

Paper 344-2012

Analyzing Survival Data with Competing Risks Using SAS® Software

Guixian Lin, Ying So, Gordon Johnston, SAS Institute Inc., Cary NC

ABSTRACT

Competing risks arise in studies when subjects are exposed to more than one cause of failure and failure due to one cause excludes failure due to other causes. For example, the cause of failure in bone marrow transplantation can be relapse, death in remission, or other causes. The standard Kaplan-Meier method for survival analysis does not yield valid results for a particular risk if failures from other causes are treated as censored. A useful quantity is the cumulative incidence function, which is the cumulative probability of failure from a specific cause over time. The SAS macro %CIF implements appropriate nonparametric methods for estimating cumulative incidence functions. The macro also implements Gray's method (Gray 1988) for testing differences between these functions in multiple groups. This paper discusses these methods and illustrates the macro.

INTRODUCTION

Standard survival analysis focuses on failure-time data that have a single type of failure. Competing risks arise when a failure can result from one of several causes and one cause precludes the others (Marubini and Valsecchi 1995; Andersen et al. 1995; Klein and Moeschberger 2003; Pintilie 2006; Klein 2010). Examples of competing risks can occur in many fields, including public health, reliability analysis, and demography. For example, in medical studies of treatment effect on death from cancer or other causes, death from other causes must be taken into account.

For competing risks data, a useful quantity is the cumulative incidence function, which is the cumulative probability of failure from a specific cause over time. A common mistake is to treat failures due to other causes as censored events and estimate the cumulative incidence function as the complement of the Kaplan-Meier estimate of the survival function. This leads to biased estimation unless it can be assumed that competing risks are independent. This approach also leads to the inappropriate use of the log-rank test for the equality of survival functions across treatment groups.

Another issue in competing risks analysis is the use of regression analysis to assess the effect of covariates on failure time. Regression methods can be based directly on the cause-specific hazard function or on the cumulative incidence function. For example, you can use the regression method to model the cause-specific hazard function under the proportional hazards assumption, and you can carry this out with the PHREG procedure by treating occurrences from other causes as censored observations. Fine and Gray (1999) and Klein and Andersen (2005) proposed methods that directly model the effect of covariates on the cumulative incidence functions. A discussion of these methods is outside the scope of this paper.

This paper describes the SAS macro %CIF, which implements nonparametric methods for estimating cumulative incidence functions. You can also use the macro to test for equality of cumulative incidence functions across treatment groups.

The paper is organized as follows: The next section describes the data set used as an example in this paper. The subsequent section describes estimation of cumulative incidence functions. Following those sections is a section that illustrates how to use the %CIF macro to estimate the cumulative incidence functions. Then two sections present a test for equality of cumulative incidence functions among groups and the use of the macro for this purpose. The appendix provides the syntax for the macro.

BONE MARROW TRANSPLANT DATA

To illustrate the methods and the application of the macro, this paper uses the bone marrow transplant (BMT) data studied by Klein and Moeschberger (2003). In this study, failure occurs from one of two competing risks: leukemia relapses or death in remission.

There were 137 patients in the study. Among them, 54 patients were diagnosed with acute myelocytic leukemia low-risk remission (AML low-risk), 15 patients were diagnosed with acute myelocytic leukemia high-risk remission (AML high-risk), and 38 patients were diagnosed with acute lymphoblastic leukemia (ALL).

The following SAS statements create a data set BMT that is used to illustrate the methods in this paper:

```
proc format;  
  value grpLabel 1='ALL' 2='AML low-risk' 3='AML high-risk';  
run;
```

```

data BMT;
  input Diagnosis Ftime Status@@;
  label Ftime="Days";
  format Diagnosis grpLabel.;
datalines;
1      2081      0      1      1602      0
1      1496      0      1      1462      0
1      1433      0      1      1377      0
1      1330      0      1      996      0
1      226      0      1      1199      0
1      1111      0      1      530      0
1      1182      0      1      1167      0
1      418      2      1      383      1

... more lines ...

3      625      1      3      48      1
3      273      1      3      63      2
3      76      1      3      113      1
3      363      2      1      383      1
;
run;

```

The data set BMT contains the following variables:

- Ftime: the follow-up time after bone marrow transplant in days
- Status: an indicator with the value 0 for censoring, 1 for relapse, or 2 for death during remission
- Diagnosis: AML low-risk, AML high-risk, or ALL.

For example, the first observation represents a patient diagnosed as ALL and has a failure time that is censored at 2,081 days.

NONPARAMETRIC ESTIMATION OF CUMULATIVE INCIDENCE FUNCTIONS

Let T and C denote the failure and censoring times, respectively. For data with K competing risks, the pair (X, δ) is observed, where $X = \min(T, C)$ and $\delta = 0, \dots, K$ is an indicator with values 0 for censoring and other values that designate specific failure causes.

For data with competing risks, two useful summaries are the cause-specific hazard function and the cumulative incidence function:

- The cause-specific hazard function $h_k(t)$ at time t is the instantaneous rate of failure due to cause k conditional on survival until time t or later. It is defined as

$$h_k(t) = \lim_{\Delta t \downarrow 0} \frac{P(t < T < t + \Delta t, \delta = k | T \geq t)}{\Delta t}, \quad k = 1, 2, \dots, K$$

- The cumulative incidence function, denoted by $F_k(t)$, is the probability of failure due to cause k prior to time t . It is defined as

$$F_k(t) = P(T \leq t, \delta = k), \quad k = 1, 2, \dots, K$$

In the literature, $F_k(t)$ is also referred to as the subdistribution function because it is not a true probability distribution.

It follows from these definitions that

$$F_k(t) = \int_0^t S(u) h_k(u) du = \int_0^t S(u) dH_k(u), \quad k = 1, 2, \dots, K$$

where $H_k(t) = \int_0^t h_k(u) du$ is referred to as the cause-specific cumulative hazard function and $S(t) = P(T > t) = e^{-\sum_{k=1}^K H_k(t)}$ is the overall survival function, which is the probability of surviving beyond time t .

Let $t_1 < t_2 < \dots < t_j < \dots < t_n$ in a study be the distinct failure times from any cause. Similar to the cumulative hazard function in standard survival analysis, the cumulative hazard function $H_k(t)$ for cause k can be estimated by

the Nelson-Aalen estimator,

$$\hat{H}_k(t) = \sum_{t_j \leq t} \frac{d_{kj}}{n_j}$$

where d_{kj} is the number of failures from cause k and n_j is the number of subjects at risk at time t_j . The overall survival function $S(t)$ can be estimated from the Kaplan-Meier estimator $\hat{S}(t)$. After plugging these two estimators into the equation for $F_k(t)$, the cumulative incidence function for cause k (Marubini and Valsecchi 1995) can be estimated as

$$\hat{F}_k(t) = \sum_{t_j \leq t} \hat{S}(t_{j-1}) \frac{d_{kj}}{n_j}$$

Note that $\hat{F}_k(t)$ is a step function that changes only at failure times t_j where d_{kj} is not 0.

For survival data with only one type of failure, there is a one-to-one relationship between the hazard function and the survival function. This is not the case in the presence of competing risks. As you can see from the equation for $F_k(t)$, this function is related to all of the cause-specific hazard functions through $S(t)$. The cumulative incidence function is not defined solely by $h_k(t)$.

A common mistake in the analysis of competing risks is to compute a Kaplan-Meier estimator for failure from cause k by treating failures from other causes as censored events. This estimator is $\hat{S}_k(t) = \prod_{t_j \leq t} (1 - \frac{d_{kj}}{n_j})$. Although the complement of the Kaplan-Meier estimator consistently estimates $\int_0^t h_k(u) e^{-H_k(u)} du$ (Martinussen and Scheike 2006), the quantity has no simple interpretation as a probability. Consequently, $1 - \hat{S}_k(t)$ is not an appropriate estimator of the cumulative incidence function without assuming independence of the risks. This independence is not verifiable (Tsiatis 1975).

To obtain pointwise confidence intervals for the cumulative incidence function, you need to know the variance of the estimated cumulative incidence function. By using the delta method (Marubini and Valsecchi 1995; Hosmer, Lemeshow, and May 2008) you can derive the variance as

$$\begin{aligned} Var(\hat{F}_k(t)) = \sum_{t_j \leq t} \left\{ [\hat{F}_k(t) - \hat{F}_k(t_j)]^2 \frac{d_j}{n_j(n_j - d_j)} \right. \\ \left. + [\hat{S}(t_{j-1})]^2 \frac{n_j - d_{kj}}{n_j^3} \right. \\ \left. - 2[\hat{F}_k(t) - \hat{F}_k(t_j)][\hat{S}(t_{j-1})] \frac{d_{kj}}{n_j^2} \right\} \end{aligned}$$

where $d_j = \sum_{k=1}^K d_{kj}$. Confidence intervals can be computed directly from $\hat{F}_k(t)$ and the variance $Var(\hat{F}_k(t))$. However, this can lead to confidence interval limits that are outside the range 0 to 1. A better way to obtain confidence intervals is based on the log(-log) transformation of the estimated cumulative incidence function. Using the delta method, the standard error of $\log[-\log(\hat{F}_k(t))]$ is

$$SE(\log[-\log(\hat{F}_k(t))]) = \frac{SE(\hat{F}_k(t))}{\hat{F}_k(t) |\log \hat{F}_k(t)|}$$

The pointwise confidence interval of $\log[-\log(\hat{F}_k(t))]$ is (L, U) , where

$$\begin{aligned} L &= \log[-\log(\hat{F}_k(t))] - z_{\alpha/2} (SE(\log[-\log(\hat{F}_k(t))])) \\ U &= \log[-\log(\hat{F}_k(t))] + z_{\alpha/2} (SE(\log[-\log(\hat{F}_k(t))])) \end{aligned}$$

and z_α is the $100(1 - \alpha)$ percentile of the standard normal distribution.

By back-transformation, the pointwise confidence interval for $F_k(t)$ is

$$(e^{-e^L}, e^{-e^U})$$

See Hosmer, Lemeshow, and May (2008) for details. The %CIF macro provides an alternate method for computing the variance based on counting process theory. The formula (Andersen et al. 1995) is complicated and is not given here.

USING THE %CIF MACRO TO ESTIMATE CUMULATIVE INCIDENCE FUNCTIONS

The following call to the %CIF macro estimates cumulative incidence functions for diagnosis groups in the BMD data, and the results are shown in Figure 1 and Figure 2.

```
%CIF (DATA=bmt, TIME=Ftime, STATUS=Status, EVENT=1, CENSORED=0, GROUP=Diagnosis, OPTIONS=NOTEST) ;
```

The parameters for the %CIF macro are as follows:

- DATA=BMT specifies the data set to analyze.
- TIME=Ftime specifies the time variable.
- STATUS=Status specifies the variable that contains the codes of the follow-up time.
- EVENT=1 specifies the value for the event of interest. In this example, the event of interest is relapse.
- CENSORED=0 specifies the value used to indicate censoring. Values that are specified with the STATUS= parameter rather than those specified with the EVENT= and CENSORED= parameters are for competing risks.
- GROUP=Diagnosis requests a cumulative incidence curve for each diagnosis group (AML low-risk, AML high-risk, and ALL).
- OPTIONS=NOTEST suppresses the test for equality of cumulative incidence functions among groups.

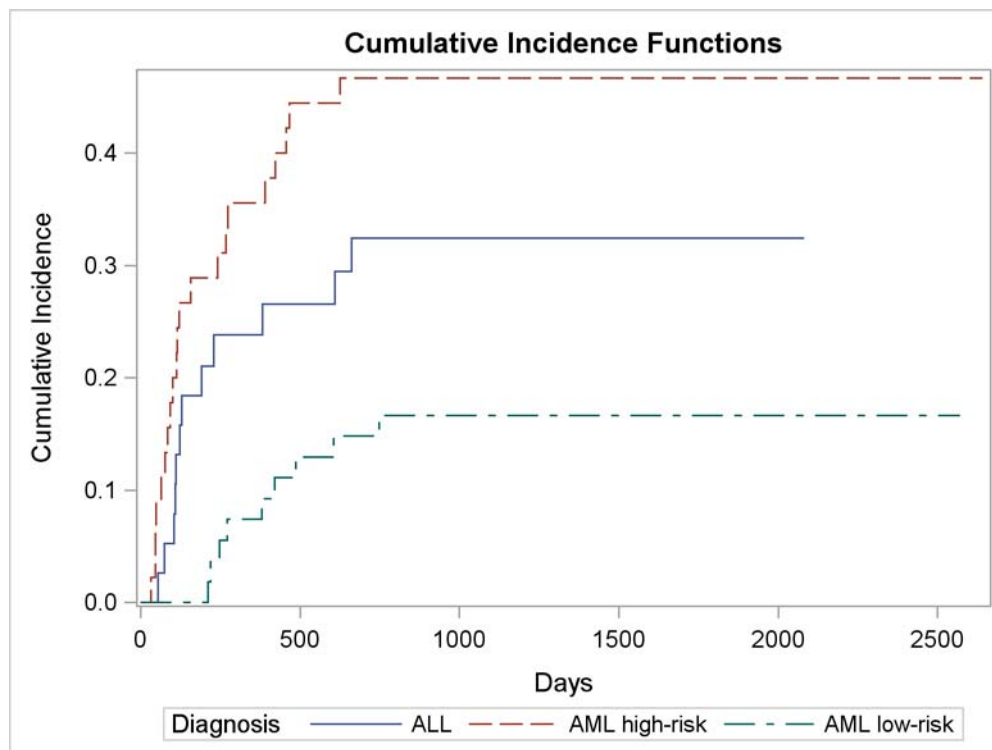
Figure 1 displays estimates of the cumulative incidence functions of relapse for each diagnosis group, the standard errors, and the 95% confidence intervals. For example, for patients in the ALL group, the estimated cumulative incidence at 662 days is 0.32429 with standard error 0.079068. The 95% confidence interval is (0.17882, 0.47869), which is computed by using the pointwise confidence interval for $F_k(t)$. You can specify other options to customize the analysis (see the Appendix). For example, you can specify the level for the confidence interval with the ALPHA= parameter.

Figure 1 Estimated Cumulative Incidence Functions of Relapse for Diagnosis Groups

Cumulative Incidence Estimates with 95% Confidence Limits					
Diagnosis	Days	CIF	StdErr	Lower 95% Limits	Upper 95% Limits
ALL	0	0.00000	0.000000	0.00000	0.00000
ALL	55	0.02632	0.026325	0.00196	0.11980
ALL	74	0.05263	0.036730	0.00923	0.15718
ALL	104	0.07895	0.044372	0.01988	0.19297
ALL	109	0.10526	0.050521	0.03274	0.22709
ALL	110	0.13158	0.055669	0.04724	0.25988
ALL	122	0.15789	0.060072	0.06300	0.29160
ALL	129	0.18421	0.063880	0.07981	0.32240
ALL	192	0.21053	0.067208	0.09748	0.35246
ALL	230	0.23799	0.070476	0.11639	0.38361
ALL	383	0.26545	0.073331	0.13601	0.41404
ALL	609	0.29487	0.076443	0.15702	0.44683
ALL	662	0.32429	0.079068	0.17882	0.47869
AML high-risk	0	0.00000	0.000000	0.00000	0.00000
AML high-risk	32	0.02222	0.022234	0.00171	0.10289
AML high-risk	47	0.06667	0.037632	0.01700	0.16533
AML high-risk	48	0.08889	0.042941	0.02790	0.19461
AML high-risk	64	0.11111	0.047433	0.04013	0.22282
AML high-risk	76	0.13333	0.051323	0.05340	0.25017
AML high-risk	84	0.15556	0.054739	0.06750	0.27681
AML high-risk	93	0.17778	0.057763	0.08231	0.30282
AML high-risk	100	0.20000	0.060453	0.09772	0.32831
AML high-risk	113	0.22222	0.062856	0.11367	0.35332
AML high-risk	115	0.24444	0.065001	0.13010	0.37790
AML high-risk	120	0.26667	0.066911	0.14697	0.40208
AML high-risk	157	0.28889	0.068607	0.16424	0.42590
AML high-risk	242	0.31111	0.070129	0.18184	0.44943
... more lines ...					
AML low-risk	0	0.00000	0.000000	0.00000	0.00000
AML low-risk	211	0.01852	0.018545	0.00147	0.08727
AML low-risk	219	0.03704	0.025980	0.00673	0.11398
AML low-risk	248	0.05556	0.031516	0.01432	0.13982
AML low-risk	272	0.07407	0.036037	0.02342	0.16459
AML low-risk	381	0.09259	0.039894	0.03360	0.18850
AML low-risk	421	0.11111	0.043266	0.04461	0.21172
AML low-risk	486	0.12963	0.046258	0.05630	0.23438
AML low-risk	606	0.14815	0.048938	0.06855	0.25655
AML low-risk	748	0.16667	0.051359	0.08127	0.27830

Figure 2 displays a plot of the estimated cumulative incidence functions for relapse for the three groups (AML low-risk, AML high-risk, and ALL). There is a clear separation of these curves: patients with AML low-risk have the lowest probability of relapse, and patients with AML high-risk have the highest probability of relapse.

Figure 2 Estimated Cumulative Incidence Functions for Diagnosis Groups



To test whether the difference are significant, you can apply Gray's method (Gray 1988), which is described in the next section.

COMPARISON OF CUMULATIVE INCIDENCE FUNCTIONS AMONG GROUPS

In medical studies, it is often of interest to test whether there is any group effect (for example, any effect due to treatment or disease) on survival. The log-rank test is widely used for this purpose in standard survival analysis. You can use the log-rank test based on hazard functions to test whether the hazard functions or survival functions are equal because there is a one-to-one relationship between the hazard function and the survival function.

However, when competing risks are present, the cumulative incidence function of an event is not defined solely by its corresponding cause-specific hazard. So the effects of treatment on the cause-specific hazard and its corresponding cumulative incidence functions can be quite different (Gray 1988; Pepe 1991; Fine and Gray 1999). The following simple example demonstrates this complication.

Suppose there are two groups (A and B) and two event types (1 and 2). Let $h_i^A(t)$ and $h_i^B(t)$ denote the cause-specific hazard functions of event $i=1, 2$ for groups A and B, respectively.

Suppose that the hazards are constant with $h_1^A(t) = h_1^B(t) = 1$, $h_2^A(t) = 1$, and $h_2^B(t) = 3$. Note that the cause-specific hazards for Event 1 are the same for Groups A and B. Let $F_1^A(t)$ and $F_1^B(t)$ denote the cumulative incidence functions of Event 1 for Groups A and B, respectively. It follows from the equation for $F_k(t)$ that

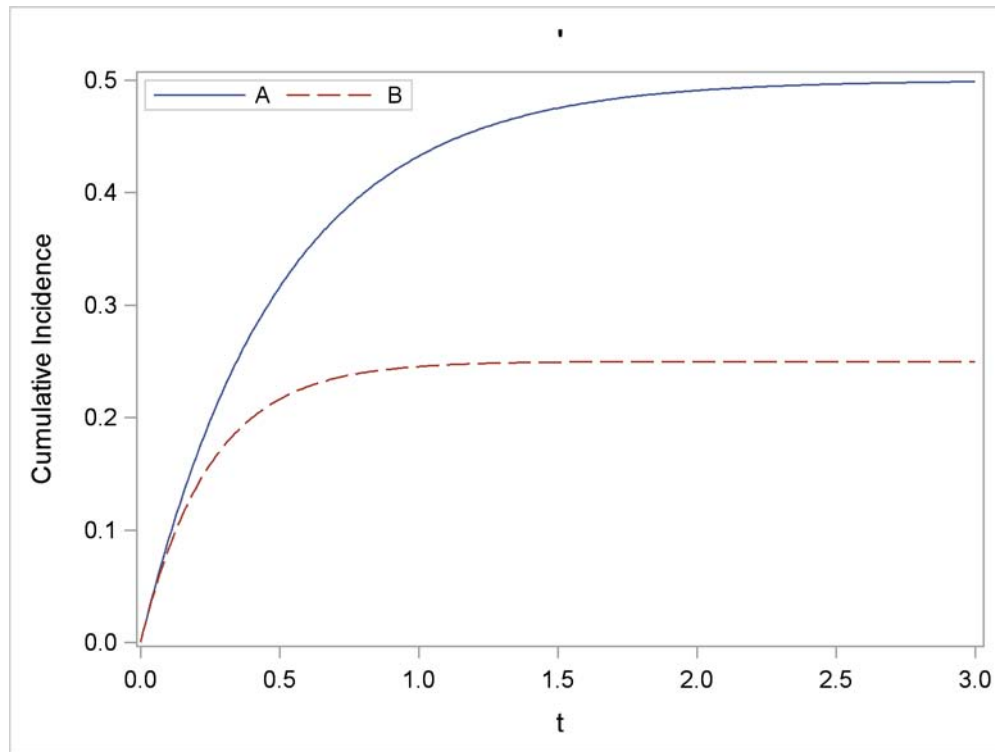
$$F_1^A(t) = \int_0^t S^A(u)h_1^A(t)du = \int_0^t e^{-2t} du = \frac{1}{2}(1 - e^{-2t})$$

$$F_1^B(t) = \int_0^t S^B(u)h_1^B(t)du = \int_0^t e^{-4t} du = \frac{1}{4}(1 - e^{-4t})$$

Although the cause-specific hazard functions of Event 1 for Groups A and B are identical ($h_1^A(t) \equiv h_1^B(t)$), the corresponding cumulative incidence functions are different ($F_1^A(t) \neq F_1^B(t)$). This demonstrates that there can be a group

effect on the cumulative incidence functions when there is no group effect on the corresponding hazard functions. Figure 3 shows that the cumulative incidence function for Event 1 in Group A is greater than the cumulative incidence function for Event 1 in Group B. Note that when t goes to infinity, the upper limits of the cumulative incidence functions $F_1^A(t)$ and $F_1^B(t)$ are $\frac{1}{2}$ and $\frac{1}{4}$, respectively.

Figure 3 Cumulative Incidence Functions from Groups A and B



You need to apply different tests for group effects depending on whether you are interested in the cause-specific hazard rate or the cumulative incidence function. When you compare treatment effects on the cumulative incidence function, the null hypothesis H_0 is that the cumulative incidence functions are identical across treatment groups. Gray (1988) proposed a test for this purpose, and you can perform this test with the %CIF macro. A detailed discussion of Gray's test is outside the scope of this paper.

When you compare treatment effects on the cause-specific hazard rate, the null hypothesis H_0 is that the cause-specific hazard functions are identical across groups. You can apply the log-rank test from PROC LIFETEST by treating the failure times from causes other than the cause of interest as censored observations.

USING THE %CIF MACRO TO COMPARE CUMULATIVE INCIDENCE FUNCTIONS

In Figure 2, it is apparent that the estimated cumulative incidence functions are markedly different, but in general a formal test is needed to assess the difference. The following statement illustrates how you can use the %CIF macro to test whether the difference is significant:

```
%CIF (DATA=bmt, TIME=Ftime, STATUS=Status, EVENT=1, CENSORED=0, GROUP=Diagnosis);
```

Note that this macro call is similar to the one that estimates the cumulative incidence functions except that the NOTEST option is not specified. By default, the macro performs Gray's test when there are two or more groups. The result of the test is shown in Figure 4.

Figure 4 Test for Equality of Cumulative Incidence Functions among Disease Groups

Gray's Test for Equality of Cumulative Incidence Functions		
Chi-Square	DF	Pr > Chi-Square
11.9229	2	0.0026

The χ^2 test statistic is 11.9229 with 2 degrees of freedom, and the p -value is 0.0026.

You can also use the %CIF macro to request a stratified test (Gray 1988) by specifying the STRATA=parameter.

SUMMARY

Standard survival analysis addresses three issues: estimation of survival functions, comparison of survival functions, and assessment of the effects of covariates on survival. Survival analysis for data with competing risks is complicated by the fact that the competing events are often mutually exclusive and dependent. Two useful quantities in competing risk analysis are cause-specific hazard functions and cumulative incidence functions. The SAS %CIF macro provides methods for estimating cumulative incidence functions and for comparing these functions across treatment group.

APPENDIX: AVAILABILITY AND SYNTAX FOR THE %CIF MACRO

You can download the %CIF macro from <http://support.sas.com/fusionpreview/previewhtml/45/997.html>. Beginning with the 12.1 release of SAS/STAT, this macro will also be available through the SAS auto call library. The syntax of the %CIF macro is as follows:

```
%CIF(DATA=, OUT=, TIME=, STATUS=, EVENT=, CENSORED=, GROUP=,
      STRATA=, ALPHA=, RHO=, SE=, TITLE=, TITLE2=, OPTIONS=);
```

Only the DATA=, OUT=, TIME=, and STATUS= parameters are required; the other parameters are optional. The parameters are defined as follows:

DATA=SAS-data-set

specifies the SAS data set that contains the data to be analyzed.

OUT=SAS-data-set

specifies an output data set with estimates of cumulative incidence functions and confidence limits. If you omit the OUT=parameter, the output data set is created with a default name of cifEstimate.

TIME=variable

specifies the name of the failure-time variable.

STATUS=variable

specifies the name of the numeric variable whose values indicate whether an observation corresponds to the event of interest, the competing events, or is censored.

EVENT=(list)

specifies values of the STATUS= *variable* that identify failure times from a cause of interest. Use blanks to separate multiple values. The default value is 1.

CENSORED=(list)

specifies values of the STATUS= *variable* that identify censored observations. Use blanks to separate multiple values. The default value is 0.

GROUP=variable

specifies the name of the variable that defines the groups for comparison.

STRATA=variable

specifies the name of the variable that defines the strata for a stratified test; see Gray (1988).

ALPHA=value

specifies the level of the pointwise confidence intervals. The *value* must be between 0 and 1. The default is 0.05.

PHO=value

specifies the power of the weight function in the test; see Gray (1988). The *value* must be between 0 and 1. The default is 0.

SE=value

specifies a method for computing the standard error of the cumulative incidence function. Specify SE=1 for the delta method, and specify SE=2 for the counting process method.

TITLE=*value*

specifies the first title for the cumulative incidence function plot. Do not use quotation marks.

TITLE2=*value*

specifies the second title for the cumulative incidence function plot. Do not use quotation marks.

OPTIONS=*keyword*

specifies one or more of the following keywords separated by blanks:

- NOPLOT suppresses plots of cumulative incidence functions.
- PLOTCL plots the pointwise confidence limits for cumulative incidence functions.
- NOCIFEST suppresses the table that contains the estimated cumulative incidence functions.
- NOTEST suppresses the test for equality for cumulative incidence functions among groups.

REFERENCES

- Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1995), *Statistical Models Based on Counting Processes*, Springer.
- Fine, J. and Gray, R. (1999), "A Proportional Hazards Model for the Subdistribution of a Competing Risk," *Journal of the American Statistical Association*, 94, 496–509.
- Gray, R. (1988), "A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk," *The Annals of Statistics*, 16, 1141–1154.
- Hosmer, D., Lemeshow, S., and May, S. (2008), *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*, Second Edition, John Wiley & Sons.
- Klein, J. (2010), "Competing Risks," *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3), 333–339.
- Klein, J. and Andersen, P. (2005), "Regression Modeling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function," *Biometrics*, 61, 223–229.
- Klein, J. P. and Moeschberger, M. L. (2003), *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer-Verlag.
- Martinussen, T. and Scheike, T. (2006), *Dynamic Regression Models for Survival Data*, Springer.
- Marubini, E. and Valsecchi, M. (1995), *Analysing Survival Data from Clinical Trials and Observational Studies*, John Wiley & Sons.
- Pepe, M. (1991), "Inference for Events with Dependent Risks in Multiple Endpoint Studies," *Journal of the American Statistical Association*, 86, 770–778.
- Pintilie, M. (2006), *Competing Risks: A Practical Perspective*, Wiley.
- Tsiatis, A. (1975), "A Nonidentifiability Aspect of the Problem of Competing Risks," in *Proceedings of the National Academy of Sciences*, volume 72, 20–22.

ACKNOWLEDGMENTS

The authors thank Bingqing Teresa Zhou (Assistant Professor of Biostatistics, Yale School of Public Health) for her work on this project during her participation in the SAS Summer Fellowship in Statistics program in 2009. The authors also thank Bob Rodriguez and Anne Baxter of SAS whose assistance and comments considerably improved the manuscript.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. You can contact the authors at the following addresses:

Guixian Lin	Ying So	Gordon Johnston
SAS Institute Inc.	SAS Institute Inc.	SAS Institute Inc.
SAS Campus Drive	SAS Campus Drive	SAS Campus Drive
Cary, NC 27513	Cary, NC 27513	Cary, NC 27513
guixian.lin@sas.com	ying.so@sas.com	gordon.johnston@sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.