Using the PHREG Procedure to Analyze Competing-Risks Data

Ying So, Guixian Lin, and Gordon Johnston, SAS Institute Inc.

ABSTRACT

Competing risks arise in studies in which individuals are subject to a number of potential failure events and the occurrence of one event might impede the occurrence of other events. For example, after a bone marrow transplant, a patient might experience a relapse or might die while in remission. You can use some standard methods of survival analysis, such as the log-rank test and the Cox regression, to analyze competing-risks data, whereas other methods, such as the product-limit estimator, might yield biased results. An increasingly common practice of assessing the probability of a failure in competing-risks analysis is to estimate the cumulative incidence function, which is the probability subdistribution function of failure from a specific cause. This paper discusses two commonly used regression approaches for evaluating the relationship of the covariates to the cause-specific failure in competing-risks data. One approach models the cause-specific hazard, and the other models the cumulative incidence. The paper shows how to use the PHREG procedure in SAS/STAT® to fit these models.

INTRODUCTION

In a classical time-to-event situation, an individual can experience the event of interest or be censored. A competing-risks situation arises when an individual can experience more than one type of event, the occurrence of one event might hinder the occurrence of other types of events, and only the time to failure for the earliest of these events is observed. Examples of competing risks are found in many fields, but they are especially prevalent in clinical studies. For example, competing risks are encountered when cancer patients are followed, and their first event can be local recurrences, distant recurrences, distant metastases, onset of secondary cancer, or death, which precludes all these events.

It has often been pointed out that in the presence of competing risks, the product-limit (Kaplan-Meier) method of estimating the distribution of time to event by ignoring events of all types other than the one of interest yields biased results. The assumption that an individual will experience the event of interest if the follow-up period is long enough does not hold in competing-risks data, because the occurrence of the event of interest can be made impossible by the occurrence of an earlier competing event. A useful quantity in competing-risks analysis is the cumulative incidence function, which is the probability subdistribution function of failure from a specific cause. Lin, So, and Johnston (2012) created a SAS macro that computes the nonparametric estimate of the cumulative incidence function and provides Gray's (1988) test for group comparisons.

Several modeling approaches are available for evaluating the effects of covariates on the cause-specific outcome in competing-risks data (Prentice et al. 1978; Larson and Dinse 1985; Fine and Gray 1999). Two approaches are especially popular. One approach models the cause-specific hazard of each event separately, by applying the standard Cox regression for the event of interest and censoring all other observations. The other approach is Fine and Gray's (1999) extension of the Cox regression that models (the hazards of) the cumulative incidence function.

The next section of this paper describes the competing-risks data that are used as an example in the paper. The subsequent section presents some basic definitions of quantities of interests in competing-risks analysis. The final section discusses how to use PHREG procedure to carry out these regression analyses of competing-risks data.

AN EXAMPLE OF COMPETING-RISKS DATA

Bone marrow transplant is a standard treatment for acute leukemia. Klein and Moeschberger (2003) present a set of bone marrow transplant data for 137 patients, grouped into three disease categories based on their diagnosis at the time of transplantation: acute lymphoblastic leukemia (ALL), acute myelocytic leukemia (AML) low-risk, and AML high-risk. Among the 137 patients in the study, 38 patients were diagnosed with ALL, 54 patients were diagnosed with AML low-risk, and 45 patients were diagnosed AML high-risk. There are a number of concomitant variables in the data set; for simplicity, only the waiting time for transplant is included here.

During the follow-up period, some patients might experience a relapse of the leukemia or some patients might die while in remission. The comparison of the disease groups focuses on the occurrence of relapse.

The following statements provide the data. The variable **Group** designates the disease group of a patient, which is either ALL, AML low-risk, or AML high-risk. The variable **T** is the disease-free survival time in days, which is either the time to censoring, the time to relapse, or the time to death while in remission, whichever occurs first. The indicator variable **Status** has three values: 0 for censored observations, 1 for patients who relapse, and 2 for patients who die before experiencing a relapse. The concomitant variable **WaitTime** is the waiting time for transplant, in days. Because this variable has a very large variation, a log transform is applied to stabilize the variance.

```
proc format;
   value DiseaseGroup 1='ALL'
                      2='AML-Low Risk'
                      3='AML-High Risk';

data bmt;
   input Group T Status WaitTime @@;
   logWaittime=log(WaitTime);
   format Group DiseaseGroup.;
   datalines;
1  2081   0   98   1  1602   0  1720  1  1496   0   127   1  1462   0   168
1  1433   0   93   1  1377   0  2187  1  1330   0  1006   1   996   0  1319
1   226   0  208   1  1199   0   174  1  1111   0   236   1   530   0   151
1  1182   0  203   1  1167   0   191  1   418   2   110   1   383   1   824
1   276   2  146   1   104   1    85  1   609   1   187   1   172   2   129
1   487   2  128   1   662   1    84  1   194   2   329   1   230   1   147
1   526   2  943   1   122   2  2616  1   129   1   937   1    74   1   303
1   122   1  170   1    86   2   239  1   466   2   508   1   192   1    74
1   109   1  393   1    55   1   331  1     1   2   196   1   107   2   178
1   110   1  361   1   332   2   834  2  2569   0   270   2  2506   0    60
2  2409   0  120   2  2218   0    60  2  1857   0    90   2  1829   0   210
2  1562   0   90   2  1470   0   240  2  1363   0    90   2  1030   0   210
2   860   0  180   2  1258   0   180  2  2246   0   105   2  1870   0   225
2  1799   0  120   2  1709   0    90  2  1674   0    60   2  1568   0    90
2  1527   0  450   2  1324   0    75  2   957   0    90   2   932   0    60
2   847   0   75   2   848   0   180  2  1850   0   180   2  1843   0   270
2  1535   0  180   2  1447   0   150  2  1384   0   120   2   414   2   120
2  2204   2   60   2  1063   2   270  2   481   2    90   2   105   2   120
2   641   2   90   2   390   2   120  2   288   2    90   2   421   1    90
2    79   2   90   2   748   1    60  2   486   1   120   2    48   2   150
2   272   1  120   2  1074   2   150  2   381   1   120   2    10   2   240
2    53   2  180   2    80   2   150  2    35   2   150   2   248   1    30
2   704   2  105   2   211   1    90  2   219   1   120   2   606   1   210
3  2640   0  750   3  2430   0    24  3  2252   0   120   3  2140   0   210
3  2133   0  240   3  1238   0   240  3  1631   0   690   3  2024   0   105
3  1345   0  120   3  1136   0   900  3   845   0   210   3   422   1   210
3   162   2  300   3    84   1   105  3   100   1   210   3     2   2    75
3    47   1   90   3   242   1   180  3   456   1   630   3   268   1   180
3   318   2  300   3    32   1    90  3   467   1   120   3    47   1   135
3   390   1  210   3   183   2   120  3   105   2   150   3   115   1   270
3   164   2  285   3    93   1   240  3   120   1   510   3    80   2   780
```

```
3   677   2  150   3    64   1   180  3   168   2   150   3    74   2   750
3    16   2  180   3   157   1   180  3   625   1   150   3    48   1   210
3   273   1  240   3    63   2   360  3    76   1   330   3   113   1   240
3   363   2  180
;
```

You can use Gray's (1988) test to compare the disease groups on the occurrence of relapse; the comparison is shown in Lin, So, and Johnston (2012). However, if you also want to adjust for some concomitant variables, such as the effect of the waiting time for transplant, you need to perform a regression analysis. This paper shows how to use the PHREG procedure to fit two popular regression models for competing-risks data.

BASIC QUANTITIES IN COMPETING RISKS

Let T and C denote the failure time and censoring time, respectively. For data that have K competing risks, the pair (X, δ) is observed, where $X = \min(T, C)$ and $\delta = 1, \ldots, K$ is an indicator that has values of 0 for censoring and other values that designate specific failure causes. For competing-risks data, two useful quantities are the cause-specific hazard function and the cumulative incidence function:

- The cause-specific hazard function $h(t)$ 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

$$h_k(t) = \lim_{\Delta t -> 0} \frac{P(t < T < T + \Delta t, \delta = k | T > t)}{\Delta t}, \quad k = 1, \ldots, K$$

- 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

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

  The cumulative incidence function is also referred to as the subdistribution function, because it is not a true probability distribution.

It follows from these definitions that

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

where $H_k(t) = \int_0^t S(u) h_k(u) du$ is the cause-specific cumulative hazard function and $S(t) = \exp\left(-\sum_{k=1}^{K} H_k(t)\right)$ is the overall survival function, which is the probability of surviving beyond time t.

In the absence of competing risks (that is, if $K = 1$), the failure distribution function $F_1(t) = 1 - \exp(-H_1(t))$ is a monotone function of the hazard function $h_1(t)$. This property does not hold in the presence of competing risks, because the cumulative incidence of an event is not defined solely by its corresponding cause-specific hazard: it also depends on the cause-specific hazards of the competing events. You can easily construct an example to illustrate that two groups that have the same cause-specific hazard for an event can have very different cumulative incidence functions (Gray 1988; Lin, So, and Johnston 2012).

Analogous to the relationship between hazard function and survivor function in the absence of competing risks, Fine and Gray (1999) define the subdistribution hazard, which is the hazard of the cumulative incidence function $F_k(t)$,

$$\tilde{h}_k(t) = \frac{d}{dt} \log(1 - F_k(t))$$

In the presence of competing risks, the subdistribution hazard $\tilde{h}_k(t)$ is not the same as the cause-specific hazard $h_k(t)$. In terms of estimating these quantities, the difference is in the risk set. For the cause-specific hazard, the risk set decreases at each time point when there is a failure of a different cause. However, for the subdistribution hazard, individuals who fail from a competing cause remain in the risk set until their potential censoring time.

To study the effect of covariates on the cause-specific outcome, you can model the cause-specific hazard function or you can model the subdistribution hazard function. These two modeling approaches might yield different results.

REGRESSION MODELS FOR COMPETING-RISKS DATA

In competing-risks data, the influence of covariates can be evaluated in relation to the cause-specific hazard or the subdistribution hazard (the hazard of the cumulative incidence function) of the specific failure type. This section briefly describes each regression model and then shows how to use the PHREG procedure to fit the model by using the bone marrow transplant data.

For the ith subject, $i = 1, \ldots, n$, let X_i, δ_i, and $\mathbf{Z}_i(t)$ be the observed time, cause of failure, and covariate vector at time t, respectively. Assume that K causes of failure are observable ($\delta_i \in \{1, \ldots, K\}$); $\delta = 0$ indicates a censored observation. Consider failure from cause 1 to be the event of interest, and consider failures from other causes to be competing events.

Modeling the Cause-Specific Hazard

The standard Cox regression is applied to the cause-specific hazard of interest, and competing events are treated as censored observations. The cause-specific hazard of interest for a subject that has covariate vector $\mathbf{Z}$ follows the proportional hazards assumption

$$h_1(t|\mathbf{Z}) = h_{10}(t)\exp(\boldsymbol{\beta}'\mathbf{Z})$$

where $h_{0,1}(t)$ is the baseline of the cause-specific hazard of interest and the vector $\boldsymbol{\beta}$ represents the covariate effects on the event of interest. In the Cox regression, the parameter vector $\boldsymbol{\beta}$ is estimated by maximizing the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_i \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{Z}_i)}{\sum_{j \in \mathcal{R}_i} \exp(\boldsymbol{\beta}'\mathbf{Z}_j)} \right)^{\delta_i = 1}$$

where $\mathcal{R}_i$ is the risk set of patients who do not fail or are not censored before X_i.

The following statements use PROC PHREG to fit the cause-specific hazard model for relapse:

```
proc phreg data=bmt;
   class Group / order=internal ref=first param=glm;
   model T*Status(0,2) = Group logWaitTime;
   hazardratio 'Cause-Specific Hazards' Group / diff=pairwise;
run;
```

Patients who die while in remission (without experiencing a relapse) have a **Status** value of 2. To treat these failures as censored observations, you add the value of 2 to the list of censoring values. The HAZARDRATIO statement requests the cause-specific hazard ratios for each pair of disease groups.

Patients who die without experiencing a relapse are treated as censored observations. As a result, there are 42 patients who have a relapse and a total of 95 censored observations (Figure 1) in the data. Figure 2 shows a significant effect (p = 0.0003) of **Group** on relapse. The cause-specific hazard ratio estimates of one risk group relative to the other are displayed in Figure 3.

Figure 1 Modeling a Cause-Specific Hazard

The PHREG Procedure

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
137	42	95	69.34

Figure 2 Cause-Specific Hazard Regression

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Group	2	16.1850	0.0003
logWaittime	1	1.8557	0.1731

Figure 3 Pairwise Cause-Specific Hazard Ratios

Cause-Specific Hazards: Hazard Ratios for Group			
Description	Point Estimate	95% Wald Confidence Limits	
Group ALL vs AML-Low Risk	2.977	1.208	7.340
Group AML-Low Risk vs ALL	0.336	0.136	0.828
Group ALL vs AML-High Risk	0.573	0.281	1.169
Group AML-High Risk vs ALL	1.745	0.856	3.560
Group AML-Low Risk vs AML-High Risk	0.192	0.086	0.431
Group AML-High Risk vs AML-Low Risk	5.195	2.321	11.630

The hazard of relapse for the AML high-risk patients is 5.2 times that for the AML low-risk patients (95% CI 2.32 to 11.63) and is 1.7 times that for the ALL patients (95% CI 0.86 to 3.56). The hazard of relapse for the ALL patients is 3.0 times that for the AML low-risk patients (95% CI 1.21 to 7.34).

Prediction of the cumulative incidence function from the cause-specific regression model has been studied by Cheng, Fine, and Wei (1998). It requires estimating the cumulative cause-specific hazards for both the cause of interest and the competing causes. The variance estimation is especially complicated. Special software is needed for the estimation (Rosthoj, Andersen, and Abidstrom 2004).

Modeling the Cumulative Incidence

Fine and Gray (1999) introduce a way to model the cumulative incidence function by defining the hazard of the cumulative incidence function, known as the subdistribution hazard, and impose the proportional hazards assumption on the subdistribution hazards,

$$\tilde{h}_1(t|\mathbf{Z}) = \tilde{h}_{10}(t)\exp(\boldsymbol{\beta}'\mathbf{Z})$$

where $\tilde{h}_{1,0}(t)$ is the baseline of the subdistribution hazard of cause 1. The partial likelihood of this proportional subdistribution hazards model is given by

$$\tilde{L}(\boldsymbol{\beta}) = \prod_i \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{Z}_i)}{\sum_{j\in\tilde{\mathcal{R}}_i} w_{ij}\exp(\boldsymbol{\beta}'\mathbf{Z}_j)} \right)^{\delta_i=1}$$

The modified risk set $\tilde{\mathcal{R}}_i$ at X_i includes patients who are still at risk for the event of interest and also patients who experience a competing event before X_i. The weights w_{ij} are needed as soon as censoring occurs. Patients who experience no event of interest before X_i are given a weight $w_{ij} = 1$, whereas patients who experience competing events before X_i are given a weight w_{ij} that reduces with time,

$$w_{ij} = \frac{\hat{G}(X_i)}{\hat{G}(\min(X_j, X_i))}$$

where $\hat{G}(t)$ is the Kaplan-Meier estimate of the survival function of the censoring distribution, which is the cumulative probability that a patient is still being followed at time t.

The regression coefficients $\boldsymbol{\beta}$ are obtained by maximizing the partial likelihood $\tilde{L}(\boldsymbol{\beta})$, and the covariance matrix of the parameter estimator is computed as a sandwich estimate. The methodology has been implemented in the PHREG procedure in SAS/STAT 13.1.

You fit this proportional subdistribution hazards model in PROC PHREG by using the MODEL statement, where you specify the failure cause of interest in the EVENTCODE= option.

The following statements use PROC PHREG to fit the proportional subdistribution hazards model:

```
proc phreg data=bmt;
   class Group / order=internal ref=first param=glm;
   model T*Status(0) = Group logWaitTime / eventcode=1;
   hazardratio 'Subdistribution Hazards' Group / diff=pairwise;
run;
```

To designate relapse (**Status** = 1) as the event of interest, you specify EVENTCODE=1 in the MODEL statement. The HAZARDRATIO statement requests the subdistribution hazard ratio for each pair of the disease groups.

Figure 4 shows a significant effect on relapse between the disease groups (p = 0.0009). The subdistribution hazard ratio estimates of one disease group relative to another are displayed in Figure 5. The hazard of relapse for the AML high-risk patients is 4.5 times that of the AML low-risk patients (95% CI 2.04 to 9.76) and is 1.6 times that of the ALL patients (95% CI 0.77 to 3.24). The hazard of relapse of the ALL patients is 2.8 times that of the AML low-risk patients (95% CI 1.22 to 6.56).

Figure 4 Subdistribution Hazard Regression

The PHREG Procedure

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
Group	2	14.0980	0.0009
logWaittime	1	2.8132	0.0935

Figure 5 Pairwise Subdistribution Hazard Ratios

Subdistribution Hazards: Hazard Ratios for Group

Description	Point Estimate	95% Wald Confidence Limits	
Group AML-Low Risk vs AML-High Risk	0.224	0.103	0.489
Group AML-High Risk vs AML-Low Risk	4.464	2.043	9.755
Group AML-Low Risk vs ALL	0.354	0.152	0.823
Group ALL vs AML-Low Risk	2.823	1.215	6.559
Group AML-High Risk vs ALL	1.581	0.772	3.240
Group ALL vs AML-High Risk	0.632	0.309	1.296

The Breslow estimator of the baseline cumulative subdistribution hazard function, denoted by $\hat{\Lambda}_{10}(t)$, incorporates the modified risk sets and the gradual reduction of weights for those artificially retained in the risk set, as in the partial likelihood of the subdistribution model. With time-invariant covariates $\mathbf{Z}$, the cumulative incidence function can be estimated by

$$\hat{F}_1(t|\mathbf{Z}) = 1 - \exp[-\hat{\Lambda}_{10}(t)\exp(\hat{\boldsymbol{\beta}}'\mathbf{Z})]$$

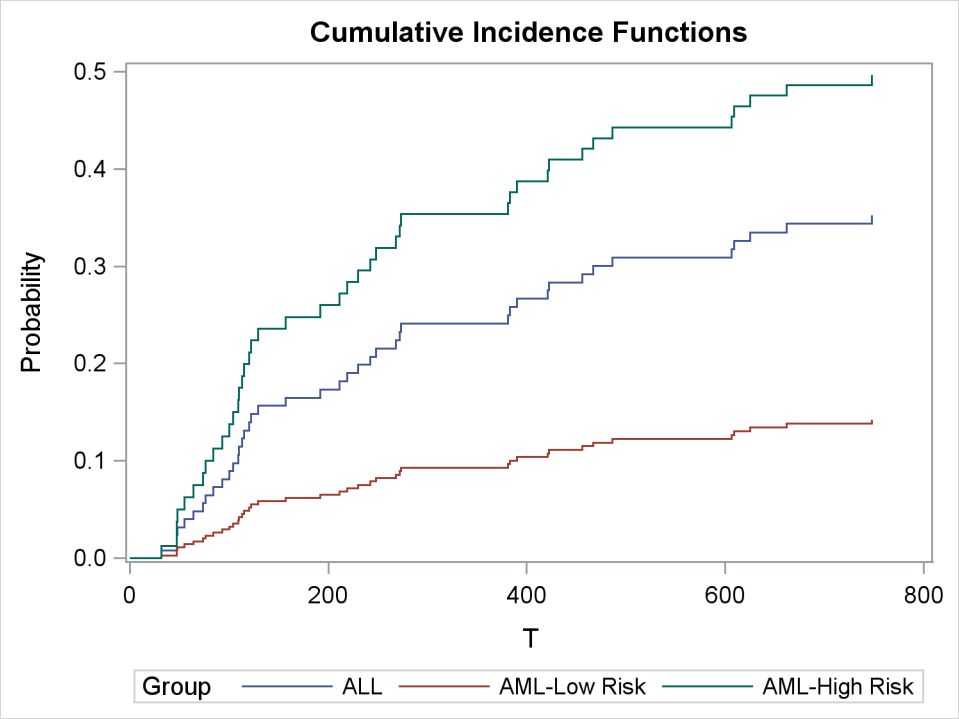
You can predict the cumulative incidence by using the BASELINE statement in PROC PHREG. You use the COVARIATES= option in the BASELINE statement to specify a data set that contains settings for predicting the cumulative incidence function or for displaying the cumulative incidence curves, which are requested here by using the PLOTS=CIF option in the PROC PHREG statement. The following statements use the PHREG procedure to plot the predicted cumulative incidence function for each disease group at **logWaitTime** = 5.2, the median value of the log of the waiting times:

```
Data Risk;
   logWaitTime=5.2;
   Group=1; output;
   Group=2; output;
   Group=3; output;
   format Group DiseaseGroup.;
   run;

proc phreg data=bmt plots(overlay=strata)=cif;
   class Group / param=glm order=internal ref=first;
   model T*Status(0) = Group logWaitTime / eventcode=1;
   baseline covariates=risk out=_null_ / rowid=Group;
run;
```

Figure 6 displays the predicted cumulative incidence for all three disease groups. At any given time after the transplant, an AML high-risk patient is more likely to relapse than an ALL patient, and an ALL patient is more likely to relapse than an AML low-risk patient.

Figure 6 CIF of the Three Disease Groups at **logTimeWait** = 5.2



You do not need to issue two separate calls to PROC PHREG, one for estimating the hazard ratios and one for predicting the cumulative incidence functions, as presented earlier. The following statements produce the earlier results when you issue a single call to PROC PHREG:

```
Data Risk;
   logWaitTime=5.2;
   Group=1; output;
   Group=2; output;
   Group=3; output;
   format Group DiseaseGroup.;
   run;

proc phreg data=bmt plots(overlay=strata)=cif;
   class Group / param=glm order=internal ref=first;
   model T*Status(0) = Group logWaitTime / eventcode=1;
   hazardratio 'Subdistribution Hazards' Group / diff=pairwise;
```

```
   baseline covariates=risk out=_null_ / rowid=Group;
run;
```

SUMMARY

This paper illustrates the use of the PHREG procedure to fit the two popular models for competing-risks data. The cause-specific hazard model is fitted as a Cox regression model by censoring all individuals who did not experience the event of interest, but prediction of the cumulative incidence function is difficult. On the other hand, the implementation of Fine and Gray's (1999) regression in the PHREG procedure in SAS/STAT 13.1 enables you to fit the subdistribution hazard model and to predict the cumulative incidence function easily.

Although these two models yield very similar hazard ratios between the disease groups for the bone marrow transplant data discussed in the paper, that might not be the case for other data. Both approaches work for their respective purposes, and each might provide useful insights about the covariates (Pintilie 2006; Dignam, Zhang, and Kocherginsky 2012). Covariate effects in the cause-specific hazard model pertain to the event of interest only, without regard to how the covariates act on the competing risks. If you are interested in the pure effect (for example, the biological mechanism) of how a specific characteristic affects an event outcome, the cause-specific hazard model is preferred. But this model is of little use to patients who must make decisions in the real world, where death from other causes plays a big role. In such cases, you should consider using the subdistribution hazard model.

REFERENCES

Cheng, S. C., Fine, J. P., and Wei, L. J. (1998), "Prediction of the Cumulative Incidence Function under the Proportional Hazards Model," *Biometrics*, 54, 219–228.

Dignam, J. J., Zhang, Q., and Kocherginsky, M. (2012), "The Use and Interpretation of Competing Risks Regression Models," *Clinical Cancer Research*, 18, 2301–2308.

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.

Gray, R. J. (1988), "A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk," *Annals of Statistics*, 16, 1141–1154.

Klein, J. P. and Moeschberger, M. L. (2003), *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd Edition, New York: Springer-Verlag.

Larson, M. G. and Dinse, G. E. (1985), "A Mixture Model for Regression Analysis of Competing Risks Data," *Applied Statistics*, 34, 201–211.

Lin, G., So, Y., and Johnston, G. (2012), "Analyzing Survival Data with Competing Risks Using SAS Software," in *Proceedings of the SAS Global Forum 2012 Conference*.

Pintilie, M. (2006), *Competing Risks: A Practical Perspective*, Statistics in Practice, Chichester, UK: John Wiley & Sons.

Prentice, R. L., Kalbfleish, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978), "The Analysis of Failure Times in the Presence of Competing Risks," *Biometrics*, 34, 541–544.

Rosthoj, S., Andersen, P. K., and Abidstrom, S. Z. (2004), "SAS Macros for Estimation of the Cumulative Incidence Functions Based on a Cox Regression Model for Competing Risks Survival Data," *Computer Method and Programs in Biomedicine*, 74, 69–75.

ACKNOWLEDGMENTS

The authors thank Bob Rodriguez and Ed Huddleston, whose assistance and comments considerably improved the manuscript.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Ying So	Guixian Lin	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
ying.so@sas.com	guixian.lin@sas.com	gordon.johnston@sas.com