Survival Analysis in the Presence of Competing Risks

Benjamin Russoniello

December 13 2019

1 Introduction

In this paper, I will review some of the current methodology available for the analysis of survival data in the presence of competing risks. These methods have been drawn from textbooks and journal articles published in the past 20 years. I demonstrate the use of these methods with hypothetical scenarios of made up data, and modelling of data drawn from a real data set. All code examples (except the one in the Appendix) will be done in SAS, and will be my original code.

2 Estimate of the Survival Function With No Competing Risks

In order to motivate the more complicated case, consider the situation when a cohort of patients is being observed after septic shock until the time of their death. Let the value of the random variable T be the number of hours until a patient has either died (experienced the event of interest), or has been discharged from the hospital, lost to follow-up, etc. Any case in which a patient stops being observed due to those events described previously is referred to as "censored". Our goal is to provide a reasonable estimate for the survival function, $S(t)$, which can be used to estimate how long a patient will be alive after experiencing septic shock. In conventional probability notation,

$$S(t) = P(T > t) = 1 - F(t)$$

, where t is any arbitrary time given in hours (1,3.1,2.175,etc) and $F(t)$ is the cumulative distribution function of T.

The product-limit estimate of $S(t)$ is given by Kaplan and Meier (1958)[1]. For this approach, consider we have a set of k patients. Suppose i of these patients have experienced the event of interest (death), and j of the patients did not experience the event before follow-up ended and are therefore censored. Let $t_1, t_2, ..., t_i$ be the values of T at which the i patients experienced the event of interest. Then, Kaplan and Meier showed that $S(t)$ can be estimated by $\hat{S}(t)$, where:

$$\hat{S}(t) = \Pi_{t_r<t}(1 - \frac{i_r}{k_r})$$

where i_r represent the number of patients who experienced the event at time t_r. Note that if

$$t_p \neq t_q$$

for any integers p and q, then it follows that i_r is always 1.

k_r is the number of patients who were at risk of experiencing the event at time t_r. This number excludes patients who have already experienced the event or are censored.

One assumption made by the Kaplan-Meier esimator of the survival function is that all subjects at risk during a specific time interval have the same instantaneous rate of failure (Putter et al)[2]. This is referred to as the hazard, and one can identify the hazard at any time t given the definition of the hazard function $h(t)$ of a random variable T:

$$h(t) = \lim_{\delta \to 0} \frac{P(t < T < t + \delta | T > t)}{\delta}$$

3 Definition of a Competing Risk and the Impact of Competing Risks on Survival Function Estimation

The Kaplan-Meier estimate of the survival function is a useful tool in obtaining a nonparametric estimate of the survival function under a given set of circumstances. These circumstances being that there is only a single event to measure, and nothing can preclude that event from happening in the duration of the study. Unfortunately, this situation is rarely realistic when studies become more specific. For example, instead of the event of interest being death, consider the case when the event of interest under study is death due to breast cancer. Certainly, death due to any other cause, such as organ failure, would prevent death due to breast cancer from occurring. In our hypothetical scenario, death due to organ failure would be referred to as a *competing risk*.

A precise definition of competing risk reads: An event whose occurrence precludes the occurrence of the primary event of interest (Austin et al)[3]. Note that this can apply to different situations than just death. One such scenario is a case when a study's primary event of interest is the first event that occurs. For example, suppose diabetic patients are being observed and the primary event of interest is that eye complications arise before any other diabetic complications arise. Then, if the first complications from diabetes are related to feet, then those complications are a competing risk. It is important to note that a patient can still experience both complications, but can only experience one first.

In a competing risks scenario, a typical approach is to treat those who have experienced a competing risk as being censored (Klein and Kleinbaum)[4], and proceeding to use the Kaplan-Meier approach for estimating the survival function at a particular time t. This approach, while simple, results in biased estimates for the survival function (Putter et al)[2].

Recall from the previous section is that one assumption made in constructing the Kaplan-Meier estimate of the survival function is that all subjects who have not yet experienced the event of interest have the same hazard. When you consider the definition of the competing risk, the issue with applying Kaplan-Meier to a competing risks scenario becomes evident. Returning to a previous example, if the event of interest is death from breast cancer, treating a patient who has experienced death from breast organ failure as censored is akin to saying that the patient still has a nonzero probability of experiencing death from breast cancer.

Gooley Et al[5] remarks that, while methods to provide a more accurate estimation of the survival function in a competing risks setting have been around since the 1970s, it is still common to see the Kaplan-Meier estimate being used inappropriately. The approach of treating competing risk events as censored, is, unfortunately, more than uncommon in medical literature. One study[6] by the University of Ottawa searched through a database of prominent medical journals and found those that conducted a Kaplan-Meier analysis to estimate the survival function of their subjects. The study found that 46% of these journals were susceptible to some sort of bias due to competing risk, and may have overestimated the actual risk of the event under study. Thus, it is important that measures be taken by statisticians and clinical researchers to incorporate the methods available to account for competing risks in order to publish accurate medical studies.

4 Cause Specific Hazard and the Cumulative Incidence Function

In order to accurately measure the risk of an event in the presence of competing risks, one nonparametric method is to use the cumulative incidence function. If T is a random variable that denotes the time to an event, then the cumulative incidence function for an event j can be defined as:

$$C_j(t) = P(T < t, D = j)$$

, where D is a set containing the possible indexes for competing risks (e.g. if there is an event of interest and two competing risk events, then $D =$ 1,2,3, where 1, 2, or 3 could be the event of interest).

It is worth noting that the cumulative incidence function is not exactly the same to $\hat{S}(t)$, the Kaplan-Meier estimate for the survival function. The cumulative incidence gives the probability that a patient will encounter the event of interest before time t, whereas $\hat{S}(t)$ measures the probability of surviving past time t. Thus, $C_j(t)$ is more often compared to 1-$\hat{S}(t)$, as we will see shortly.

In order to estimate $C_j(t)$ for some event j, it is first necessary to define the *cause-specific* hazard function. The cause-specific hazard function gives the instantaneous rate of failure from some cause j, and is denoted as:

$$h_j(t) = \lim_{\delta \to 0} \frac{P(t < T < t + \delta, D = j | T > t)}{\delta}$$

Let $\hat{C}_j(t)$ be the estimate for the cumulative incidence function for an event j. Then, Putter et al[2] offers a complete derivation of this estimate as follows:

$$\hat{C}_j(t) = \sum_{r: t_r < t} p_j(t_r)$$

where

$$p_j(t_r) = \hat{h}_j(t_r)\hat{S}(t_{r-1})$$

$\hat{h}_j(t_r)$ is the estimate for the cause-specific hazard at time t_r, and is defined as:

$$\hat{h}_j(t_r) = \frac{i_{jr}}{k_r}$$

where i_{jr} is the number of subjects who failed at cause j at time t_r. k_r is the number of subjects at risk from event j at time t_r. k_r will always exclude those who have been censored or those who have failed from a competing risk event.

$\hat{S}(t)$ denotes overall survival, and is estimated as discussed in section 2. In that formula, i_r represents those who have failed from any cause, and k_r only excludes those who have been censored up to time t.

A simplistic way to think of the cumulative incidence function, is that it is a sum of the probabilities of failing at all times before your time of interest. These probabilities are estimated by taking the probability of surviving up to just before a time of interest, and then experiencing the event. The probability of experiencing the event at a specific time is given by the estimate of the cause specific hazard.

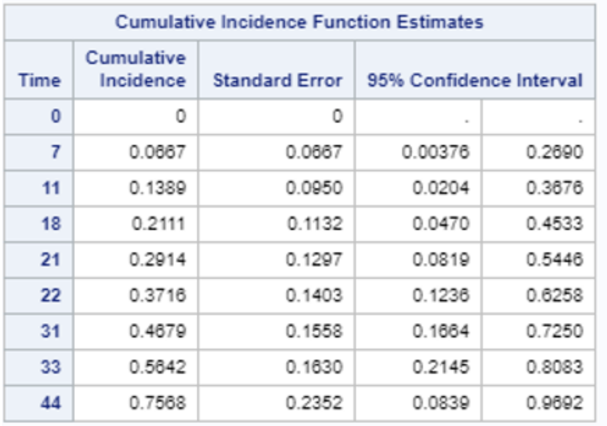
The appendix of this paper contains an R code example that computes the nonparametric estimate of the cumulative incidence function using simple logic. In SAS, the PROC LIFETEST procedure can be used to obtain estimates cumulative incidence function and its graph for certain events.

For example, consider a hypothetical scenario where 15 subjects were under study to examine the risk of experiencing some event of interest (Event 1). There exists one competing risk event (Event 2) that some patients experienced over the course of the study. Then, the following SAS code can be used to generate a cumulative incidence function estimate for times of the occurrence of the event of interest (Note that Event=0 denotes a censoring event in this scenario):

```
PROC LIFETEST method=KM data=Example plots=CIF;

time Time * Event (0)/failcode=1;
run;
```

The key difference between this code and the code used to generate a Kaplan-Meier estimate for a survival function is the introduction of the **failcode** option in the **time** statement. This specifies that I would like to generate the cumulative incidence function for the event coded as 1. The code above generates some important pieces of output, most notably, the cumulative incidence function estimates (with interval estimates):



Cumulative Incidence Function Estimates				
Time	Cumulative Incidence	Standard Error	95% Confidence Interval	
0	0	0	.	.
7	0.0667	0.0667	0.00376	0.2690
11	0.1389	0.0950	0.0204	0.3676
18	0.2111	0.1132	0.0470	0.4533
21	0.2914	0.1297	0.0819	0.5446
22	0.3716	0.1403	0.1236	0.6258
31	0.4679	0.1558	0.1664	0.7250
33	0.5642	0.1630	0.2145	0.8083
44	0.7568	0.2352	0.0839	0.9692

Figure 1: CIF estimate from SAS

and a graph of the cumulative incidence curve:

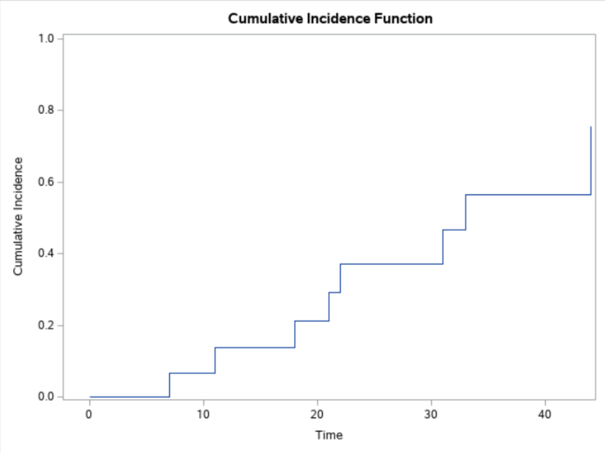
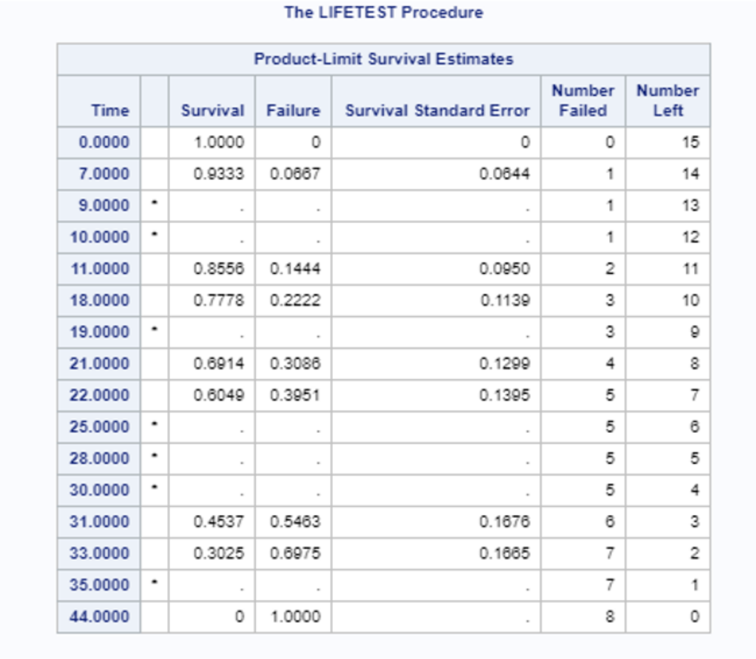


Figure 2: The CIF graphed from SAS

Suppose that, instead of calculating the cumulative incidence function, we simply treating competing risks as censored observations, and computed a Kaplan-Meier estimate of the survival function. Then, what estimates would we have gotten for $P(T < t)$ for the values of t at which the event of interest occurred? That is given by the following output:



The LIFETEST Procedure

Product-Limit Survival Estimates

Time		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000		1.0000	0	0	0	15
7.0000		0.9333	0.0667	0.0644	1	14
9.0000	*	.	.	.	1	13
10.0000	*	.	.	.	1	12
11.0000		0.8556	0.1444	0.0950	2	11
18.0000		0.7778	0.2222	0.1139	3	10
19.0000	*	.	.	.	3	9
21.0000		0.6914	0.3086	0.1299	4	8
22.0000		0.6049	0.3951	0.1395	5	7
25.0000	*	.	.	.	5	6
28.0000	*	.	.	.	5	5
30.0000	*	.	.	.	5	4
31.0000		0.4537	0.5463	0.1676	6	3
33.0000		0.3025	0.6975	0.1665	7	2
35.0000	*	.	.	.	7	1
44.0000		0	1.0000	.	8	0

Figure 3: Kaplan Meier estimate of the survival function for event 1

We see that, if we let

$$P(T < t) = 1 - \hat{S}(t)$$

, then $P(T < 30)$ is .3951. However, if instead we let

$$P(T < t) = \hat{C}_1(t)$$

, then $P(T < 30)$ is .3716. While this isn't a large amount of bias, our estimate with 1-$\hat{S}(t)$ tends to get less and less accurate as more data points are added to the analysis (Putter et al[2]).

In addition to doing these basic analyses, one can also investigate the covariate impact on survival risk. This can be done in SAS by comparing two CIF curves that are generated based on different stratum (sex, age group, etc.). This is an analogous procedure to the log-rank test described in Woodward[7] for testing the hypothesis:

$$H_0 : S_1(t) = S_2(t)$$

$$H_a : S_1(t) \neq S_2(t)$$

Where the $S_i(t)$ represent different survival functions for different stratum. In SAS, we can conduct a test of the following hypothesis:

$$H_0 : C_{j1}(t) = C_{j2}(t)$$

$$H_a : C_{j1}(t) \neq C_{j2}(t)$$

which tests the equality of the cumulative incidence function for cause j across two stratified sets of subjects. This is known as Gray's test, and its details are given in Gray[8]. The following SAS code can be used in order to conduct Gray's test:

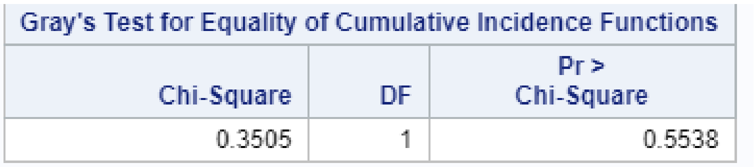
```
proc lifetest method=KM data=Example plots=CIF;

   time Time * Event (0)/failcode=1;
   strata Sex;

run;
```

In this simple example, we had 30 subjects who experienced one of two different competing events (or were censored before an event could be observed). The subjects were stratified by sex. We were interested in seeing if the cumulative incidence function for event 1 was statistically different between the two strata.

The important piece of output given by this SAS code is the test-statistic and p-value associated with Gray's test:



Gray's Test for Equality of Cumulative Incidence Functions

Chi-Square	DF	Pr > Chi-Square
0.3505	1	0.5538

Figure 4: The results of Gray's Test in SAS

Based on this p-value, we do not reject the null hypothesis. We do not have enough evidence to conclude that the cumulative incidence functions for event 1 are not statistically different across the two sexes.

5 Regression Modeling of Hazards to Assess Covariate Impact

Just as we used the Cox proportional hazards model in the case of modelling a single event (details in Woodward[7]), it is also possible to model the cause-specific hazard as a function of some covariates of interest. Putter et al[2] models the cause-specific hazard as:

$$h_j(t|\mathbf{Z}) = h_{j0}(t)e^{\beta_j^\top \mathbf{Z}}$$

, where h_j is the cause-specific hazard function for event j, h_{j0} is the baseline cause-specific hazard function of event j, $\mathbf{Z}$ is the vector of covariates, and β_j is the vector of coefficients for the covariates. Modeling this in SAS is simple enough, but the interpretation requires more effort.

For some motivation, I will apply cause-specific hazard regression to a real data set. This data consisted of 177 patients who received a stem cell transplant for acute leukemia. This transplant is done in an effort to restore healthy bone marrow in the patients. Our event of interest is relapse, and a competing risk is transplant-related death. The data set contains several covariates, including sex, age, disease (ALL vs. AML), the type of transplant, and the phase of cancer. This data set was taken from Scrucca et al[9].

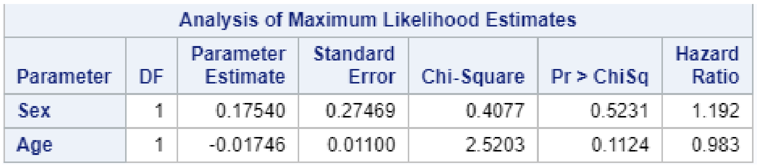
In SAS, a cause-specific hazard regression can be fit using the PROC PHREG procedure, just as if we were modelling a regular Cox Proportional Hazards Model. The key addition is that when modeling a cause-specific hazards model, the competing risk events are treated as censoring events.

In order to model the hazard of relapse as a function of Sex and Age, the following SAS code can be used after importing the dataset. A Status of 0 represents a censored observation, and a Status of 2 represents a competing event (such as transplant-related death):

```

proc phreg data=Transplant plots=survival;
model ftime * Status (0,2) = Sex Age;
run;
```

A key piece of the output produced by SAS is estimates of the coefficients for the variables, along with their standard errors:



Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Sex	1	0.17540	0.27469	0.4077	0.5231	1.192
Age	1	-0.01746	0.01100	2.5203	0.1124	0.983

Figure 5: Cause-Specific Hazard Model Parameter Estimates

A helpful interpretation of the covariate coefficients is provided by Austin and Fine[10]. Austin and Fine explain that e raised to the coefficient of a covariate equals the impact of a one unit increase in the covariate on the cause-specific hazard function. Then, in the model fit above, a one unit increase in age would account for a .983 increase in the instantaneous rate of experiencing relapse among those patients who have not experienced a relapse.

One drawback of the cause-specific hazard regression model is that it cannot assess covariate impact on the cumulative incidence function. That is, the cause-specific hazard regression model cannot measure covariate impact on overall incidence of the event of interest. In order to assess covariate impact on the cumulative incidence function, one needs to look at the *subdistribution* hazard. The subdistribution hazard for event j is defined as:

$$h_j^{sd}(t) = \lim_{\delta \to 0} \frac{P(t < T < t + \delta, D = j | T > t \cup (T < t \cap D \neq j)}{\delta}$$

The subdistribution hazard gives the instantaneous rate of failure among those who have not yet experienced an event of type j (Austin et al)[3]. The key difference between the subdistribution hazard and the cause-specific hazard is that those who experienced an event other than event j remain in the risk set. Defining the subdistribution hazard in this way allows it to relate to the cumulative incidence function of event j. That is,

$$h_j^{sd}(t) = -\frac{dlog(1 - C_j(t))}{dt}$$

Then, imposing a proportional hazards assumption on the subdistribution hazard and modelling it like the cause-specific hazard also allows one to measure covariate impact on the cumulative incidence function. A model for the subdistribution hazard function can be written as:

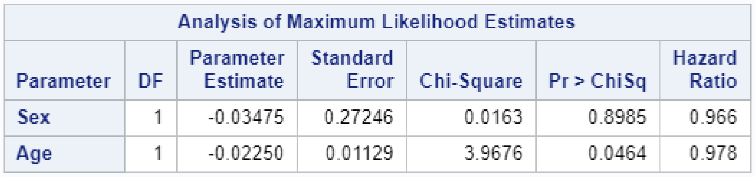
$$h_j^{sd}(t|\mathbf{Z}) = h_j^{sd0}(t)e^{\beta_j^\top \mathbf{Z}}$$

where h_j^{sd} is the subdistribution hazard function for event j, h_j^{sd0} is the baseline subdistribution hazard function of event j, $\mathbf{Z}$ is the vector of covariates, and β_j is the vector of coefficients for the covariates.

The following SAS code can be run to model the subdistribution hazard function:

```
proc phreg data=Transplant plots=CIF;
model ftime * Status (0) = Sex Age/failcode=1;
run;
```

Unlike the cause-specific hazard, modelling the subdistribution hazard does not involve treating the competing risk events as censoring events. Instead, the **failcode** option in the **model** statement allows us to specify which event's subdistribution hazard we want to model. SAS gave the following output for the parameter estimates in the model:



Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Sex	1	-0.03475	0.27246	0.0163	0.8985	0.966
Age	1	-0.02250	0.01129	3.9676	0.0464	0.978

Figure 6: Subdistribution Hazard Model Parameter Estimates

Austin and Fine[10] stress that interpreting the parameter estimates as an effect on the cumulative incidence function requires care. The sign of the parameter estimate can tell us which the direction the impact goes, but the magnitude of the rth covariate's impact on the cumulative incidence function cannot simply be interpreted as e^{β_r}. In our example, since the coefficient for the Age variable is negative, we can conclude that an increase in age results in a reduction in incidence for relapse.

6 Conclusions

The analysis of time to event data requires careful thinking, especially when competing risks are involved. Treating competing risks as censored and using a Kaplan-Meier estimate, while a simple procedure, biases results and can misrepresent risks of serious diseases. It is extremely important that statisticians working with medical data gain familiarity with the methods discussed, as they can only benefit from their implementation in analysis of time to event data. Clinical researchers would benefit greatly from being shown how to do these analyses in SAS. They are simple to implement, and with the help of a statistician, able to be interpreted accurately.

While the methods discussed in this paper are by no means exhaustive, they should be able to get anyone started with competing risk survival analysis. A future learning opportunity for myself is in competing risk analysis of multi-state models, where there are multiple transitions between a beginning state and a final state. This methodology is presented in Putter et al[2], but unfortunately I did not have time to explore this topic further. However, I have learned a great deal about survival analysis, both from this course and researching this topic.

7 References

1. Kaplan, E. L., and Paul Meier. "Nonparametric Estimation from Incomplete Observations." Journal of the American Statistical Association, vol. 53, no. 282, 1958, pp. 457–481. JSTOR, www.jstor.org/stable/2281868.

2. Putter, H., Fiocco, M. and Geskus, R.B. (2007), Tutorial in biostatistics: competing risks and multi-state models. Statist. Med., 26: 2389-2430.

doi:10.1002/sim.2712

3. Austin, Peter C et al. "Introduction to the Analysis of Survival Data in the Presence of Competing Risks." Circulation vol. 133,6 (2016): 601-9. doi:10.1161/CIRCULATIONAHA.115.017719

4. Kleinbaum, David G., and Mitchel Klein. Survival Analysis: a Self-Learning Text. Springer, 2012.

5. Gooley, T.A., Leisenring, W., Crowley, J. and Storer, B.E. (1999), Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Statist. Med., 18: 695-706. doi:10.1002/(SICI)1097-0258(19990330)18:6¡695::AID-SIM60¿3.0.CO;2-O

6. Walraven, C. and McAlister, F. (2016), Competing risk bias was common in Kaplan–Meier risk estimates published in prominent medical journals. Journal of Clinical Epidemiology, 2016-01-01, Volume 69, Pages 170-173.e8 6

7. Woodward, Mark. Epidemiology: Study Design and Data Analysis. Taylor amp; Francis, 2014.

8. Gray, Robert J. "A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk." The Annals of Statistics, vol. 16, no. 3, 1988, pp. 1141–1154. JSTOR, www.jstor.org/stable/2241622.

9. Scrucca, L., Santucci, A. Aversa, F. Regression modeling of competing risk using R: an in depth guide for clinicians. Bone Marrow Transplant 45, 1388–1395 (2010) doi:10.1038/bmt.2009.359

10. Austin, PC, Fine, JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. Statistics in Medicine. 2017; 36: 4391–4400. https://doi.org/10.1002/sim.7501

A Cumulative Incidence Function From Scratch

The following is an R script for a function that allows a user to compute cumulative incidence function estimates. The output is similar to SAS, in that each estimate will be shown for each time that an event occurred.

This is a rudimentary script that assumes that all times are different. It also assumes only right-censored data.

```

#survivalData - Data Frame of events and censoring
#eventColumn - Name of Column containing events/censoring
#timeColumn - Name of column containing time until event recorded
#censoredCode - What represents a censored observation
#Failcode - What represents the event you're interested in calculating CIF for




CIFCalc <- function(survivalData,timeColumn,eventColumn,censoredCode,failCode) {
attach(survivalData)

  sortedSurvivalData <- survivalData[order(timeColumn),]


atRisk <- nrow(sortedSurvivalData)  #risk set is entire group at time 0
survAtTime0 <- 1 #S(0) = 1
survVec <- c(survAtTime0) #initialize vector storing survival at time j
timeVec <- c(0) #initialize vector storing time
failEvents <- 0 #initialize counting number of failure events to help build vector below
failVec <- rep(0,nrow(sortedSurvivalData))
timeJVec <-vector("numeric")
CIFatJ <- vector("numeric")

for (j in 1:nrow(sortedSurvivalData)) {
  timej <- timeColumn[j]
  eventj <- eventColumn[j]

  if (eventj==censoredCode) { #censoring occurs at time j
    survVec[j+1] <- survVec[j] #survival does not   change
    atRisk <- atRisk-1
  }

  if (eventj!=failCode & eventj != censoredCode) { #competing risk occurs at time j, so calculate survival for time j

    survj <- 1-(1/atRisk)
    survVec[j+1] <- survj *survVec[j]
    atRisk <- atRisk-1



    }
  if (eventj==failCode) { #event of interest occurs at time j
    failEvents<-failEvents + 1
    hazardj <- 1/atRisk
    survj <- 1-(1/atRisk)
    survVec[j+1] <- survj *survVec[j]
    failVec[failEvents] <- hazardj*survVec[j]
    atRisk <- atRisk-1
    #Probability of failing at time j is the probability of
    #not having experienced an event by time j-1 and then experiencing the event at time j
    timeJVec[failEvents] <-timej
    CIFatJ[failEvents] <- sum(failVec)

  }




}

CIF <- sum(failVec)
timesAndCIF <- as.data.frame(cbind(timeJVec,CIFatJ))
colnames(timesAndCIF) <- c("Time","CIF")
return(timesAndCIF)
}
```