# Competing Risks

Germán Rodríguez
grodri@princeton.edu

Spring, 2001; revised Spring 2005

In this unit we consider the analysis of multiple causes of failure in the framework of competing risk models. An excellent reference on this material is Chapter 8 in Kalbfleisch and Prentice (2002), or Chapter 7 in the 1980 edition.

## 1 Introduction and Notation

Consider an example involving multiple causes of failure. Women who start using an intrauterine device (IUD) are subject to several risks, including accidental pregnancy, expulsion of the device, removal for medical reasons and removal for personal reasons. K-P discuss three areas of interest in the analysis of competing risks such as IUD discontinuation:

1 Studying the relationship between a vector of covariates $x$ and the rate of occurrence of specific types of failure; for example the covariates of IUD expulsion.

2 Analyzing whether people at high risk of one type of failure are also at high risk for others, even after controlling for covariates; for example are women who are at high-risk of expelling an IUD also at high risk of accidental pregnancy while wearing the device?

3 Estimating the risk of one type of failure after removing others; for example how long would we expect women to use an IUD if we could eliminate the risk of expulsion?

It turns out that we can answer the first of these questions, but the other two are essentially intractable. The third question can be answered under the strong assumption that the competing risks are independent, which essentially assumes away the second question.

We start by introducing some notation. Let $T$ be a continuous r.v. representing survival time. We assume that when failure occurs it may be one of $m$ distinct types indexed by $j \in \{1, 2, \ldots, m\}$, and we let $J$ be a r.v. representing the type of failure. Also, we let $x$ be a vector of covariates.

### 1.1 Cause-Specific Hazards

We define the overall hazard rate as usual:

$$\lambda(t, x) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt | T \geq t, x\}}{dt}.$$

We will also define a *cause-specific hazard rate*, representing the instantaneous risk of dying of cause $j$:

$$\lambda_j(t, x) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt, J = j | T \geq t, x\}}{dt}.$$

In words, we calculate the conditional probability that a subject with covariates $x$ dies in the interval $[t, t + dt)$ *and* the cause of death is the $j$-th cause, given that the subject was alive just before time $t$. We turn the probability into a rate dividing by $dt$ and then take the limit as $dt \to 0$.

By the law of total probability, we have

$$\lambda(t, x) = \sum_{j=1}^{m} \lambda_j(t, x),$$

because failure must be due to one (and only one) of the $m$ causes. If two types of failure can occur simultaneously we define the combination of the two as a new type of failure, so we can maintain this assumption.

### 1.2 Integrated Hazard and Survival

The overall survival function can be defined as usual:

$$S(t, x) = e^{-\Lambda(t,x)},$$

where $\Lambda(t, x)$ is the cumulative risk obtained by integrating the overall hazard

$$\Lambda(t, x) = \int_0^t \lambda(u, x) du.$$

We have assumed that the covariates are fixed to keep the notation simple. Extension to time-varying covariates is fairly straightforward, but calculation of the survival function requires specifying the trajectory of time-varying covariates.

The function $S(t, x)$ has a clear meaning as the probability of surviving *all* types of failure up to time $t$.

We will also define, by analogy with $S(t, x)$, the function

$$S_j(t, x) = e^{-\Lambda_j(t,x)},$$

where $\Lambda_j(t, x)$ is the integrated or cumulative hazard for case $j$;

$$\Lambda_j(t, x) = \int_0^t \lambda_j(u, x) du.$$

Note, however, that

**Note 1** *$S_j(t, x)$ will not, in general, have a survivor function interpretation if $m > 1$.*

### 1.3 Cause-Specific Densities

We can also define a *cause-specific density* of failures at time $t$, say

$$\begin{aligned} f_j(t, x) &= \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt, J = j|x\}}{dt} \\ &= \lambda_j(t, x) S(t, x). \end{aligned}$$

This density represents the unconditional risk that a subject dies at time $t$ of cause $j$. By the law of total probability, the overall density of deaths at time $t$ is

$$f(t, x) = \sum_{i=1}^{m} f_j(t, x).$$

## 2 Estimation: One Sample

Consider first the homogeneous case with no covariates.

### 2.1 Kaplan-Meier

The Kaplan-Meier estimator can easily be generalized to include competing risks. Let

$$t_{j1} < t_{j2} < \ldots < t_{jk_j}$$

denote the $k_j$ distinct failure times for failures of type $j$. Let $n_{ji}$ denote the number of subjects at risk just before $t_{ji}$ and let $d_{ji}$ denote the number of deaths due to cause $j$ at time $t_{ji}$. Then the same arguments used to derive the usual K-M estimator lead to

$$\hat{S}_j(t) = \prod_{i:t_{ji}<t} \left(1 - \frac{d_{ji}}{n_{ji}}\right).$$

It is interesting to note that $\hat{S}(t)$ is exactly the same as the standard K-M estimator that one would obtain if all failures of type other than $j$ were treated as censored cases.

If there are no ties between different types of failure, then

$$\hat{S}(t) = \prod_{j=1}^{m} \hat{S}_j(t),$$

so the K-M estimator of the overall survival is the product of the K-M estimators of the cause-specific survivor-like functions.

### 2.2 Nelson-Aalen

The Nelson-Aalen estimator of the cause-specific cumulative hazard is

$$\hat{\Lambda}_j(t) = \sum_{i:t_{ji}<t} \frac{d_{ji}}{n_{ji}},$$

and corresponds to an estimate of the cause-specific hazard $\lambda_j(t)$ that takes the value $d_{ji}/n_{ji}$ at $t_{ji}$ and 0 elsewhere. An alternative estimate based on K-M is

$$\hat{\Lambda}_j(t) = -\log \hat{S}_j(t).$$

Of course one could also exponentiate minus the Nelson-Aalen integrated hazard to obtain an alternative estimator of the cause-specific survivor-like function $S_j(t)$.

## 3 Estimation: Regression Models

Suppose we have $n$ observations consisting of four pieces of information each:

$$(t_i, d_i, j_i, x_i),$$

where $t_i$ is the observation time, $d_i$ is a death indicator (1 if dead, 0 if censored), $j_i$ is a cause of death index (takes a value between 1 and $m$ for deaths and is undefined for censored cases), and $x_i$ is a vector of covariates.

### 3.1 The Likelihood Function

Under non-informative censoring the likelihood function can be written as

$$L = \prod_{i=1}^{n} \lambda_{j_i}(t_i, x_i)^{d_i} S(t_i, x_i).$$

This likelihood is constructed in the usual manner. A subject censored at time $t_i$ contributes the probability of being alive at that time :

$$S(t_i, x_i).$$

A subject observed to die at time $t_i$ of cause $j_i$ contributes the density of deaths of that cause at that time $f_{j_i}(t_i, x_i)$, which can be written in terms of the hazard and survivor functions as

$$\lambda_{j_i}(t_i, x_i) S(t_i, x_i).$$

Introducing the indicator $d_i$ allows us to write the two types of terms in a compact way.

Recalling that $S(t_i, x_i) = \prod S_j(t_i, x_i)$, we can write the likelihood as

$$L = \prod_{i=1}^{n} \lambda_{j_i}(t_i, x_i)^{d_i} \prod_{j=1}^{m} e^{-\Lambda_j(t_i, x_i)}.$$

Let $d_{ij}$ indicate whether subject $i$ died of cause $j$. Clearly $d_i = \sum_j d_{ij}$, because you can die of at most one cause. We can then write

$$L = \prod_{i=1}^{m} \prod_{j=1}^{m} \lambda_j(t_i, x_i)^{d_{ij}} e^{-\Lambda_j(t_i, x_i)}.$$

Because the order in which we multiply is immaterial, we can state two important results:

**Note 2** *The overall likelihood function is a product of $m$ likelihoods, one for each type of failure.*

This means that we can estimate the $\lambda_j(t, x)$ by maximizing separate likelihoods, as longs as they do not depend on the same parameters. Moreover,

**Note 3** *The likelihood involving a specific type of failure is exactly the same likelihood you would obtain by treating all other types of failures as censored observations.*

In other words, each of these likelihoods has exactly the same form that we have studied before. Fitting models is thus a question of applying what we already know.

### 3.2 Weibull Regression

Suppose the $j$-th hazard function follows a proportional hazards model with Weibull baseline, say

$$\lambda_j(t, x) = \lambda_{j0}(t) e^{x'\beta},$$

where the baseline hazard is

$$\lambda_{j0}(t) = \lambda_j p_j (\lambda_j t)^{p_j - 1}.$$

In view of the above results, we can estimate the parameters $(p_j, \lambda_j, \beta_j)$ using the techniques discussed before, simply by treating failures for causes other than $j$ as censored cases.

Note that we have allowed *all* parameters to depend on the cause of death. We could, if we wanted, use different $x$'s for each type of failure.

If we wanted to restrict all the Weibulls to have the same index, for example, so $p_j = p \forall j$, then the overall likelihood function would not factor out and we would not be able to use this simplification. The same would be true if we wanted to force some $\beta$'s to be equal across causes (but why?). In either case one would have to maximize the full likelihood.

### 3.3 Cox Regression and Partial Likelihood

We can also fit a proportional hazards model without any assumptions about the baseline hazards $\lambda_{j0}(t)$. The standard Cox argument leads to a partial likelihood

$$L = \prod_{j=1}^{m} \prod_{i=1}^{k_j} \frac{e^{x'_{ji(j)}\beta_j}}{\sum_{k \in R(t_{ji})} e^{x'_{jk}\beta_j}},$$

where $k_j$ is the number of distinct times of death due to cause $j$, $t_{ji}$ denotes the $i$-th such time, $R(t_{ji})$ is the risk set at time $t_{ji}$ and $i(j)$ is the index of the case that died at $t_{ji}$. Again:

**Note 4** *The overall partial likelihood is a product of $m$ partial likelihoods, one for each type of failure, and each identical to the partial likelihood one would obtain by treating all other causes of death as censored cases.*

If you wanted to restrict the $m$ baseline hazards so they are in turn proportional to a super-baseline, say

$$\lambda_{j0}(t) = \lambda_0(t) e^{\gamma_j},$$

then a different partial likelihood would be obtained; see Equations 8.15-8.16 in K-P.

### 3.4 Piece-wise Exponential Survival

Here's my favorite model in the context of competing risks. Following the standard argument in Holford or Laird and Olivier, we define intervals with breakpoints $0 = \tau_1 < \tau_2 < \ldots < \tau_{k_1} = \infty$, and assume that the baseline hazard for the $j$-th type of failure is a step function with a constant value in each interval:

$$\lambda_{j0}(t) = \lambda_{jk}, \quad \text{for} \quad t \in [\tau_k, \tau_{k+1}).$$

Then the factor in the likelihood function corresponding to failures of type $j$ is identical to the kernel of a Poisson likelihood that treats the number of deaths of cause $j$ in interval $k$ to people with covariate values $x_i$ as Poisson with mean

$$\mu_{ijk} = E_{ik}\lambda_{jk}e^{x_i'\beta_j},$$

where $E_{ik}$ is the total exposure of people with covariates $x_i$ in interval $k$. (Note that the exposure is not cause-specific, at any time each subject is at risk of dying from any cause.)

Thus, we can fit competing risk models by running a series of Poisson regressions where we treat the number of deaths due to each cause as the outcome and the exposure to all causes as the offset.

A nice feature of this model concerns the conditional probability of dying due to cause $j$ at time $t$ given that the subject dies (of some cause) at time $t$. The probability of dying of cause $j$ at time $t$ given survival to just before $t$ is

$$\lambda_j(t, x) = \lambda_{j0}(t)e^{x'\beta_j} = e^{\alpha_{jk}+x'\beta_j},$$

say, where $\alpha_{jk} = \log \lambda_{jk}$ is the log baseline risk for cause $j$ in interval $k$. Now the overall risk at that instant is

$$\lambda(t, x) = \sum_{j=1}^{m} \lambda_j(t, x) = \sum_{j=1}^{m} e^{\alpha_{jk}+x'\beta_j}.$$

The conditional probability of interest can then be obtained as

$$\pi_{jk} = \frac{e^{\alpha_{jk}+x'\beta_j}}{\sum_{r=1}^{m} e^{\alpha_{rk}+x'\beta_r}},$$

and can be seen to follow a multinomial logit model with the same parameters $\beta_j$ as the competing risk model!

This means that we can break down the analysis of competing risks into two parts, using

1 a standard hazards model to get the overall risk, and

2 a multinomial logit model on cause of death.

The results would be exactly the same as fitting separate Poisson models to failures of each type.

# 4 The Identification Problem

So far we have focused on observable quantities. The literature on competing risks defines *latent* failure times

$$T_1, T_2, \ldots, T_m$$

where $T_j$ is the time when the subject would fail due to the $j$-th cause.

The problem, of course, is that we only observe the *shortest* of these, as well as an index which tells us which of the $T$'s we have observed. Formally, the data are realizations of two r.v.'s

$$\begin{aligned} T &= \min\{T_1, T_2, \ldots, T_m\} \\ J &= \{j : T_j \le T_k \forall\ k\} \end{aligned}$$

## 4.1 Multivariate and Marginal Survival

Let us introduce a joint survivor function, also called the multiple decrement function

$$S_M(t_1, t_2, \ldots, t_m) = \Pr\{T_1 \ge t_1; T_2 \ge t_2; \ldots; T_m \ge t_m\}.$$

Extension to covariates is trivial, so I will keep the notation simple by omitting them.

To be alive at time $t$ all of these potential failure times have to exceed $t$. This gives us a key identity relating the multivariate and marginal survival functions:

$$S(t) = S_M(t, t, \ldots, t).$$

This shows, incidentally, that $S(t)$ is well defined.

We can also define the cause-specific hazards in terms of partial derivatives of the log of the multivariate survival function:

$$\begin{aligned} \lambda_j(t) &= \lim_{dt \to 0} \frac{\Pr\{t \le T_j < t + dt | T > dt\}}{dt}, \\ &= -\frac{\partial}{\partial t_j} \log S_M(t, t, \ldots, t). \end{aligned}$$

The multivariate survival function is a function of $t_1, t_2, \ldots, t_m$. The last line refers to the log of the partial derivative w.r.t. $t_j$ evaluated at $t_1 = t_2 = \ldots = t_m = t$. The calculation is conditional on the overall survival time $T$ being at least $t$. As a result, the event $t \leq T_j < t + dt$ is equivalent to the events $t \leq T < t + dt$ and $J = j$. Thus, the result is the same as the cause-specific hazard introduced earlier.

**Note 5** *Because the likelihood of the data depends only on the cause-specific hazards $\lambda_j(t)$, it follows that only these hazards or functions of them can be estimated. Other quantities are not identifiable.*

For example the marginal distributions of the latent failure times are not identifiable. Let

$$\begin{aligned} S_j^*(t) &= \Pr\{T_j \geq t\} \\ &= S_M(0, \ldots, 0, t, 0, \ldots, 0), \end{aligned}$$

where the $t$ appears as the $j$-th argument to $S_M()$, denote the marginal distribution of $T_j$. From this marginal survival we can define a marginal hazard

$$\lambda_j^*(t) = -\frac{d}{dt} \log S_j^*(t).$$

This marginal hazard $\lambda_j^*(t)$ is not in general the same as the cause-specific hazard $\lambda_j(t)$. In fact, it cannot be written as a function of $\lambda_j(t)$ without further assumptions. It is therefore not identifiable.

Here comes the big exception. If the latent times $T_j$ are *mutually independent* then

$$S_M(t_1, \ldots, t_m) = \prod_{j=1}^{m} S_j^*(t_j).$$

It then follows that

$$S_j^*(t) = S_j(t),$$

so the marginal survival is the same as the cause-specific survivor-like function we introduced before, and

$$\lambda_j^*(t) = \lambda_j(t),$$

so the marginal hazard is the same as the cause-specific hazard (and is therefore identified).

There is a catch, however. Because the multivariate survival function cannot be estimated, the hypothesis of independence cannot be tested.

In other words, when all you observed is the minimum of the latent times, you cannot distinguish between independent competing risks and infinitely many dependent competing risks that produce exactly the same cause-specific hazards.

## 4.2 A Bivariate Example

In case you still believe that competing risks models are identified, here is a counter example. The following equation shows a bivariate survival model

$$S(t_1, t_2) = \exp\{1 - \alpha_1 t_1 - \alpha_2 t_2 - e^{\alpha_{12}(\alpha_1 t_1 + \alpha_2 t_2)}\},$$

where $\alpha_1, \alpha_2 > 0$ and $\alpha_{12}$ measures the dependence between $T_1$ and $T_2$.

Taking logs and differentiating w.r.t. $t_j$ we find the cause-specific hazards to be

$$\lambda_j(t) = \alpha_j(1 + \alpha_{12} e^{\alpha_{12}(\alpha_1+\alpha_2)t})$$

and it is clear that all three parameters can be estimated.

Consider, however, a model of independent competing risks, where the marginal (and cause-specific hazards) are given by the above equation. Integrating the marginal hazards we obtain the marginal cumulative hazards and exponentiating minus those gives the marginal survival functions. Multiplying the two survivor functions together we obtain the joint survivor function

$$S(t_1, t_2) = \exp\{1 - \alpha_1 t_1 - \alpha_2 t_2 - \frac{\alpha_1 e^{\alpha_{12}(\alpha_1+\alpha_2)t_1} + \alpha_2 e^{\alpha_{12}(\alpha_1+\alpha_2)t_2}}{\alpha_1 + \alpha_2}\},$$

and clearly $\alpha_{12}$ is not a measure of association because by construction $T_1$ and $T_2$ are independent!

The point here is that the two bivariate survivor functions are different—moreover, in one case the latent times are correlated while in the other they are independent—yet they lead to the same cause-specific hazards and thus have the same observable consequences.

Thus, if you use the first model and interpret $\alpha_{12}$ as a measure of association between the causes you are relying on untestable assumptions.

## 4.3 Discussion

The identification problem does not arise if one can observe more than one $T_j$, but this is usually not feasible. An exception is attrition in panel studies, where one can treat death and attrition as competing risks. It may be possible to have special follow-up studies of attriters to determine if death has occurred. Having both time to attrition and time to death allows estimation of the correlation between these outcomes.

Heckman has proposed identifying the marginal survival functions by introducing covariates that are supposed to affect one of the latent times but not the others. The problem, again, is that these assumptions themselves are not testable. You cannot check whether a covariate really has no effect on a given type of failure, you have to assume it.

Regrettably, this means that we cannot achieve objective 2 at all:

**Note 6** *Data on time to death and cause of death do not permit studying the relationship among failure modes, or even testing for independence.*

It also means that we can achieve the third objective in only a limited sense:

**Note 7** *We can only estimate survival following cause-removal under the untestable assumption that the competing risks are independent.*

Of course we are talking about independence given the observed covariates $x$, so if you have measured every conceivable covariate the assumption of independence would not be unreasonable.

A final note on terminology. The overall probability of failure due to cause $j$ in some interval $A$ is

$$\int_A \lambda_j(t,x)e^{-\Lambda(t,x)}dt;$$

the subject survives all causes up to time $t$, then dies of cause $j$.

The same probability if only cause $j$ was operating is, under the assumption of independence

$$\int_A \lambda_j(t,x)e^{-\Lambda_j(t,x)}dt;$$

the subject survives cause $j$ up to time $t$, then dies of cause $j$.

In the statistical literature these are called crude and net probabilities, respectively. The demographic literature is not consistent. To avoid confusion it is best to refer to the latter as cause-deleted. In this example all causes other than $j$ have been deleted.