-Meier estimator when the $C(t)$ value is equal to 0. These estimators overestimated the true survival function.

When the $C(t)$ value was correctly defined or when the assigned $C(t)$ value was greater than the true maximum delay value, the Hu and Tsiatis estimator and the Van der Laan and Hubbard estimator were the same. These estimators underestimated the true survival function.

When the assigned value was less than the true maximum delay value, the Hu and Tsiatis estimator overestimated the true survival function, and it was more biased than the Van der Laan and Hubbard estimator.

As demonstrated in Tables 1 - 4, the differences between the estimators in the magnitude of bias decreased at later time points. In general, the Van der Laan and Hubbard estimator was less biased than the Hu and Tsiatis estimator and the Kaplan-Meier estimator.

As shown in Tables 5 - 8, all estimators were more consistent at later time points. The Van der Laan and Hubbard estimator was more likely to be consistent than other estimators at later time points. As we expected, the estimators were more consistent when the sample size was larger.

Lastly, Table 9 showed that the Hu and Tsiatis estimator values were changed when the follow-up time, rather than the last available monitoring time, was used as the censoring time. The Hu and Tsiatis estimators with different $C(t)$ values were no longer the same as the Kaplan-Meier estimator or the Van der Laan and Hubbard estimator. When there is no delay, the Hu and Tsiatis estimator was less biased than other estimators. The Hu and Tsiatis estimator was more biased than the Kaplan-Meier estimator when the assigned $C(t)$ value was smaller than the true maximum delay value. When the assigned value was equal to or larger than the true maximum delay value, the Hu and Tsiatis estimator was less biased than other estimators at earlier time points.

Table 1. Sample Size = 60, Mean Failure Rate = 1 year.

Delay	t	S(t)	KM	1 – $G_1(t,t)$	1 – $G_1(t,t+1)$	VH
No Delay	1	0.3679	0.3465	0.3465	0.3465	0.3465
	2	0.1353	0.1075	0.1075	0.1075	0.1075
	3	0.0498	0.0333	0.0333	0.0333	0.0333
UNIF (0, 0.5)	1	0.3679	0.4475	0.4475	0.3408	0.3408
	2	0.1353	0.1424	0.1424	0.1040	0.1040
	3	0.0498	0.0418	0.0418	0.0331	0.0331
UNIF (0, 1.0)	1	0.3679	0.6103	0.6103	0.3519	0.3519
	2	0.1353	0.1945	0.1945	0.1127	0.1127
	3	0.0498	0.0520	0.0520	0.0352	0.0352
UNIF (0, 2.0)	1	0.3679	0.7999	0.7999	0.4783	0.3725
	2	0.1353	0.3855	0.3855	0.1533	0.1278
	3	0.0498	0.1116	0.1116	0.0535	0.0535

Table 2. Sample Size = 60, Mean Failure Rate = 2 years.

Delay	t	S(t)	KM	1 – $G_1(t,t)$	1 – $G_1(t,t+1)$	VH
No Delay	1	0.6065	0.5860	0.5860	0.5860	0.5860
	2	0.3679	0.3227	0.3227	0.3227	0.3227
	3	0.2231	0.1635	0.1635	0.1635	0.1635
UNIF (0, 0.5)	1	0.6065	0.6655	0.6655	0.5798	0.5798
	2	0.3679	0.3667	0.3667	0.3177	0.3177
	3	0.2231	0.1852	0.1852	0.1590	0.1590
UNIF (0, 1.0)	1	0.6065	0.7683	0.7683	0.5842	0.5842
	2	0.3679	0.4242	0.4242	0.3212	0.3212
	3	0.2231	0.2134	0.2134	0.1606	0.1606
UNIF (0, 2.0)	1	0.6065	0.8815	0.8815	0.6826	0.6145
	2	0.3679	0.5795	0.5795	0.3748	0.3491
	3	0.2231	0.2899	0.2899	0.1942	0.1942

Table 3. Sample Size = 100, Mean Failure Rate = 1 year.

Delay	t	S(t)	KM	$1 – G_1(t,t)$	$1 – G_1(t,t+1)$	VH
No Delay	1	0.3679	0.3454	0.3454	0.3454	0.3454
	2	0.1353	0.1115	0.1115	0.1115	0.1115
	3	0.0498	0.0323	0.0323	0.0323	0.0323
UNIF (0, 0.5)	1	0.3679	0.4488	0.4488	0.3386	0.4488
	2	0.1353	0.1455	0.1455	0.1106	0.1455
	3	0.0498	0.0369	0.0369	0.0286	0.0369
UNIF (0, 1.0)	1	0.3679	0.6080	0.6080	0.3497	0.3497
	2	0.1353	0.1975	0.1975	0.1101	0.1101
	3	0.0498	0.0533	0.0533	0.0322	0.0322
UNIF (0, 2.0)	1	0.3679	0.7980	0.7980	0.4822	0.3749
	2	0.1353	0.3817	0.3817	0.1492	0.1229
	3	0.0498	0.1082	0.1082	0.0435	0.0435

Table 4. Sample Size = 100, Mean Failure Rate = 2 years.

Delay	t	S(t)	KM	$1 – G_1(t,t)$	$1 – G_1(t,t+1)$	VH
No Delay	1	0.6065	0.5850	0.5850	0.5850	0.5850
	2	0.3679	0.3295	0.3295	0.3295	0.3295
	3	0.2231	0.1611	0.1611	0.1611	0.1611
UNIF (0, 0.5)	1	0.6065	0.6658	0.6658	0.5799	0.5799
	2	0.3679	0.3768	0.3768	0.3271	0.3271
	3	0.2231	0.1868	0.1868	0.1601	0.1601
UNIF (0, 1.0)	1	0.6065	0.7684	0.7684	0.5859	0.5859
	2	0.3679	0.4223	0.4223	0.3190	0.3190
	3	0.2231	0.2121	0.2121	0.1583	0.1583
UNIF (0, 2.0)	1	0.6065	0.8841	0.8841	0.6820	0.6106
	2	0.3679	0.5816	0.5816	0.3777	0.3512
	3	0.2231	0.2933	0.2933	0.1930	0.1930

Table 5. MSE. Sample Size = 60, Mean Failure Rate = 1 year.

Delay	t	KM	$1 – G_1(t,t)$	$1 – G_1(t,t+1)$	VH
No Delay	1	0.0670	0.0670	0.0670	0.0670
	2	0.0511	0.0511	0.0511	0.0511
	3	0.0372	0.0372	0.0372	0.0372
UNIF (0, 0.5)	1	0.0729	0.0729	0.0719	0.0719
	2	0.0569	0.0569	0.0517	0.0517
	3	0.0378	0.0378	0.0361	0.0361
UNIF (0, 1.0)	1	0.0691	0.0691	0.0720	0.0720
	2	0.0665	0.0665	0.0560	0.0560
	3	0.0463	0.0463	0.0406	0.0406
UNIF (0, 2.0)	1	0.0601	0.0601	0.0817	0.0816
	2	0.0830	0.0830	0.0660	0.0623
	3	0.0690	0.0690	0.0576	0.0576

Table 6. MSE. Sample Size = 60, Mean Failure Rate = 2 years.

Delay	t	KM	$1 - G_1(t,t)$	$1 - G_1(t,t+1)$	VH
No Delay	1	0.0751	0.0751	0.0751	0.0751
	2	0.0727	0.0727	0.0727	0.0727
	3	0.0757	0.0757	0.0757	0.0757
UNIF (0, 0.5)	1	0.0739	0.0739	0.0772	0.0772
	2	0.0735	0.0735	0.0734	0.0734
	3	0.0898	0.0898	0.0892	0.0892
UNIF (0, 1.0)	1	0.0643	0.0643	0.0740	0.0740
	2	0.0838	0.0838	0.0872	0.0872
	3	0.0874	0.0874	0.0885	0.0885
UNIF (0, 2.0)	1	0.0489	0.0489	0.0777	0.0842
	2	0.0865	0.0865	0.0910	0.0909
	3	0.1038	0.1038	0.1066	0.1066

Table 7. MSE. Sample Size = 100, Mean Failure Rate = 1 year.

Delay	t	KM	$1 - G_1(t,t)$	$1 - G_1(t,t+1)$	VH
No Delay	1	0.0545	0.0545	0.0545	0.0545
	2	0.0396	0.0396	0.0396	0.0396
	3	0.0273	0.0273	0.0273	0.0273
UNIF (0, 0.5)	1	0.0577	0.0577	0.0555	0.0555
	2	0.0411	0.0411	0.0377	0.0377
	3	0.0300	0.0300	0.0282	0.0282
UNIF (0, 1.0)	1	0.0576	0.0576	0.0584	0.0584
	2	0.0549	0.0549	0.0445	0.0445
	3	0.0376	0.0376	0.0316	0.0316
UNIF (0, 2.0)	1	0.0434	0.0434	0.0626	0.0584
	2	0.0649	0.0649	0.0515	0.0502
	3	0.0479	0.0479	0.0410	0.0410

Table 8. MSE. Sample Size = 100, Mean Failure Rate = 2 years.

Delay	t	KM	$1 - G_1(t,t)$	$1 - G_1(t,t+1)$	VH
No Delay	1	0.0546	0.0546	0.0546	0.0546
	2	0.0570	0.0570	0.0570	0.0570
	3	0.0609	0.0609	0.0609	0.0609
UNIF (0, 0.5)	1	0.0508	0.0508	0.0576	0.0576
	2	0.0624	0.0624	0.0600	0.0600
	3	0.0618	0.0618	0.0630	0.0630
UNIF (0, 1.0)	1	0.0467	0.0467	0.0607	0.0607
	2	0.0653	0.0653	0.0631	0.0631
	3	0.0690	0.0690	0.0716	0.0716
UNIF (0, 2.0)	1	0.0364	0.0364	0.0533	0.0590
	2	0.0696	0.0696	0.0731	0.0748
	3	0.0795	0.0795	0.0821	0.0821

Table 9. When the follow-up time is used as the censoring time
(Sample Size = 60, Mean Failure Rate = 1 year)

Delay	t	S(t)	KM	1 – $G_1(t,t)$	1 – $G_1(t,t+1)$	VH
No Delay	1	0.3679	0.3349	0.3661	0.3661	0.3349
	2	0.1353	0.0959	0.1367	0.1367	0.0959
	3	0.0498	0.0359	0.0685	0.0685	0.0359
UNIF (0, 0.5)	1	0.3679	0.4437	0.4799	0.3681	0.3309
	2	0.1353	0.1244	0.1758	0.1383	0.0914
	3	0.0498	0.0437	0.0929	0.0929	0.0437
UNIF (0, 1.0)	1	0.3679	0.5952	0.6297	0.3638	0.3337
	2	0.1353	0.1661	0.2311	0.1402	0.0914
	3	0.0498	0.0531	0.1104	0.1104	0.0531
UNIF (0, 2.0)	1	0.3679	0.7904	0.8145	0.4978	0.3783
	2	0.1353	0.3437	0.4335	0.2090	0.1295
	3	0.0498	0.0803	0.1840	0.1840	0.0803

4.0 DISCUSSION

In this thesis, we investigated right-censored survival data when it has delayed ascertainment. We investigated new estimators proposed by Hu and Tsiatis and Van der Laan and Hubbard, and compared them with the Kaplan-Meier estimator to find the most accurate estimator for right-censored survival data with delayed ascertainment. These new estimators are valid when patients are monitored regularly. If there is no delay, the Hu and Tsiatis estimator and the Van der Laan and Hubbard estimator both simplify to the Kaplan-Meier estimator. If recording of patients' status is delayed and not up to date at the time of analysis, both new estimators are less biased than the Kaplan-Meier estimator. When the assigned $C(t)$ value in the Hu and Tsiatis estimator is equal to or greater than the true maximal delay time, the Hu and Tsiatis estimator and the Van der Laan and Hubbard estimator are the same. The differences between the estimators in the magnitude of bias decreased at later time points and as the duration of delay decreased. Our study shows that the Van der Laan and Hubbard estimator performs more accurately than the Hu and Tsiatis estimator when the ascertainment of the event is delayed.

In many clinical trials, we want to compare survival distributions between different treatment groups. Different treatment groups might have different delayed ascertainment processes. Fine and Tsiatis (2000) showed that the lagged logrank test performs well in spite of delayed reporting if the ascertainment processes are not different. If there is a difference in ascertainment, they found that the lagged logrank test can be biased and that their new estimator can be less biased. In the future, more sophisticated methods to test differences in survival distributions when treatment groups have different patterns of check-ups should be developed.

Another issue is that in many clinical trials, the event of interest is not a terminal event. Events such as a hospitalization or having a car accident can occur several times; that is, they could be recurrent events. Furthermore, there are other clinical studies which involve time to event outcomes related to competing causes of disease or death. For example, we may be interested in deaths caused by breast cancer but a breast cancer patient might die due to a disease other than breast cancer. Hence, we need to develop new estimators which account for delayed ascertainments with recurrent or competing events.

APPENDIX A

PARTIAL PROOF OF HU AND TSIATIS ESTIMATOR

A.1 SURVIVAL DISTRIBUTION FUNCTION

(Assumption 1) $\text{pr}\{V(t) \leq t + C(t) | R(t) = 1\} = 1$

$$\leftrightarrow \frac{\text{pr}\{V(t) \leq t + C(t),\ \ R(t) = 1\}}{\text{pr}\{R(t) = 1\}} = 1$$

$$\leftrightarrow \text{pr}\{V(t) \leq\ t\ +\ C(t), R(t) = 1\} = \text{pr}\{R(t)\ =\ 1\}$$

$\therefore$ $S(t) = \text{pr}(T \geq t) = 1 - \text{pr}\{R(t) = 1\} = 1 - \text{pr}\{V(t) \leq t + C(t), R(t) = 1\}$

A.2 CAUSE-SPECIFIC HAZARD FUNCTIONS FOR V(t)

If there is no censoring, the cause-specific hazard functions for V(t) is defined as:

$$\lambda_j(t, u) = \lim_{h \to 0} h^{-1}\, \text{pr}\,\{u \leq \text{V}(t) <\ u + h\, , \text{R}(t) =\ j | \text{V}(t) \geq\ u\}\ \ (j\ =\ 0{,}1).$$

If there is censoring from incomplete follow-up, the cause-specific hazard functions for V(t) is expressed only using the observable random variables {$X_i(t)$, $\Delta_i(t)$, $R_i^*(t)$; $t \geq 0$}:

$$\lambda_j^*(t,u) \;=\; \lim_{h \to 0} h^{-1}\,\mathrm{pr}\,\{u \leq \mathrm{X}(t) < \; u + h\,, \Delta(t) \,=\, 1, \mathrm{R}^*(t) \,=\, j|\mathrm{X}(t) \,\geq\, u\,\}\;\; (j \,=\, 0{,}1)\,.$$

Since X(t)= V(t) and $R^*(t)$= R(t) when $\Delta(t)$=1,

$$\lambda_j^*(t,u) \;=\; \frac{\lim_{h \to 0} h^{-1}\,\mathrm{pr}\{\mathrm{u}{\leq}\mathrm{V}(t){<}\,u{+}\mathrm{h}\,,\;\; C \geq t,\;\; \mathrm{R}(t){=}\;j\,\}}{pr\{\mathrm{V}(t){\geq}\,u,\;\; C \geq u\}}\;(j \,=\, 0{,}1)\,.$$

Under the Hu and Tsiatis second assumption,

$$\lambda_j^*(t,u) = \frac{\lim_{h \to 0} h^{-1}\,\mathrm{pr}\{u{\leq}\mathrm{V}(t){<}\,u{+}h\,,\;\; \mathrm{R}(u){=}\;\mathrm{j}\,\}{*}pr\{C \geq u\}}{pr\{\mathrm{V}(t){\geq}\;u\}{*}pr\{C \geq u\}}\;(j \,=\, 0{,}1)$$

$$= \lambda_j(t, u)$$

A.3 COUNTING PROCESS

Suppose a sample of n subjects has the observable random variables {$X_i(t)$, $\Delta_i(t)$, $R_i^*(t)$; $t \geq 0$} (i = 0,…, n). Define the counting process as: $N_{ji}(t, u) = I\{X_i(t) \leq u, \Delta_i(t) = 1\,, R_i^*(t) = j\}$ (j = 0,1, i = 0,…, n, $u \geq 0$), and the at-risk process as: $Y_i(t, u) = I\{X_i(t) \geq \mathrm{u}\}$. Let $N_j(t, u) = \sum_i N_{ji}(t,u)$, $N(t, u) = N_0(t, u) + N_1(t, u)$, and $Y(t, u) = \sum_i Y_i(t,u)$. We also define $dN_j(t,u)$ be the change of in the counting process over a short time interval [u, u+h).

We can substitute $\lambda^*_j(t,x)$ in (2.1.1) with $dN_j(t,x)/Y(t,x)$ and derive the following estimator:

$$\hat{\mathrm{S}}(t) \,=\, 1 - \int_0^{t+c(t)} exp - \{\int_0^x \frac{dN_1(t,u)+dN_0(t,u)}{Y(t,u)}\} \frac{dN_1(t,x)}{Y(t,x)}$$

$$= 1 - \int_0^{t+c(t)} exp - \{\int_0^x \frac{dN(t,u)}{Y(t,u)}\} \frac{dN_1(t,x)}{Y(t,x)}$$

$$= 1 - \int_0^{t+c(t)} \hat{E}(t, x^-) \frac{dN_1(t,x)}{Y(t,x)} \qquad \text{(A.3.1)}$$

, where $\hat{E}(t, x)$ is the Kaplan-Meier estimator of $E(t, x) = \mathrm{pr}(V(t) \geq x)$ and it is asymptotically equivalent to $exp - \{\int_0^x \frac{dN(t,u)}{Y(t,u)}\}$. The estimator (A.3.1) is equivalent to (2.1.2).

APPENDIX B

PARTIAL R CODE

```
install.packages("survival")
library(survival)

est <- function(n,los,mean_fail,inter_visit,delayed_time,x,C_x){
      #n=sample size
      #los=length of study;x=time point of interest
      #mean_fail=mean failure time
      #inter_visit=mean inter-arrival time
      #delayed_time
      #x=time point of interest

#----------------#
# Create dataset #
#----------------#
E1 <- runif(n,min=0,max=los)             #entering time
T1 <- rexp(n,1/mean_fail)                #failure times
delay <- runif(n,min=0,max=delayed_time)
Ak <- T1 + delay                  #final time at the failure time is recorded
F1 <- los - E1                           #follow-up times

# Times to hospital visits: a Possion process
U.matrix <- matrix(NA, nrow=n, ncol=30)
for (i in 1:n) {
      U.matrix[i,] <- cumsum(rexp(30,1/inter_visit))    #30 is random number
  }
for (i in 1:n) {
      if (U.matrix[i,1]>=Ak[i]){
            U.matrix[i,]<-Ak[i]} else
U.matrix[i,c(which(U.matrix[i,]>=Ak[i]))]<- Ak[i]
  }
U.matrix <- cbind(U.matrix,Ak)

# censoring time: the last monitoring time before follow-up for those
# individuals who are censored
C1 <- NULL
for (i in 1:n) {
      if (U.matrix[i,1]<=F1[i]&&F1[i]<Ak[i]){
```

```
C1[i]<-U.matrix[i,max(which(U.matrix[i,]<=F1[i]))]) } else C1[i] <- F1[i]
  }

#death indicator at time x
R <- NULL
  for (i in 1:n) {
      R[i]<-ifelse(T1[i]<=x,1,0)
  }

# First time at which an individual's vital status at time x is known
V <- NULL
  for (i in 1:n) {
      V[i] <- ifelse(R[i]==0,U.matrix[i,min(which(U.matrix[i,]>=x))],Ak[i])
  }

# For the right censored data
X <- pmin(V,C1)
delta <- ifelse(X==V,1,0) #indicate whether vital status at x is known
R_c <- NULL               #death indicator, defined only if vital status is
known
  for (i in 1:n) {
      if(delta[i]==1){
            R_c[i]<-R[i]} else R_c[i]<-NA
  }

## END OF DATA GENERATION ##
#-----------------------------------------------#

#-----------------------------------------------#
#       SURVIVAL DISTRIBUTION ESTIMATORS        #
#-----------------------------------------------#

## True survival distribution, S(x)
true.S <- exp(-x/mean_fail)

#----------------------------#
# Kaplan-Meier estimator, KM #
#----------------------------#
status1 <- as.numeric(Ak<C1)                    #status
surv.data <- Surv(pmin(Ak,C1), status1)         #right-censored data
KM.time <- summary(survfit(surv.data~1))$time
if(max(KM.time)>=x){
   KM.S<-c(summary(survfit(surv.data~1),times=x)$surv) } else KM.S<-
c(summary(survfit(surv.data~1),times=c(max(KM.time)))$surv)

#------------------------#
# Hu and Tsiatis, HT=1-G #
#------------------------#
G_ind.1 <- NULL     #C_x value=0 (assuming no delay)
  for (i in 1:n){
      G_ind.1[i] <-ifelse(X[i]<=x & delta[i]==1 & R_c[i]==1,1,0)
   }
G_ind.2 <- NULL     #C_x value is given
  for (i in 1:n){
      G_ind.2[i] <-ifelse(X[i]<=x+C_x & delta[i]==1 & R_c[i]==1,1,0)
   }
```

```
denom <- NULL
Y <- matrix(NA, nrow=n, ncol=n)
  for (i in 1:n){
    for (l in 1:n) {
      Y[i,l] <- ifelse(X[l]>=X[i],1,0)
      }
    denom[i] <- sum(Y[i,])
   }

X.surv.data <- Surv(X,delta)
X.KM.time <- summary(survfit(X.surv.data~1))$time
num <- NULL
  for (i in 1:n){
      num[i]<-
summary(survfit(X.surv.data~1),times=ifelse(X.KM.time[1]<X[i],X.KM.time[max(w
hich(X.KM.time<X[i]))],0))$surv
   }
#combine
G.1 <- NULL
  for (i in 1:n){
      G.1[i] <- G_ind.1[i]*num[i]/denom[i]
   }
HT.1 <- 1 - sum(G.1)  #when C_x value=0

G.2 <- NULL
  for (i in 1:n){
      G.2[i] <- G_ind.2[i]*num[i]/denom[i]
   }
HT.2 <- 1 - sum(G.2)  #when C_x value is given

#-----------------------#
# Laan and Hubbbard, LH #
#-----------------------#
#combine
LH <- NULL
  for (i in 1:n){
      LH[i] <- R[i]*delta[i]*num[i]/denom[i]
   }

LH <-1- sum(LH)

#---------------------#
# Report the result   #
#---------------------#
out <- c(true.S,KM.S,HT.1,HT.2,LH)
return(out)
}

#------------------------------------------------------------------#
#-------------------------#
# Simulation              #
#-------------------------#

#change for different settings
n=60
mean_fail=1
```

```
#constant settings
los=4
inter_visit=0.5
C_x=1

### Pattern I (No delay)
delayed_time=0

#12months
x=1
P1 <- matrix(NA, nrow=NT, ncol=5)
p1.stime1<- system.time(
for(sim in 1:NT){
      P1[sim,] <- est(n,los,mean_fail,inter_visit,delayed_time,x,C_x)
      })
mse.KM <- sd(P1[,2])                #MSE of KM
mse.HT1 <- sd(P1[,3])               #MSE of HT
mse.HT2 <- sd(P1[,4])               #MSE of HT
mse.LH <- sd(P1[,5])                #MSE of LH
P1.row1 <- c(colMeans(P1),mse.KM,mse.HT1,mse.HT2,mse.LH)


### Pattern II (Delay ~ UNIF(0,1))
delayed_time=1

#12months
x=1
P2 <- matrix(NA, nrow=NT, ncol=5)
p2.stime1<- system.time(
for(sim in 1:NT){
      P2[sim,] <- est(n,los,mean_fail,inter_visit,delayed_time,x,C_x)
      })
mse.KM <- sd(P2[,2])                #MSE of KM
mse.HT1 <-sd(P2[,3])                #MSE of HT w/C(x)=0
mse.HT2 <-sd(P2[,4])                #MSE of HT w/C(x)=1
mse.LH <-sd(P2[,5])                 #MSE of LH
P2.row1 <- c(colMeans(P2),mse.KM,mse.HT1,mse.HT2,mse.LH)


### Pattern III (Delay ~ UNIF(0,2))
delayed_time=2

#12months
x=1
P3 <- matrix(NA, nrow=NT, ncol=5)
p3.stime1<- system.time(
for(sim in 1:NT){
      P3[sim,] <- est(n,los,mean_fail,inter_visit,delayed_time,x,C_x)
      })
mse.KM <- sd(P3[,2])                #MSE of KM
mse.HT1 <-sd(P3[,3])                #MSE of HT w/C(x)=0
mse.HT2 <-sd(P3[,4])                #MSE of HT w/C(x)=1
mse.LH <-sd(P3[,5])                 #MSE of LH
P3.row1 <- c(colMeans(P3),mse.KM,mse.HT1,mse.HT2,mse.LH)


### Pattern IV (Delay ~ UNIF(0,0.5))
```

```
delayed_time=0.5

#12months
x=1
P4 <- matrix(NA, nrow=NT, ncol=5)
P4.stime1<- system.time(
for(sim in 1:NT){
      P4[sim,] <- est(n,los,mean_fail,inter_visit,delayed_time,x,C_x)
      })
mse.KM <- sd(P4[,2])                #MSE of KM
mse.HT1 <-sd(P4[,3])                #MSE of HT w/C(x)=0
mse.HT2 <-sd(P4[,4])                #MSE of HT w/C(x)=1
mse.LH <-sd(P4[,5])                 #MSE of LH
P4.row1 <- c(colMeans(P4),mse.KM,mse.HT1,mse.HT2,mse.LH)


##Note: Codes for other time points are omitted.
```

BIBLIOGRAPHY

Fine, J. and Tsiatis, A. (2000). Testing for differences in survival with delayed ascertainment. *Biometrics* **56**, 145-153.

Fleming, T. and Harrington, D. (1991). Counting Processes and Survival Analysis. New York: Wiley.

Hu, P. and Tsiatis, A. (1996). Estimating the survival function when ascertainment of vital status is subject to delay. *Biometrika* **83**, 371-379.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 5**3**, 457-481.

Klein, J. and Moeschberger, M. (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.

Miller, R. (1981). Survival Analysis. New York: John Wiley and Sons.

Robins, J. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate makers. *Proceedings of the Biopharmaceutical Section, American Statistical Association*, 24-33.

Van der Laan, M. and Hubbard, A. (1998). Locally efficient estimation of the survival distribution with right-censored data and covariates when collection of data is delayed. *Biometrika* **85**, 771-738.