5-2015

Survival Analysis for Truncated Data and Competing Risks

Michael Steelman
Utah State University

SURVIVAL ANALYSIS FOR TRUNCATED DATA
AND COMPETING RISKS

by

Michael Steelman

A research project submitted in partial fulfillment
Of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

Dr. Chris Corcoran
Major Professor

Dr. John Stevens
Committee Member

Dr. Adele Cutler
Committee Member

UTAH STATE UNIVERSITY
Logan, Utah

2015

Survival Analysis for Truncated Data and Competing Risks

Abstract

The purpose of this project is to consider the problems of left truncation and competing risks in analyzing censored survival data, and to compare and contrast various approaches for handling these problems. The motivation for this work comes from an analysis of data from the Cache County Memory Study. Study investigators were interested in the association between early-life psychologically stressful events (e.g., parental or sibling death, or parental divorce, among others) and late-life risk of Alzheimer's disease (AD). While conventional methods for censored survival data can be applied, the presence of left truncation and competing risks (i.e., other adverse events such as death that may lead to censoring with respect to AD) may require some consideration in order to avoid potential bias in terms both of estimation and inference. In this paper we briefly summarize the issues of truncation and competing risks in the context of survival analysis, and apply and compare several approaches suggested in the literature to the Cache County Data.

Introduction

Data taken from the Cache County Memory Study (CCMS) can be used to illustrate survival analysis in the context of left truncation and competing risk. Left truncation describes the setting where at least some sampled subjects are not at risk for the event of interest during some period following study entry. Without appropriate adjustment, the inclusion of some individuals not at risk for the event of interest can result in potential bias. Competing risks arise when there are other possible events that will prevent the event of interest from being observed. There are researchers who concern themselves primarily with the event of interest and will delete observations in a dataset that have experienced a different event. This has the potential of negatively biasing inferences made about the population. Reality is rarely so neat as to allow only one possible outcome to be observed. For example, when measuring survival times between treatment and cure, the patient may die before the endpoint of interest is observed. The CCMS data are incomplete; that is to say, they exhibit both left truncation and right censoring. CCMS data also contain a variety of outcomes, thus necessitating some consideration of competing risks. We will examine these issues in an analysis of CCMS data to study the effects of psychological stressors on Alzheimer's disease.

Description of Data

The data for this study were collected as part of the Cache County Memory Study. Participants in the CCMS were 65 and older and living in Cache County as of Jan. 1, 1995. The initial sample contained 5092 subjects, comprising 90% of the eligible population. Descriptive information for this sample is found in *Table 1.* The sample is 52.5% female. The average age at time of entry in the sample is 76.34 (sd = 7.34). The average age at last contact is 80.82 (sd = 6.71). This variable indicates age of AD onset or age of last contact for non-cases. The sample also is 65.93% married; 29.27% widowed; with the rest being divorced, separated, or never married. The average years of education is 13.18 (sd = 2.89).

The stressor factors considered were based on general lifetime stressors and interval specific stressors. Lifetime stressors were multiple marriages, no marriages, no college education, low socioeconomic status, no children, many (>8) children, and long-term unskilled employment. Time intervals used were birth to four years old, five to eleven, twelve to seventeen, eighteen to 30, 31 to 50, and 50 to the date of the baseline interview. So-called interval stressors (i.e. experienced during various intervals of childhood or adulthood) included death of a father, mother, sibling, spouse, or child, premature or low birth weight child, stillbirth, single parenthood, and divorce. While there is value in examining individual stressors or subsets of related stressors, for the purpose of this analysis we consider cumulative stress, measured as the total number of individual stressors. The average individual number of stressors experienced was 5.96 (sd = 3.21).

Survival Analysis

A survival analysis uses longitudinal time-to-event data (Allison, 2010). With respect to the distribution of event times, data analysis focuses generally on the hazard and survivor functions. The hazard function for time-to-event data characterizes instantaneous risk, defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t},$$

where *t* is a time point, *T* is the time of the event, and *Δt* is the smallest possible amount of time we can calculate. It is important to note that hazard functions are based on individual risk, not population risk. This means that each hazard is calculated for the individual at a specific time.

In practice, hazards are generally time-dependent. The hazard for a person getting in an accident is greater if they are driving during rush hour than if they are driving in the middle of the day.

The survival curve *S*(*t*) is the complement of the cumulative distribution function (cdf), or P(T > t). The survival curve is the probability that a person is still alive at time *t* given that they have survived to that point. Survivor functions are typically estimated using the Kaplan-Meier (K-M) curve. Every time there is an event, the survival probability decreases (Rich, et al., 2010). These changes are not continuous and can be better explained as a stepwise function. The probabilities represent the proportion of people who have survived the event given they are still at risk. Each new proportion is multiplied by the former cumulative proportion of survivors to obtain the new probability. This probability will hold constant until the next event time and so forth.

There are times when some people in a study may not have an event. It is possible that they had it earlier, or the study did not last long enough to observe the event occurring. This is called censoring. Observations are considered right censored when either a subject drops out of the study or the study ends before the event of interest is observed. Data are considered left censored when the event occurs at some unknown time before the study begins. They are considered interval censored if all we know is that the event occurred between two different times. This study took efforts to adjust for right censoring. Left censoring was ignored because the timing of events in the past were known. Interval censoring was ignored because while the study was collected in waves, the timing of the event was recorded based on time of diagnosis, not time of report.

Regression Models

A subset of parametric models can be formulated in this regression setting, meaning that there are assumptions about the distribution of the hazards across time. There is also a subset of non-parametric models that can be formulated when we do not know what the distribution of the hazards are. Out of the numerous techniques that exist, David Cox's proportional hazards model is the most popular (Cox, 1972). The hazard function for subject *i* at time *t* is written as:

$$h_i(t) = \lambda_0(t)\exp(\beta_1 x_{i1} + \cdots + \beta_k x_{ik}).$$

In this model, $\lambda_0(t)$ represents an unknown, nonnegative function equal to the hazard function when all covariates are zero. Also, β_k represents the average change in risk for the event for a unit change in covariate *k*. X_{ik} represents the value of the covariate for person *i*. The linear model in the exponent represents a set of covariates. As an exponentiated linear model is not easily interpreted, it will be more feasible to take the log of both sides of the equation. This results in a new regression model:

$$\log h_i(t) = \ \alpha(t) + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}.$$

In this model, $\alpha(t)$ is equal to log $\lambda_0(t)$. The results of a statistical model fit in this way now represent the log hazards.

Cox's regression model is considered to be semi-parametric. The model is able to accommodate all different hazard functions when predicting the hazards, but does not rely on any particular distribution needed (Allison, 2010). This makes it a non-parametric test. It also relies on the idea that hazards for the values of the covariates are constant (linear) across time. This is a parametric assumption and shows how Cox regression relies on both a parametric and a non-parametric test.

Model Assumptions

There are some assumptions that were made when fitting the models. A failure of the data to follow these assumptions would end with the results being invalidated. There are three primary assumptions to address: no collinearity, functional form of continuous covariates, and proportional hazards.

Collinearity describes when multiple covariates are actually measuring the same thing. This is a problem if it occurs because the significance of the effect will be masked by the other covariate. Checking for collinearity can be done by looking at the proportion of variance that each covariate accounts for. In *Table 2,* I looked at the proportion of variances for the covariates and concluded that none of the components showed collinearity.

Functional form is the idea that continuous covariates are linear across time. This is the assumption that was mentioned earlier when discussing the parametric part of Cox regression. If the covariates do not follow functional form, then transformations will have to be performed on them. The results of the check for functional form can be seen in *Table 3*. There were no violations of this assumption in this model.

I also tested whether or not the hazards were proportional for the covariates. Without proportional hazards, the Cox regression model would be invalidated under traditional methods. The results of this diagnostic shown in *Table 4* indicate there were no violations of this assumption in the model.

Log-Rank Test

A log-rank test can be used to determine if independent groups have different survival curves. I used it to determine differences in AD-free survival by gender. While not perfectly identical in every case, log-rank tests can be approximated by fitting a Cox regression model with a single, binary covariate (SAS Institute Inc., 1999**)**. The results for this test are seen in

Table 5. The log-rank test was not statistically significant, so I did not stratify other models by gender.

Left Truncation

Whereas censoring is an outcome that a researcher cannot control, truncation is a direct result of the researcher choosing to exclude a certain group of people from the study because of their event times. Left truncation is a result of a researcher only including people who have not yet experienced the event. This may sound like an ideal situation for survival analysis, but bias is introduced because there may be people who have had an event at earlier times. For example, in the population the CCMS dataset was sampled from, it is reasonable to assume that some people may have died from AD before the study had a chance to measure them. If these people are from the same cohort as the sampled population, the estimates of survival times will have become larger than what they should be.

To best demonstrate why left truncation is a problem and how to fix it, more information will be needed on how survival analysis works. Survival is measured on a certain time-scale. This time-scale must have a value where time is zero, also known as a baseline. In most cases of survival analysis, the baseline is considered to be at the start of the study. This makes the assumption that the probability of the event does not change with age. To control for this, a Cox regression model can include age at entry into the study as a covariate. An alternative approach, according to Canchola, et al. (2003), is to treat age as the time-scale.

Treating age as the time-scale comes with its disadvantages. One disadvantage is that the procedure is computationally expensive, meaning it takes longer for a computer to finish the analysis. Another issue that requires more attention is that when age is the time-scale, the baseline becomes birth. This helps explain the source of the bias with left truncation. If people truly were observed since birth, then survival estimates would be bias free. Since this is rarely the case, the assumption can be made that survival estimates are higher than what they should be, since data for people who have experienced the event before the study might not be available. The solution is to restrict survival times from the age the participant entered the study to the age they left the study. This technique restricts survival estimates to apply only to the people who are the same age during the study. The result is a lower survival estimate (Canchola, et al., 2003).

When selecting which time-scale to use, it is important to consider what you are trying to find out. Wanting to know what happens to a person across the course of a study is different from wanting to know what happens as a person ages. While there may not be a large difference statistically, using age as the time-scale when studying aging effects does lead to easier interpretation of estimates since they are being made over the course of aging and not time (Lamarca, et al., 1998).

Results of Left Truncation

A comparison of the survival estimates before and after left truncation is demonstrated in *Figure 1* and, as can be seen, the estimates are reduced by the adjustment. The results in *Table 6* demonstrate what happens to the Cox regression model before and after left truncation is adjusted for. Before the data were adjusted for left truncation, age at time of entry was considered a significant covariate ($\beta = -.136$, $p < .001$). After the adjustment, there were no significant covariates.

Competing Risk

There are times in survival analysis where something happens that prevents the observation of the event of interest. These events are called competing risks. In the CCMS dataset, it has been noted when a person is diagnosed with AD and when they were diagnosed with another form of dementia. There is a choice to just ignore this event because it is not the primary concern, or we can account for it in the analysis. The general consensus tends to be the more data we have the better.

The two most popular methods for handling competing risks are cause-specific hazards and the cumulative incidence function (So, Lin, & Johnston, 2014). Cause-specific hazards are simpler than the cumulative incidence function. Cause-specific hazards fit a model that only cares about a single event of interest at a time. When using traditional survival analysis, a status variable is used to indicate whether participants experienced the event or were censored. With cause-specific hazards, the status variable can indicate the event, censoring, or some competing event. When researchers are only interested in a single event, they indicate when the event occurs with the status variable and treat all other outcomes as right censored observations. If the researcher is interested in all events, then multiple models can be fit with each competing event having a turn as the primary event of interest. This method is useful because it gives a better idea of how the covariates affect the hazard of the event of interest. If the goal of the study is to know how variables are associated with an event then cause-specific hazards is the most appropriate approach.

The strength of this approach can also be a weakness. Consider the example above of fitting a different cause-specific hazard model for each event when there are multiple events. In theory, it would make sense for the sum of the probabilities from all possible outcomes to equal one. The problem is that because the other events are being censored, the interpretation becomes the probability of the event of interest at time *t* in the absence of other events (Allison, 2010). When we treat events as absent, they are counted in each model we fit. This results in lower survival estimates on the survival curve and a cumulative probability of events occurring to be greater than one.

The most common alternate approach is called the cumulative incidence function. This method can be used to determine the probability of a particular event occurring in the presence of other events. The function is fixed so that the sum of the probabilities of any event and survival at a particular time will be equal to one. In other words, all possible scenarios are represented. This technique is best used in applied settings. The estimates do not work as well when it comes to understanding how a particular event is affected by covariates, but it is extremely useful when an inference is made about the patient's actual likelihood of survival of any kind of event death. The way hazard estimates are calculated in the cumulative incidence function for competing events is similar to the hazard estimates calculated for single events. The only difference is that for each event that occurs, the survival probability decreases, but the probability of experiencing a certain event at a certain time only increases when that event occurs (Kim, 2007). By using this technique, the sum of all possible outcomes (survival versus all possible events) equals one across time.

Results of Cause-Specific Hazards

Because the object of this study was to determine how the risk for Alzheimer's is associated with psychological stressors, cause-specific hazards would be the most appropriate method. In addition to AD, observations were also made when participants experienced a form of dementia that was not AD. As an alternate technique, I deleted any observation that experienced a competing risk. The results in *Table 7* show the differences between the two models. It should also be noted that left truncation and competing risks can be accounted for at the same time. That is why the model that accounts for left truncation in *Table 6* and the model that accounts for competing risks in *Table 7* are the same.

Results for the Cumulative Incidence Function

Even though the cumulative incidence function was not the best choice for this study, I still thought it would be beneficial to get a deeper understanding of what happens with this model. In *Figure 2*, I drew the cdf for both of the competing risks using cause-specific hazards, AD and non-AD dementia. The graph shows that the sum of all possible outcomes was 1.72.   In *Figure 3*, I drew the cdf for the competing risks using the cumulative incidence function. This time the sum of all possibilities was one.

The results of a regression model with the cumulative incidence function is shown in *Table 8*. In this model, it was shown that gender ($\beta = .228$, $p < .05$) and age of entry ($\beta = -.052$, $p < .001$) were both significant covariates. It should be noted that the method used to get the estimates was not able to adjust for left truncation. Being able to adjust for left truncation would likely change the results like they did in *Table 6*.

Conclusions

There are many different techniques within survival analysis. The difficulty lies in discovering which methods are the most appropriate to use. While the sampling procedure from the CCMS dataset gives us a unique opportunity to see the event times of participants before the study began, it is still possible that there were people who developed AD then died before the study began. This results in left truncation still being present and requiring adjustment. Since this study was more concerned with associations between stressors and AD, I decided that using cause-specific hazards to control for competing risks would be the better option.

One additional final issue is the possibility of competing risks to be dependent on each other. When this occurs, bias is introduced into the analysis. Unfortunately, there is currently no way to ensure that the competing risks are independent. There has been work on methods to adjust for the bias, but the assumptions needed to use those adjustments are still untestable (Geloven, et al., 2014).

References

Allison, P. D. (2010). *Survival Analysis Using SAS® : A Practical Guide Second Edition*. Cary, NC: SAS Institute Inc.

Canchola, A. J., Stewart, S. L., Bernstein, L., West, D. W., Ross, R. K., Deapen, D., … Horn-Ross, P. L. (2003). Cox Regression Using Different Time-Scales. Retrieved from www.lexjansen.com/wuss/2003/DataAnalysis/i-cox_time_scales.pdf

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34 (2),* 187-220. Retrieved from http://www.jstor.org/stable/2985181

Geloven, N. v., Geskus, R. B., Mol, B. W., & Zwinderman, A. H. (2014). Correcting for the Dependent Competing Risk of Treatment Using Inverse Probability of Censoring Weighting and Copulas in the Estimation of Natural Conception Chances. *Statistics in Medicine*, *33,* 4671-4680. doi: 10.1002/sim.6280

Kim, H. T. (2007). Cumulative Incidence in Competing Risks Data and Competing Risks Regression Analysis. *Clin Cancer Res*, *13(2),* 559-565. Doi: 10.1158/1078-0432.CCR-06-1210

Lamarca, R., Alonso, J., Gomez, G., & Munoz, A. (1998). Left-Truncated Data with Age as Time Scale: An Alternative for Survival Analysis in the Elderly Population. *Journal of Gerontology: MEDICAL SCIENCES, 53A (5)*, M337-M343.

SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

Table 1. Descriptive Summary for Subjects by Status

Status Variables	No Dementia	Alzheimer’s	Non-Alzheimer’s Dementia
N = 5092	4150	645	297
Entry Age	75.46 (7.05)	80.64 (7.19)	79.34 (7.59)
Education	13.22 (2.86)	13.03 (3.04)	12.94 (2.94)
Female	56.58%	66.05%	51.85%
Married	67.74%	54.11%	63.33%
Widowed	27.14%	42.68%	33.33%
Other married status	5.12%	3.21%	3.33%
Total Stressors	5.81 (3.14)	6.62 (3.50)	6.60 (3.29)

Table 2. Collinearity Diagnostics

			Proportion of Variation						
Number	Eigenvalue	Condition Index	Intercept	Education	Gender	Entry Age	Widowed	Other Marital	Total Stressors
1	4.86	1.00	2 E-4	.001	.011	2 E-4	.009	.002	.006
2	1.00	2.20	4 E-6	7 E-5	3 E-4	7 E-7	.064	.791	3 E-4
3	.61	2.81	6 E-4	.006	.048	5 E-4	.488	.160	.004
4	.32	3.89	3 E-5	3 E-4	.831	1 E-4	.210	.046	.05
5	.18	5.18	.001	.044	.050	7 E-4	.107	6 E-5	.623
6	.02	14.54	.040	.857	.037	.080	.008	3 E-4	.300
7	.003	38.96	.958	.092	.023	.918	.113	.001	.017

Table 3. Supremum Test for Functional Form

Variable	Education	Entry Age	Total Stressors
Max Abs. Value	12.04	17.46	13.08
P-value	.32	.21	.31

Table 4. Supremum Test for Proportional Hazards Assumption

Variable	Education	Gender	Entry Age	Widowed	Other Marital Status	Total Stressors
Max Abs. Value	1.14	.80	1.05	1.20	.51	.84
P-value	.18	.46	.13	.15	.78	.46

Table 5. Log-Rank Tests for Gender

Model	DF	Estimate (SE)	Chi-Square	P-value
Log-rank test	1	.169 (.105)	2.595	.107

Table 6. Cox Regression for Left Truncation

Variables	Education	Gender	Entry Age	Widowed	Other Marital Status	Total Stressors
No LT	.008 (.019)	.180 (.116)	-.136*** (.013)	.112 (.124)	-.148 (.312)	-.012 (.019)
LT	.009 (.020)	.182 (.116)	-.024 (.017)	.100 (.124)	-.185 (.312)	-.012 (.019)

*** - significant at α=.001

Table 7. Cox Regression with Competed Risks

Variables	Education	Gender	Entry Age	Widowed	Other Marital Status	Total Stressors
Deleted	.010 (.020)	.173 (.116)	-.018 (.017)	.088 (.123)	-.204 (.312)	-.011 (.019)
Competing	.009 (.020)	.182 (.116)	-.024 (.017)	.100 (.124)	-.185 (.312)	-.012 (.019)

*** - significant at α=.001

Table 8. Cumulative Incidence Function with Regression

Variables	Education	Gender	Entry Age	Widowed	Other Marital Status	Total Stressors
CI Model	.011 (.016)	.228* (.096)	-.052*** (.007)	.077 (.103)	-.099 (.244)	-.009 (.015)

* - significant at α = .05, *** - significant at α = .001

Figure 1. Survival Curves Before and After Adjusting for Left Truncation With Censored Competing Risks.

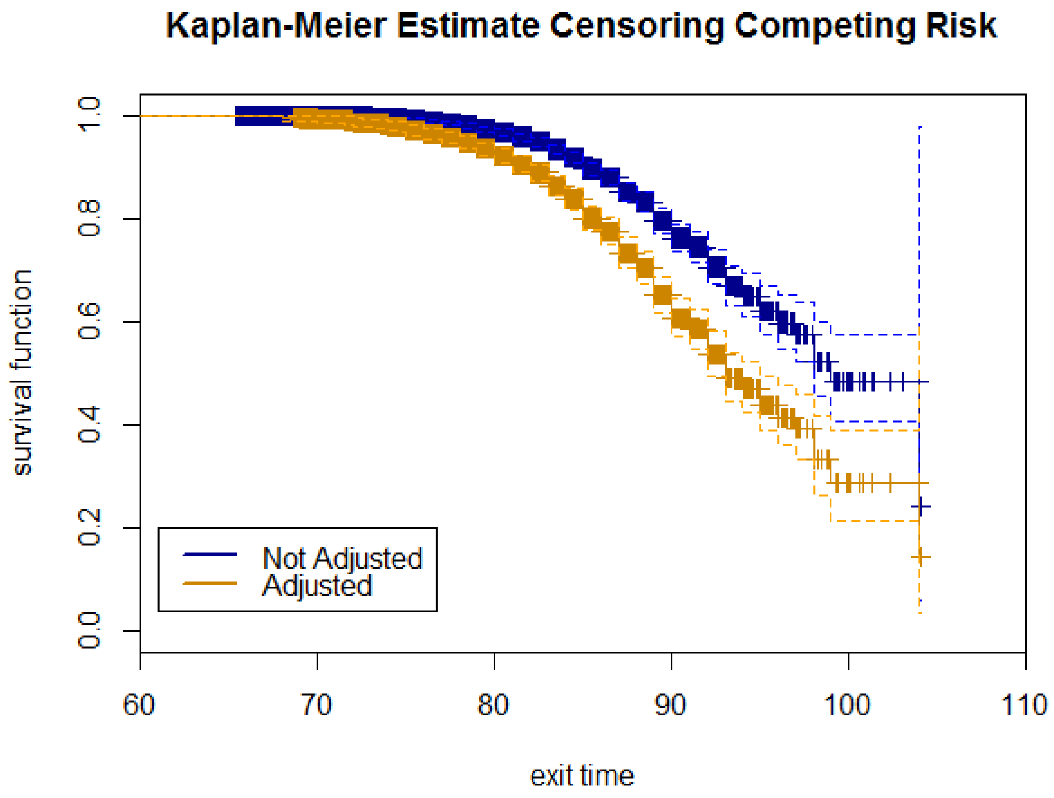


Figure 2. CDF of Hazards for Cause-Specific Hazards Method

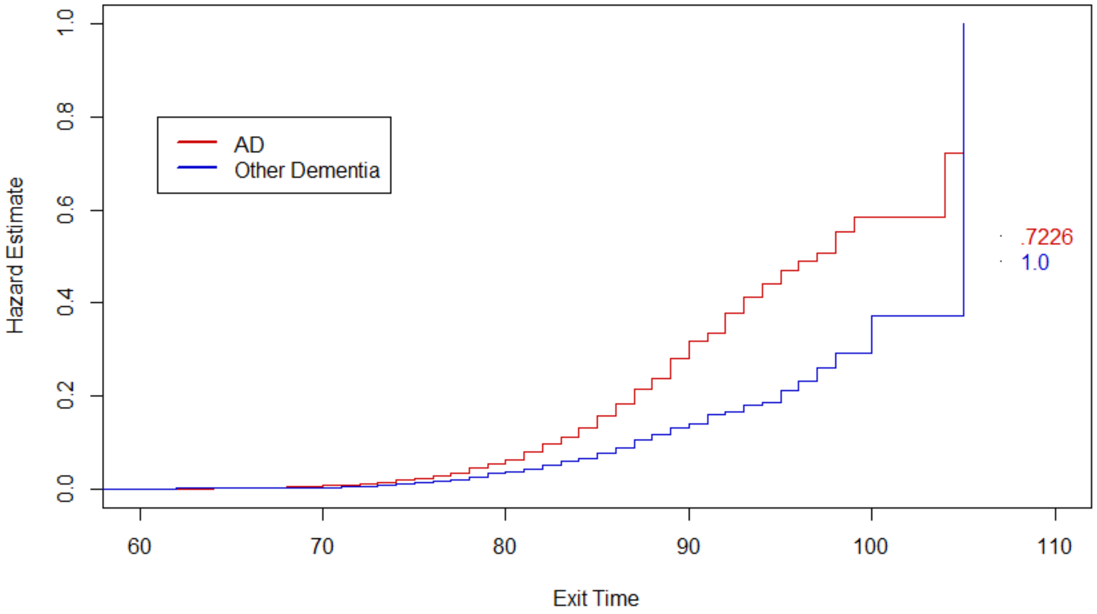
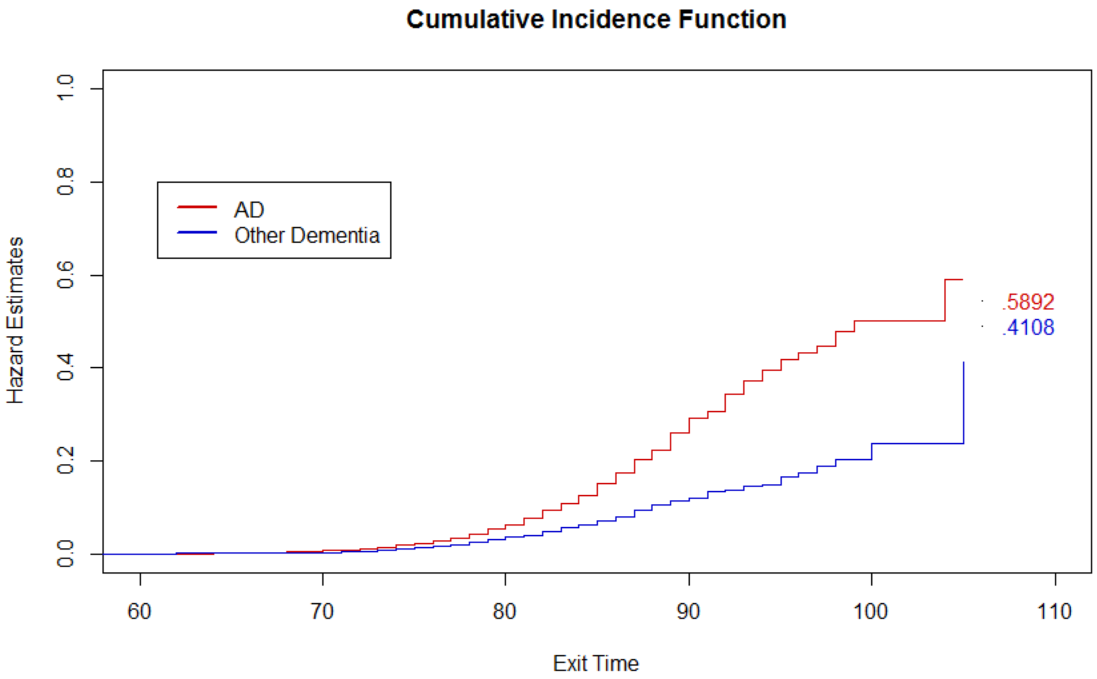


Figure 3. CDF of Hazards for Cumulative Incidence Function Method


Cumulative Incidence Function
AD
Other Dementia
Hazard Estimates
Exit Time