SAS® GLOBAL FORUM 2021

Paper 1117-2021

A SAS Macro to Perform Kaplan-Meier Multiple Imputation for Survival Analyses with Competing Events

Jialin Han, Sai Liu, Margaret R. Stedman, Maria E. Montez-Rath, Division of Nephrology, Department of Medicine, Stanford University School of Medicine, Palo Alto, California, USA.

ABSTRACT

Analysis of cumulative incidence functions requires special attention when there are competing events. A competing event is an event that either precludes or changes the probability of the event of interest from occurring. There are different approaches to handling competing events in survival analysis. One approach, developed by Fine and Gray (1999), for example, cannot be applied directly to stratified data or to data with time-varying covariates. For these situations, Ruan and Gray (2008) proposed an alternative approach (Kaplan-Meier multiple imputation (KMI)) which recovers the missing censoring times for those who experienced a competing event. The missing censoring times are imputed from a non-parametric multiple-imputation approach based on the Kaplan-Meier estimator. In this paper, we introduce a user-friendly SAS Macro (%SASKMI) to implement this approach. %SASKMI generates a dataset with a new event status and imputed times that can be analyzed using a regular Cox model (PROC PHREG). The output dataset follows PROC MI data standards so PROC MIANALYSIS can be used to summarize the results. To demonstrate the effectiveness of the new macro using a real data example, we compare the effect estimates, standard errors, and run times obtained after applying %SASKMI to that of a standard Fine and Gray model as implemented in PROC PHREG. We find that %SASKMI performs similarly to the Fine and Gray option in PROC PHREG but with significant run time reduction.

INTRODUCTION

Survival analysis is the analysis of time to event data such as the time to a cardiovascular event or time to death. Sometimes, in medical studies, observation of the event of interest for a patient is not possible due to the study ending or the patient failing to return for a follow-up visit prior to the event occurring. These situations that prevent researchers from observing an individual from start to finish are known as censoring events. Under the assumption that censoring is non-informative, meaning that it is not imbalanced across exposure groups, regular survival models are sufficient for the correct analysis of the data.

Statisticians must pay special attention to the existence of competing events. A competing event is a separate event that either changes the probability of the event of interest from occurring or precludes the event of interest from occurring (Bakoyannis and Touloumi, 2012; Gooley et al, 1999; Pintilie, 2007; Lau et al, 2009). For example, death is considered a competing event since it prevents the outcome of interest from happening. In survival analysis, competing events are problematic because they can cause overestimation of the probability of survival if they are ignored (Schuster et al, 2020).

Competing risks methodology has gained popularity in the medical literature because it allows for the calculation of "real-world" probabilities for the event of interest (Pintilie, 2007). It accomplishes this by keeping individuals who have had a competing event in the risk set – instead of dropping them from the risk set (as in a censoring approach) (Lau et al, 2009). Fine and Gray (1999) proposed a proportional hazards model for the subdistribution of a competing risk that allows one to make statements related to the cumulative incidence function and risks (Fine and Gray, 1999). The model can be implemented in the PHREG procedure in SAS where the competing event is specified in the MODEL statement by adding the EVENTCODE= option.

The Fine and Gray model is not suitable in many situations such as stratified analyses, time-varying covariates, weighted analysis of case-cohort samples and clustered survival data analysis (Ruan and Gray, 2008). There have been several extensions to the Fine and Gray model to accommodate clustered and stratified data (Zhou et al, 2010, 2011; Satsahiam et al, 2006), but implementation of these models is complex.

Ruan and Gray (2008) proposed a Kaplan–Meier multiple imputation approach (KMI) that recovers potential censoring times for failures from competing causes (Arthur et al, 2010). Their approach was based on the fact that standard survival regression analysis methods can be applied to "censoring complete" data where potential censoring times are observed for all the individuals who have failed from a competing event (Fine and Gray, 1999). The approach can be applied in most of the settings mentioned above and the "complete" data can be analyzed using techniques and software developed for ordinary right censored survival data.

In this paper we introduce an innovative and user-friendly SAS Macro (%SASKMI) that implements the KMI approach. To demonstrate the effectiveness of the approach, we use an example dataset to perform a comparison of the effect estimates, standard errors, and run times after application of the %SASKMI macro and the Fine and Gray model as implemented in PROC PHREG.

THE KAPLAN-MEIER MULTIPLE IMPUTATION (KMI) APPROACH

The KMI approach reformulates competing risks as a missing data problem, meaning that the potential censoring time for those people who experience the competing event is missing or unobserved. The procedure then imputes the missing data under the assumption that the Kaplan-Meier survival estimator of the conditional censoring distribution can be applied to those with competing events. Additionally, a model for the censoring distribution can be adjusted for covariates if they are found to be correlated with the censoring mechanism. In this case, the imputed censoring times are randomly drawn from the estimated conditional censoring distribution. When the largest time is an event time (main or competing event), an additional censoring time ε needs to be added to the censoring distribution. In this macro, we specify ε as 1, meaning that we will add one day to the largest observed event time. As a result, the imputed censoring time will be larger than the observed time for an individual with the competing event.

Each observed competing event time is then replaced with the imputed potential censoring time rendering censoring-complete data (Fine and Gray, 1999). This process is repeated to form *k* imputed data sets. After the imputation, we can apply a regular Cox proportional hazards regression model to estimate the effect of the variable of interest within each imputation *k*. The final estimate is then obtained by combining the coefficients from each of *the k* analyses using the Rubin's rule (Little and Rubin, 1987). See **Figure 1**, for a diagram of the process of estimating one coefficient using 5 imputed datasets. The process can be expanded to the estimation of multiple coefficients.

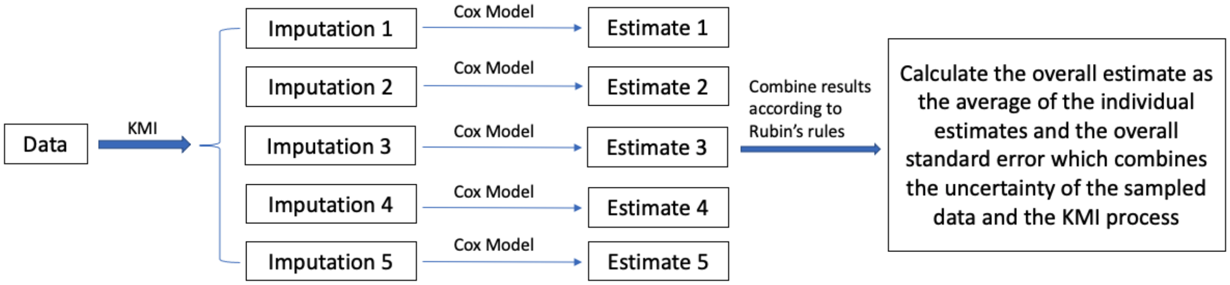

Data
KMI
Imputation 1
Imputation 2
Imputation 3
Imputation 4
Imputation 5
Cox Model
Cox Model
Cox Model
Cox Model
Cox Model
Estimate 1
Estimate 2
Estimate 3
Estimate 4
Estimate 5
Combine results according to Rubin's rules
Calculate the overall estimate as the average of the individual estimates and the overall standard error which combines the uncertainty of the sampled data and the KMI process

Figure 1. Example flow chart of the implementation of the KMI approach with 5 imputations.

THE %SASKMI MACRO

The %SASKMI macro implements the KMI approach to impute potential censoring times for those individuals with a competing event. The macro is called with the parameters included in **Table 1**. We recommend including covariates, such as age, sex and race, that potentially correlate with the censoring distribution. When covariates are specified, censoring times are imputed from a proportional hazard model, otherwise, the censoring times are imputed from the KM estimator of the observed censoring distribution. The user specifies the code number for the main event and the censoring event, and all other codes will be considered as competing events.

Macro parameters	Required/ Optional	Definition
DATA=	Required	SAS input data set name.
EVENT=	Required	The variable name of the outcome of interest. This is usually coded as 0, 1, or 2 where 0 denotes a censoring event, 1 denotes the outcome of interest and 2 denotes the competing event. The macro allows for more than one competing event.
EVENTCODE=	Optional	The code number for the outcome of interest in the EVENT variable, default is 1.
CENSCODE=	Optional	The code number for a censored outcome in the EVENT variable, default is 0.
TIME=	Required	The time to event variable.
ADJVAR=	Optional	List of covariates to be used to model the censoring distribution.
CLASS=	Required if category variable exist in ADJVAR	All categorical covariates from ADJVAR must be included here.
NIMP=	Required	The number of imputations.
SEED=	Optional	The random control seed. The default is 123. Since the KMI approach is based on randomly drawing censoring times from the estimated conditional censoring distribution, we recommend specifying a SEED in order to replicate the results.

OUT=	Optional	The name of the output dataset. The default is “OUT”.

Table 2. Parameters for Macro %SASKMI

The macro generates an output dataset with three new variables: NEWTIME, NEWEVENT and _IMPUTATION_. NEWTIME contains the imputed time. NEWEVENT is coded so that individuals with the competing event are now coded as censored. _IMPUTATION_ contains the imputation number. The macro also generates two figures, one containing the model information and the other providing summary statistics for event time, censoring time and each imputed time.

The macro is available upon request from the author (Jialin Han) and is posted to GitHub website (https://github.com/hjlhanjialin/SASKMI). Once downloaded, use the %INCLUDE statement to specify the directory where the “SASKMI” macro is stored.

```
%include "Directory/SASKMI.SAS";
```

EXAMPLE

DATA

As an example, we used a random sample of 30,000 individuals from a study investigating the effect of glomerular disease subtypes on cardiovascular death in patients on dialysis (O’Shaughnessy et al, 2018). The original data was obtained from the United States Renal Data System (USRDS), a national registry of US patients with end stage renal disease. The data contain information on 62,874 adult patients with ESRD attributed to one of six glomerular disease subtypes (IgA nephropathy (IgAN), focal segmental glomerulosclerosis (FSGS), membranous nephropathy (MN), membranoproliferative glomerulonephritis (MPGN), lupus nephritis (LN), vasculitis (VAS)) who initiated dialysis in the continental US between January 1st, 1997 and October 1st, 2014.

The following variables are included in the dataset:

- ID: The system generated identification.
- CV_DEATH: Outcome, cardiovascular death at 5 years. Patients are censored for loss insurance coverage or end-of-study (31 December 2014). Kidney transplantation and non-cardiovascular death were treated as competing events. The variable is coded as 0 (censored event), 1 (cardiovascular death) or 2 (competing event).
- TIME2CV_DEATH: Time to cardiovascular death in days.
- GN: Main exposure variable indicating the six glomerular disease subtypes.
- AGE: Baseline age.
- RACE: Race.
- REGION: Region.

The following code inputs the data:

```
Proc format;
 value race 1 = “White” 2 = “Black” 3 = “Asian” 4 = “Other”;
 value region 1 = “Northeast” 2 = “Midwest” 3 = “South” 4 = “West”;
 value GN 1=”FSGS” 2=”MN” 3=”MPGN” 4=”VAS” 5=”LN” 6=”IgAN”
;
run;
Data GNstudy;
input ID CV_DEATH TIME2CV_DEATH AGE RACE REGION GN;
 datalines;
1      2      285    66     1      2      5
```

```
2      0      1826   57     3      1      1
3      1      435    46     4      4      6
4      2      132    51     1      3      1
5      0      1826   56     2      3      1
6      2      36     71     2      4      6
7      0      1826   32     1      1      1
8      2      171    75     2      2      2
9      0      1826   62     1      2      1
10     0      1826   46     1      4      5
…… more lines ……
30000 0      298    45     3      4      6
;
run;
```

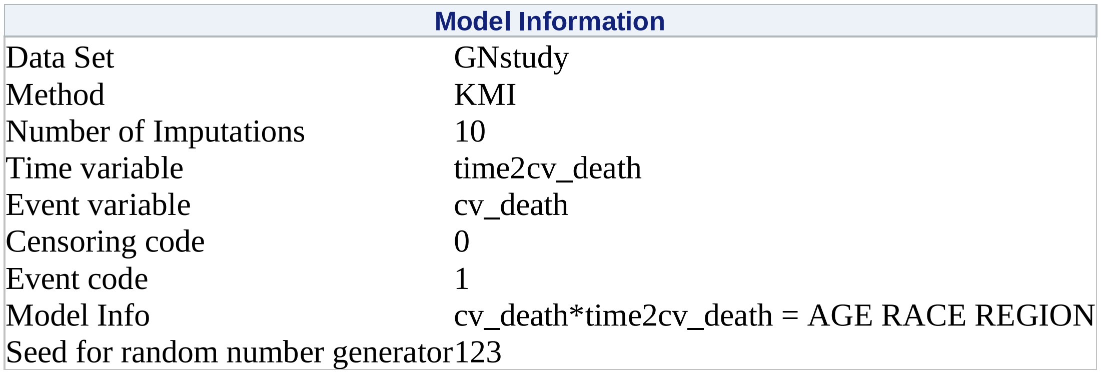
MACRO OUTPUT

The following code runs the %SASKMI macro with 10 imputations. The KMI model includes the patient characteristics: AGE, RACE, and REGION.

```
%SASKMI(
      DATA          = Gnstudy,
      ADJVAR        = AGE RACE REGION,
      CLASS         = RACE REGION,
      EVENT         = CV_DEATH,
      EVENTCODE     = 1,
      CENSCODE      = 0,
      TIME          = TIME2CV_DEATH,
      NIMP          = 10,
      SEED          = 123,
      OUT           = OUT
);
```

Once the macro is finished, it outputs a dataset named “OUT” with the three new variables: NEWTIME, NEWEVENT and _IMPUTATION_.

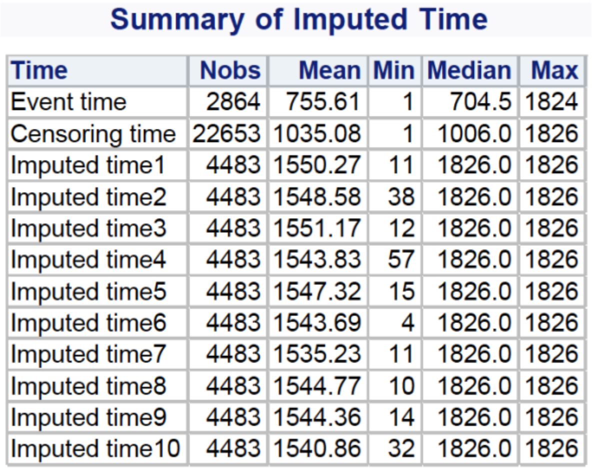
The macro outputs a summary of the model specifications as shown in **Figure 2**.



Model Information	
Data Set	GNstudy
Method	KMI
Number of Imputations	10
Time variable	time2cv_death
Event variable	cv_death
Censoring code	0
Event code	1
Model Info	cv_death*time2cv_death = AGE RACE REGION
Seed for random number generator	123

Figure 2. KMI Model Information

Figure 3 shows the output generated summarizing the time to event variable. It includes summary statistics among those with the main event of cardiovascular death (Event time), among those originally censored (Censoring time) and among those with the competing event for each imputation data set (Imputation time 1-10). Note that the imputed times are stable, with a median imputed time to event of 1826 days in all imputed data sets.



Summary of Imputed Time

Time	Nobs	Mean	Min	Median	Max
Event time	2864	755.61	1	704.5	1824
Censoring time	22653	1035.08	1	1006.0	1826
Imputed time1	4483	1550.27	11	1826.0	1826
Imputed time2	4483	1548.58	38	1826.0	1826
Imputed time3	4483	1551.17	12	1826.0	1826
Imputed time4	4483	1543.83	57	1826.0	1826
Imputed time5	4483	1547.32	15	1826.0	1826
Imputed time6	4483	1543.69	4	1826.0	1826
Imputed time7	4483	1535.23	11	1826.0	1826
Imputed time8	4483	1544.77	10	1826.0	1826
Imputed time9	4483	1544.36	14	1826.0	1826
Imputed time10	4483	1540.86	32	1826.0	1826

Figure 3. Summary of Imputation Time

ESTIMATION OF THE FINAL EFFECT USING %SASKMI

Using the NEWTIME and NEWEVENT variables created in the %SASKMI output dataset, we can then perform Cox proportional hazards regression to estimate the effect of the main exposure on the outcome. The first step is to run PROC PHREG to fit the model to each imputed dataset. The final estimate is obtained by running PROC MIANALYZE. The following statements demonstrate this process.

```
/*
Using OUT data with NEWTIME NEWEVENT
Run regular Cox model on each imputed dataset
*/
PROC PHREG DATA=OUT;
 BY _imputation_;
 CLASS RACE REGION GN(REF = “IgAN”)/PARAM=REFERENCE;
 MODEL NEWTIME*NEWEVENT(0) = GN AGE RACE REGION;
 ODS OUTPUT ParameterEstimates = EST;
RUN;
/*Process the output data to create a new variable with labels that refer to
the various levels of the exposure variable*/
DATA EST1;
 SET EST;
 Var = Parameter||Label;
RUN;
PROC SORT DATA=est1;
 BY var;
RUN;
/*Combine results to get overall estimate*/
PROC MIANALYZE DATA=EST1;
 BY var;
 MODELEFFECTS estimate;
 STDERR StdErr;
 ODS OUTPUT ParameterEstimates = est_out;
run;
```

ESTIMATION OF THE EFFECT USING THE FINE AND GRAY MODEL

For comparison, we also applied the Fine and Gray model as implemented in PROC PHREG with the EVENTCODE option. The estimated subdistribution coefficients are provided directly in the PROC PHREG output.

```
Proc phreg data=GNStudy;
CLASS RACE REGION GN(REF = “IgAN”)/PARAM=REFERENCE;
MODEL TIME2CV_DEATH*CV_DEATH(0) = GN AGE RACE REGION/eventcode=1;
run;
```

RESULTS

We found the estimates between the two models to be very similar (**Table 2**). For example, the estimated subdistribution coefficient (standard error) for LN vs IgAN is 0.25 (0.09) in the Fine and Gray model and 0.24 (0.09) when using the KMI approach. However, we found that the %SASKMI macro uses a fraction (4%) of the time it takes to run the Fine and Gray model (8.1 seconds vs. 3.1 minutes, respectively).

	Fine and Gray	KMI
	(PROC PHREG)	(*%SASKMI*)
Total run time	3.1 minutes	8.1 seconds
Exposure coefficient estimate (SE)		
FSGS vs IgAN	0.32(0.07)	0.32(0.07)
LN vs IgAN	0.24(0.09)	0.25(0.09)
MN vs IgAN	0.24(0.10)	0.24(0.10)
MPGN vs IgAN	-0.07(0.09)	-0.07(0.09)
VAS vs IgAN	0.67(0.08)	0.66(0.08)

Table 2. Total run time and coefficient estimate (standard error) comparison between the Fine and Gray model in PROC PHREG and the KMI approach in %*SASKMI.*

CONCLUSION

Our macro successfully implements the KMI approach in SAS and provides a more flexible and efficient tool to model time to event data in the presence of competing events when compared to the application of the Fine and Gray model. Parameter estimates and standard errors from our macro are equivalent to those obtained using a Fine and Gray model. In addition, our macro significantly reduces the run time in comparison with fitting the Fine and Gray model in PROC PHREG.

REFERENCES

Bakoyannis, G. and Touloumi, G., 2012. “*Practical methods for competing risks data: A review.*” Statistical methods in medical research, 21(3), pp.257-272.

Gooley, TA., Leisenring, W., Crowley, J., Storer, BE. 1999. “*Estimation of failure probabilities in the presence of competing risks: new representations of old estimators.*” Statistics in Medicine; 18:695–706.

Pintilie, M. 2007. “*Analyzing and interpreting competing risk data.*” Statistics in Medicine; 26(6), pp.1360-1367.

Lau, B., Cole, S.R. and Gange, S.J. 2009. “*Competing risk regression models for epidemiologic data.*” American journal of epidemiology, kwp107.

Schuster, NA., Hoogendijk, EO., Kok, A.L.A., Twisk, JWR., Heymans, MW. 2020. “*Ignoring competing events in the analysis of survival data may lead to biased results: a* *nonmathematical illustration of competing risk analysis."* Journal of Clinical Epidemiology, Volume 122, Pages 42-48.

Fine, JP., Gray, RJ. 1999. *"A proportional hazards model for the subdistribution of a competing risk."* Journal of the American Statistical Association. 1999; 94:496–509.

Zhou, B., Fine, J., Latouche, A., Labopin, M. 2011. *"Competing risks regression for clustered data."* Biostatistics (Oxford, England). 13. 371-83. 10.1093/biostatistics/kxr032.

Zhou, B., Latouche, A., Rocha, V., Fine, Jason. 2010. *"Competing Risks Regression for Stratified Data."* Biometrics. 67. 661-70. 10.1111/j.1541-0420.2010.01493.x.

Ruan, PK., Gray, RJ. (2008). *"Analyses of cumulative incidence functions via non-parametric multiple imputation."* Statistics in Medicine: volumn 27, Issue 27, Pages 5709-5724.

Satsahian, S., Resche-Rigon, M., Chevret, S., Porcher, R. 2006. *"Analysing multicentre competing risks data with a mixed proportional hazards model for the subdistribution."* Statistics in Medicine 2006; 25:4267–4278.

Arthur, Allignol., Jan, Beyersmann. 2010 *"Software for fitting nonstandard proportional subdistribution hazards models."* Biostatistics, Volume 11, Issue 4, October 2010, Pages 674–675.

Little, RJA., Rubin, DB. 1987. Statistical Analysis with Missing Data. New York: John Wiley & Sons.

O'Shaughnessy, MM., Liu, S., Montez-Rath, ME., Lafayette, RA., Winkelmayer, WC. *"Cause of kidney disease and cardiovascular events in a national cohort of US patients with end-stage renal disease on dialysis: a retrospective analysis."* Eur Heart J. 2019 Mar 14;40(11):887-898.

ACKNOWLEDGMENTS

We would like to thank Dr. Michelle M O'Shaughnessy for allowing us to use her study cohort in our example.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jialin Han
Stanford University
Colin-Jialin.han@stanford.edu or hjlhanjialin@gmail.com