The role of survival functions in competing risks

Núria Porta[(1)], Guadalupe Gómez[(1)] and M.Luz Calle[(2)]

[(1)] *Dept. of Statistics and Operations Research*
Universitat Politècnica de Catalunya, Barcelona (Spain)
[(2)] *Dept. of Systems Biology, Universitat de Vic, Vic (Spain)*

DR 2008/06
28 May 2008

Copies of this report may be downloaded at `http://www-eio.upc.es/~nporta`.

corresponding author: Núria Porta
Dept. Statistics and Operations Research,
UPC, Campus Nord, C5-224,
c. Jordi Girona 1-3, 08034 Barcelona,
tel. +34 934054095 fax +34 934015855
email: nuria.porta-bleda@upc.edu

The role of survival functions in competing risks[1]

N. Porta, G. Gómez and M.L. Calle

Abstract: *Competing risks data usually arises in studies in which the failure of an individual may be classified into one of k ($k > 1$) mutually exclusive causes of failure. When competing risks are present, there are two main differences with classical survival analysis: (i) survival functions are not mainly used to describe cause-specific failures and, (ii) classical estimation techniques may provide biased results. The main goal of this paper is to review, clarify and present the formulation of a competing risks model and the basic nonparametric estimation methods. We show why the use of survival functions in the competing risks framework may mislead the user, and we illustrate the presented methodologies by developing two examples from real data. The methods presented here can be implemented with several statistical packages, including R, SPSS and SAS: we give some highlights on how to perform a competing risks analysis with these software packages.*

Keywords: *Cause-specific hazard; cumulative incidence function, survival-like function.*

1 Introduction

Standard survival analysis endpoints measure the time that takes an individual to fail due to a particular event, measured from an origin of time. This time of interest is usually characterized by means of the hazard function, representing the rate of occurrence of the event at a given time t, but mostly via the survival function, representing the probability of *surviving* up to time t, that is, the probability that the event has not yet occurred before time t. In the presence of non-informative right censoring, and given a random sample of observed individuals, both functions are empirically estimable through consistent quantities such as the Nelson-Aalen estimator for the hazard function, or the Kaplan-Meier estimator for the survival function.

While in the classical setting there is a single cause for the event occurring, there are situations where several causes of failure are possible, but only the occurrence of the first of them can be observed. This situation is known as competing risks. As an example, consider an individual which is at risk of dying from cardiovascular disease (CV), which is the clinical endpoint of the study. The individual is also at risk of dying due to causes distinct from CV. In this setup dying from any other cause is a competing risk for dying due to CV. Often the competing risk event is ignored, treated as right-censored observation, and classical survival methods are used for inference. However, some standard methods, such as Kaplan-Meier, would provide biased estimators for the different probabilities of interest unless the competing risks event is independent of the main event of interest. Furthermore, as discussed in Tsiatis (1975), the independence between distinct causes of failure cannot be checked on the basis of the competing risks observed data.

Specific methods are hence needed. The distinguishing feature of a competing risks setting is that for each individual, besides a lifetime T, there is a cause of failure C. In this situation a joint model for T and C is needed, and their joint distribution can be completely specified through the cause-specific hazard, that is, the instantaneous risk of failing at a given time from a given cause, among all individuals at risk at that time. Given a random sample of competing risks data, cause-specific

[1]This work was partially supported by the grant 050831 from La Marató de TV3 Foundation and grant MTM2005-0886 of the Spanish Department for Science and Technology. Núria Porta is a recipient of a doctoral research fellowship from the Catalan Ministry of Innovation, Universities and Enterprise.

hazards are directly estimable. For each cause, if other failures are treated as censored observations, the Nelson-Aalen method provides a consistent estimate of the cause-specific hazards (Prentice *et al.*, 1978). The joint distribution can also be specified by means of the cumulative incidence functions, representing the probability of failing from a given cause before a specific time. In this case, all causes of failure are involved to estimate the cumulative incidence function of a given cause, and thus other failures cannot be treated as censored observations. The multiple decrements method proposed by Aalen (1978) provides an appropriate way to estimate the cumulative incidence functions, since the Kaplan-Meier method is not valid under competing risks (Pepe and Mori, 1993).

Two examples of competing risks data are used to illustrate the methodologies developed in this paper. The first example comes from industrial engineering reliability analysis, about life testing of a small electrical appliance during its development. For each unit, data consists of the number of cycles it takes the unit to failure or to removal from test, together with its failure code (Nelson, 1982). For this small appliance, the characterization and frequency of the most common causes would allow changes in design to develop a more reliable appliance. The second example correspond to a cohort of patients diagnosed with follicular cell lymphoma (Pintilie, 2006). The goal of the study was to assess the long-term outcome to the treatment given after diagnosis. The outcome included non-response to treatment, first relapse and disease-related death. Patients diagnosed between 1967 and 1996 were recorded in the study. This cohort experiences not disease-related deaths as a consequence of becoming older during the long follow-up. These deaths are considered as a competing risk to disease-related failures.

Both cause-specific hazards and cumulative incidence functions describe the time to the first failure T, and the cause of it, C. They must be interpreted taking into account the presence of other causes. For instance, the cause specific hazard for the j^{th} cause is not the risk of failing by cause j at t: it is the risk of failing *first* by cause j *than* by other causes. Therefore, these two functions cannot be interpreted as marginal functions, as if other causes of failure were absent. As a matter of fact, marginal probabilities are not estimable from observed competing risks data (Cox, 1959; Tsiatis, 1975). These are subtle issues that have confused many users when interpreting these functions, and efforts to clarify them are numerous in the literature, polemics included (Prentice *et al.*, 1978; Pepe and Mori, 1993; Gooley *et al.*, 1999; Llorca and Delgado-Rodríguez, 2004; Pintilie, 2007a; Wolbers and Koller, 2007; Latouche *et al.*, 2007; Pintilie, 2007b; Putter *et al.*, 2007). To further obscure the problem, a question naturally arises: why, in the competing risks setting, cumulative incidence functions are used to characterize the time of interest, instead of some kind of cause-specific *survival function* just as in classical survival analysis? In this paper we define three different survival-like functions and discuss their interpretation as well as the relations among them.

In this paper we review, clarify and present the formulation of a competing risks model and the basic nonparametric estimation methods. We show why the use of survival functions in the competing risks framework may mislead the user, and we illustrate the presented methodologies by developing two examples from real data. The methods presented here can be implemented with several statistical packages, including `R`, SPSS and SAS: we give some highlights on how to perform a competing risks analysis with these software packages. The paper is organized as follows: in Section 2 we develop the formulations of the competing risks model, pointing out the differences with classical survival analysis, with special interest in the interpretation of the functions involved and their characteristic features. A simulated illustration is presented to clarify the role of the functions of interest. In Section 3, the problem of estimating cause-specific hazards, cumulative incidence and survival-like functions is tackled, and results of the two real data examples are presented. Finally, Section 4 provides step-by-step guidelines on how to obtain such estimates using the statistical packages `R`, SPSS and SAS.

2 Competing risks formulation

In the following subsection, the competing risks model is specified through the characterization of the joint distribution of (T, C). In section 2.2, we will focus on distinct specifications of the survival function in the competing risks framework. A simulated illustration is developed in section 2.3 to clarify the role of the functions of interest.

2.1 Model specification

Define, for each individual, the pair (T, C), T being the failure time, and C the failure cause. T is assumed to be a continuous and positive random variable, and C takes values in the finite set $\{1, \ldots, k\}$. It is considered that the individual fails from one and only one cause. For instance, in the appliance data set, C takes values in $\{1, 2\}$, corresponding to mode 1 and mode 2 failures. In the follicular cell lymphoma data set also two causes of failure are possible, 1 for disease-related failures and 2 for deaths not disease-related. The joint distribution of (T, C) is completely specified through either the cause-specific hazards, $\lambda_j(t)$, or through the cumulative incidence functions, $F_j(t)$ (Lawless, 2003).

The cause-specific hazard function for the j^{th} cause is defined as

$$\lambda_j(t) = \lim_{\Delta t \to 0} \frac{Pr\big(T < t + \Delta t, C = j | T \geq t\big)}{\Delta t} \qquad j = 1, \ldots, k \tag{2.1}$$

and represents the rate of occurrence of the j^{th} failure. On the other hand, the cumulative incidence function from type j failure, also referred to as subdistribution function, is defined by

$$F_j(t) = Pr(T \leq t, C = j) \qquad j = 1, \ldots, k,$$

and corresponds to the probability of a subject failing from cause j in the presence of all the competing risks.

The total hazard $\lambda(t)$ and the overall survival function $S(t)$ for T are defined, respectively, in terms of the specific hazards as follows:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{Pr(T < t + \Delta t | T \geq t)}{\Delta t} = \sum_{j=1}^{k} \lambda_j(t),$$
$$S(t) = Pr(T > t) = e^{-\int_0^t \lambda(u)du} = e^{-\sum_{j=1}^{k} \int_0^t \lambda_j(u)du}.$$

The distribution function for T is obtained from the cumulative incidence functions through $F(t) = P(T \leq t) = \sum_{j=1}^{k} F_j(t)$. In addition, the subdensity functions $f_j(t)$ from cause j and the marginal distribution of C are respectively given by:

$$f_j(t) = \frac{d}{dt} F_j(t) = \lambda_j(t) S(t), \text{ and}$$
$$\pi_j(t) = Pr(C = j) = \lim_{t \to \infty} F_j(t) \qquad j = 1, \ldots, k.$$

The cumulative incidence function for cause j, $F_j(t)$, can be obtained as well from the cause specific hazard $\lambda_j(t)$ and the overall survival function $S(t)$ from the relationship:

$$F_j(t) = \int_0^t \lambda_j(u) S(u) du \qquad j = 1, \ldots, k. \tag{2.2}$$

The survival function can be factorized into the k following functions $S_j^*(t) = e^{-\int_0^t \lambda_j(u)du}$:

$$S(t) = e^{-\sum_{j=1}^k \int_0^t \lambda_j(u)du} = \prod_{j=1}^k e^{-\int_0^t \lambda_j(u)du} = \prod_{j=1}^k S_j^*(t). \tag{2.3}$$

Caution is needed when interpreting functions $S_j^*(t)$. Despite having the mathematical properties of continuous survivor functions, they are not the survivor functions of any observable random variable (Lawless, 2003). Moreover, $S_j^*(t) \neq 1 - F_j(t)$. In the next section, we will discuss the problem of interpreting these probabilities, among other survival-like functions which are easily defined and intuitively appealing.

2.2 Cause-specific survival-like functions

In the competing risks framework, compared to classical survival analysis where the survival function is often used to describe T, it seems odd to use the cumulative incidence function $F_j(t)$ instead of some type of *cause-specific survival function* for cause j. The first candidates would be the above-defined $S_j^*(t)$ functions. We have noted, though, that these functions do not have the usual meaning of a survival function in the classical approach. In addition, it can be seen that they do not correspond to the joint probability of failing from cause j after t, $P[T > t, C = j]$.

These considerations lead us to define two more functions that may play the role of cause-specific survivals. On one hand, we define

$$S_j(t) = 1 - F_j(t),$$

as the complement of the cumulative incidence function. On the other hand, and by analogy with the definition of $F_j(t)$, we define

$$\tilde{S}_j(t) = P[T > t, C = j].$$

These three functions $S_j^*(t)$, $S_j(t)$ and $\tilde{S}_j(t)$ are, for each j, survival-like functions, that is, functions which satisfy some of the mathematical properties of a survival function. In the following, we will deepen on their interpretation, we will see why they are not proper survival functions, and which is the relationship among them. To this aim it is worthwhile to keep in mind that a function $S(t)$ is a survival function if it is defined in $[0, \infty)$, it is non-negative and non-increasing, it is right-continuous, it satisfies $S(0) = 1$, and $\lim_{t\to\infty} S(t) = 0$. In addition, $S(t)$ is a survival function of a random variable T if $S(t) = P[T > t]$.

We will first focus in the interpretation of function $S_j(t) = 1 - F_j(t)$. It represents the probability of not failing from cause j before t. It is not a proper survivor function because

$$\lim_{t\to\infty} S_j(t) = 1 - \lim_{t\to\infty} F_j(t) = 1 - P[C = j],$$

which is strictly positive if there are at least two causes of failure. Moreover,

$$S_j(t) = 1 - F_j(t) = 1 - F(t) + \sum_{\ell\neq j} F_\ell(t) = S(t) + \sum_{\ell\neq j} P[T \leq t, C = \ell],$$

that is, it is the sum of the probability of having not failed for any cause by t plus the probability of having failed before t from other causes than j. Function $S_j(t)$ is the relevant piece to build the risk set for Fine and Gray's (1999) regression model for the cumulative incidence function.

The second survival-like function of interest, $\tilde{S}_j(t) = P[T > t, C = j]$, represents the probability of failing from cause j after t, and it is defined by analogy with the cumulative incidence function F_j. It is not a proper survivor function because

$$\tilde{S}_j(0) = P[C = j]$$

which is strictly below 1 if there are at least two causes of failure. The relationship with $F_j(t)$ is given by

$$\begin{aligned}\tilde{S}_j(t) &= P[T > t, C = j] = P[T > j|C = j]P[C = j] = (1 - P[T \leq t|C = j])\, P[C = j] \\ &= P[C = j] - P[T \leq t, C = j] = P[C = j] - F_j(t).\end{aligned}$$

Hence, it behaves like a complementary probability for $F_j(t)$, complementary on the probability of failing from cause j, $P[C = j]$. Note as well that the overall survival function $S(t)$ could be decomposed in terms of $\tilde{S}_j(t)$ as follows:

$$S(t) = 1 - F(t) = 1 - \sum_{j=1}^{k} F_j(t) = 1 - \sum_{j=1}^{k} P[C = j] + \sum_{j=1}^{k} P[T > t, C = j] = \sum_{j=1}^{k} \tilde{S}_j(t).$$

The expression of $S(t)$ as a sum of $\tilde{S}_j(t)$ is indeed different from the alternative decomposition $S(t) = \prod_{j=1}^{k} S_j^*(t)$ (see (2.3)), and shows that $\tilde{S}_j(t)$ and $S_j^*(t)$ are different. These functions are estimable consistently based on observed competing risks data, but in the literature they have been less used than their counterparts $F_j(t)$ (for example, as in Peterson, 1976).

The critical point about interpreting survival-like functions, though, correspond to functions $S_j^*(t)$. These functions, encountered in the factorization of the survival function $S(t)$ (2.3), correspond to the survival functions that would be obtained from the cause-specific hazard functions $\lambda_j(t)$, when failure times from other causes are treated as censoring observations. In doing so, the assumption of independence between failure time and censoring time is possibly violated. Thus, only when distinct causes of failure are assumed to be independent, $1 - S_j^*(t)$ is fully interpretable as the probability of failing from cause j if the other causes of failure were removed (Gooley *et al.*, 1999). Hence, only under independence, $S_j^*(t)$ and $S_j(t)$ are equal.

To clarify this interpretation, assume that in the appliance data, several testing procedures are performed in order to assess the reliability of different designs for the small electrical appliances. At a selected stage of the developing process, two causes of failure are possible, mode 1 and mode 2, so $S_j^*(t)$, $j = 1, 2$ can be estimated. Now assume that, given these estimates, a change in design would result in mode 1 being eliminated, and that it would not affect failures due to mode 2. Under this scenario, new appliances manufactured under the new design will fail only due to mode 2 with a probability of $1 - S_2^*(t)$. This probability is here fully interpretable as a marginal probability, though the underlying assumption was that mode 1 and mode 2 failures were independent. Otherwise, a change in design to remove mode 1 would have modified the probability of failing due to mode 2. In competing risks arising from medical studies, the inherit risk of failing from a specific cause of a given individual cannot be easily removed. Such an interpretation of $S_j^*(t)$ necessarily implies that being at risk of cause 1 is independent of being at risk of cause 2. Many situations arise, though, where such an assumption is untenable.

$S_j^*(t)$ can be estimated from observed data using the Kaplan-Meier methodology. The availability of software to obtain this estimate has lead to the incorrect use of $1 - S_j^*(t)$ to estimate the cumulative

probability of failure from type j, $F_j(t) = 1 - S_j(t)$. With this procedure, a biased estimate of $F_j(t)$ is obtained (Pepe and Mori, 1993). This is clear intuitively since $S_j^*(t)$ only depends on the cause-specific hazard $\lambda_j(t)$, whereas $F_j(t)$ depends on all cause-specific causes $\hat{\lambda}_\ell(t)$, $\ell \in \{1, \ldots, k\}$ through the survival function $S(t)$ (see 2.2). Specifically, an estimate of $1 - S_j^*(t)$ would overestimate $F_j(t)$ (Putter *et al.*, 2007). This is reasonable, because if an individual failing from other causes is treated as a censored observation, it is assumed that the individual will fail from the cause of interest j at any point in the future, which in some situations may be unfeasible: if an individual dies due to cancer, he/she would not certainly die (again) due to a heart attack. By censoring individuals, we expect a higher incidence of failures.

In effect, there always exist $s > 0$ such as $F_j(s) < 1 - S_j^*(s)$. Indeed, it always exists $\ell \neq j$ and $s > 0$ such as $\Lambda_\ell(s) = \int_0^s \lambda_\ell(u)du > 0$. That is, there exists at least one other cause of failure with at least one failure. Otherwise, there would not be competing risks in our data. Therefore, $\Lambda(s) = \sum_{m=0}^{k} \Lambda_m(s) > \Lambda_j(s)$. Being $g(u) = e^{-u}$ non-increasing and $\Lambda_j(u)$ non-negative,

$$S_j^*(s) = e^{-\Lambda_j(s)} > e^{-\Lambda(s)} = S(s),$$

and thus,

$$F_j(s) = \int_0^s S(u)\lambda_j(u)du < \int_0^s S_j^*(u)\lambda_j(u) = 1 - S_j^*(s).$$

2.3 Illustrating a competing risks model

An alternative model for the competing risks situation is the latent failure times approach. A failure time for each cause is defined, $T_1, T_2, \ldots, T_k$, and each represents the hypothetical failure time if other causes of failure were not present. Therefore, an underlying joint multivariate distribution $F(t_1, \ldots, t_k)$ rules the model. However, when all risks are present, only $T = \min(T_1, \ldots, T_k)$ together with $C = j$ with $T = T_j$ can be observed. Then, the joint distribution is not identifiable based uniquely on the observed data -the identifiability problem defined as such by Tsiatis (1975)-. Indeed, two distinct joint distributions $F_1(t_1, \ldots, t_k)$ and $F_2(t_1, \ldots, t_k)$ may result into the same marginal for (T, C). Only under strong assumptions such as independence the multivariate distribution is identifiable, but the independence assumption cannot be tested with the competing risks observed data. The latent failure approach, then, has little practical use. See Kalbfleisch and Prentice (2002), Lawless (2003) and Andersen *et al.* (2002), for example, for detailed discussion and further references on this issue.

To better comprehend the difficulties arising from a competing risks analysis, we present an illustration based on simulated competing risks data and we graphically represent the functions describing the joint distribution (T, C): the cause-specific hazards and the cumulative incidence functions. In addition, the marginal distributions of the latent times from which competing risks data is simulated are also plotted, in order to highlight the differences between the estimable and not estimable functions under a competing risks analysis. Finally, survival-like functions presented in the previous section are plotted to highlight their properties.

To obtain a competing risks model we first specify a bivariate time distribution model, which have been chosen following Klein and Moeschberger (1988) and Zheng and Klein (1995). Assume T_1 and T_2 are taken as Weibull distributions with shape parameters 2 and 4 respectively, and identical scale equal to 1. Denote by $H_1(t)$ and $H_2(t)$ their corresponding marginal survival functions, and define

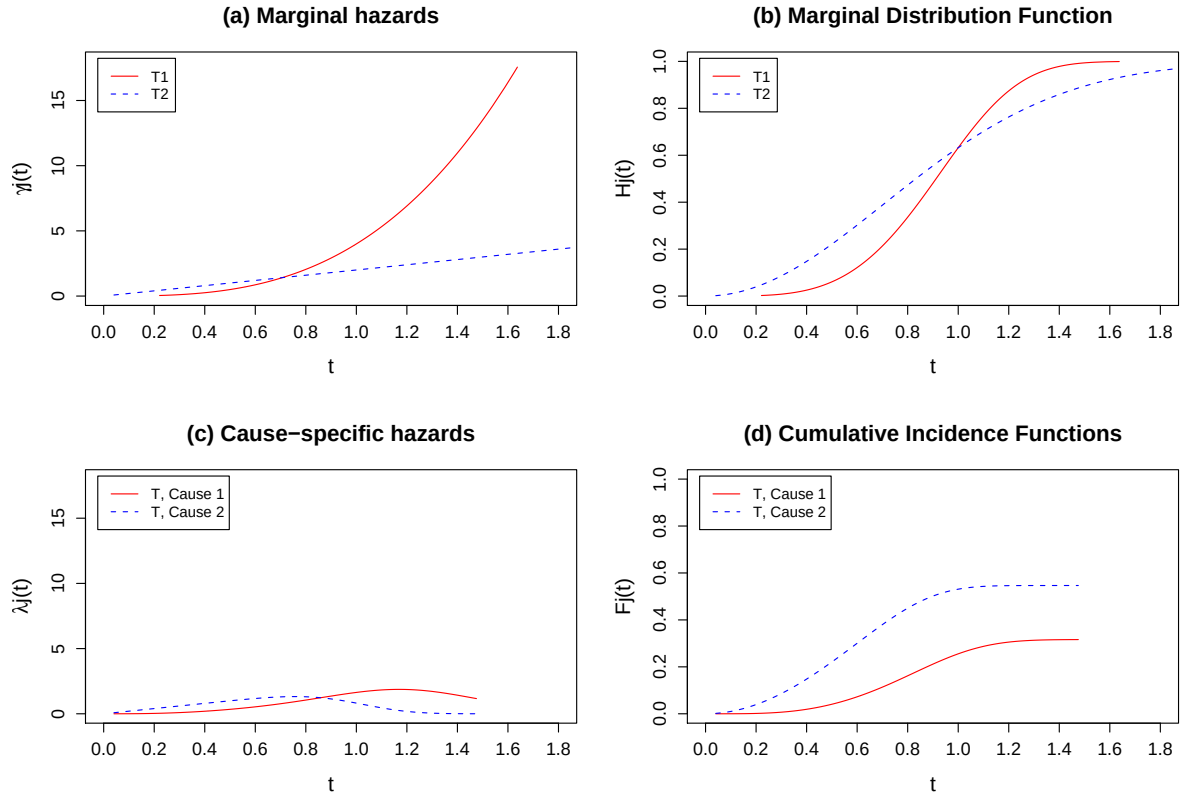

Figure 1: (a) Marginal hazard functions for T_1 and T_2 (not observed). (b) Marginal distribution functions for T_1 and T_2 (not observed). (c) Cause-specific hazards for cause 1 and cause 2 (observed). (d) Cause-specific cumulative incidence functions for cause 1 and cause 2 (observed).

the joint survival function $H(t_1, t_2)$ via a Clayton copula (Clayton, 1978) as follows:

$$H(t_1, t_2) = P(T_1 > t, T_2 > t) = \left[\left(\frac{1}{H_1(t_1)} \right)^{\theta-1} + \left(\frac{1}{H_2(t_2)} \right)^{\theta-1} - 1 \right]^{-\frac{1}{\theta-1}} \qquad \theta \geq 1, \qquad (2.4)$$

where $\theta \geq 1$ represents the positive association between T_1 and T_2. For the purpose of this illustration, θ is taken equal to 3, which corresponds to a value of 0.5 for Kendall's τ. Figure 1 (a) and (b) show the marginal hazard and distribution functions of T_1 and T_2, respectively.

In a competing risks framework, only the minimum between T_1 and T_2 would be observed, $T = \min(T_1, T_2)$ together with a label identifying the random variable that achieves the minimum, $C = j$ if $T = T_j$. To simulate competing risks data, it suffices to generate bivariate time data coming from model (2.4) and obtain (T, C) from these data. The relationship between the bivariate distribution and the competing risks model -for cause $j = 1$, for instance- is given by the following expressions:

$$S(t) = P[T > t] = P[\min(T_1, T_2) > t] = P[T_1 > t, T_2 > t] = H(t, t),$$

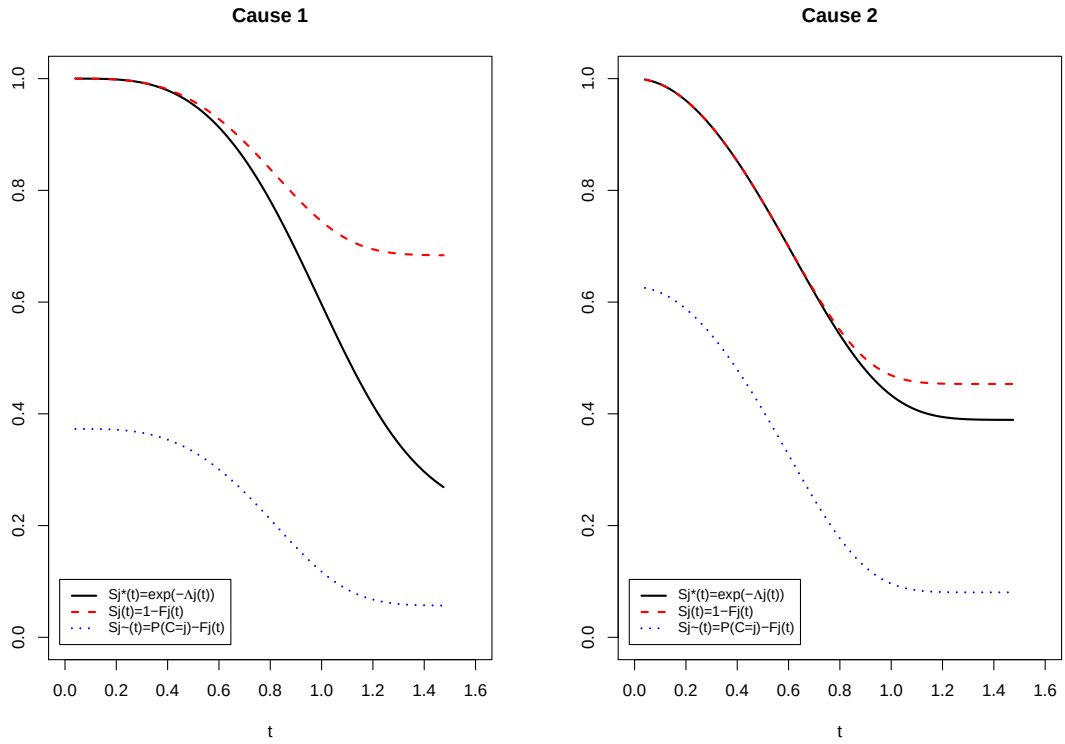

Figure 2: Survival-like functions for each cause of failure.

$$F_1(t) = P[T \leq t, C = 1] = P[T_1 \leq t, T_1 < T_2] = \int_0^t dt_1 \int_{t_1}^{\infty} \frac{\partial^2 H(t_1, t_2)}{\partial t_1 \partial t_2} dt_2$$

$$\lambda_1(t) = \frac{f_1(t)}{S(t)}$$

Cause-specific hazards and cumulative incidence functions for each of the two causes are depicted in Figure 1 (c) and (d), respectively, showing a difference between the competing risk model and the marginal distributions T_1 and T_2. It must be reminded here that in the presence of real competing risks data, marginal distributions cannot be identified, and thus never recovered, from observed data.

Figure 2 shows the behavior of the three survival-like functions $S_j(t)$, $\tilde{S}_j(t)$ and $S_j^*(t)$ defined in the previous section in the scenario simulated above. Some of their properties are clearly seen in the figure. For example, function $\tilde{S}_j(t)$, plotted with a dotted line, starts at 0 with value 0.4 for cause 1 and 0.6 for cause 2, representing the probabilities $P[C = 1]$ and $P[C = 2]$ respectively. Indeed, these are the proportion of each cause simulated from our data. On the other hand, $S_j(t) = 1 - F_j(t)$ -dashed line- is systematically over $S_j^*(t)$ -solid line-. Therefore, the Kaplan-Meier methodology providing an estimate for $1 - S_j^*(t)$ would systematically overestimate $F_j(t)$.

3 Estimation

In this section we first propose non-parametric estimators for all the functions introduced in section 2. Next we illustrate the methods for the appliance data set and the follicular cell lymphoma study, emphasizing the differences between cause-specific survival-like functions $S_j^*(t)$, S_j, and $\tilde{S}_j(t)$.

3.1 Nonparametric estimation

Consider a random sample of n individuals, $(T_1, C_1), \ldots, (T_n, C_n)$, where T_i is the time of failure and C_i is the cause of failure for subject i. For each individual, there exists a non-negative right censoring time V_i, independent of (T_i, C_i). Let $\delta_i = I(T_i \leq V_i)$ be the censoring indicator, and denote by $\tilde{C}_i = \delta_i C_i$ the cause of failure for failing individuals or 0 for censored individuals. The observed data for the i^{th} individual are given by

$$\{Y_i = \min(T_i, V_i), \delta_i, \tilde{C}_i\}.$$

Let $0 < y_1 < \cdots < y_N$ be the ordered distinct observed time points. We denote by d_{ij} the number of subjects failing from cause j at time y_i. The number of subjects failing at time y_i from any cause is obtained by the sum of subjects failing for each cause at y_i, $d_i = \sum_{j=1}^{k} d_{ij}$. Finally, we define n_i as the number of individuals at risk at y_i, that is, alive and uncensored just prior to this time. We note that $n_i = \sum_{\ell=1}^{n} I_\ell(y_i)$, where $I_\ell(y_i) = I(y_\ell \geq y_i)$ is an indicator function which takes value 1 if $y_\ell \geq y_i$, 0 otherwise.

An estimate of the cause-specific hazard for cause j (2.1) at time y_i is given by $\hat{\lambda}_j(y_i) = \frac{d_{ij}}{n_i}$, and it is 0 at any other time. Hence, the Nelson-Aalen estimator for the cumulative cause-specific hazard function, $\Lambda_j(t) = \int_0^t \lambda_j(u)du$, is given by

$$\hat{\Lambda}_j(t) = \sum_{i:y_i \leq t} \frac{d_{ij}}{n_i} \qquad j = 1, \ldots, k. \tag{3.1}$$

The overall survival function for T can be estimated either by the Kaplan-Meier estimate:

$$\hat{S}(t) = \prod_{i:y_i < t} \left(1 - \frac{d_i}{n_i}\right)^{\delta_i},$$

or as a function of the Nelson-Aalen estimate, that is, $\tilde{S}(t) = \exp\left[-\sum_{j=1}^{k} \hat{\Lambda}_j(t)\right]$.

A natural non-parametric estimate of the cumulative incidence function $F_j(t) = \int_0^t \lambda_j(u)S(u)du$ (2.2) is given by

$$\hat{F}_j(t) = \int_0^t \hat{\lambda}_j(u)\hat{S}(u)du = \sum_{i:y_i \leq t} \frac{d_{ij}}{n_i}\hat{S}(y_i^-) \qquad j = 1, \ldots, k, \tag{3.2}$$

where, given that $\hat{S}(t)$ is a step function jumping at y_i, $\hat{S}(y_i^-)$ is the value of $\hat{S}$ at the left limit of y_i. The probability of not failing from cause j before t, that is, $S_j(t) = 1 - F_j(t)$, is straightforwardly

estimated by $\hat{S}_j(t) = 1 - \hat{F}_j(t)$, with $\hat{F}_j$ given in (3.2). The probability of failing from cause j after time t, $\tilde{S}_j(t)$, is consistently estimated by

$$\hat{\tilde{S}}_j(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i > t, C_i = j) \qquad j = 1, \ldots, k \tag{3.3}$$

(Peterson, 1976). Finally, functions $S_j^*(t)$ are estimated using the Kaplan-Meier methodology restricted to specific failures for each cause treating failures from other causes as right-censored observations,

$$\hat{S}_j^*(t) = \prod_{i:y_i<t} \left(1 - \frac{d_{ji}}{n_i}\right)^{\delta_{ij}}.$$

In the following two sections, previous estimation procedures are illustrated by applying them to the appliance and the follicular cell lymphoma data sets.

3.2 Example: The appliance data

The appliance data set is described in Nelson (1982). The original experiment consisted of several life tests applied on 407 units to determine the number of cycles a unit would work until failure or removal from test, and its failure code. Units were tested at various stages in their development program, and divided by date of manufacture into five groups. Our illustration is restricted to 106 units corresponding to groups 2 and 3 for which a manual test was performed. For these groups, only six causes of failure exhibit. For the purpose of our example we distinguish the most frequent cause 11 labeled as Mode 2 from the other 5 causes grouped together as Mode 1.

Table 1 presents the nonparametric estimates proposed in section 3.1 for the appliance data set. The first column contains a subset of observed time-cycles at which a failure of either type has occurred, y_j. The first failure occurred in the 45^{th} time-cycle, and the last event in the time-cycle 1198. The second and third column show, respectively, the number n_j of individuals at risk of failing right before time y_j, and the number d_j of individuals failing at y_j. The fourth column represents the estimation of the overall survival, that is, the survival function of the time T to the first event happening. The next three columns summarize the number of individuals failing d_{1j}, the estimated cause-specific hazards $\hat{\lambda}_1(y_j)$, and the cumulative incidence function $\hat{F}_1(y_j)$ at time y_j for Mode 1. The last three columns summarizes the same quantities for Mode 2. Figure 3 graphically represents the estimated cumulative hazard functions and the estimated cumulative incidence functions for each cause. It can be observed that the behavior of both failure modes is similar up to 400 time-cycles approximately. From this point on, mode 2 failures are more frequent.

Results for the estimation of survival-like functions $S_j(t)$, $S_j^*(t)$ and $\tilde{S}_j(t)$ are shown in Table 2. For each mode of failure, the table contains the number of failures at a given time and estimates for the three survival-like functions. They are also plotted in Figure 4. Characteristic features of these survival-like functions, as described in section 2.2, are distinguished in this example. For example, $\tilde{S}_j(0)$ equals to the proportion of failures due to cause j. In our data set, 12 units fail due to mode 1, thus $12/106 = 0.1132$, which is exactly the value of $\tilde{S}_1(0)$ (see the first row of Table 2). There are 34 mode 2 failures, therefore, $\tilde{S}_2(0) = 34/106 = 0.3208$. Notice, in addition, that $S_j(t)$ is slightly greater than $S_j^*(t)$ for all t, confirming the existent bias in the Kaplan-Meier estimation in the presence of competing risks.

Time	**No. at risk**	**Total no. of failures**	**Estimated overall survival**	**Mode 1**			**Mode2**		
				No. of failures	**Estimated failure rate**	**Estimated cumulative incidence**	**No. of failures**	**Estimated failure rate**	**Estimated cumulative incidence**
y_j	n_j	d_j	$\hat{S}(y_j)$	d_{1j}	$\hat{\lambda}_1(y_j)$	$\hat{F}_1(y_j)$	d_{2j}	$\hat{\lambda}_2(y_j)$	$\hat{F}_2(y_j)$
0	106	0	1.0000	0	0.0000	0.0000	0	0.0000	0.00006
45	106	1	0.9906	1	0.0094	0.0094	0	0.0000	0.0000
73	104	1	0.9717	0	0.0000	0.0094	1	0.0096	0.0189
190	94	1	0.9143	1	0.0106	0.0381	0	0.0000	0.0476
241	89	1	0.9040	1	0.0112	0.0484	0	0.0000	0.0476
281	87	1	0.8835	1	0.0115	0.0587	0	0.0000	0.0578
410	82	3	0.8314	1	0.0122	0.0795	2	0.0244	0.0892
485	78	1	0.8103	0	0.0000	0.0900	1	0.0128	0.0997
571	72	1	0.7671	0	0.0000	0.1008	1	0.0139	0.1321
635	47	1	0.6610	0	0.0000	0.1008	1	0.0213	0.2382
658	46	2	0.6323	1	0.0217	0.1152	1	0.0217	0.2525
838	23	1	0.5338	0	0.0000	0.1152	1	0.0435	0.3511
1198	5	2	0.2475	1	0.2000	0.1977	1	0.2000	0.5549
1300	3	0	0.2475	0	0.0000	0.1977	0	0.0000	0.5549

Table 1: Non-parametric estimated survival function, cause-specific hazards and cumulative incidence functions for the appliance data.

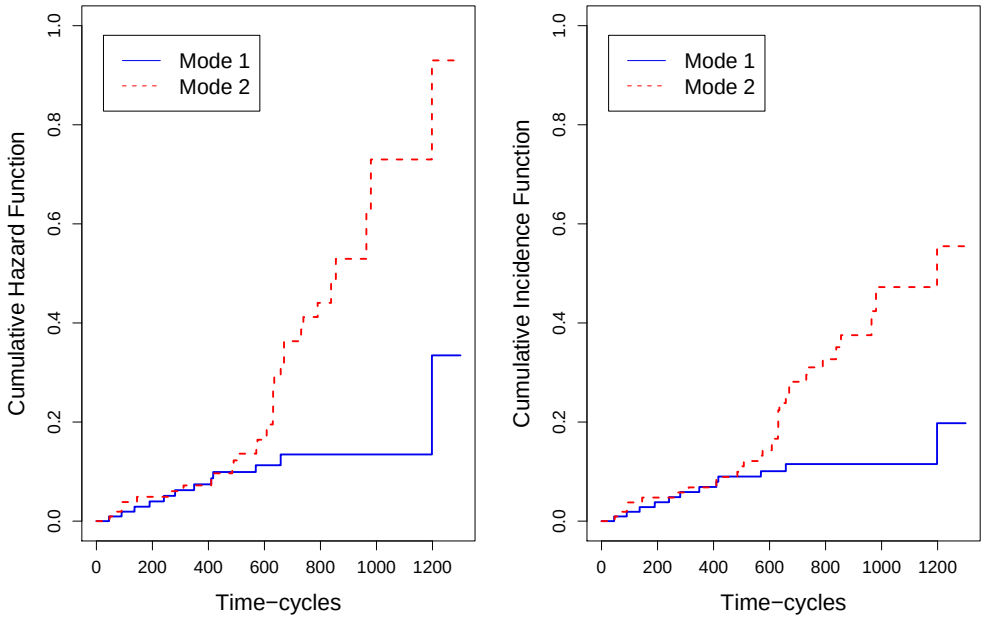

Figure 3: Appliance data set: (a) Cause-specific cumulative hazard functions. (b) Cumulative incidence functions.

	Mode 1				**Mode2**			
y_j	d_{1j}	$\hat{S}_1(y_j)$	$\tilde{\hat{S}}_1(y_j)$	$\hat{S}_1^*(y_j)$	d_{2j}	$\hat{S}_2(y_j)$	$\tilde{\hat{S}}_2(y_j)$	$\hat{S}_2^*(y_j)$
0	0	1.0000	0.1132	1.0000	0	1.0000	0.3208	1.0000
45	1	0.9906	0.1038	0.9906	0	1.0000	0.3208	1.0000
73	0	0.9906	0.1038	0.9906	1	0.9811	0.3019	0.9810
190	1	0.9619	0.0755	0.9608	0	0.9524	0.2736	0.9518
241	1	0.9516	0.0660	0.9500	0	0.9524	0.2736	0.9518
281	1	0.9413	0.0566	0.9391	0	0.9422	0.2642	0.9410
410	1	0.9205	0.0377	0.9167	2	0.9108	0.2358	0.9073
485	0	0.9100	0.0283	0.9051	1	0.9003	0.2264	0.8957
571	0	0.8992	0.0189	0.8927	1	0.8679	0.1981	0.8597
635	0	0.8992	0.0189	0.8927	1	0.7618	0.1226	0.7408
658	1	0.8848	0.0094	0.8733	1	0.7475	0.1132	0.7247
838	0	0.8848	0.0094	0.8733	1	0.6489	0.0566	0.6118
1198	1	0.8023	0.0000	0.6987	1	0.4451	0.0000	0.3782
1300	0	0.8023	0.0000	0.6987	0	0.4451	0.0000	0.3782

Table 2: Estimates of survival-like functions $S_j(t)$, $\tilde{S}_j(t)$ and $S_j^*(t)$ in the appliance data.

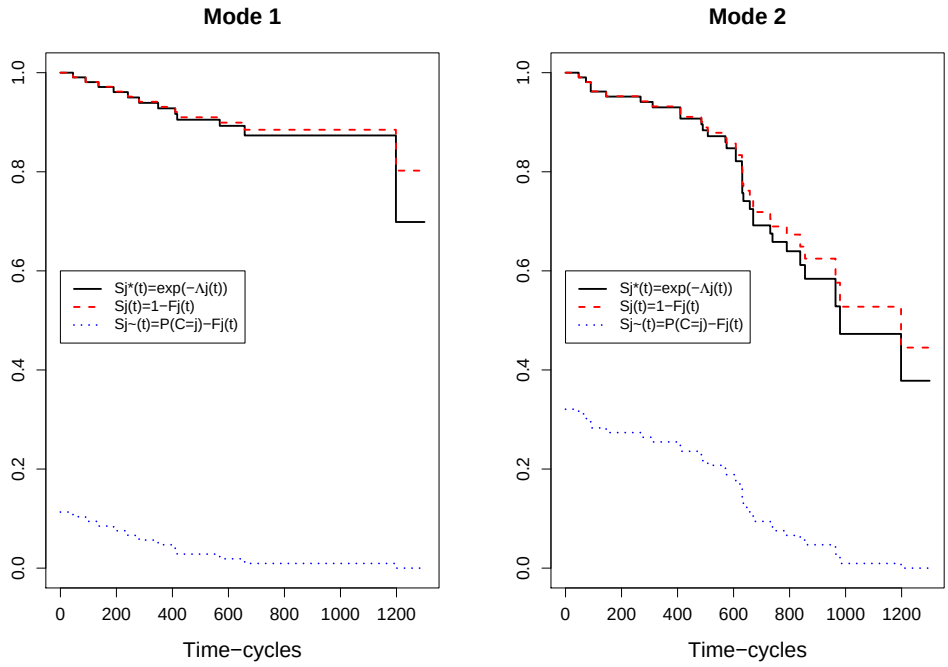

Figure 4: Survival-like functions in the appliance data.

3.3 Example: The follicular cell lymphoma study

The follicular cell lymphoma data set is found as an example in Pintilie (2006), and it is available at `http://www.uhnresearch.ca/hypoxia/People_Pintilie.htm`. The database was created at the Princess Margaret Hospital in Toronto, and contains data from 541 patients having follicular cell type lymphoma registered at the hospital between 1967 and 1996, with early stage disease (I or II), and treated with radiation alone (RT) or with radiation and chemotherapy (CMT). The goal of this study was to report the long-term outcome in this group of patients, including the following disease-related events: non-response to treatment, relapse and death. Time to the first failure is computed in years from the date of diagnosis. In addition, deaths not disease-related are frequent in this cohort, and must be considered as a competing event to disease-related failures.

Table 3 contains the non parametric estimates for the overall survival function, representing the probability of surviving without failures at each time point, as well as the cause-specific estimates for the hazard and the cumulative incidence functions. From Figure 5 it seems clear that disease failures are more frequent than non disease-related deaths, mostly for early times, but after 20 years of following, the risk of dying increases and deaths are more frequent.

Figure 6 represents the survival-like functions S_j, $\tilde{S}_j$ and S_j^* and punctual estimates can be found in Table 4. S_j^* -estimated by the Kaplan-Meier methodology- and $\tilde{S}_j = 1 - F_j$ -correctly estimated by the multiple decrements method- are remarkably different in this example, showing how the Kaplan-Meier method would overestimate the risk of failing due to the disease.

Time	No. at risk	Total no. of failures	Estimated overall survival	Disease Failure			Death without disease failure		
				No. of failures	Estimated failure rate	Estimated cumulative incidence	No. of failures	Estimated failure rate	Estimated cumulative incidence
y_j	n_j	d_j	$\hat{S}(y_j)$	d_{1j}	$\hat{\lambda}_1(y_j)$	$\hat{F}_1(y_j)$	d_{2j}	$\hat{\lambda}_2(y_j)$	$\hat{F}_2(y_j)$
0.000	541	0	1.0000	0	0.0000	0.0000	0	0.0000	0.0000
0.559	491	1	0.9057	1	0.0020	0.0943	0	0.0000	0.0000
0.988	461	1	0.8503	1	0.0022	0.1405	0	0.0000	0.0092
1.418	433	1	0.8003	1	0.0023	0.1868	0	0.0000	0.0129
1.963	404	1	0.7466	1	0.0025	0.2331	0	0.0000	0.0204
2.697	371	1	0.6854	0	0.0000	0.2868	1	0.0027	0.0278
3.452	343	1	0.6371	1	0.0029	0.3296	0	0.0000	0.0333
4.857	292	1	0.5723	1	0.0034	0.3754	0	0.0000	0.0524
5.908	261	1	0.5415	0	0.0000	0.3999	1	0.0038	0.0586
7.083	213	1	0.4954	0	0.0000	0.4306	1	0.0047	0.0740
9.607	157	1	0.4257	0	0.0000	0.4798	1	0.0064	0.0946
11.732	124	1	0.3683	0	0.0000	0.5139	1	0.0081	0.1178
21.265	31	1	0.2496	0	0.0000	0.5618	1	0.0323	0.1886
31.102	1	0	0.0830	0	0.0000	0.5718	0	0.0000	0.3452

Table 3: Non-parametric estimated survival function, cause-specific hazards and cumulative incidence functions for the follicular lymphoma data.

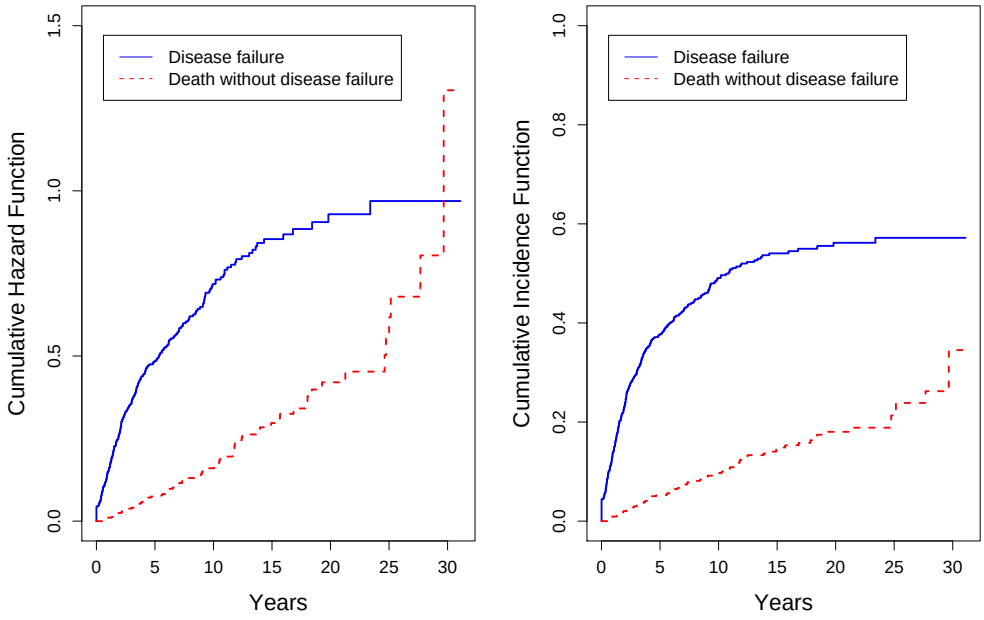

Figure 5: Follicular cell lymphoma data set: (a) Cause-specific cumulative hazard functions. (b) Cumulative incidence functions.

	Disease Failure				**Death without disease failure**			
y_j	d_{1j}	$\hat{S}_1(y_j)$	$\tilde{\hat{S}}_1(y_j)$	$\hat{S}_1^*(y_j)$	d_{2j}	$\hat{S}_2(y_j)$	$\tilde{\hat{S}}_2(y_j)$	$\hat{S}_2^*(y_j)$
0.000	0	1.0000	0.5028	1.0000	0	1.0000	0.1405	1.0000
0.559	1	0.9057	0.4085	0.9057	0	1.0000	0.1405	1.0000
0.988	1	0.8595	0.3623	0.8593	0	0.9908	0.1312	0.9895
1.418	1	0.8132	0.3161	0.8125	0	0.9871	0.1275	0.9850
1.963	1	0.7669	0.2699	0.7652	0	0.9796	0.1201	0.9756
2.697	0	0.7132	0.2163	0.7100	1	0.9722	0.1128	0.9654
3.452	1	0.6704	0.1738	0.6654	0	0.9667	0.1072	0.9574
4.857	1	0.6246	0.1294	0.6167	0	0.9476	0.0887	0.9279
5.908	0	0.6001	0.1072	0.5902	1	0.9414	0.0832	0.9175
7.083	0	0.5694	0.0813	0.5563	1	0.9260	0.0702	0.8906
9.607	0	0.5202	0.0462	0.4997	1	0.9054	0.0555	0.8518
11.732	0	0.4861	0.0240	0.4590	1	0.8822	0.0407	0.8025
21.265	0	0.4382	0.0018	0.3935	1	0.8114	0.0111	0.6343
31.102	0	0.4282	0.0000	0.3778	0	0.6548	0.0000	0.2197

Table 4: Estimates of survival-like functions $S_j(t)$, $\tilde{S}_j(t)$ and $S_j^*(t)$ in the follicular cell lymphoma data.

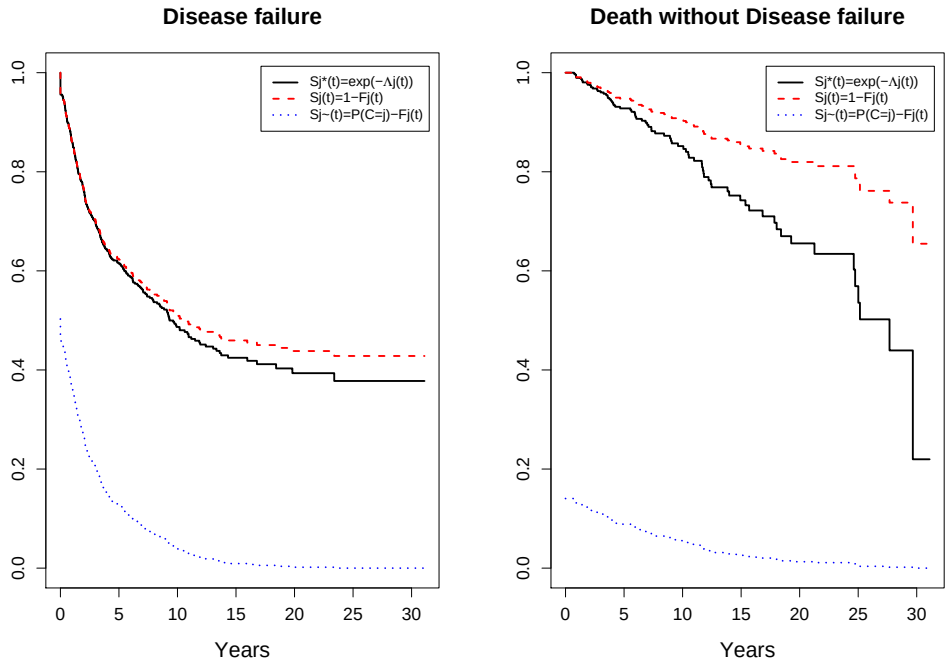

Figure 6: Survival-like functions in the follicular cell lymphoma data set.

4 Software

4.1 Competing risks analysis with R

In the following, we describe the code in the free software R[2] used to implement the methodology explained in this paper. Two additional packages are needed: `survival` and `cmprsk`. The former is included by default with the software, but needs to be loaded to access its functions by the command:

```
library(survival).
```

The `cmprsk` needs first to be downloaded from R's web site, and loaded similarly. Assume we have a data frame containing at least two columns `time` and `cens`, being, respectively, the vector with observed times for each individual, and the vector of failing causes. The `cens` vector equals 0 when individuals are censored at their observed time, or takes value j among the distinct possible causes of failure. For this illustration, assume there are only two causes of failure, and therefore, `cens` takes values in $\{0, 1, 2\}$.

Number of individuals at risk and failing at any time t

Two simple functions can be implemented to obtain, at any given time `t`, the number of individuals at risk of failing for any cause and the number of individuals failing from each cause:

[1]R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`

```
risk<-function(t=0,vT){
    val<-sum((vT>=t),na.rm=T)
    return(val)
    }
```

```
fail<-function(t=0,vT,vC,c=1){
    val<-sum((vT==t)*(vC==c),na.rm=T)
    return(val)
    }.
```

The `risk` function provides, at any time `t`, the number of individuals at risk, based on the information given by the vector of times `vT`. The `fail` function provides the specific number of failures at time `t` from cause `c`, based on the information given by the vector of times `vT` and the vector of causes `vC`.

Now we apply functions `risk` and `fail` to each element of vectors `time` and `cens`, in order to obtain vectors of the same length containing the number of individuals at risk (`ni`), and the number of individuals failing from each cause (`d1` and `d2`):

```
ts<-c(0,unique(sort(time)))
ni<-numeric(length(ts))
d1<-numeric(length(ts))
d2<-numeric(length(ts))
    for(i in 1:(length(t))){
        ni[i]<-risk(ts[i],time)
        d1[i]<-fail(ts[i],time,cens,1)
        d2[i]<-fail(ts[i],time,cens,2)
    }
```

Cause-specific hazards and cumulative incidence functions

Estimates for the cause-specific hazards at any observed time are easily obtained by:

```
lam1<-d1/ni
lam2<-d2/ni.
```

The Kaplan-Meier estimate of the survival function for `time`, without taking into account distinct causes of failure, is obtained by the `survfit` function. We need to define a censoring indicator for any of the two events: `delta=1` when `cause=1` or `2`, `0` otherwise.

```
delta<-as.integer(cens!=0)
sur<-survfit(Surv(time,delta))
S<-c(1,sur$surv)
```

The cumulative incidence functions can be obtained using the `cuminc` function from the `cmprsk` package (Gray, 2004):

```
cif<-cuminc(time,cens,cencode=0)
```

From this object `cif` we can extract the cumulative incidence function from each cause, `cif1` and `cif2`.

Survival-like functions

Estimates for $S_j^*(t)$ are then obtained by

```
S1.est<-1-cif1
S2.est<-1-cif2.
```

A possible implementation of expression (3.3) to estimate $\tilde{S}_j(t)$ is:

```
S1.td<-numeric(length(ts))
S2.td<-numeric(length(ts))
for (i in:length(ts)){
    S1.td[i]<-mean((time>ts[i])*(cens==1))
    S2.td[i]<-mean((time>ts[i])*(cens==2))
}.
```

Finally, $S_j(t)$ are obtained by taking as failures only those of type j, and treating other causes as censored observations. The Kaplan-Meier estimate is obtained from such data:

```
delta1<-as.integer(cens==1)
delta2<-as.integer(cens==2)
S1<-survfit(Surv(time,delta1))$surv
S2<-survfit(Surv(time,delta2))$surv.
```

4.2 Competing risks analysis with **SPSS**

Methods to deal with competing risks analysis are not implemented in the mainstream statistical software SPSS[3] but can be obtained in a few simple steps. In this section, we describe in some detail those steps, and pieces of `SPSS` syntax are given. Assume, as in the previous section, that the time to failure of interest is denoted by `time`, and the variable `cause` contains the causes of failure as well as the censoring indicator coded by 0.

Cause-specific hazards and cumulative incidence functions

Firstly, we estimate the overall survival function $S(t)$ by the Kaplan-Meier method from the time variable `time` and the censoring indicator defined by `cause` $\neq$ `0`, that is, events are failures due to any cause. In the SPSS windows, the steps to follow are:

1. Analyze ▶ Survival ▶ Kaplan-Meier...
2. Select `time` for the 'Time' box in the opened window, and `cause` for the 'Status' box. Press the 'Define Event...' button.
3. The event of interest is defined by the codes of any kind of failure. Select the second checkbox 'Range of values' and introduce the code for the first failure and the code for the last failure. For example, if there were three causes of failure in your data set, specify "from 1 to 3". Press the 'Continue' button.
4. Press the 'Save' button, and specify the survival function to be saved. Press the 'Continue' bottom.
5. Now you can press the 'OK' button to obtain the estimates or you can press the 'Paste' button. Code will be generated and pasted into a syntax file, which is useful to keep trace of the work done or when a specific procedure must be run several times.

The syntax generated in the past steps is:

```
KM
  time  /STATUS=cause(1 THRU 2)
  /PRINT TABLE MEAN
  /SAVE SURVIVAL .
```

Estimates for $S(t)$ function will be added as a new column in our data set. Estimates are only computed for those cases corresponding to a failure time. Therefore, cases corresponding to censored observations are assigned a missing value. We rename this new variable in our data set so as not to get confused in the next steps. The variable can be directly renamed in the Variables Window, but it can be saved in the syntax script:

[2]SPSS Inc. SPSS 15.0 for Windows (2006) v.15.0.2, Chicago IL. URL `http://www.spss.com`

```
rename variables (SUR_1=Surv).
```

There are two procedures providing estimates for the cause-specific hazards. On one hand, we can estimate the survival-like functions $S_j^*(t)$ by a Kaplan-Meier analysis for each cause of failure, and then obtain estimates for the cumulative cause-specific hazards by $\hat{\Lambda}_j(t) = -\log S_j^*(t)$. Hence, follow the steps above for each cause of failure separately. To do so, inside the 'Define Event...' button, select the first option 'Single value' and define the code for the cause of interest. For example, in the follicular cell lymphoma data, disease-related relapse is coded by 1. To determine the hazard specific to relapse, we will define the event by the value 1 in the variable `cause`. In addition to the survival function, also the cumulative hazard function is saved. Since the survival functions obtained this way estimate $S_j^*(t)$, we will rename these variables accordingly. For example, if only two causes of failure are possible:

```
KM
  time  /STATUS=cause(1)
  /PRINT TABLE MEAN
  /SAVE HAZARD SURVIVAL.

KM
  time  /STATUS=cause(2)
  /PRINT MEAN
  /SAVE HAZARD SURVIVAL.
RENAME VARIABLES (SUR_1 SUR_2=S_star_1 S_star_2).
```

The second procedure to obtain estimates for the cumulative cause-specific hazards refers to the method of Nelson-Aalen. There is no specific procedure to apply the method in `SPSS`. Therefore, we may save the necessary information and compute the hazard manually, as we did in `R`, or we can obtain them indirectly by fitting a Cox model without covariates, by saving the hazard function in the 'Save' menu. In doing so, Cox-Snell residuals are computed, which corresponds to the Nelson-Aalen estimate for the cumulative hazard function when no covariates are present. The two procedures provide similar estimates for the cumulative cause-specific hazards.

We will need some data manipulation to obtain the point cause-specific estimates. We have been working on the original data set, with one register for each individual. Now we need a data set containing just distinct failure times. To obtain it, delete variables not involved in the analysis and keep `time, cause, Surv, S_star_1, S_star_2, HAZ_1` and `HAZ_2`, and create with this variables a new data set. There may be duplicate registers due to ties if data are aggregated. There exists an option in the 'Data' menu to identify those duplicates:

1. Data ▶ Identify duplicate Cases...
2. In the first box, 'Define matching cases', include `time, cause` and `Surv`.

For each set of duplicate registers, SPSS chooses the last one -by default, it can be changed but there's no need in this analysis-. A new variable `FirstLast` is created to indicate the selected records. Now we apply the following filter to our data:

```
FILTER OFF. USE ALL.
SELECT IF(FirstLast=1  &  cause ~= 0).  EXECUTE.
```

This filter can be applied directly through the Windows Menu by:

1. Data ▶ Select Cases ...
2. Select the second radio button 'If condition is satisfied...'. Press the 'If...' button.
3. Write the condition `FirstLast=1 & cause ~= 0` within the expression menu. Press 'Continue'.

4. In the 'Non-selected cases are...' box, select the third radio button 'Deleted'. Press the 'Accept' button.

Cause-specific hazards are obtained from the cumulative cause-specific hazards estimated previously by $\hat{\lambda}_j(y_i) = \hat{\Lambda}_j(y_i) - \hat{\Lambda}_j(y_{i-1})$. We can use the `DIFF` function, which produces new variables based on the differences between elements of existing variables. For example, for cause 1 failures:

```
SORT CASES BY
  cause (A) time (A) .
CREATE h01=DIFF(HAZ_1,1).

DO IF ($CASENUM=1).
    COMPUTE h1 = HAZ_1.      *the first failure time
ELSE.
    COMPUTE h1 = h01.
END IF.  EXECUTE.
```

Finally, cumulative incidence functions are estimated using expression 3.2, first obtaining, for each specific failure time y_i, the product $\lambda_j(y_i)S(y_i)$ then adding up to the previous terms:

```
COMPUTE sh1 = Surv * h1 .  EXECUTE .
COMPUTE sh2 = Surv * h2 .  EXECUTE.
CREATE cif1 cif2 = CSUM(sh1 sh2).
```

Function `CSUM` produces a new variable based on the cumulative sums of an existing variable.

Survival-like functions

We have previously find out how to estimate $S_j^*(t) = \exp\{-\int_0^t \lambda_j(u)du\}$, by performing several Kaplan-Meier analysis or each cause of failure. $S_j(t) = 1 - F_j(t)$ is easily obtained in SPSS by the transformation:

```
COMPUTE Sur_1 = 1-cif1 .  EXECUTE.
COMPUTE Sur_2 = 1-cif2 .  EXECUTE.
```

Finally, to estimate the $\tilde{S}_j(t)$ functions we need to reopen the original data set. We will estimate this function at every observed time in our data. At this point, we need to know how many failures due to each cause there are in our data. For example, in the follicular cell lymphoma data set, there are 272 disease-related relapses and 76 non disease-related deaths. Now we simply compute the cumulative number of events for each cause, which can be obtained by performing Kaplan-Meier analysis for every cause, and inside the 'Save...' button, select 'Cumulative events':

```
KM
  time  /STATUS=cause(1)
  /PRINT NONE
  /SAVE CUMEVENT .
KM
  time  /STATUS=cause(2)
  /PRINT NONE
  /SAVE CUMEVENT .
```

We need the complementary number of events, that is, at each time t, the number of failures due to a specific cause that will occur after t. Since we know the total number of events, we can define:

```
COMPUTE I_CUM_1 = 272-CUM_1 .
EXECUTE .
COMPUTE I_CUM_2 = 76-CUM_2 .
EXECUTE .
```

These vectors contain, for each observed time, the number of individuals failing after this time. The desired estimates are obtained by dividing each element by the total number of cases, 541 in the follicular data set:

```
COMPUTE S_til_1 =I_CUM_1 / 541 .
EXECUTE .
COMPUTE S_til_2 =I_CUM_2/ 541 .
EXECUTE .
```

4.3 Competing risks analysis with SAS

There are no SAS[4] procedures specifically designed to perform a competing risks analysis. However, we can use existing procedures and web-available macros to implement the methodology. The Kaplan-Meier estimate of the overall survival function is obtained by the LIFETEST procedure. The OUTSURV option permits to create a new data frame containing the estimates. Version 9.2 of this software provides the Nelson-Aalen estimates for the cumulative hazard function (see 3.1), but it is not possible with earlier versions, where they must be constructed manually (follow the example available at http://www.ats.ucla.edu/stat/sas/examples/asa/asa2.htm).

Cumulative incidence functions can be obtained manually from the previous estimates, but there are macros available to construct them. See for example, the book by Pintilie (2006), where the author describes some useful macros for non-parametric and regression analysis (available at the author's web site). Other macros can be found at http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm and http://www.biostat.mcw.edu/software/SoftMenu.html (Gichangi and Vach, 2005). When it comes to survival-like functions, $S_j^*(t)$, through the Kaplan-Meier estimates, and $S_j(t)$, from the cumulative incidence functions, are straightforward to obtain with the above-mentioned procedures. To obtain $\tilde{S}_j(t)$ at each observed failure time point, the LIFETEST procedure can also be used. For example, consider the follicular cell lymphoma data: there are 541 patients, 272 for whom a disease-related relapse -cause 1- is observed. To estimate $\tilde{S}_1(t)$ associated to these relapses, apply the LIFETEST procedure to failures from cause 1, that is, treating failures from cause 2 as censored observations. By the ODS statement, information on number of failures at each time point is saved (variable Failed in the new data set), which is enough information to implement the estimator (3.3):

```
*compute the number of individuals failed from cause 1 at each time;
proc lifetest data=data1;
time dftime*cause(0,2);
ods output ProductLimitEstimates=cause1_0;
run;

*compute the number of individuals remaining to fail;
data cause1_1;
set cause1_0;
where Survival^=.;
ind=272-Failed;
keep time failed left ind;
run;

*obtain S~ for cause 1, for each time;
data cause1_2;
set cause1_1;
Surv_tilde_1=ind/541;
run;
```

[3]SAS Institute Inc. SAS for Windows (2001) v.8.2, Cary, NC, USA. URL http://www.sas.com

References

Aalen, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics*, **6**(3), 534–545.

Andersen, P. K., Abildstrom, S. Z., and Rosthøj, S. (2002). Competing risks as a multi-state model. *Stat Methods Med Res*, **11**(2), 203–215.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**(1), 141–151.

Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society. Series B (Methodological)*, **21**(2), 411–421.

Fine, J. and Gray, R. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94**(446), 496–509.

Gichangi, A. and Vach, W. (2005). Competing risks: A guided tour.

Gooley, T. A., Leisenring, W., Crowley, J., and Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat Med*, **18**(6), 695–706.

Gray, R. (2004). *The cmprsk package*. The Comprehensive R Archive network. `http://cran.r-project.org/src/contrib/Descriptions/cmprsk.html`.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Klein, J. P. and Moeschberger, M. L. (1988). Bounds on net survival probabilities for dependent competing risks. *Biometrics*, **44**(2), 529–538.

Latouche, A., Beyersmann, J., and Fine, J. P. (2007). Comments on 'analysing and interpreting competing risk data'. *Stat Med*, **26**(19), 3676–9; author reply 3679–80.

Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Llorca, J. and Delgado-Rodríguez, M. (2004). [survival analysis with competing risks: estimating failure probability.]. *Gac Sanit*, **18**(5), 391–397.

Nelson, W. (1982). *Applied Life Data Analysis*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons.

Pepe, M. S. and Mori, M. (1993). Kaplan-meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Stat Med*, **12**(8), 737–751.

Peterson, A. V. (1976). Bounds for a joint distribution function with fixed sub-distributions functions: Applications to competing risks. *Proceedings of the National Academy of Sciences of the USA*, **73**, 11–13.

Pintilie, M. (2006). *Competing risks: a practical perspective*. Wiley.

Pintilie, M. (2007a). Analysing and interpreting competing risk data. *Stat Med*, **26**(6), 1360–1367.

Pintilie, M. (2007b). Author's reply. *Stat Med*, **26**(19), 3679–3680.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., J., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, **34**(4), 541–554.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*, **26**(11), 2389–2430.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proc Natl Acad Sci U S A*, **72**(1), 20–22.

Wolbers, M. and Koller, M. (2007). Comments on 'analysing and interpreting competing risk data' (original article and author's reply). *Stat Med*, **26**(18), 3521–3; author reply 3523.

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, **82**(1), 127–138.