Survival Modeling of Competing Risk Using R: An Undemanding Simulation Approach

[1]Ponnuraja C, [2]Valarmathi S, [3]Babu C Lakshmanan, [4]Venkatesan P

[1,4]Dept. of Statistics, National Institute for Research in Tuberculosis (ICMR), Chennai, Tamil Nadu India
[2]Dept. of Epidemiology, The Tamilnadu Dr. MGR Medical University, Chennai, Tamil Nadu India
[3]Professional-1, Computer Science Corporation (Chennai), India

Abstract

Simulating survival data are necessary for considerate and to evaluate for statistical models. Additionally, inadequate to have real data and also want to know the real status, it leads for simulation. We simulate Competing Risks (CR) survival data with the intention to understand the key concepts. Simulation can be viewed as the practical aspect of probabilistic task of constructing CR process. Simulation done using R and its add-on packages of Scrucca et al. (2007) and analyze them to observe whether the proposed methodology works well. It illustrates with R which allows the user to simulate survival times from parametric models. Standard parametric distributions are used to generate Survival times by Bender et al. (2005), Burton et al. (2006) and Beyersmann et al. (2009). Finally it accomplished with few highlights using simulated data on how to execute competing risk regression analysis with R.

Keywords

Competing Risks, Cumulative Incidence, Cause-Specific Hazards, Sub-Distribution Hazard, Model Selection

Navigation by Section:

I. Introduction

In recent years different approaches for the analysis of time-to-event data in the presence of competing risks (patients can fail from one of two or more mutually exclusive types of event) were introduced (Haller and Ulm, 2013). Competing risks data usually arises in studies in which the failure of an individual may be classified into one of k mutually exclusive causes of failure. When competing risks are present, there are two main differences with classical survival analysis that they are survival functions are not mainly used to express cause-specific failures and classical estimation procedures may present biased results.

The most important approaches are hazard of cause-specific and hazard of subdistribution rates for the analysis of competing risks data. Simulation studies often replace analytical comparisons when other approaches became complication and also simulation can be performed more easily and allow investigation of nonstandard scenarios. We present an approach to generate competing risks data following flexible pre-specified sub distribution hazards.

Survival function represents the probability that an individual survives from the time origin (for example, time of the study enrollment or disease diagnosis) to sometime beyond t. The hazards function or hazard rate, h(t), is the probability that an individual dies at time t, conditional on having survived to that time, which is defined as:

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} \right\}$$

The hazard function, therefore, represents the instantaneous death rate for an individual surviving up to time t and provides a full characterization of the distribution ofT . (Collett, 2003). The main concern with this approach is how to study the impact of important covariates on the distribution ofT. To do this, we assume the variation in the distribution of event and censoring times can be characterized by a vector of observed explanatory covariates, x, which can be either time-invariant or time-dependent covariates. Under the Cox proportional hazards model, the hazard function for the event time T associated with the covariates x is defined as:

$$h(t) = h_0(t)\, e^{\beta X}$$

A. Competing Events Cause-Specific and Sub Distribution Hazards

In a classical way of defining survival analysis is considered a time T until one single possible event. That is for example, time until an event occur (death). However, a combined endpoint is considered, like a case in point that in any clinical studies often investigate 'disease-free survival', specifically time until (occurrence of a) disease or death (without prior ailment), whatever comes first. The aim of a competing risks model is to distinguish between the possible types or causes of that first event.

Competing risks in survival analysis refer to a situation where subjects under investigation are exposed to more than one possible type of events. Thus, each subject is associated with a pair (T,R) where T is the time-to-event (event time or failure time) and R is the reason of the event for that subject. Here we assume that the possible causes are numbered from1,...,K . The cause-specific hazard function in the competing risks model is the hazard of failing from a given cause k in the presence of the competing events

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t \le T < t + \Delta t, R = k \mid T \ge t)}{\Delta t} \right\} \quad with\ R = 1, \dots, K$$

With covariates, the regression model on cause-specific hazards is

$$h_k(t; X) = h_{0k}(t)\, e^{\beta X}$$

The total hazard h(t;X) equals the value of its corresponding hazards function summed up to time t. It is then

$$h(t; X) = \sum_{k=1}^{K} h_k(t)$$

This equation means that the all-cause hazard rate is the sum of K hazards. Here, it is stated that the cause-specific hazards completely determine the stochastic behavior of the competing risks process. The aim is to generate competing risks data for a pair of cause-specific hazards between h_{o1} (t) and h_{o2} (t). The simulation algorithm is the key structure for simulating more complex multistate data. The causespecific hazard measures the instantaneous failure rate due to one risk at a time. It is routinely estimated by constructing the Cox models on cause-specific hazards (Lim et al. 2011).

Fine and Gray(1999) projected a model for the sub-distribution hazard of the CIF. The sub-distribution hazard is a core concept in this approach, and it is defined as the hazard of failing from a given cause in the presence of competing events, given that a subject has survived or has already failed due to different causes. We can write the subdistribution hazard for cause r as:

$$h_k(t) = \lim_{\Delta_t \to 0} \frac{\Pr(t \le T < t+\Delta_t, R=k \mid T \ge t \bigcup (T \le t \bigcap R \ne k))}{\Delta_t}$$

$$= -\frac{d}{dt}\log(1 - I_k(t))$$

Where $I_k(t) = \Pr(T \le t, R= k)$ is the CIF for cause k

Fine and Gray(1999) adopted a semiparametric proportional hazards model for the sub-distribution hazard of cause k for a subject with covariate vector X as follows:

$$h_k(t / X) = h_{k0}(t)\exp(\beta_k^T, X)$$

where $h_{k0}(t)$ is the baseline subdistribution hazard of cause k, and β_k is the vector of coefficients for the covariates.

II. Simulating Competing Risks Data

R is open source software, distributed under the General Public License. Sources and other additional packages for R software can be obtained through the CRAN (Comprehensive R Archive Network), at http://cran.R-project.org. The R software comes with a set of manuals. It is suggested for beginners to read the handbook on "An introduction to R". There are other books available for knowing more about R as an introductory text and data analysis books. R software is compatible with all operating systems. The installing binary for Windows 95, 98, ME, NT4, 2000, and XP is available at http://cran.r-project.org/bin/windows/base/. After downloading the file, install as usual on the user's computer and advised to set the CRAN mirror at your nearest place.

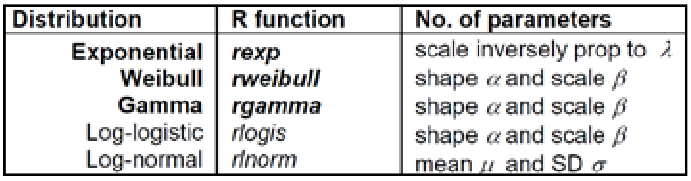
Competing risk analysis is available in an add-on package called cmprsk. R prompt the symbol always ">" and then it expects input commands. Installing the R package cmprsk through online for windows users:

select 'Packages', from the main menu, select 'Install package(s)', choose a CRAN site, (always to choose nearest places of users' area) select the cmprsk package to download and install.

Other information and details of how to install packages for other operating systems are available in the R Installation and Administration manual.

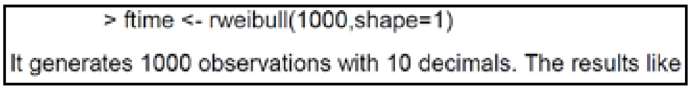
Generally the "library" function lists all available packages in the libraries. It is necessary to ensure that the installed packages "cmprsk" and "timereg" are available or not. The purpose of "timereg" is to fit any regression models and specifically semi parametric model for the cause-specific quantities using survival data.

The common parametric models for survival data with related R functions for simulating survival times and the associated failure time distributions are listed above.



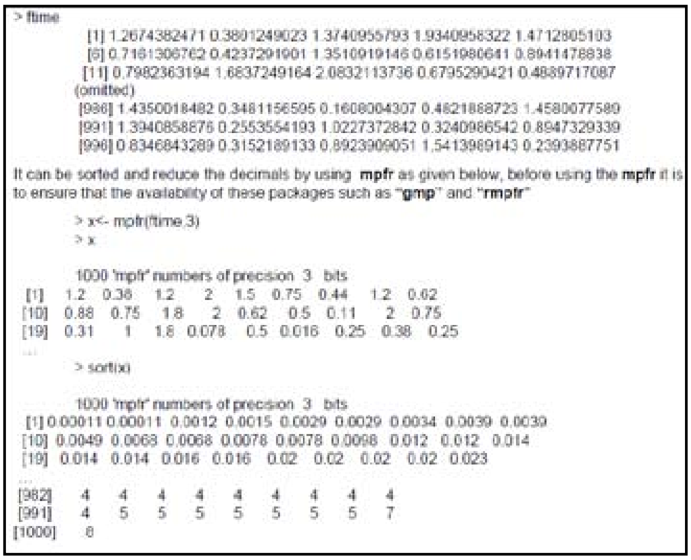
Distribution	R function	No. of parameters
Exponential	*rexp*	scale inversely prop to λ
Weibull	*rweibull*	shape α and scale β
Gamma	*rgamma*	shape α and scale β
Log-logistic	*rlogis*	shape α and scale β
Log-normal	*rlnorm*	mean μ and SD σ

From the above table, we are using the first three distributions to simulate the competing risk data and the last two are in progress. The ftime generated using all the above listed out distributions. But it is not possible to generate at a time. The ftime will be generated initially for Weibull distribution using "rweibul"l for 1000 observations with shape as one. The ftime will also be generated for remaining distribution namely exponential(rexp) and Gamma (rgamma) for comparison purpose



```
          > ftime <- rweibull(1000,shape=1)
```

It generates 1000 observations with 10 decimals. The results like

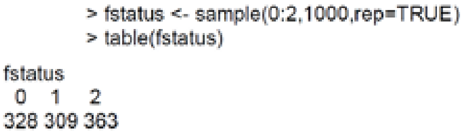


```
> ftime
          [1] 1.2674382471 0.3801249023 1.3740955793 1.9340958322 1.4712805103
          [6] 0.7161306762 0.4237291901 1.3510919146 0.6151980641 0.8941478838
         [11] 0.7982363194 1.6837249164 2.0832113736 0.6795290421 0.4889717087
        (omitted)
        [986] 1.4350018482 0.3481156595 0.1608004307 0.4821868723 1.4580077589
        [991] 1.3940858876 0.2553554193 1.0227372842 0.3240986542 0.8947329339
        [996] 0.8346843289 0.3152189133 0.8923909051 1.5413969143 0.2393887751
```

It can be sorted and reduce the decimals by using **mpfr** as given below, before using the **mpfr** it is to ensure that the availability of these packages such as "**gmp**" and "**rmpfr**"

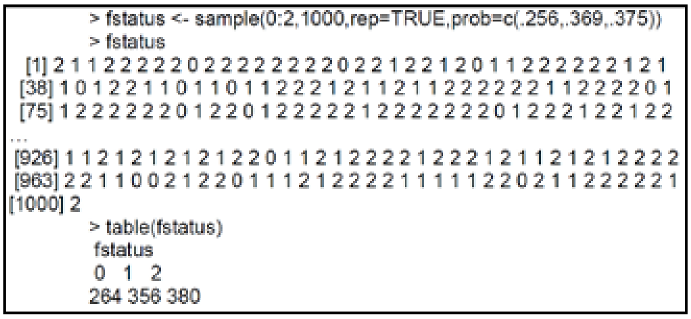
```
        > x<- mpfr(ftime,3)
        > x

        1000 'mpfr' numbers of precision  3   bits
 [1]    1.2   0.38    1.2     2   1.5   0.75   0.44    1.2   0.02
[10]   0.88   0.75    1.8     2  0.62    0.5   0.11      2   0.75
[19]   0.31      1    1.6 0.078   0.5  0.016   0.25   0.38   0.25
...
        > sort(x)

        1000 'mpfr' numbers of precision  3   bits
 [1] 0.00011 0.00011  0.0012  0.0015  0.0029  0.0029  0.0034  0.0039  0.0039
[10]  0.0049  0.0063  0.0068  0.0078  0.0078  0.0098   0.012   0.012   0.014
[19]   0.014   0.014   0.016   0.016    0.02    0.02    0.02    0.02   0.023
...
[982]    4    4    4    4    4    4    4    4    4
[991]    4    5    5    5    5    5    5    5    7
[1000]   8
```

The "sort" is used to identify the range of "ftime". The status command generated by default and resulted as:



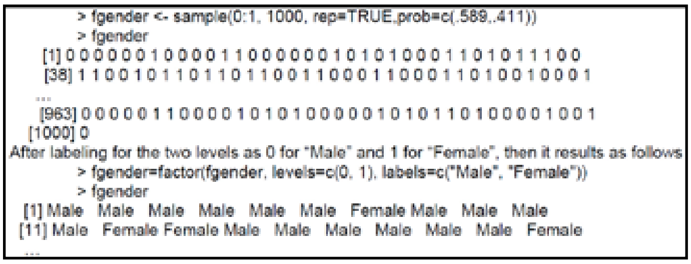
```
          > fstatus <- sample(0:2,1000,rep=TRUE)
          > table(fstatus)
fstatus
  0   1   2
328 309 363
```

Now, we need to generate a failure status (fstatus), either 1 or 2, for each of the 1000 event times. The following code generates 1000 observations, each of which decides on failure types with specified probability values for all possible causes including censoring cases is as follows:



```
          > fstatus <- sample(0:2,1000,rep=TRUE,prob=c(.256,.369,.375))
          > fstatus
  [1] 2 1 1 2 2 2 2 2 0 2 2 2 2 2 2 2 2 2 0 2 2 1 2 2 1 2 0 1 1 2 2 2 2 2 2 1 2 1
 [38] 1 0 1 2 2 1 1 0 1 1 0 1 1 2 2 2 1 2 1 1 2 1 1 2 2 2 2 2 2 1 1 2 2 2 2 0 1
 [75] 1 2 2 2 2 2 2 2 0 1 2 2 0 1 2 2 2 2 2 1 2 2 2 2 2 2 2 2 0 1 2 2 2 1 2 2 1 2 2
...
 [926] 1 1 2 1 2 1 2 1 2 1 2 2 0 1 1 2 1 2 2 2 2 1 2 2 2 1 2 1 1 2 1 2 1 2 2 2 2
 [963] 2 2 1 1 0 0 2 1 2 2 0 1 1 1 2 1 2 2 2 2 1 1 1 1 1 2 2 0 2 1 1 2 2 2 2 2 1
[1000] 2
          > table(fstatus)
           fstatus
           0   1   2
          264 356 380
```

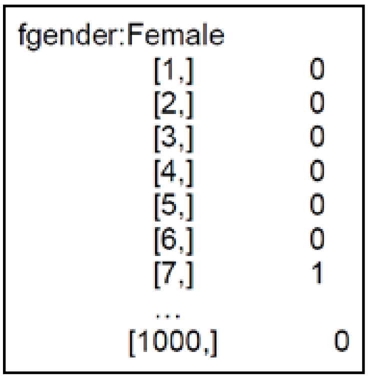
There ar e 356 events owing to the event of interest, 380 competing risk events and 264 censored individuals. The event times are denoted as ftime. The time variable ftime gives the distinct event times for all causes as well as censored cases. The status variable "fstatus" was created with specific "fstatus", as a function of the predictors x1, x2, x3 as continuous variables and “fgender” as a categorical variable. The “fgender” is to be created as a categorical variable and it has two levels male and female. The categorical variable “fgender” to be coded them numerically. Several coding of a factor is ‘baseline’ codification. For a factor or categorical variable made of N levels or categories, we must create N-1 indicator variables. The variable is coded as 1 in the presence of a given category and 0 otherwise. reason that the rate of censoring is not to exceed 40 percent, the rate for the first cause of the event of interest is approximately (0.4) and for the competing risk events the rate is (0.4). We aim to model the failure time “ftime”, with censoring and competing events provided by



```
        > fgender <- sample(0:1, 1000, rep=TRUE,prob=c(.589,.411))
        > fgender
  [1] 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 1 0 1 0 1 0 0 0 1 1 0 1 0 1 1 1 0 0
  [38] 1 1 0 0 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 0 0 1 1 0 0 0 1 1 0 1 0 0 1 0 0 0 1
 ...
 [963] 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 0 0 0 0 1 0 0 1
[1000] 0
After labeling for the two levels as 0 for "Male" and 1 for "Female", then it results as follows
        > fgender=factor(fgender, levels=c(0, 1), labels=c("Male", "Female"))
        > fgender
 [1] Male   Male   Male   Male   Male   Male   Female Male   Male   Male
 [11] Male   Female Female Male   Male   Male   Male   Male   Male   Female
 ...
```

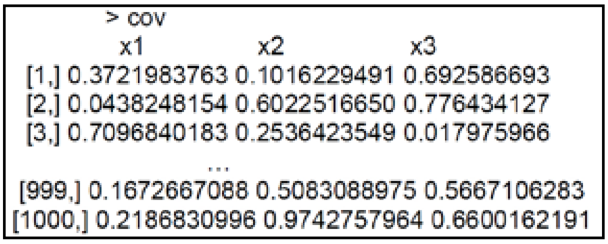
R function “factor2ind()”, which creates a matrix of indicator variables from a factor (Scrucca et al. (2010)). This is for to obtain the indicator variable for “fgender” using ‘Male’ as baseline we use:

> factor2ind(fgender, "Male") # it is required to run “crr-addsonR”: Refer (Scrucca et al). (2010)). The results will appear as below:



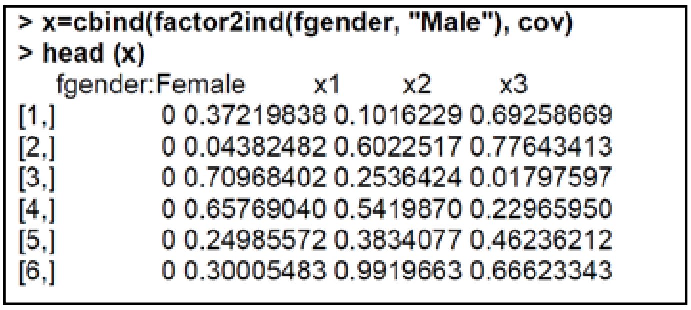
```
fgender:Female
          [1,]      0
          [2,]      0
          [3,]      0
          [4,]      0
          [5,]      0
          [6,]      0
          [7,]      1
          ...
       [1000,]      0
```

The runif is an arguments, generates random deviates using uniform distribution. The length of the result is determined by “n” for runif. The other three independent covariates to be generated using the following code. > cov <- matrix(runif(3000),nrow=1000)
. > cov <- matrix(runif(3000),nrow=1000)
> dimnames(cov)[[2]] <- c("x1","x2","x3")
Running the above codes and it gives us the desired 1000 independent observation on each variable under so called “x1”, “x2” and “x3”. We note that simulations are generated on a covariate data structure as given below:



```
      > cov
        x1            x2            x3
  [1,] 0.3721983763 0.1016229491 0.692586693
  [2,] 0.0438248154 0.6022516650 0.776434127
  [3,] 0.7096840183 0.2536423549 0.017975966
                      ...
 [999,] 0.1672667088 0.5083088975 0.5667106283
[1000,] 0.2186830996 0.9742757964 0.6600162191
```

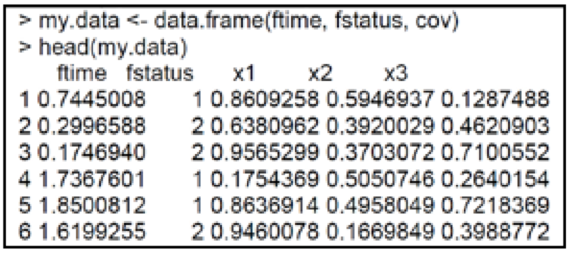
Now, we use the function cbind() to concatenate by columns all variables x1, x2,x3, and the indicator variable for fgender. The first rows of the design matrix are:



```
> x=cbind(factor2ind(fgender, "Male"), cov)
> head (x)
   fgender:Female        x1        x2         x3
[1,]            0 0.37219838 0.1016229 0.69258669
[2,]            0 0.04382482 0.6022517 0.77643413
[3,]            0 0.70968402 0.2536424 0.01797597
[4,]            0 0.65769040 0.5419870 0.22965950
[5,]            0 0.24985572 0.3834077 0.46236212
[6,]            0 0.30005483 0.9919663 0.66623343
```

III. Analysis of Simulated Competing Risk Survival Data

Simulation for ftime based on Weibull distribution and the related data combined with fstatus along with three independent covariates x1, x2 and x3 under the common file name called my.data



```
> my.data <- data.frame(ftime, fstatus, cov)
> head(my.data)
      ftime  fstatus        x1        x2        x3
1 0.7445008        1 0.8609258 0.5946937 0.1287488
2 0.2996588        2 0.6380962 0.3920029 0.4620903
3 0.1746940        2 0.9565299 0.3703072 0.7100552
4 1.7367601        1 0.1754369 0.5050746 0.2640154
5 1.8500812        1 0.8636914 0.4958049 0.7218369
6 1.6199255        2 0.9460078 0.1669849 0.3988772
```

Individual 1 experiences competing event 1 at time 0.7445008, individual 2 experiences competing event 2 at time 0.2996588, and so on.

We first estimate the cumulative incidence curve and competing risks regression models using the crr() which is contained in the cmprsk for model selection as well as comparison purpose. Fine and Gray (1999) and Grey (2010) proposed a model for the sub distribution hazard of the CIF with the sub distribution hazard as a key concept. We identify the event time and the censoring variable for competing risk as Surv(ftime,fstatus == 0). The regression model contains only an intercept term (+ 1). The fstatus variable gives the causes associated with the different events. Cause S = 1 specifies that we consider type 1 events, and the censoring code is given by the fstatus variable. The times at which the estimates are computed based on the argument times = ftime, the default is to use all cause “1” time points that are numerically stable.

> fgender=factor(fgender, levels=c(0, 1), labels=c("Male", "Female"))

> fstatus=factor(fstatus, levels=c(0,1,2), labels=c("censored","D eath","competing risk"))

A. Cumulative Incidence Function

The cumulative incidence curve estimations based on the cmprsk's CumIncidence() function . All the cumulative incidence curves are presented in Figure 1 a, b and c. Figure 1 (a) shows the cumulative incidence curves for fgender between two groups and ftime was simulated using Weibull distribution. Figure 1 (b) shows the cumulative incidence curves for fgender between two groups and ftime was simulated using exponential and figure 1(c) show the same concept but the time variable ftime was simulated using gamma distribution. The R packages etm (Allignol et al. 2011) and mstate (de Wreede et al. 2010, 2011) can also be used to compute the cumulative incidence curve with 95% confidence intervals

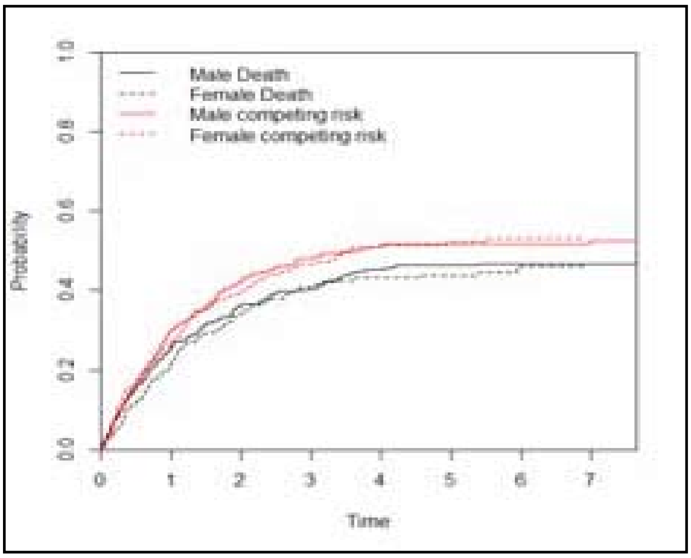


Fig. 1(a):

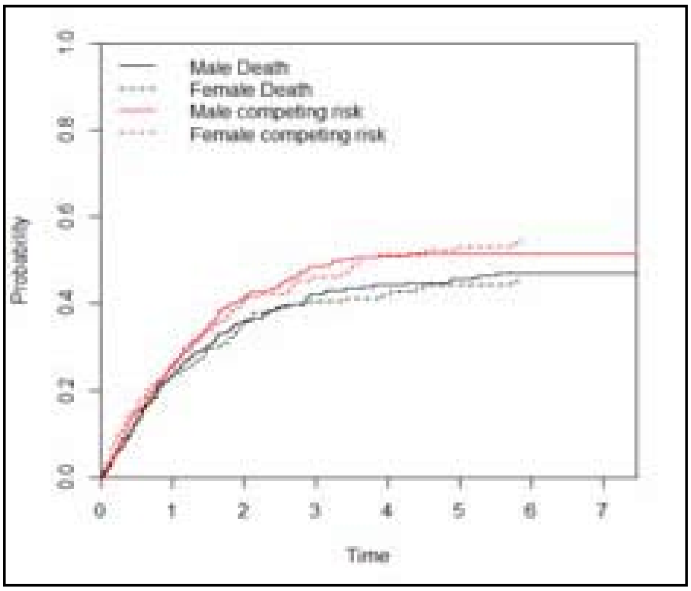


Fig. 1(b):

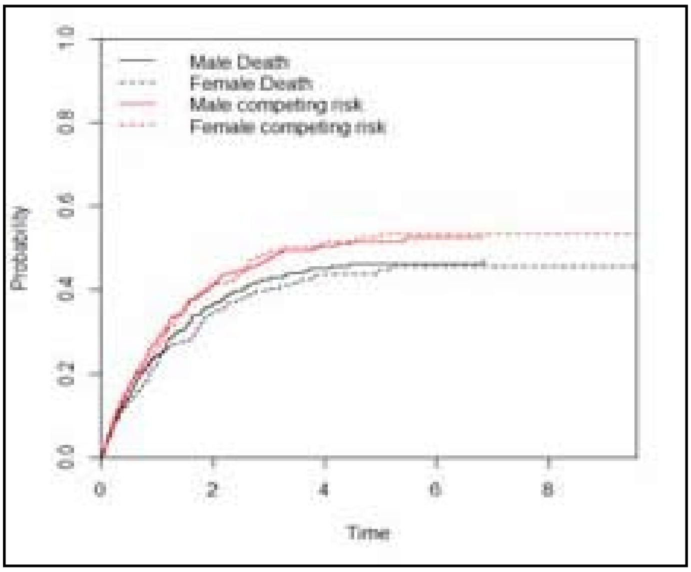


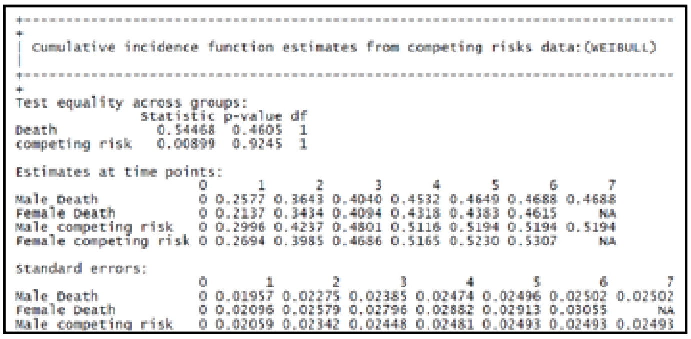
Fig. 1(c):

Fig. 1: Cumulative Incidence Curves Presented in fig. 1(a), (b) and (c) for **fgender** with three different parametric distributions namely Weibull, Exponential and Gamma as fig. 1(a), 1(b), 1(c) Respectively

Cumulative incidence function estimates from competing risks data for fgender using the function

CumIncidence. All three tables are given below as per the distribution which is shown in the above figure.

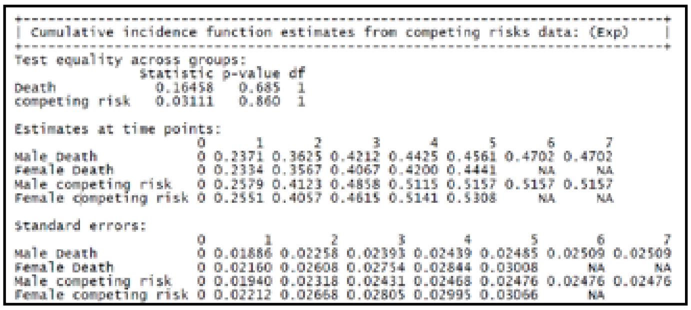
> fit=CumIncidence(ftime, fstatus, fgender, cencode="censored", xlab="Time")



```
+--------------------------------------------------------------------------
+
| Cumulative incidence function estimates from competing risks data:(WEIBULL)
|
+--------------------------------------------------------------------------
+
Test equality across groups:
               Statistic p-value df
Death            0.54468  0.4605  1
competing risk   0.00899  0.9245  1

Estimates at time points:
                      0      1      2      3      4      5      6      7
Male Death            0 0.2577 0.3643 0.4040 0.4532 0.4649 0.4688 0.4688
Female Death          0 0.2137 0.3434 0.4094 0.4318 0.4383 0.4615     NA
Male competing risk   0 0.2996 0.4237 0.4801 0.5116 0.5194 0.5194 0.5194
Female competing risk 0 0.2694 0.3985 0.4686 0.5165 0.5230 0.5307     NA

Standard errors:
                      0       1       2       3       4       5       6       7
Male Death            0 0.01957 0.02275 0.02385 0.02474 0.02496 0.02502 0.02502
Female Death          0 0.02096 0.02579 0.02796 0.02882 0.02913 0.03055      NA
Male competing risk   0 0.02059 0.02342 0.02448 0.02481 0.02493 0.02493 0.02493
```

> ftime <- rexp(1000, 0.9)

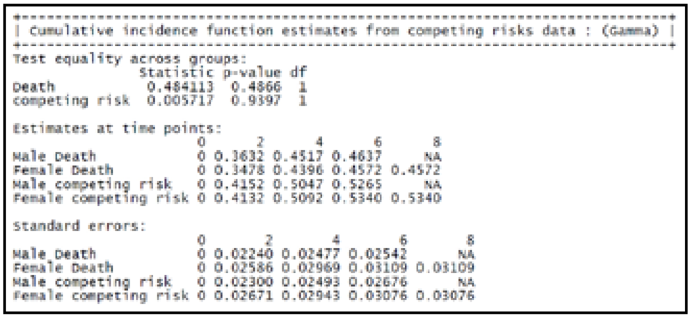
> fit=CumIncidence(ftime, fstatus, fgender, cencode="censored", xlab="Time")



```
+---------------------------------------------------------------------------+
| Cumulative incidence function estimates from competing risks data: (Exp)    |
+---------------------------------------------------------------------------+
Test equality across groups:
               Statistic p-value df
Death            0.16458   0.685  1
competing risk   0.03111   0.860  1

Estimates at time points:
                      0      1      2      3      4      5      6      7
Male Death            0 0.2371 0.3625 0.4212 0.4425 0.4561 0.4702 0.4702
Female Death          0 0.2334 0.3567 0.4067 0.4200 0.4441     NA     NA
Male competing risk   0 0.2579 0.4123 0.4858 0.5115 0.5157 0.5157 0.5157
Female competing risk 0 0.2551 0.4057 0.4615 0.5141 0.5308     NA     NA

Standard errors:
                      0       1       2       3       4       5       6       7
Male Death            0 0.01886 0.02258 0.02393 0.02439 0.02485 0.02509 0.02509
Female Death          0 0.02160 0.02608 0.02754 0.02844 0.03008      NA      NA
Male competing risk   0 0.01940 0.02318 0.02431 0.02468 0.02476 0.02476 0.02476
Female competing risk 0 0.02212 0.02668 0.02805 0.02995 0.03066      NA
```

> ftime<-rgamma(1000,shape=1)

> fit=CumIncidence(ftime, fstatus, fgender, cencode="censored", xlab="Time")



```
+---------------------------------------------------------------------------+
| Cumulative incidence function estimates from competing risks data : (Gamma) |
+---------------------------------------------------------------------------+
Test equality across groups:
               Statistic p-value df
Death           0.484113  0.4866  1
competing risk  0.005717  0.9397  1

Estimates at time points:
                      0      2      4      6      8
Male Death            0 0.3632 0.4517 0.4637     NA
Female Death          0 0.3478 0.4396 0.4572 0.4572
Male competing risk   0 0.4152 0.5047 0.5265     NA
Female competing risk 0 0.4132 0.5092 0.5340 0.5340

Standard errors:
                      0       2       4       6       8
Male Death            0 0.02240 0.02477 0.02542      NA
Female Death          0 0.02586 0.02969 0.03109 0.03109
Male competing risk   0 0.02300 0.02493 0.02676      NA
Female competing risk 0 0.02671 0.02943 0.03076 0.03076
```

B. Competing Risks Regression

The first competing risk regression model for a specified reason for competitive event can be produced by typing

> mod1=crr(ftime,fstatus,x,failcode="censored")

Recollect the source code defining categorical variable for "fgender" before executing the competing risks regression analysis as follows

> factor2ind (fgender,"Male") (#source code…..)

> x= cbind (factor2ind (fgender,"Male"), cov)

Male is the reference category

1. Model-1-Weibull

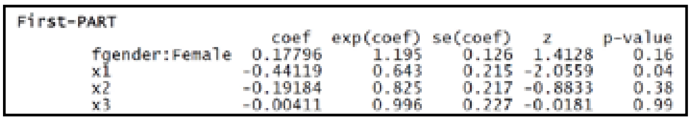
> mod1=crr(ftime,fstatus,x,failcode="censored")

> Summary (mod1)

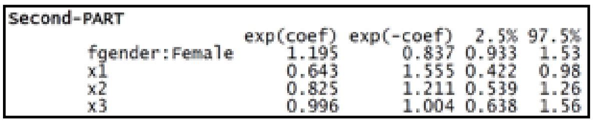
Competing Risks Regression

Call:

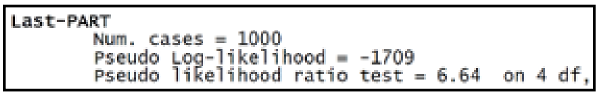
crr(ftime = ftime, fstatus = fstatus, cov1 = x, failcode = "censored")



```
First-PART
                      coef  exp(coef) se(coef)     z     p-value
     fgender:Female  0.17796    1.195    0.126  1.4128    0.16
     x1             -0.44119    0.643    0.215 -2.0559    0.04
     x2             -0.19184    0.825    0.217 -0.8833    0.38
     x3             -0.00411    0.996    0.227 -0.0181    0.99
```



```
Second-PART
                    exp(coef) exp(-coef)  2.5% 97.5%
     fgender:Female     1.195      0.837 0.933  1.53
     x1                 0.643      1.555 0.422  0.98
     x2                 0.825      1.211 0.539  1.26
     x3                 0.996      1.004 0.638  1.56
```

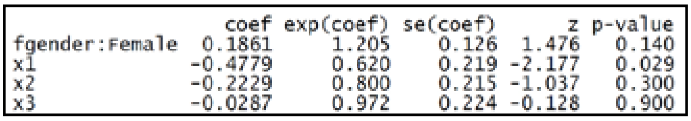


```
Last-PART
        Num. cases = 1000
        Pseudo Log-likelihood = -1709
        Pseudo likelihood ratio test = 6.64  on 4 df,
```

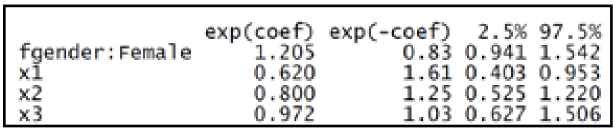
The "ftime" follows Weibull distribution. The output of Competing Risks Regression analysis consists of three parts. The first part of the output shows for each term in the design matrix the estimated coefficient $\hat{\beta}$, the relative risk exp ($\hat{\beta}$), the standard error, the z-value and the corresponding P-value for assessing significance. In this model, gender is not significant, followed by x2 and x3, whereas x1 is the only continuous significant. The second part of the output for competing risks regression shows the relative risk for each term, exp ($\hat{\beta}$), and a 95% confidence interval. The sub distribution hazard ratio for a categorical covariate is the ratio of sub distribution hazards for the actual group with respect to the baseline. If the covariate is continuous then the hazard risk refers to the effect of a one unit increase in the covariate, with all other covariates being equal. In our simulated data, exp (0.17796) = 1.195 is the risk of a female with respect to a male, and exp (-0.44119) = 0.643 is the risk for covariate x1, exp (-0.19184) = 0.825 is the relative for covariate x2 and exp (-0.00411) =0.996 is the risk for covariate x3. The last part of the output shows the pseudo log likelihood at maximum and the pseudo likelihood ratio test is based on the difference in the objective function at the global null and at the final estimates. Since this objective function is not a true likelihood, this test statistic is not asymptotically chisquare.

2. Model-2-Exponential

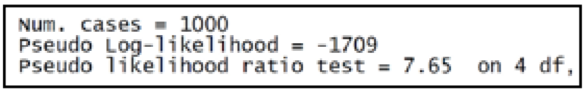
```
> mod2=crr (ftime,fstatus,x,
        failcode ="censored")
> summary(mod2)
Competing Risks Regression
Call: crr(ftime = ftime, fstatus = fstatus, cov1
                = x, failcode = "censored")
```



```
                  coef exp(coef) se(coef)      z p-value
fgender:Female  0.1861     1.205    0.126  1.476   0.140
x1             -0.4779     0.620    0.219 -2.177   0.029
x2             -0.2229     0.800    0.215 -1.037   0.300
x3             -0.0287     0.972    0.224 -0.128   0.900
```



```
               exp(coef) exp(-coef)  2.5% 97.5%
fgender:Female     1.205       0.83 0.941 1.542
x1                 0.620       1.61 0.403 0.953
x2                 0.800       1.25 0.525 1.220
x3                 0.972       1.03 0.627 1.506
```

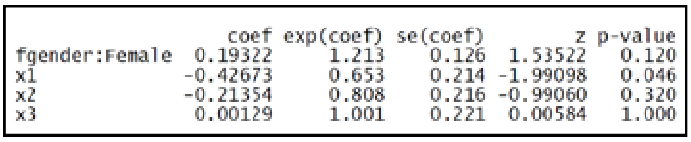


```
Num. cases = 1000
Pseudo Log-likelihood = -1709
Pseudo likelihood ratio test = 7.65  on 4 df,
```

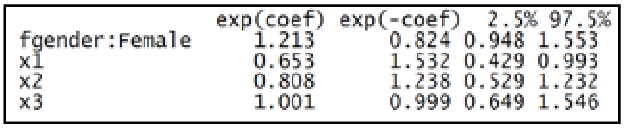
In our simulated data, exp (0.1861) = 1.205 is the risk of a female with respect to a male, and exp (-0.4779) = 0.620 is the relative risk for covariate x1, exp (-0.2229) = 0.800 is the risk for covariate x2 and exp (-0.0287) =0.972 is the risk for covariate x3.

3. Model-3-Gamma

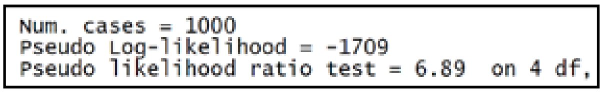
```
> mod3=crr(ftime,fstatus,x,
        failcode="censored")
> summary(mod3)
Competing Risks Regression Call:
crr(ftime = ftime, fstatus = fstatus, cov1 = x,
failcode = "censored")
```



```
                   coef exp(coef) se(coef)        z p-value
fgender:Female  0.19322     1.213    0.126  1.53522   0.120
x1             -0.42673     0.653    0.214 -1.99098   0.046
x2             -0.21354     0.808    0.216 -0.99060   0.320
x3              0.00129     1.001    0.221  0.00584   1.000
```



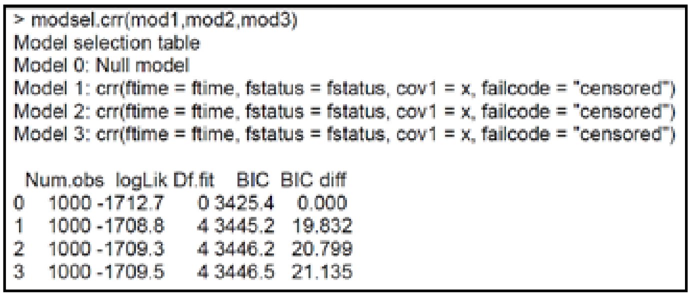
```
               exp(coef) exp(-coef)  2.5% 97.5%
fgender:Female     1.213      0.824 0.948 1.553
x1                 0.653      1.532 0.429 0.993
x2                 0.808      1.238 0.529 1.232
x3                 1.001      0.999 0.649 1.546
```



```
Num. cases = 1000
Pseudo Log-likelihood = -1709
Pseudo likelihood ratio test = 6.89  on 4 df,
```

IV. Model Selection

The likelihood of the data for a given model is a measure of the goodness of fit. However, the likelihood is increased when the number of parameters in the model is also increased and it leads over fitting. To avoid this over fitting, information criteria penalize the likelihood on the basis of the number of estimated parameters. Such criteria can be used for the selection of a model among a set of candidate models. Two of the most commonly used information criteria are the Akaike Information Criteria (AIC) (Akike, 1974) and the Bayesian information criteria (BIC) (Schwartz, 1978). The AIC is defined asAIC= -2l + 2d, where l is the maximized value of the loglikelihood for a given model and d is the number of free parameters to be estimated. For a regression model, d is usually equal to the number of estimated coefficients. Thus, AIC includes a penalty, which is an increasing function of the number of estimated parameters. In contrast, BIC is defined as BIC= - 2l + log(n)d where n is the number of observations. Both AIC and BIC are not executing to provide a test on the model in the sense of hypothesis testing, rather they provide a tool for ranking the competing models according to the criterion. In this data analysis, after fitted the preliminary model, It is realized that some covariates appeared not to be significant or only marginally significant, therefore these covariates removal from the model. This problem can be recast as a model selection problem using one of the information criterion concepts. By fitting a set of candidate models for which it pursues model selection after removal of all non significant and marginally significant covariates. The function modsel.crr() allows model selection on a list of models. It can be executed as follows:



```
> modsel.crr(mod1,mod2,mod3)
Model selection table
Model 0: Null model
Model 1: crr(ftime = ftime, fstatus = fstatus, cov1 = x, failcode = "censored")
Model 2: crr(ftime = ftime, fstatus = fstatus, cov1 = x, failcode = "censored")
Model 3: crr(ftime = ftime, fstatus = fstatus, cov1 = x, failcode = "censored")

  Num.obs  logLik Df.fit    BIC  BIC diff
0    1000 -1712.7      0 3425.4    0.000
1    1000 -1708.8      4 3445.2   19.832
2    1000 -1709.3      4 3446.2   20.799
3    1000 -1709.5      4 3446.5   21.135
```

For each model, it has included an argument in the call to the function; the output provides the sample size, the maximized loglikelihood, the number of estimated parameters (Df.fit), the BIC value and the BIC difference with respect to the minimum value observed from the set of candidate models. The null model which is labeled as Model 0 in the output is automatically included. Moreover this is the model with no covariates, so it serves as a reference for the inclusion of any of the available predictors. The smallest BIC value is achieved by the null model; all others are almost closely with each other. However Model1 is the next closest to the reference model.

V. Conclusion

This paper concludes how to perform a flexible competing risks regression analysis for simulated data using add-on packages available for the R statistical software. It presents the straightforward way to simulate this kind of competing risks survival data with different parametric distributions. This paper also presented a typical competing risks regression model analysis, in which the cumulative incidence of a specific reason in the presence of the competitive event. These models are useful for a detailed analysis of how covariate effects predict the cumulative incidence; the same was illustrated and demonstrated with how to fit these models in R.

VI. Acknowledgments

The authors gratefully acknowledge helpful feedback from the Reviewing Editor and all anonymous reviewers.

References

[1] Akaike, H.,"A new look at the statistical model identification. IEEE Trans Automat Contr; 19: pp. 716–723, 1974.

[2] Allignol, A., Schumacher, M., and Beyersmann, J. (2011). Empirical Transition Matrix of Multistate Models: The etm Package." Journal of Statistical Software, 38(4), 1{15. [Online] Availbale: http: //www.jstatsoft.org/v38/i04/.

[3] Bender, R., Augustin, T., Blettner, M.,"Generating survival times to simulate Cox proportional hazards models", Statistics in Medicine; 24, pp. 1713–1723.

[4] Beyersmann, J., Latouche, A., Buchholz, A., Schumacher, M., "Simulating Competing Risks Data in Survival Analysis", Statistics in Medicine; 28, pp. 956-971, 2009.

[5] Burton, A., Altman, D., Royston, P., Holder, R.,"The design of simulation studies in medical statistics", Statistics in Medicine; 25, pp. 4279– 4292, 2006.

[6] De Wreede, L.C., Fiocco, M., Putter, H.,"The mstate Package for Estimation and Prediction in Non- and Semi-Parametric Multi-

[7] State and Competing Risks Models", Computer Methods and Programs in Biomedicine, 99, pp. 261-274, 2010.

[8] Fine, J.P., Gray, R.J.,"A Proportional Hazards Model for the Subdistribution of a Competing Risk", Journal of the American

[9] Statistical Association, 94: pp. 496-509, 1999.

[10] Gray, R.J.,"A class of k-sample tests for comparing the cumulative incidence of a competing risk. Ann Stat; 16, pp. 1141–1154, 1988.

[11] Gray, R.J. (2010),"CMPRSK: Sub distribution Analysis of Competing Risks", R package version 2.2- 1, [Online] Available: http://CRAN.Rproject. org/package=cmprsk.

[12] Haller, B., Ulm, K.,"Flexible simulation of competing risks data following prespecified sub distribution hazards. Journal of Statistical Computation and Simulation, 2013.

[13] Schwartz, G.,"Estimating the dimension of a model", Ann Stat; 6: pp. 31–38, 1978.

[14] Scrucca, L., Santucci, A., Aversa, F.,"Competing risk analysis using R: An easy guide for clinicians", Bone Marrow Transplantation; 40, pp. 381–387, 2007.

[15] Scrucca, L., Santucci, A., Aversa, F.,"Regression Modeling of Competing Risk Using R: An in Depth Guide for Clinicians", Bone Marrow Transplantation; 45, pp. 1388-1395, 2010.

[16] Van Houwelingen, H.C., Putter, H.,"Dynamic predicting by landmarking as an alternative for multi-state modeling: An application to acute lymphoid leukemia data. Lifetime Data Anal 14, pp. 447-463, 2008.