Advanced Methods in Epidemiology

Practical Session on Competing Risks

Prof. dr. Steven Abrams

Master of Epidemiology | Academic year: 2022-2023

Background

In survival analysis, one is interested in studying the time to a specific event, potentially in the presence of censoring (or truncation). Censored observations arise when the event of interest has not yet occurred prior to, for example, the end of the study period (i.e., referred to as administrative right censoring). Other reasons for (right) censoring are individuals that are lost to follow-up before the event time could have been recorded or the occurrence of another event that prevents the one of interest to occur. In the latter case, the other event can be a competing one if individuals experiencing the competing event are different from those that do not with regard to the occurrence of the event of interest. For example, in clinical oncology cancer-related mortality may be of primary interest, but other causes of death can prevent its occurrence and deaths caused by reasons other than cancer are typical examples of competing risks.

Survival analysis in the presence of competing risks poses additional challenges for the correct estimation of the hazard function (i.e., the rate or instantaneous probability of experiencing the event of interest) and the cumulative incidence function (i.e., the risk of experiencing the event up to time t). In standard survival analysis, there exists a direct one-to-one relationship between the cumulative incidence function (CIF) and the hazard function (see below). However, such a relationship does not necessarily exists in a competing risks setting. In this practical session, we demonstrate the use of different R functions to estimate the CIF and to study the effect of covariates on either the cause-specific hazard functions and/or CIF. Both the use of the Cox proportional hazard model for the estimation of covariate effects on the cause-specific hazards as well as the Fine and Gray model for the direct estimation of such effects on the CIF are considered hereunder in more detail.

Key concepts in survival analysis

Survival data can be characterized by means of the so-called hazard function $\lambda(t)$ which provides a dynamic description of the instantaneous risk of failing given survival until time t. Alternatively, one could consider the cumulative hazard function $\Lambda(t)$, which in discrete time, simplifies to the sum of $\lambda(t)$ from time zero up to time t (in a finite number of discrete time steps). The cumulative hazard function is defined as the integral of the hazard function on the interval $[0, t]$ in continuous time. In the absence of competing risks, the specification of the (cumulative) hazard function defines the underlying time-to-event distribution uniquely (for a continuous random variable T^*), implying a one-to-one relation with the CIF or survival function (i.e., 1 - CIF) $S(t)$ defined as:

$$S(t) = \exp[\Lambda(t)] = \exp(-\int_0^t \lambda(u)du). \quad (1)$$

The survival function $S(t)$ can be estimated non-parametrically using the Product-Limit estimator, also and more commonly referred to as the *Kaplan-Meier* estimator. Alternatively, one can rely on the Nelson-Aalen estimator to estimate the cumulative hazard function $\Lambda(t)$.

The cumulative distribution function (or cumulative incidence function in a non-competing risk setting), characterizing the risk of experiencing the event of interest, is defined as

$$\text{CIF}(t) = 1 - S(t) = 1 - \exp[\Lambda(t)] = 1 - \exp(-\int_0^t \lambda(u)du). \tag{2}$$

In the presence of competing risks, however, the CIF cannot be directly linked to the hazard function $\lambda(t)$ anymore. Moreover, the marginal cumulative incidence estimated using the Kaplan-Meier estimator is always larger than the sum of the cause-specific cumulative incidences for the different competing risks. The estimated marginal cumulative incidence, by means of the Kaplan-Meier estimator, represents the probability of having experienced the event of interest at time t in a hypothetical world in which other competing events are absent, at least if censoring induced by competing events is independent from the event time process of interest. The latter assumption, however, is untestable based on the event time data at hand.

Data application

The data we consider to demonstrate competing risks analysis is contained in the *riskRegression* package in R. A total of 205 patients with melanoma were followed after a surgical operation until the end of 1977. The data can be retrieved as follows:

```
# Import the data
#-----------------
#install.packages("riskRegression")
library(riskRegression)
data(Melanoma)
```

1. Study the variables in the dataset.
2. Graphically depict the event time distribution for the event of interest (i.e., death as a result of melanoma) and distinguish between censored and uncensored observations.

Competing risks analysis - Combined analysis

A first approach towards analysing competing risks survival data is to combine all events. This combined analysis provides insights into the overall hazard function.

1. Estimate the overall cumulative incidence function using the Kaplan-Meier estimator for the survival function.
2. Interpret this result. What is the disadvantage of this approach?

Non-parametric estimation of the (cause-specific) hazard function

1. Estimate the marginal cumulative distribution function (CIF). More specifically, ignore the competing risk and consider such observations as being (right-)censored for the event of interest.
2. How do you interpret the aforementioned result?

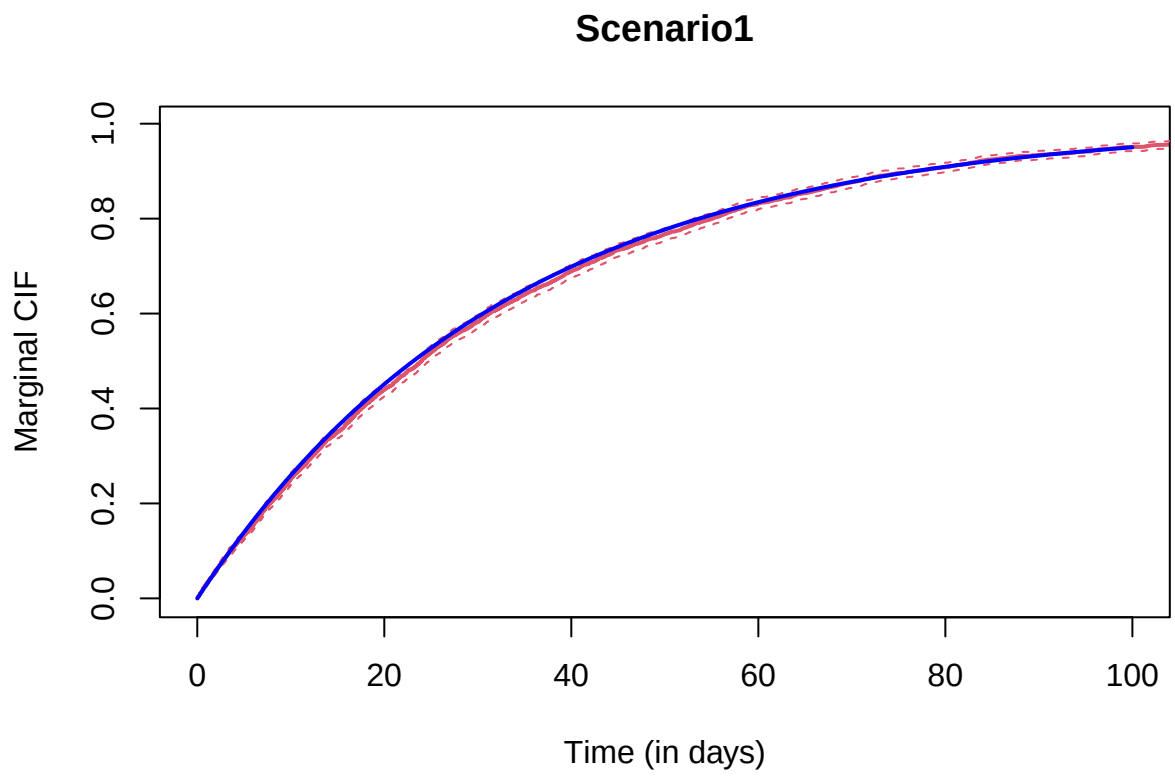
ADDITIONAL INFORMATION: Hereunder we demonstrate that bias is introduced in the estimation of the marginal CIF in case of dependence between event times T_1 and T_2. In the simulation approach considered below, we impose association using a copula approach (i.e., a copula is a function that can be used to join two marginal distributions into a joint bivariate distribution for two random variables) in which a single copula parameter measures the strength of association (theta) between the two random variables. The marginal distributions for T_1 and T_2 are assumed to be exponential with exponential rates $lambda_1$ and $lambda_2$. Independent censoring (exponential process with parameter cparam) is generated, for example, induced by administrative censoring. We consider different scenarios and investigate the impact on the estimated marginal CIF.

- In *scenario 1*, we assume that the rate of occurrence of event 1 is larger than the rate of occurrence of event 2 (the competing risk).
- In *scenario 2*, we assume that the rate of occurrence of event 1 is smaller than the rate of occurrence of event 2 (the competing risk).

```
library(copula)

# Generating dependent competing risks (T1, T2) and independent censoring time C
#-------------------------------------------------------------------------------
clayton.cmprsk.simul <- function(sample_size, hazard = "EXP",
                                          param1, param2, theta_parm,
                                          cens_hazard = "EXP", cparam, seednr){
  set.seed(seednr);
  random_samples <- rCopula(copula=claytonCopula(theta_parm, dim = 2), n=sample_size)
  u = random_samples[,1]; v = random_samples[,2];

  if (hazard == "EXP"){
    # message("Data generation: param-vector should include rate parameter");
    t1 = -log(1-u)/param1[1];
    t2 = -log(1-v)/param2[1];
  }

  if (hazard == "WEIBULL"){
    # message("Data generation: param-vector should include shape and decay parameters,
    # resp.");
    t1 = (-log(1-u)/(param1[2]))**(1/param1[1]);
    t2 = (-log(1-v)/(param2[2]))**(1/param2[1]);
  }

  if (cens_hazard == "EXP"){
    w = runif(sample_size, 0, 1);
    ct = -log(w)/cparam[1];
    status = rep(0, sample_size)
    for (i in 1:sample_size){
      if (t1[i] <= t2[i]){
        if (t1[i] <= ct[i]){
          status[i] = 1
        }
      }
      if (t1[i] > t2[i]){
        if (t2[i] <= ct[i]){
          status[i] = 2
        }
      }
```

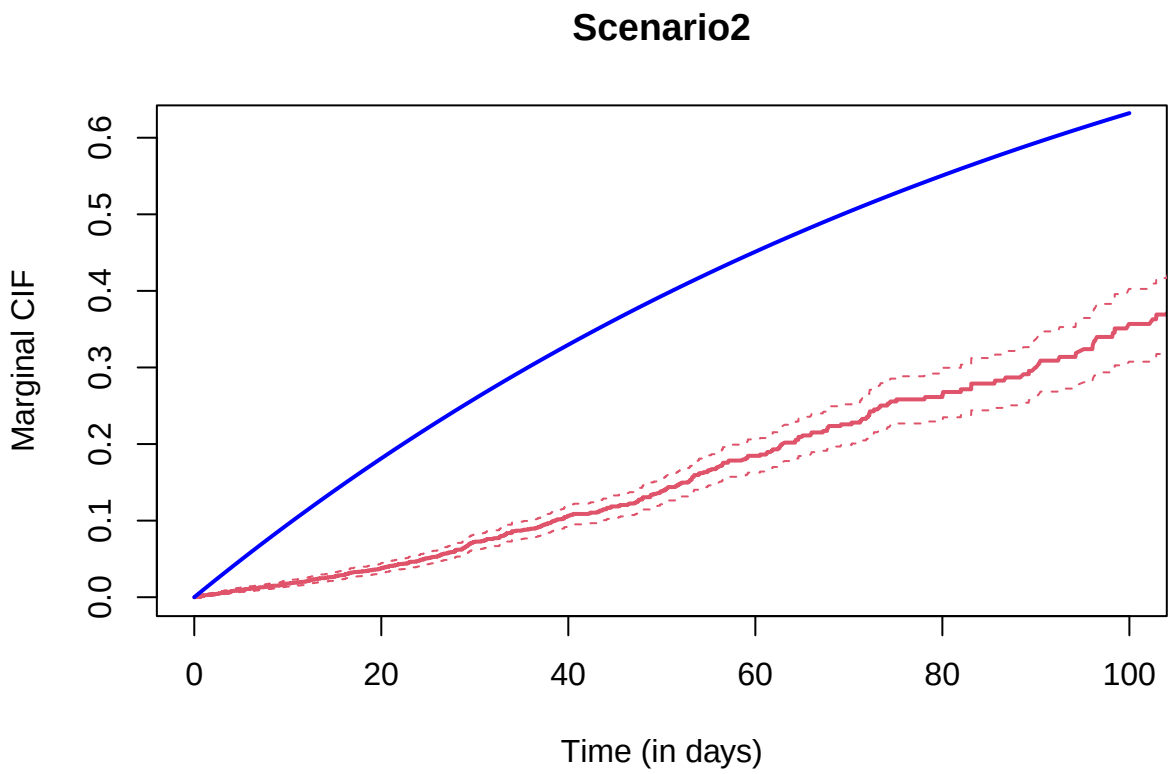
```
    }
    t = pmin(pmin(t1, t2), ct);
  }

  if (cens_hazard == "WEIBULL"){
    w = runif(sample_size, 0, 1)
    ct = (-log(w)/(cparam[2]))**(1/cparam[1]);
    status = rep(0, sample_size)
    for (i in 1:sample_size){
      if (t1[i] <= t2[i]){
        if (t1[i] <= ct[i]){
          status[i] = 1
        }
      }
      if (t1[i] > t2[i]){
        if (t2[i] <= ct[i]){
          status[i] = 2
        }
      }
    }
    t = pmin(pmin(t1, t2), ct);
  }

  simul.data = data.frame(t1 = t1, t2 = t2, ct = ct, t = t, status)
  return(simul.data)
}

# Scenario 1: T1 preceeds T2 most frequently
#-------------------------------------------
dat = clayton.cmprsk.simul(sample_size = 5000, hazard = "EXP",
                           param1 = 0.03, param2 = 0.01,
                           theta_parm = 2,
                           cens_hazard = "EXP",
                           cparam = 0.005,
                           seednr = 12345)
# head(dat)

# True marginal CIF
#------------------
lambda = 0.03
grid = seq(0,100,1)
true_cif = 1 - exp(-lambda*grid)

# Estimate the marginal CIF
#--------------------------
# library(survival)
surv_object <- Surv(time = dat$t,
                    event = as.numeric(dat$status == 1))

# plot(surv_object, main = "Kaplan-Meier estimator for S(t)",
#      xlab = "Time (in days)", ylab = "S(t)")

fit_sim1 <- survfit(surv_object ~ 1)
```

```r
par(mfrow = c(1,1))
plot(fit_sim1$time, 1-fit_sim1$surv, lwd = 2, col = 2, type = "l",
     main = "Scenario1",
     xlab = "Time (in days)", ylab = "Marginal CIF",
     xlim = c(0, 100))
lines(fit_sim1$time, 1-fit_sim1$lower, lty = 2, col = 2)
lines(fit_sim1$time, 1-fit_sim1$upper, lty = 2, col = 2)
lines(grid, true_cif, lwd = 2, col = "blue")
```




```r
## The estimated marginal CIF is close to the true marginal CIF (blue) due to
## the fact that P(T2 < T1) is small. Consequently, the event of interest T1
## is almost always directly observed and not preceded by the competing risk.

# Scenario 2: T2 preceeds T1 most frequently
#-------------------------------------------
dat = clayton.cmprsk.simul(sample_size = 5000, hazard = "EXP",
                           param1 = 0.01, param2 = 0.03,
                           theta_parm = 2,
                           cens_hazard = "EXP",
                           cparam = 0.005,
                           seednr = 12345)
# head(dat)

# True marginal CIF
#------------------
lambda = 0.01
grid = seq(0,100,1)
true_cif = 1 - exp(-lambda*grid)
```

```r
# Estimate the marginal CIF
#--------------------------
# library(survival)
surv_object <- Surv(time = dat$t,
                    event = as.numeric(dat$status == 1))

# plot(surv_object, main = "Kaplan-Meier estimator for S(t)",
#      xlab = "Time (in days)", ylab = "S(t)")

fit_sim2 <- survfit(surv_object ~ 1)

plot(fit_sim2$time, 1-fit_sim2$surv, lwd = 2, col = 2, type = "l",
     main = "Scenario2",
     xlab = "Time (in days)", ylab = "Marginal CIF",
     xlim = c(0, 100))
lines(fit_sim2$time, 1-fit_sim2$lower, lty = 2, col = 2)
lines(fit_sim2$time, 1-fit_sim2$upper, lty = 2, col = 2)
lines(grid, true_cif, lwd = 2, col = "blue")
```




```r
## The estimated marginal CIF provides a biased estimate of the true underlying
## marginal CIF (blue curve) due to the presence of the dependent competing risk.

# Scenario 3: Independence between T1 and T2
#-------------------------------------------

## How would the figure look like in case of independence between T1 and T2?
```

3. The cumulative incidence functions (CIFs) for different causes, i.e., cause-specific CIFs, are used for the statistical description of survival data with competing risks. In order to do so, the Kaplan-Meier (KM) estimator can be used in case there are no competing risks OR when competing risks are assumed to be

independent (see above). However, the KM method provides biased estimates in case the independence assumption is violated (i.e., the event time distributions for individuals that do experience or do not experience the competing event are different), hence, one should estimate cause-specific CIFs to gain additional insights into the nature of the survival data under study. The *cuminc()* function is the R package *cmprsk* can be used to estimate the cause-specific CIFs. Estimate the cause-specific CIF for death related to melanoma and for death related to another cause.

4. The *cuminc()* function also allows for the estimation of the cause-specific CIFs for different groups of patients, for example, males and females. Look in the documentation of the function to estimate the cause-specific CIFs for females and males separately.

NOTE: The following arguments need to be specified in cuminc(): (1) a failure time variable, (2) a variable with distinct code for different causes of failure and optionally (3) a group argument that takes a variable specifying distinct groups. In this example, patients are divided in two groups by sex (males and females).

5. Interpret the results in question Q4.

6. In order to perform a formal statistical test to check whether the difference between males and females is different, a modified χ^2-statistic can be considered (Gray, 1988). Perform this test (see documentation *cuminc()* function). Interpret the result.

Cause-specific hazard regression

A cause-specific hazard regression model is used to investigate the effect of different covariates directly on the cause-specific hazard function (i.e., the rate of occurrence of a specific cause/event). A cause-specific hazard regression model can be fitted using the standard Cox proportional hazards regression model while treating failures from the cause of interest as events and failure from other causes as censored observations. The effect of covariates on the cause-specific hazard can be estimated using the partial likelihood method as proposed by Cox. Such a model is fit using the *coxph()* function in the *survival* package in R.

1. Assess the impact of gender, age and invasion on the cause-specific hazard function for dying as a result of the melanoma.

NOTE: The first argument of the coxph() function takes an object of class Surv, where the "status==1" indicates that only status value of 1 is considered as event and other values are considered as censored. The ***summary output*** *shows the coefficients and corresponding hazard ratios (HRs). The last five lines display statistics for the fitness of the model.*

2. Interpret the output of the Cox proportional hazards model fitted in question Q1 above.

NOTE: Alternatively, cause-specific regression can also be performed using the CSC() function contained in the riskRegression package in R.

3. Interpret the results of the competing risk analysis obtained in Q2 above.

Subdistribution hazard regression

A subdistribution hazard regression model is also referred to as the Fine and Gray model. The motivation for the use of the Fine and Gray model is that the effect of a covariate on the cause-specific hazard function may be quite different from that on the cumulative incidence. In other words, a covariate may have strong influence on the cause-specific hazard, but no effect on the CIF. The difference between cause-specific hazard regression and subdistribution hazard regression is that both hazard functions rely on a different treatment of the competing risk events. The former considers competing risk events as non-informative censoring, whereas the latter takes into account the informative censoring nature of the competing risk events.

The Fine and Gray model can be fitted using the *FGR() function* in the *riskRegression* package in R.

1. Consider the Fine and Gray model for the example at hand. More specifically, fit a subdistribution hazards regression model with gender, age and invasion as covariates.
2. Interpret the findings in question Q1 above.
3. Discuss the differences between the cause-specific hazard regression and subdistribution hazard regression models in terms of interpretation and formulation.

References

- Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. Annals of Statistics, 16:1141-54.
- Zhang, Z. (2017). Survival analysis in the presence of competing risks. Annals of Translational Medicine, 5(3):47.