Exploring the Method for Analyzing Interval Censored Data Using Imputation in Competing Risks Model

Tiandong Li[1], Ulrike Luderer[2], Dean Baker[2],
[1]Westat, 1600 Research Blvd, Rockville, MD 20850
[2]University of California at Irvine, Irvine, CA 92697

Abstract
We consider the problem of analyzing interval censored data comparing cumulative incidence functions by demographic variables in the presence of competing risks. In this paper, we explore two methods based on imputation, the EM-type method and Multiple Imputation. Basically, we imputed the exact event time for interval censored data and take advantage of standard estimation methods for right censored data. We analyzed data from the National Children's Study to estimate cumulative risks of transition between Probability of Pregnancy Statuses and to examine the effect of major demographic variables.

Keywords: Competing risks, Cumulative Incidence Function, EM, Interval Censoring, Imputation, Cause-Specific Hazards.

1. Introduction

In this article, we explore methods based on imputation to analyze interval censored time-to event data in the presence of competing risks. Wei and Tanner (1990) described two imputation methods for analyzing interval censored data: the EM-type method and Multiple Imputation. To estimate the survival function, Pan (2000) applied a general semi-parametric Cox model on the Multiple Imputation algorithm with interval-censored data. We are extending these methods to competing risks models. We considered both methods and extend them to a semi-parametric proportional subdistribution hazards model which incorporate competing risks in the data.

The statistical analyses were applied to the data from the National Children's Study (NCS) to assess changes in women's pregnancy intention over time. The Probability of Pregnancy Groups (PPG) was created to measure pregnancy intention.

Specifically, the research question is whether time to the first transition from one PPG status to another varies by race/ethnicity, age, and marital status. Although the PPG status could change after the first transition, it is not the concern of this study. This paper focuses on the ongoing investigation of appropriate statistical methods to use to answer this question.

This paper is organized as follows. Section 2 describes the data structure from the National Children's Study (NCS). In section 3, the competing risks survival models for interval censored data is illustrated. Section 4 presents some analysis results and, finally, section 5 discusses what has been learned from this exploration and provides some directions for further research.

2. Data

This study analyzed the data from NCS, an ongoing longitudinal cohort study designed to evaluate the influence of environment exposures on child health and development in the United States. The Initial Vanguard Study of NCS is a pilot study conducted in 7 locations. Pregnant women and women who might become pregnant were invited to enroll. After the initial screening, non-pregnant and eligible women were followed up by telephone until they became pregnant or until the recruitment period ended. The data collection period was from early 2009 to Dec. 2010, by the end of which, 35,000 eligible women enrolled. About 13,000 eligible women had at least one follow-up and is the sample for the current analysis.

To assess changes in pregnancy intention over time, recruited women were classified into PPG status, which includes five categories, low, moderate, high-nontryer, high-tryer, and pregnant. These statuses are defined by women's age, the number months they had been trying to conceive, and sexual behaviors. For example, if women were "currently trying" to conceive for less than 5 months, they were placed in the high-tryer PPG; if between 5 and 11 months and age 18-34, they were placed in the moderate PPG; if 12 months or more or 5-11 months and age 35-49 they were placed in the low PPG, etc.

As known from the definition, these statuses are not sequential. For example, women starting from the low status can change to a high-tryer without going through the moderate status. Thus, from one initial status, the potential first transition can be one of four resulting statuses[1]. For example, if the initial status is low PPG, the potential resulting statuses can be moderate, high-nontryer, high-tryer and pregnant. Each woman is at risk of multiple transitions at the same time point. This situation is commonly referred to as a competing risks problem.

The PPG status was evaluated at the initial screening and each follow-up call. The frequency of the follow-up calls after the initial screening varied according to their initial PPG status. Women classified as having a moderate PPG were called every 3 months, and women classified as having a low PPG were contacted every 6 months from May 2009 to December 2009 and every 3 months thereafter. The high-tryer and high-nontryer PPG women were contacted at one, two, and four months after being so identified. Therefore, the observed time-to-transition is considered as interval censored, in that the timing of any transition within the interval is not known. In addition, the time interval between assessments is not constant, due to the screening schedule, nonresponding follow-up calls and variations of the call time in field operations. As mentioned previously, the protocol of the screening schedule for women with the low initial status changed from 6 months to 3 months at December 2009. A woman may not be located or may refuse to respond for one screening and come back to the study in the next screening, hence the time interval is wider than requested in the protocol. Furthermore, it may take several telephone calls to get a response, so that the exact screening time may vary around the protocol schedule.

[1] Pregnant is not considered as an initial status.

3. Method

To accommodate the special data features discussed above, competing risks and interval censored, we are considering imputation models to fill in the exact time within the time interval and take advantage of existing competing risks survival models. This section starts with introducing the existing competing risks models and the imputation method for interval censored data, and then demonstrates the combined models considered in the analysis of the current data.

3.1 Competing Risks

A simple model called Cause-Specific Hazards (CSH) model (e.g. Lee and Wang, 2003) has been widely used to analyze the competing risks data. This type of model utilizes regular survival models, such as Cox proportional hazards model, by treating any transitions other than the one of interest as right censored. This treatment assumes that the competing risks are independent, which means that individuals censored at time t should be representative for those still at risk at that time.

However, this assumption is not satisfied in our study. For example, a woman who changes from moderate to low PPG status is less likely to change to high-tryer, compared to the rest of the sample.

The Cause-Specific Subdistribution Hazards (CSSH) model from Fine and Gray (1999) is for competing risks models where the “independence” assumption does not hold. The hazard function is defined as below:

$$\gamma_k(t, X) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr\{t \leq T \leq t + \Delta t,\ j = k | \mathrm{T} \geq t \cup (T \leq t \cap j \neq k), x)$$

where $\gamma_k(t, x)$ represents the subdistribution hazard, which is the limiting function for the probability of transition k within the time interval Δt after time t, conditioning on no transition k before time t and transitions other than k happening before t. x represents the covariates.

In this model, competing transitions ($j \neq k$) are not treated as censored, but are still kept in the risk set. The assumption for this proportional hazards model is

$$\gamma_k(t, X) = \gamma_{k0}(t) e^{X\beta}$$

where γ_{k0} denotes the baseline hazard of the subdistribution for transition k. As summarized by Kohl and Heinze (2013), for such a model the partial likelihood of the subdistribution hazards model is defined as

$$L(\beta) = \prod_{l=1}^{r} \frac{\exp(x_l \beta)}{\sum_{i=R_l} w_{li} \exp(x_i \beta)}$$

where r is the number of all time points ($t_1 < t_2 < ... < t_r$) where transition k occured, and x_l is the covariate row vector of the subject experiencing transition k at t_l. The denominator for the likelihood at each time point is constructed based on the risk set R_l, which is defined as

$$R_l = \{i;\ t_i \geq t \cup (\ t_i \leq t \cap, j \neq k)\}$$

At each time point t_l, the set of individuals at risk R_l includes those who are still at risk of any transitions and those who have had a competing transition before time point t_l . The weights w_{li} are attached to each case in the risk set. The weight w_{li} = 1 for the cases without any transition prior to t_l. For cases with competing events before t_l, time-dependent weights are defined as

$$w_{li} = \frac{\hat{G}(t_i)}{\hat{G}(min(t_i, t_l))}$$

where $\hat{G}(\cdot)$ denotes the Kaplan-Meier estimate of the survival function of the censoring random variable, i.e., the probability of still being followed-up at t. These latter individuals have weights $w_{li} \leq 1$, which decrease with time. For the purpose of simplicity, no ties in event times are assumed.

Similar to the Cox proportional hazards model, the subdistribution hazards model requires no assumed distribution for the baseline hazard function, which is empirically estimated. SAS Macro %PSHREG is readily available to implement this model.

3.2 Interval Censored Data

Another feature of the NCS data in this study is interval censoring. Non-pregnant eligible women were periodically followed up to assess the current PPG status. If the status for one follow-up is different to that of the previous screening, a transition can be identified. But the timing of the transition within the interval between the two screenings is not known.

If the time interval is constant, a regular survival analysis can be conducted simply by using the time interval as the analysis time unit. However, this is not the case in this study as mentioned in Section 2.

Another approach to deal with interval censored data is to impute the exact transition time and then apply the existing analysis method for right censored data. This method is preferred in this study, as the theory for the subdistribution hazards model and the corresponding SAS macro are readily available.

Two algorithms are discussed in the literature, Multiple Imputation (MI) and the EM-type method. Wei and Tanner (1990) and Pan (2000) discuss the MI method for the interval censored data. Wei and Tanner (1990) briefly mention the EM-type method, but no full discussion is available. However, for both methods, the application to competing risks data is never explored.

We will consider both methods in this paper and details are provided in the next section.

3.3 Combined Models

Diagram 1 summarizes the data features of this study and the corresponding statistical models available. This paper explores the combined models to analyze interval censored data in a competing risks model through imputation methods.

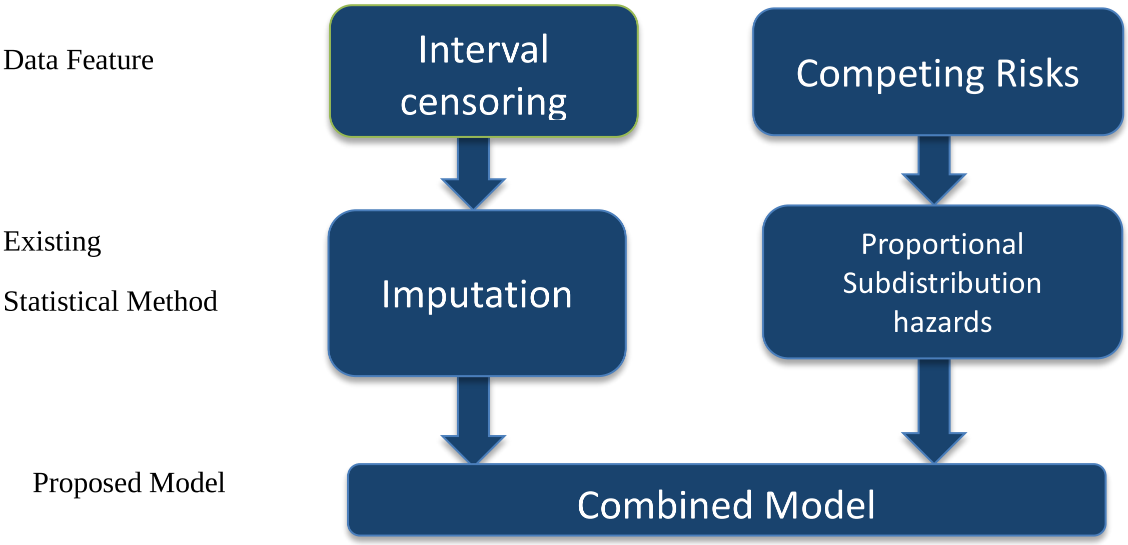


Diagram 1. Data features, existing statistical methods and the combined model

3.3.1 Combined Model using Multiple Imputation

The combined model starts with imputing an exact time in the time interval and then applies the subdistribution hazards model to estimate the parameter and the baseline survival function. As the estimated density function (or the survival function) is needed for the imputation, an iteration procedure is conducted between the imputation and the estimation of parameters and the baseline survival function. Multiple imputed values are created to account for uncertainty within the time intervals. Furthermore, for all the transitions based on the same initial status, separate survival models are fitted in sequence to impute the time for the corresponding transitions. Note that this method may not be considered as a proper MI method as it is not based on a Bayesian framework and no prior distribution is specified for the parameter. Multiple imputed values were used to account for uncertainty within time intervals, but may not have the same feature as the MI method discussed by Rubin (1987). Specifically, the MI method takes the following 6 steps:

1. Initial imputation: For all transitions, generate *j* sets of right-censored observations as initial time points for modeling, where *j*=1,…, *m*. This study used 10 imputations.

2. Model fitting: For transition outcome *k*, fit the Proportional Subdistribution Hazards model to estimate $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}(\boldsymbol{k})}^{(\boldsymbol{i})}$ and then baseline survival function $\hat{S}_{j(k)0}^{(i)}$, where *j*=1,…, *m*, *i* is the iteration number and *k* is the transition type.

3. Combine estimates: calculate estimates based on Rubin’s (1987) formula $\widehat{\boldsymbol{\beta}}_{(\boldsymbol{k})}^{(\boldsymbol{i+1})} = \frac{1}{m}\sum_j \widehat{\boldsymbol{\beta}}_{\boldsymbol{j}(\boldsymbol{k})}^{(\boldsymbol{i})}$, $\hat{S}_{(k)0}^{(i+1)} = \frac{1}{m}\sum_j \hat{S}_{j(k)0}^{(i)}$ and

$Var(\widehat{\boldsymbol{\beta}}_{(\boldsymbol{k})}^{(\boldsymbol{i+1})}) = \frac{1}{m}\sum_j Var\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{j(k)}}^{(\boldsymbol{i})}\right) + (1+\frac{1}{m})(\sum_j[\widehat{\boldsymbol{\beta}}_{\boldsymbol{j(k)}}^{(\boldsymbol{i})} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{k})}^{(\boldsymbol{i+1})}]^2)/(m - 1)$, where the first term in the variance formula represents the within imputation variance and the second term represents the between imputation variance.

4. Update Imputed values: repeat the imputation in step 1, based on estimated $\widehat{\boldsymbol{\beta}}_{(\boldsymbol{k})}^{(\boldsymbol{i+1})}$ and $\hat{S}_{(k)0}^{(i+1)}$.

5. Repeat steps 2-4 for each transition outcome

6. Repeat steps 2-5 until $\widehat{\boldsymbol{\beta}}_{(\boldsymbol{k})}^{(\boldsymbol{i})}$ converges

3.3.2 Combined Model using EM-type of Method

The EM-type method uses a similar iteration procedure as the MI method. The differences include the following features: 1) only one set of imputed data is produced, 2) initial imputations can be purposefully set to diverse values, such as interval mid-point, end point, etc., and 3) Imputed length of time is the expectation of all time points within the interval weighted by the density distribution, rather than random draws from the distribution.

Under the framework of the EM-type method, the expectation step is the imputation of the exact time using the expectation of time given estimated $\widehat{\boldsymbol{\beta}}_{(\boldsymbol{k})}^{(\boldsymbol{i})}$ and $\hat{S}_{(k)0}^{(i)}$. Then in the maximization step, $\widehat{\boldsymbol{\beta}}_{(\boldsymbol{k})}^{(\boldsymbol{i+1})}$ and $\hat{S}_{(k)0}^{(i+1)}$ are estimated using the updated imputation of time and maximum likelihood method in the subdistribtuion hazards model.

After $\widehat{\boldsymbol{\beta}}_{(\boldsymbol{k})}^{(\boldsymbol{i})}$ converges, we added one step to account for uncertainty within the time interval by drawing multiple imputed values based on the estimates from the last iteration $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{last})}$ and ${\hat{S}_0}^{(last)}$. These multiple draws are not strictly multiple imputation as Rubin (1987) discussed and may not fully account for the uncertainty within the time interval.

4. Analysis Results

Both the MI method and the EM-type method were implemented to all the transitions. As an example, the analysis results for the transition from low to moderate is presented to illustrate the assessment of the methods. In the survival models, covariates include age (<25 (reference), 25-29, 30-34, 35+), race/ethnicity (Hispanic, Non-Hispanic White (reference), NonHispanic Black, NonHispanic Asian, NonHispanic other), and marital status (currently married (reference), currently not married). For the purpose of simplicity, no time varying covariates were considered.

Table 1 shows the transition rates for the low initial status. Among the 7,504 women with a low initial status, 30.2% were observed to change to moderate in the first transition, 5.5% changed to other status and 64.3% did not have any transition until the end of the observation period.

Table 1. Transition rates for the low initial status

Category	*Transition Rate (%)*
Low to Moderate	30.2
Low to Other transitions	5.5
Right Censored	64.3
Total	100
Sample size (Initial status=LOW)	7,504

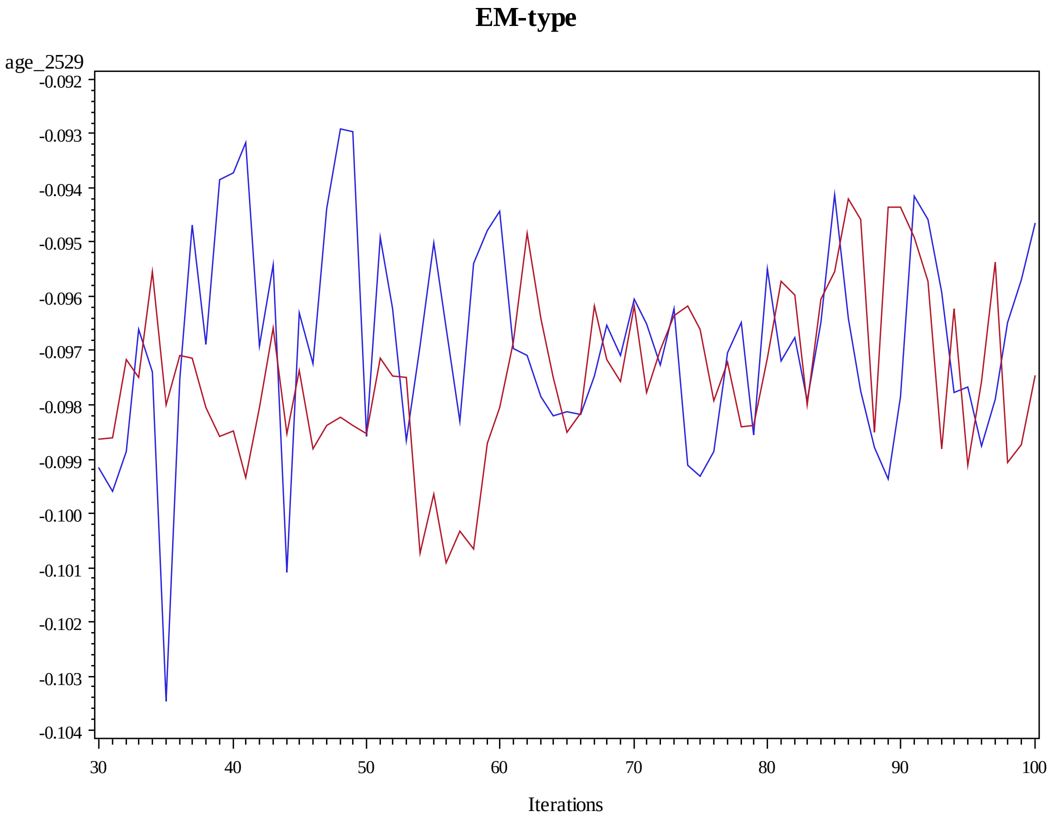
4.1 Convergence

As both models involve iteration processes, convergence was evaluated. The convergence criteria that Pan (2000) used are 1) the difference between parameter estimates in adjacent iterations less than 0.01 or 2) the number of iterations larger than 50. As multiple models are fitted for the competing risks, convergence may be harder to achieve than in Pan (2000)'s paper.

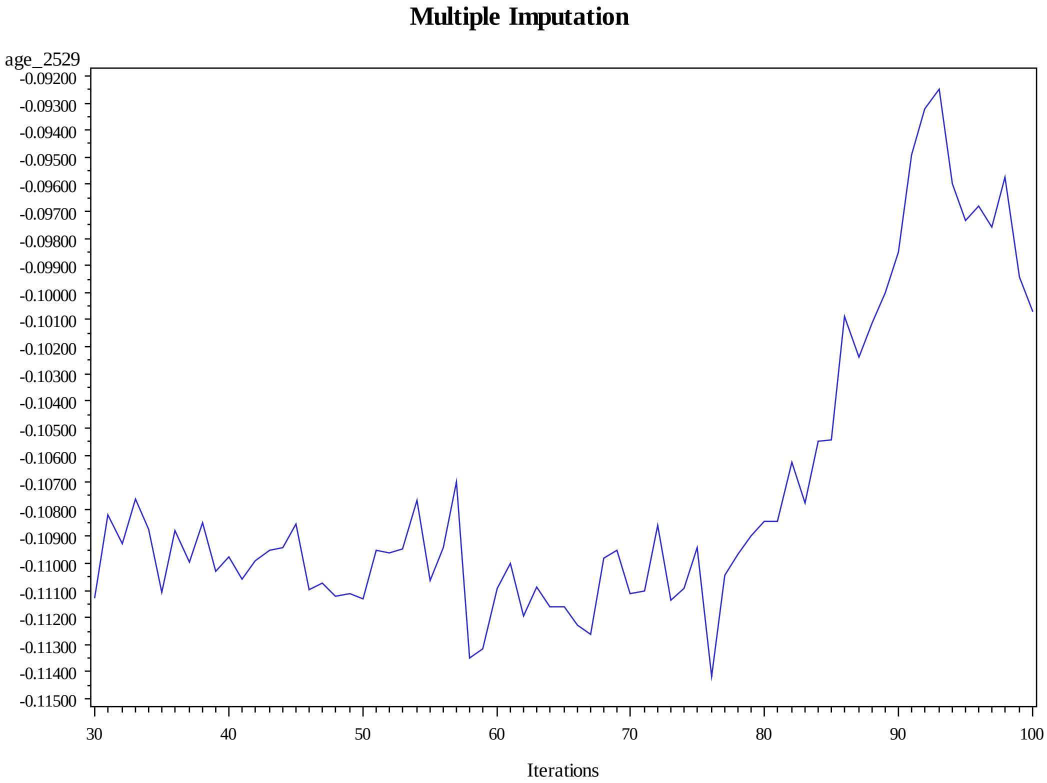
To fully evaluate our models, we ran over 100 iterations, each of which contains a run across all the transition types. The parameter estimates were then plotted to evaluate the convergence.

As an example, graph 1 shows the parameter estimates of age 25-29 vs. age <25 across iterations for the EM-type method, controlling for race/ethnicity and marital status. The x-axis shows the iteration number after the 30th iteration, where the estimates stabilize, and the y-axis shows the parameter estimates. The blue line is for the model using mid-time point in the interval as the initial length of time and the red line using the upper end time. Both lines converged to the same neighborhood, although there is still a variation between -0.093 and -0.103. We found similar patterns for other parameter estimates for EM-type method.

Graph 2 shows the same parameter estimates for the MI method. According to the graph, the parameter estimates didn't stabilize within 100 iterations. This instability is also found for other parameter estimates using the MI method.



Graph 1. Parameter estimates of age 25-29 vs. age <25 across iterations for the EM-type method



Graph 2. Parameter estimates of age 25-29 vs. age <25 across iterations for the MI method

4.2 Shrinkage for EM Estimates

In the EM-type method, the exact time is imputed with the weighted average of all the time points within the interval. Using this algorithm, the imputed values never took the values on the two ends of the intervals, and hence will never be imputed to the extreme small time values or the extreme large values in the possible range of the event time. If we compare the imputed time to event based on the EM algorithm and the imputed valued based on random draws, the distribution based on the EM algorithm shrinks to the middle. As the estimate of interest is a regression type of parameter, which is not a simple average of the times, and is in an exponential form, the estimate could be biased due to this shrinkage. Table 2 compares the parameter estimates from the last iteration of the EM-type method and those from the random imputation based on the density function estimated from the last iteration of the EM-type method. For the runs using different initial values, a shift of parameter estimates was observed and confirmed the potential bias based on the EM-type method. Similar shifts were observed for other parameter estimates as well.

Table 2. Comparing the parameter estimates from the last iteration of the EM-type method and that from the random imputation based on the density function estimated from the last iteration

Age 25-29 vs. Age <25	*Point Estimate at the last iteration*	*Point Estimate based on multiple draws*
EM-type Midpoint	-0.099	-0.109
EM-type Upper Endpoint	-0.097	-0.107

4.3 Variance Estimates using EM-type Method

After the convergence of the EM-type method, multiple draws from the time interval were conducted to form multiple imputed data sets and take into account the uncertainty within the time interval. Table 3 shows the variance estimates from this step. The estimates for age 25-29 vs. age <25 is presented for two runs using different initial time points, the midpoint and the upper end point.

As shown in the table, the between imputation variance accounts for a small portion of the overall variance, which would not be reflected by single imputation. The last column shows that the remaining variation of the estimate across iterations from the EM-type method accounts for an ignorable amount of variation compared to the overall variance.

Table 3. Variance Estimates using the EM-type method

Age 25-29 vs. Age <25	*Point Estimate*	*Between Imputation Variance (in 10^{-4})*	*Within Imputation Variance (in 10^{-4})*	*Overall Variance (in 10^{-4})*	*Variance among the last 25 iterations (in 10^{-4})*
EM-type Midpoint	-0.109	4.6	86.8	92.3	0.02
EM-type Upper Endpoint	-0.107	4.8	86.8	92.6	0.03

5. Discussion and Future Study

This paper explored two imputation algorithms to analyze interval censored data under competing risks model. Both algorithms demonstrate some drawbacks when analyzing the NCS pre-pregnancy data.

- Using the EM-type method, the point estimates and variance estimates converge using different initial values. The between imputation variance accounts for a small portion of the overall variance. However, the parameter estimates may be biased due the shrinkage of the imputed event time.
- The MI method did not converge within the 100 iterations, using 10 imputations.

For the next stage of this research, we will consider the following steps.

- The exploration will be extended to reflect some recent developments in the literature. A parametric model for interval censored data and competing risks (Hudgens, Li, and Fine, 2014) has been published and will be considered.

- Simulation studies based on a simplified data structure may be helpful to understand the convergence issue of the MI method.

- If a proportional hazards model can be used for this analysis, the extension to include time varying variables should be evaluated. When the proportional hazards assumption is not satisfied, interaction terms with time may be needed to extend the model.

References

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association,* 94, 496-509.

Hudgens, M. G., Li, C. and Fine, J. P. (2014). Parametric likelihood inference for interval censored competing risks data. *Biometrics,* 70, 1-9.

Kohl, M. and Heinze, G. (2013). %PSHREG: A SAS Macro for Proportional and Nonproportional Subdistribution Hazards Regression for Survival Analyses with Competing Risks. *Technical Report.*

Lee E. T. and Wang J. W. (2003) *Statistical Methods for Survival Data Analysis*. 3rd ed. New York, NY, John Wiley & Sons, Inc.

Pan W. (2000). A multiple imputation approach to Cox regression with interval—censored data. *Biometrics,* 56,199-203.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms. *Journal of the American Statistical Association,* 85, 699–704.