Nomogram for survival analysis in the presence of competing risks

Zhongheng Zhang¹, Ronald B. Geskus²,³, Michael W. Kattan⁴, Haoyang Zhang⁵, Tongyu Liu⁶

¹Department of Emergency Medicine, Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, China; ²Wellcome Trust Major Overseas Programme, Oxford University Clinical Research Unit, Ho Chi Minh city, Vietnam; ³Nuffield Department of Clinical Medicine, Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, UK; ⁴Department of Quantitative Health Sciences, Cleveland Clinic Foundation, Cleveland, Ohio, USA; ⁵Division of Biostatistics, J.C. School of Public Health and Primary Care, The Chinese University of Hong Kong, Shatin, Hong Kong, China; ⁶Department of Gynecological Oncology, Fujian Cancer Hospital & Fujian Medical University Cancer Hospital, Fuzhou 350000, China

Correspondence to: Zhongheng Zhang. Department of Emergency Medicine, Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, No. 3, East Qingchun Road, Hangzhou 310016, China. Email: zh_zhang1984@zju.edu.cn.

Abstract: Clinical research usually involves time-to-event survival analysis, in which the presence of a competing event is prevalent. It is acceptable to use the conventional Cox proportional hazard regression to model cause-specific hazard. However, this cause-specific hazard cannot directly translate to the cumulative incidence function, and the latter is usually clinically relevant. The subdistribution hazard regression directly quantifies the impact of covariates on the cumulative incidence. When estimating the subdistribution hazard, subjects experiencing competing event continue to contribute to the risk set, and censoring weights are assigned to them after the competing event time. The weights are the conditional probability that a subject remains uncensored, and can be modelled to depend on the covariates of a subject. The first option to perform regression on the subdistribution hazard was the crr() function in the cmprsk package. However, it is not straightforward to draw a nomogram, which is a user-friendly tool for risk prediction, with the crr() function. To overcome this problem, we show an alternative method to use a nomogram function based on result of subdistribution hazard modeling.

Keywords: Nomogram; survival analysis; competing risks; subdistribution

Submitted Jun 28, 2017. Accepted for publication Jul 12, 2017.
doi: 10.21037/atm.2017.07.27
View this article at: http://dx.doi.org/10.21037/atm.2017.07.27

Introduction

Clinical research usually involves time-to-event survival analysis. It is common in clinical medicine that a cohort of patients under observation can have one of multiple mutually exclusive types of outcome. For patients with sarcoma, they may die from sarcoma-related death, or non-sarcoma death. These two outcomes are mutually exclusive because a patient can never experience both of the events. In epidemiology, this phenomenon has been extensively studied under the term competing risks analysis. The occurrence of the competing even precludes the occurrence of the event of interest; it is censored. When the censoring event is independent to the distribution of the event of interest, it can be considered as non-informative censoring. This assumption cannot be tested empirically. Then the standard methods of survival analysis still apply. However, a competing risks is usually associated with the future risk of the event of interest: the competing event is usually not independent (1). The traditional Cox proportional hazard model can still be applied directly to the competing-risks situation, which results in the fitting of a cause-specific hazards model. Although the coefficients estimated from such a model can reflect the effect of a covariate on the cause-specific hazard, it will not necessarily reflect the effect on cumulative incidence. The later quantity is often more
clinically relevant. In this context, Fine et al. developed a regression model on the subdistribution hazard that provides a one-to-one correspondence between parameter estimates and cumulative incidence (2). The subdistribution hazard ratio estimated from the Fine and Gray model has no direct clinical interpretation for subject matter audience, but it reflects the impact of covariate on cumulative incidence.

The 

function provided in the cmprsk package was the first tool to perform regression analysis on the subdistribution hazard (3). However, the object returned by 

cannot be passed directly to the nomogram function in the rms package to draw a nomogram for survival analysis in the presence of competing risks. While investigators are interested in using nomogram to show cumulative incidence of a specific cause (4), there is no package for this purpose. This tutorial aims to provide a step-by-step approach to create a competing-risk nomogram. The dataset is reshaped to long format (5). Each subject experiencing competing event is expanded to several rows in the long format, and he or she continues to contribute to the risk set after the time when the competing event occurs.

**Working example**

Due to the complexity of the structure of the survival data with competing risks, a special R package has been developed for simulation of such a dataset (6). We will employ 

function in the survsim package for this purpose. Readers not interested in how to generate such a dataset can skip this section.

```r
> install.packages("survsim")
> library(survsim)
> set.seed(10)
> df <- crisk.sim(n=500, foltime=10,
dist.ev=rep("lnorm",2),
anc.ev=c(1.48, 0.53),
beta0.ev=c(3.80, 2.54),
dist.cens="lnorm",
anc.cens=3.5,beta0.cens=5.42,
z==NULL,
beta=list(c(0.21,0.017),c(0.37,0.016)),
x=list("normal",0,1),
c("bern", 0.564)), nsit=2)
```

In the example, we generated a data frame called df as shown above. The nid is the subject id that is unique for each individual subject. The variable “cause” is the cause of event corresponding to follow-up time, and it is missing if no event is observed. There are two levels (1 and 2) corresponding to the event of interest and competing event, respectively. The variable “time” is the observed analysis time. The variable “status” contains logical values 1 and 0, indicating whether corresponding event has been observed or not. For individuals with status =0, their events fall beyond the maximum follow-up time and are considered as censored. The variables start and stop specify the time at which the follow-up begins and ends, respectively. Heterogeneity is indicated by the variable z (there is no heterogeneity here). The last two columns are covariates x and x.1. The former is a numeric variable and the latter is a
Understanding the censoring weights

Conventionally, the cause-specific cumulative incidence can be estimated by the cause-specific hazard, which is a special case of the Aalen-Johansen estimator of the transition probability in multi-state models. Geskus showed that the cause-specific subdistribution (cumulative incidence) can be estimated by a product-limit estimator in the form of (8):

\[ \hat{H}_R(t) = \prod_{t \in \mathcal{T}_n} \left\{ 1 - \frac{d_i(t_{ij})}{r_i(t_{ij})} \right\} \]

The superscript PL indicates it is a product-limit form. It is similar in form to the Kaplan-Meier, but now the terms contain an estimate of the subdistribution hazard. In contrast to the cause-specific hazard that can be estimated using conventional Cox proportional hazard model, individuals experiencing a competing event remain included in the denominator. It means that the occurrence of a competing event is ignored and such subjects remain in the risk set until they would have been censored (9).

The modified risk set is expressed as:

\[ r^*(t) = \sum_{i=1}^{n} w_i(t) \]

\[ w_i(t) = \begin{cases} 1 & \text{if } \delta_i \in [0,1] \\ \frac{G_i(t)}{G_i[\min\{C_i, T_i\}]} & \text{if } \delta_i = 2 \end{cases} \]

where \( n \) is the number of subjects. The event times are \( \{T_1, T_2, T_3, \ldots, T_n\} \), and censoring times are \( \{C_1, C_2, C_3, \ldots, C_n\} \). But we can only observe \( Z_i = \min\{C_i, T_i\} \) and \( \delta_i \) indicates the type of event with 0 indicating censoring, 1 indicating event and 2 for competing event. The censoring survival probability can be modeled as dependent on the unique covariate vector for a given subject (10).

Transforming dataset format to include censoring weights

The `crprep()` function in the mstate package is a good tool to create the weighted data set for competing risks analysis (5). Alternatively, the survival package has a function `finegray()` that can also be used to create the weighted data set.

```R
> install.packages("mstate")
> library(mstate)
> df$cause<-ifelse(is.na(df$cause), 0, df$cause)
> df.w <- crprep("time", "cause", data=df, trans=c(1,2), cens=0, id="nid", keep=c("x","x.1"))
```

The first argument of `crprep()` function is a character string “time” indicating the column name in the data frame “df” that contains the stop time of follow-up. The start time is 0 by default, but can be specified using the `Tstart` argument. The “cause” is the name of the variable indicating the status at the end of follow-up. The argument “trans” specifies values of the status for which weights are to be calculated. Here, weights are calculated for event type 1, there is no need to compute them for event type 2. The value 0 in the “cause” column is considered to be censoring. The id argument generates a character string indicating a column containing the subject identifier. The covariates x and x.1 are retained in the new data set by using the keep argument. Now let’s take a look at the data set in long format.

```R
> round(head(df.w,15),2)
tid Tstart Tstop status weight. cens x x.1 count failcode
1 1 1.00 7.66 2 1.00 0.02 1 1 1
2 1 1.00 7.66 2 1.00 0.02 1 2 1
3 1 1.00 7.66 2 1.00 0.02 1 3 1
4 1 1.00 7.66 2 1.00 0.02 1 4 1
5 1 1.00 7.66 2 1.00 0.02 1 5 1
6 1 1.00 7.66 2 1.00 0.02 1 6 1
7 1 1.00 7.66 2 1.00 0.02 1 7 1
8 2 1.00 7.66 2 1.00 0.02 1 8 1
```

© Annals of Translational Medicine. All rights reserved.
Subjects experiencing the competing event (status = 2) take more than one row. The first row represents the follow-up until the observed event; the follow-up continues after 7.66 for subject 1 and there are censoring weights assigned for them. Censoring weights represent the conditional probability of remaining uncensored. The count variable contains information on the counting of rows within a subject. The failcode is the event type under consideration.

> with(df.w,table(failcode,status))

<table>
<thead>
<tr>
<th>status</th>
<th>failcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>84</td>
</tr>
<tr>
<td>1</td>
<td>115</td>
</tr>
<tr>
<td>2</td>
<td>2482</td>
</tr>
</tbody>
</table>

When cause = 1 in the df dataset is considered as the event of interest, the crprep() function generates a long dataset with failcode = 1. There are 115 subjects who experience the event of interest, and they are not expanded in the long format df.w. Subjects with cause = 2 are considered to experience the competing event, and there are 301 such cases in the original dataset df. They are expanded to 2,482 rows in the long dataset. When cause = 2 in the df dataset is considered as the event of interest, the crprep() function generates a long dataset with failcode = 2. There are 301 subjects that experience that event, and they are not expanded in the long format df.w. Subjects with cause = 1 are considered to experience the competing event, and there are 115 such cases in the original dataset df. They are expanded to 1,789 rows in the long dataset. Obviously, we will perform analysis by restricting to the subset with failcode = 1.

**Draw a nomogram**

With the long format dataset, the subdistribution hazard model can be fitted using conventional functions for survival analysis. The cph() function from the rms package will be employed to fit a proportional subdistribution model. The resulting object can be passed to the nomogram() function in the rms package (11).

> library(rms)
> ddist <- datadist(df.w)
> options(datadist='ddist')
> mod <- cph(Surv(Tstart,Tstop,status==1)~x+x.1, data=df.w, weight=weight.cens, subset=failcode==1, surv=T)
> surv <- Survival(mod)
> nom.sur<- nomogram(mod, fun=list(function(x) 1-surv(3,x), function(x) 1-surv(5,x), function(x) 1-surv(7,x), function(x) 1-surv(9,x)), funlabel=c("3-year event1 Prob.", "5-year event1 Prob.", "7-year event1 Prob.", "9-year event1 Prob."), lp=F)

The function surv() creates a survival function, which receives an argument of time at which cause-specific
cumulative incidence will be calculated. The function 1-surv() is to create the cause-specific cumulative incidence. The fun.side argument specifies the side on which the tick marks are positioned, with 1 indicating below the axis and 3 above the axis (Figure 1). This argument aims to avoid potential overlap between axis labels.

**Comparison with the result from crr() function**

To make sure that the result obtained from the subdistribution hazards model as described above is in consistent with the result from the crr() function, we make a comparison in this section. The parameter estimates of the two methods are the same except for differences that are explained by the stopping criterion in the likelihood maximization. However, the estimates of the standard errors are slightly different. The coefficients of the subdistribution hazard model can be obtained as follows:

```r
> mod$coef
x  x.1
-0.04711762 -0.23700849
```

Regression modeling of the subdistribution function can be performed using the crr() function as well (3):

```r
> library(cmprsk)
> mod.crr<-crr(df$time,df$cause,
cov1=df[,c("x","x.1")],
failcode=1,
cencode=0)
```

The regression coefficient obtained from the crr() function can be examined by the following way:

```r
> mod.crr
convergence: TRUE
coefficients:
x  x.1
-0.04712  -0.23700
standard errors:
[1] 0.08586 0.18610
two-sided p-values:
x  x.1
0.58 0.20
```

The coefficients obtained from crr() function match well to that obtained from cph() function.

**Conclusions**

This tutorial provides a step-by-step approach to the generation of a nomogram for survival data in the presence of a competing event. The tutorial only considered right censored data, but the same approach can be used for left truncated data. While the function crr() contained in the cmprsk package can be used for subdistribution hazard modeling, which has the one-to-one correspondence to the cumulative incidence, it cannot be directly passed to the nomogram() function for drawing a nomogram. Alternatively, the subdistribution can be estimated by a product-limit estimator, which is shown to be equivalent to that estimated by the Aalen-Johansen method. The proportional subdistribution hazards model is not a product-limit structure. However, the same idea of weights is used, because the subdistribution is estimated. This can be realized by setting weights for subjects experiencing competing events in crprep() function. Then the result
can be passed to cph() function in the rms package. The object returned by the cph() function can be passed to the nomogram() function to draw a nomogram.

**Acknowledgements**

**Funding:** The study was funded by Zhejiang Engineering Research Center of Intelligent Medicine (2016E10011) from the First Affiliated Hospital of Wenzhou Medical University.

**Footnote**

**Conflicts of Interest:** The authors have no conflicts of interest to declare.

**References**
