UNIVERSITY OF MILAN

Doctoral School in Biomedical, Clinical and Experimental Sciences

Department of Clinical Sciences and Community Health

Doctorate of Research in Biomedical Statistics (XXVII cycle)

SURVIVAL ANALYSIS AND REGRESSION MODELS IN THE PRESENCE OF COMPETING AND SEMI-COMPETING RISKS

MED/01

Doctorate student: Annalisa Orenti (R09590)

Supervisor: Prof. Patrizia Boracchi

Coordinator of the PhD program: Prof. Adriano Decarli

Academic Year 2013-2014

Abstract

Evaluation of a therapeutic strategy is complex when the course of a disease is characterized by the occurrence of different kinds of events. Competing risks arise when the occurrence of specific events prevents the observation of other events. Different survival or incidence functions can be defined in the presence of competing risks and a relevant issue is an adequate knowledge of the methodological background in order to apply a suitable statistical analysis for the study aims. This work aims at presenting different estimates of survival or incidence probabilities used in this framework. From clinical application, it emerges that crude cumulative incidence is widely diffuses both to estimate incidence probabilities and to evaluate covariate effects. On the contrary net survival functions, although of clinical interest, are not diffused because of more difficult model structure and lack of software availability.

If the independence assumption between different events is tenable, Kaplan-Meier method can be used to estimate net survival. Otherwise multivariate distribution of times based on Copulas can be adopted. In the case of different causes of death, relative survival can be interpreted as net survival only under specific assumptions on the mortality pattern.

A particular case on competing risks arises when only fatal events can prevent the observation of the non fatal ones, but not vice versa (semi-competing risks). The estimate of an interpretable measure of association between times to non fatal and fatal event is often of biological interest, in order to understand the disease progression. In the statistical literature some approaches have been proposed to estimate the association between two independently doubly censored failure times, but more specific approach have to be applied in the presence of semi-competing risks. After estimating the association parameter, the survival function of a non terminal event can be estimated after fixing a copula structure by means of the semi parametric methods proposed by Fine, Jiang and Chappell or the copula graphic estimator.

Furthermore when the interest is to evaluate the effect of different therapeutic strategies or covariates on the occurrence of a non terminal event in a semi-competing risks setting, specific regression model have to be adopted. I propose here to adopt the methodology

based on pseudo-observations, having the advantage to be implemented by generalized linear models approaches.

Simulation studies are performed to compare the performances of methods to estimate the association between events, of methods based on Copulas models to estimate net survival and of regression method for net survival in the presence of semi-competing risks.

A case series of breast cancer patients is used to illustrate different methods of estimating net survival functions on the causes of deaths and on the severe non fatal events in the presence of competing and semi-competing risks framework.

Index

1. Introduction

The evaluation of the effect of a therapy or the impact of prognostic factors are particularly complex when the course of a disease is characterized by the occurrence of different events during the follow-up time. For example with regard to the neoplastic disease we can observe “hard endpoints”, such as death, which prevent the subsequent occurrence of other “non fatal events”, such as relapses in loco regional sites, distant metastases or malignancies elsewhere.

In severe diseases the time elapsed from the beginning of the treatment and death is always an end-point of interest. Information on time to death is not always available because some patients are still alive at the study ending (administrative censoring) or they are lost to follow-up (censoring times). A bivariate distribution of the random variables time to death (T) and censoring time (C) is then of concern. The interest is on the marginal survival function of time to death, in fact results are reported in terms of survival probability during follow-up time. For each patient, the observed data is the minimum between T and C, thus, if C is observed, it prevents the observation of T (and vice-versa). T and C ”compete” one each other in being observed and this condition is named ”competing risks”. The information provided by such a kind of available data is sufficient to determine uniquely marginal survival distribution only under the assumption of independence between T and C (Zheng and Klein, 1995). Under this assumption Kaplan-Meier method (Kaplan and Meier, 1958) is used to estimate survival curves. Under the independence assumption, when clinical and pathological characteristics are recorded, univariate parametric (e.g. exponential or weibull) (Marubini Valsecchi, 1995) or semi-parametric (Cox, 1972) survival regression models on the hazard of death can be used to evaluate the prognostic role of each variable on marginal survival.

In several studies information on the causes of deaths is also considered in order to evaluate their specific impact. Available data are times to death and corresponding causes or censoring time for alive patients. An exhaustive classification is then used, the simplest being a binary one: death for causes related or not related to the disease. In this case the main interest is on the death for causes related to the disease and on its pertinent marginal survival function. In fact results are often reported in terms of “cause specific survival”. In this context times to death for causes not related to the disease censor the times to deaths

for causes related to the disease, thus a multivariate distribution is of concern and competing risks are acting. Under the assumption of independent censoring (administrative and lost to follow-up), Kaplan-Meier method can be used to estimate marginal survival probabilities and the above mentioned regression models considering cause specific hazards can be used to evaluate the impact of prognostic factors, only if the independence assumption between time to death for the cause of interest and time to death for other causes is also tenable. This approach is used in several papers, to estimate "cause specific" survival (e.g in the case of breast cancer see Tai et al., 2012 among others).

The interpretation of the estimated "cause specific" survival needs to be done in terms of "net" survival, i.e in the hypothetical situation where mortality for the cause of interest can be observed for all patients. If the independence assumption between causes of death is not tenable, the estimate of the marginal survival function requires the knowledge of the multivariate distribution. In this case, estimating problems arise because observed data do not allow to uniquely identify the multivariate distribution of times to events (non identifiability, Tsiatis, 1975).

A proposed solution is based on the assumption of a particular structure of the multivariate distributions. Several multivariate survival distributions have been proposed, most of them are based on parametric distributions of marginal survival functions (Hougaard, 1987). More flexible multivariate distributions are Copulas.

A copula is defined as a function that joins multivariate distribution functions to their univariate marginal uniform distribution functions (Kpanzou, 2007). Copula parameters are the association between marginal distributions and parameters of marginal distributions if they are parametrically defined. An advantage of copulas is that the marginal distributions need not to be parametrically defined, thus they can be non parametric as well.

Regression models for the marginal distribution can be obtained combining a defined copula structure with estimable survival functions in presence of competing risks. A disadvantage of these models is the non-direct interpretation of regression coefficients (Lo and Wilke, 2011; Lo and Wilke, 2014).

The above mentioned analysis are based on the classification of causes of death, thus, it makes sense only if a "reliable" classification is available. As an example, for the study reported in (Martelli et al., 2014), the classification of the cause of death was based on the previous neoplastic events and if there were doubts the general practitioner was contacted

in order to have further information on patient's health status. Nevertheless adequate information on cause of death is not always available. Without complete and reliable information on the cause of death we can resort to relative survival analysis for estimating net survival (Rutherford et al., 2012).

Relative survival is based on the relative survival ratio (RSR) which is the ratio between the observed survival in the patient group and the expected survival of a comparable group from the general population, matched to the patients with respect to the main demographic factors affecting patient survival (age, sex, calendar year). Relative survival is useful to evaluate the excess of mortality related to the diseases in the study sample (Ederer et al., 1961) and can be interpreted as net survival only under the following assumptions: the causes of deaths are independent, the reference population is practically free of the cause of interest and the death rate for other causes acts in the same way in the sample patients and in the reference population.

Specific regression models for relative survival have been proposed to estimate the effect of the variables on the ratio or the difference between observed hazards of death and the expected ones in the general population (Estève et al., 1990; Hakulinen and Tenkanen, 1987; Dickman et al., 2003; Andersen et al. 1985).

If incidence of mortality is of concern, the overall incidence can be decomposed in the incidences for each one of the causes of death. In the framework of competing risks, crude cumulative incidences need to be considered (Kalbfleish and Prentice, 2002). These, in fact, estimate the probability of dying for each cause in the situation where different causes of death are acting.

Semi-parametric regression models on the effect of the variables on the hazard of the crude cumulative incidence (sub-distribution hazard) have been proposed by Fine and Gray (1999).

Alternative models are based on generalized linear models on "pseudo-values" of crude cumulative incidences (Klein and Andersen, 2005), these models having the advantage to different link functions including the one giving the Fine and Gray model as a particular case.

In evaluating the efficacy of a therapy, survival from death is not the only end-point of concern. The interest is also in the time elapsed between a starting point (initiation of a therapy or date of enrolment in the study) and the onset of adverse events, which are relevant for the study aims.

In the most comprehensive end-point all possible events should be considered (e.g. in cancer: local relapses, distant metastases, other primary tumours and deaths free from every documented event). Every event which occurs during the follow-up could be considered directly or indirectly related to the failure of the therapy thus, a failure is observed at the occurrence of the first event. In this situation, the composite endpoint may be called "first failure". The failure probability is the measure of interest and can be obtained by complement to one of Kaplan-Meier estimates. Following the first evaluation, a more detailed analysis on the causes of failures is frequently considered by using some subsets of adverse events or each single event. The probability of failing for different causes of failure is the probability of observing each event as first (as an example, in breast cancer is usually the probability of failing for local relapse and/or distant metastases).

Causes of failures are submitted to an exhaustive classification thus only a cause of failure is reported for each patient. This imply that the failure for a specific cause prevents the observation for another thus, a competing risks setting is of concern and the estimate is based on crude cumulative incidence (Kalbfleish and Prentice, 2002). In this framework Fine and Gray regression model can be used to evaluate the covariate effect's on the causes of treatment failure.

In most situation results are reported in terms of probability to be free of treatment failure at a given time. Subsamples of composite end-point are usually considered, as in the case of relapse free survival. The estimate has to be interpreted in terms of marginal (net) survival function, i.e. an hypothetical situation where relapse can be observed for all patients. The analysis is frequently performed by Kaplan-Meier method, censoring times to occurrence of other events which are not included in the end-point (Moliterni et al., 2003) and by Cox regression model on cause-specific hazard.

Kaplan-Meier method implies the assumption of independence between times to causes of failure of interest and time to occurrence of other causes of failure. It has to be taken into account that this assumption is rarely tenable. Thus specific method to estimate net survival in competing risk settings need to be used.

A particular case of competing risks arises when the end-point of interest is composed by one or more non fatal events and the only "competing" event is a fatal one. This situation is usually referred as "semi-competing risks" as the occurrence of fatal event precludes the occurrence of non fatal events but not vice-versa (Fine et al., 2001).

In semi-competing risks settings, times to fatal events are always observable and the incomplete observation relies only to non fatal events, thus a more efficient estimating procedure can be used with respect to the presence of competing risks being known the "upper wedge" of the bivariate distribution (Fine et al., 2001). Specific regression models have been proposed to evaluate covariates effects on the hazard of the net distribution of the non fatal event under an assumed copula structure (Peng and Fine, 2007; Hsieh and Huang, 2012).

Considering the flexibility of models based on pseudo-values and their direct application based un standard GLM software, they can be applied for all models based on different survival/incidence functions. Thus they can be advantageous to estimate covariate effect on marginal survival in semi-competing risks.

1.1 Aims of this work

In the presence of several different events during follow-up several survival functions are of concerns. The choice of the suitable function depends on the study aim thus the characteristics of the different functions need to be exploited. Crude cumulative incidence are estimable functions in presence of competing risks and several regression models can be applied in this framework and are readily available in many statistical software. The situation is different if net survival is of concern because assumptions on unknown multivariate distributions on times to competing events are needed. The methodological background and related models are not diffused in statistical literature and proper functions in statistical software are not available. The work's aims can be summarized as follows:

i) to present the functions used in survival analysis when different kinds of events occur during the follow-up. Starting from the multivariate distribution of latent failure time, the different survival/incidence function and the pertinent hazard functions are defined and compared;
ii) to present estimates of the above mentioned survival/incidence functions;
iii) To present the main characteristics of semi parametric regression models on overall survival and crude cumulative incidence;

As our experience is on the analysis of types of cancer that allow to record both the cause of death and the event history with a long follow-up, our interest relies on the estimation of net survival for death related to breast cancer and the net event free survival. The first is in the

presence of competing risks and the second in the presence of semi-competing risks. Given that in semi-competing risks, partial information on the joint distribution is available, we concentrate the study of the properties of estimation and regression models in the semi competing risks framework

iv) to present and compare different methods for estimating the association between times to different events in a semi-competing risks framework;
v) to exploit and compare the performance of different methods to estimate net survival estimate in a semi-competing risks context;
vi) to propose an innovative regression method for evaluating the effect of covariates on survival from a non terminal event based on pseudo-values;
vii) to compare the performance of the model with that of other available regression models.

For illustrative purposes a clinical example on small breast carcinoma is used to illustrate the estimates of net survival in presence of several events during the follow-up and the presentation of regression models for net survival in the presence of semi-competing risks.

2. Materials and methods

2.1 Definition of survival distribution by latent failure times

At the beginning of follow-up each patient is considered at risk for all the K events, jointly considering the vector of "latent" or "potential" failure times to K different events (t_1,…,t_K), enables postulating the joint survival function:

$$S(y_1, \dots, y_k, \dots, y_K) = P(Y_1 > y_1, \dots, Y_k > y_k, \dots, Y_K > y_K),$$

where y_k is the potential time to event k. This is a right-sided cumulative distribution satisfying: S(0, ... , 0, ... , 0)=1 and S(∞, ... , ∞, ... , ∞)=0. An implicit assumption of the joint survival function is that every subject experiences all events sooner or later, thus if an event different form k at time t has already occurred for a subject j, he still is at risk of experiencing the event k after t. These event times are called "potential" as they are not always observed in real world.

The survival probability at time t for all events (overall survival) is :

$$S(t) = S(t, \dots, t, \dots, t) = P(Y_1 > t, \dots, Y_k > t, \dots, Y_K > t)$$

It can be shown that the marginal distribution of Y_k from S is a proper survival distribution in the hypothetical condition where the events other than k were removed:

$$S_k(t) = S(0, \dots, t, \dots, 0) = P(Y_1 > 0, \dots, Y_k > t, \dots, Y_K > 0)$$

This is the net survival function from event k (Marubini and Valsecchi, 1995).

It is worth noting that in the case of independence the overall survival equals the product of net survivals for different causes: $S(t) = \prod_k S_k(t)$.

A second approach to latent failure times interpretation focuses attention only on the time to the first event for each subject, which is always observed: $T=\min(Y_1,\dots,Y_k,\dots,Y_K)$.

Given the time and type of first event for each subject (T, J) it is always possible to estimate the probability of k as first event:

$$I_k(t) = P(Y_1 > t, \dots, Y_k \leq t, \dots, Y_K > t)$$

This is the crude cumulative incidence function of failure for event k.

2.2 Overall survival, crude cumulative incidence, net survival and related hazard function

When there is no need to distinguish among different kinds of events, the interest is focused on ***"overall" survival*** i.e. the probability of surviving from any event over time t:

$$S(t)=P(T>t)$$

This survival probability can be written in terms of the ***overall hazard*** function, or instantaneous failure rate, which enables studying the dynamic process of the disease over time:

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

where $\lambda(t) \cdot \Delta t$ is the probability of dying in the infinitesimal interval between t and t+Δt, given survival until time t. The following relationship between survival and hazard function holds: $S(t) = e^{-\Lambda(t)}$, where $\Lambda(t) = \int_0^t \lambda(u)du$ is the cumulative hazard function.

Net (or marginal) survival is the probability that the individual's occurrence time for a given event k exceeds a pre-assigned time t: $S_k(t) = P(Y_k > t)$.

The corresponding ***net (or marginal) hazard*** is the probability of dying for cause k in the infinitesimal interval between t and t+Δt, conditionally to the fact that event k has not occurred before time t, in the hypothetical situation where all patients experience event k:

$$\phi_k(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq Y_k < t + \Delta t | Y_k \geq t)}{\Delta t}$$

When the random variable of concern is T=min(Y_1,…,Y_k,…,Y_K), in the case of different events, the hazard for a specific event, called ***cause-specific hazard*** is considered:

$$\lambda_k(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t, K = k | T \geq t)}{\Delta t},$$

which is the probability of event k in the infinitesimal interval between t and t+Δt, in the presence of the remaining events acting simultaneously, given survival from all events until t.

The additive property is valid and the overall hazard can be expressed as the sum of all cause-specific hazards: $\lambda(t) = \sum_k \lambda_k(t)$.

For the survival corresponding to cause-specific hazard ($S_k^*(t) = e^{-\Lambda_k(t)}$), the property $S(t) = \prod_k S_k^*(t)$ holds. It is worth of note that $S_k^*(t)$ has no meaning, unless the different events are independent. Only in the case of independence among events, the cause-specific hazard equals the net hazard: $\lambda_k(t) = \phi_k(t)$, and thus the cause-specific survival equals the net survival.

If the interest is in the overall incidence $I(t) = P(T \leq t) = 1 - S(t)$ and decomposing it in the different events, the ***crude cumulative incidence*** is of concern, which is the estimated probability of observing event k, within time t:

$$I_k(t) = P(T \leq t, K = k)$$

The following property holds: $I(t) = \sum_k I_k(t)$. The interpretation of the "survival" probability for an event k obtained as 1-crude cumulative incidence probability is not straightforward because "surviving" to the events of interest does not imply the non occurrence of the not considered events (e.g. one can die without documented relapse and "survive" to relapse).

The corresponding ***sub-distribution hazard*** is the probability that k occurs as first event in the infinitesimal interval between t and t+dt, conditionally to the fact that no events have occurred before t or an event different from k had occurred before t [14].

$$\tilde{\lambda}_k(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t;\ K = k | T \geq t \text{ or } (T < t;\ K \neq k))}{\Delta t}$$

How it can be argued form its definition, sub-distribution hazard is a measure which is not of direct clinical interpretation.

2.3 Non parametric estimates of overall survival and crude incidence functions

If there is no need to distinguish among different events, and under the assumption of independent censoring the Kaplan-Meier method can be adopted in order to estimate overall survival probability and obtain the corresponding overall incidence.

If there is the need to distinguish among different events, in order to estimate net survival, the simplest approach is to assume that the different events are statistically independent, i.e. the time of occurrence of event k under one set of study conditions in which all K events are operative is precisely the same as under an altered set of conditions in which all events except the k^{th} have been removed. In this case net survival for the event k can be estimated by means of Kaplan-Meier method considering as censored times to occurrence of events different from k. More generally, however, the elimination of certain events may well alter the risks of other types of events. Evidently any assumption about the relationship between the observed T and times to failure for specific events, given the removal of other events will require detailed knowledge of the system under study and of the mechanism for events removal.

Otherwise, if independence cannot be assumed, the distribution of survival function from different events (multivariate joint survival distribution) is needed. This is not non-parametrically identifiable unless times to each event are known (Zheng and Klein, 1995).

In survival analysis copula models have been often used to express the joint survival distribution of times to multiple events as a function of their marginal survival distributions and parameters of their association (Marubini and Valsecchi, 1995). Properties of copula functions and pertinent estimators of net survival are detailed in the next section.

Otherwise if the incidence of different events are considered, crude cumulative incidence can be estimated by Kalbfleish and Prentice (2002) method:

$$I_k(t) = P(T \leq t, K = k) = \sum_{s=0}^{t} \lambda_k(s) \cdot S(s)$$

In this case, $\lambda_k(s)$ is the cause-specific hazard estimated by Kaplan-Meier method, considering as censored times to occurrence of other events and S(s) is the overall survival.

The estimate of sub-distribution hazard can be obtained as follows:

$$\tilde{\lambda}_k(s) = \frac{\lambda_k(s) \cdot S(s)}{1 - I_k(s)}$$

It is worth noting that 1-$S_k(t)$ obtained by Kaplan-Meier method does not provide an estimate of crude cumulative incidence of event k.

For sake of simplicity, in the case where one of the K considered events is observed for each patient, crude cumulative incidence for the event k is the proportion of patients who experience the event k, and it is less or equal to 1- $S_k^*(t)$.

In fact, comparing

$$I_k(t) = \sum_{s=0}^{t} \lambda_k(s) \cdot S(s)$$

with

$$1 - S_k^*(t) = \sum_{s=0}^{t} \lambda_k(s) \cdot S_k^*(s)$$

the overall survival is always lesser than the cause-specific survival: $S(s) \leq S_k^*(s)$.

2.4 Estimates of net survival in competing risks setting by copula functions and relative survival

Without assumption on independence of time to events, net survival is not estimable in a non parametric way. Several works deal with this problems (Peterson, 1976; Peterson 1978; Klein and Moeschberger, 1988; Dignam et al., 1995). In particular Peterson (1976) showed that the net survival probability for event k is bounded between overall survival and the complement to 1 of the crude cumulative incidence of the event of interest:

$$S(t) \leq S_k(t) \leq 1 - I_k(t).$$

For sake of simplicity, we consider a situation where only dichotomous classification is made: the event of interest and all other competing events. In the case of perfect positive correlation, the net survival probability of the event should be exactly equal to the overall survival (lower bound). Otherwise in the case of perfect negative correlation, the net survival probability of the event should be exactly equal to the complement to 1 of the crude cumulative incidence of the event of interest (upper bound).

When there is no adequate information on the causes of death, the relative survival ratio can be computed:

$$\text{Relavive survival} = \frac{S_O(t)}{S_E(t)},$$

where $S_O(t)$ is the overall survival in the sample under study and $S_E(t)$ is the expected survival of a comparable group of the general population, matched to the sample under

study with regard to the main demographic characteristics (sex, age, year). Several methods have been proposed to calculated expected mortality from the population mortality tables. The population mortality tables gives, for every calendar year (y), sex (s) and age (a), the conditional probability of death (q_{asy}). The corresponding daily hazard is:

$$\lambda_{asy} = -\frac{\log(1 - q_{asy})}{365.25}.$$

The cumulative hazard of death for each subject (Λ_j) is obtained by summing the daily hazard for the time the subject is considered under observation in the study. The corresponding expected survival is $S_{Ej}(t) = e^{-\Lambda_j(t)}$.

The expected survival of the population under study is obtained as:

$$S_E(t) = \frac{\sum_{j=1}^{n} w_j(t) \cdot S_{Ej}(t)}{\sum_{j=1}^{n} w_j(t)}$$

Where w_j is a weight, depending on the method used to estimate expected survival (Ederer and Heise, 1959; Hakulinen, 1982).

Under the additive structure, the overall hazard of death is the sum of the hazard of death due to the disease of interest and the hazard of death due to other causes (cause specific hazards), then the overall survival is the product of the corresponding cause specific survival.

$$\frac{S_O(t)}{S_E(t)} = \frac{S^*_{O\,1}(t) \cdot S^*_{O\,2}(t)}{S^*_{E\,1}(t) \cdot S^*_{E\,2}(t)}$$

i) in the presence of independence between the causes of death, the cause specific survival correspond to net survival

$$\frac{S_O(t)}{S_E(t)} = \frac{S_{O\,1}(t) \cdot S_{O\,2}(t)}{S_{E\,1}(t) \cdot S_{E\,2}(t)}$$

ii) if the contributes of the cause of interest is negligible in the general population $S_{E\,1}(t) \approx 1$

iii) if the mortality related to other causes acts in the same way in the sample under study and in the general population $S_{O\,2}(t) \approx S_{E\,2}(t)$

Then

$$\frac{S_O(t)}{S_E(t)} \approx S_{O\,1}(t),$$

Thus the relative survival can be interpreted as a net survival for the cause of interest.

Anyway, in the presence of adequate information on causes of death, in order to estimate net survival, semi-parametric copula models have often been used to express the joint

survival distribution of times to different events as a function of their marginal survival distributions and parameters of their association:

$$P(Y_1 > y_1; Y_2 > y_2) = C_\theta\{S_1(y_1); S_2(y_2)\},$$

where $S_1(y_1)$ and $S_2(y_2)$ are marginal survival function for the event of interest and for all other events, and $C_\theta(\cdot,\cdot)$ is a copula for the dependency between Y_1 and Y_2. The parameter θ measures the association between Y_1 and Y_2, and may captures all distribution free dependence.

Several structured copula functions have been proposed in the literature (Nelsen, 1999), but in competing or semi-competing risks setting Archimedean Copulas are often used because they can be expressed in a closed form through a copula generator function and their association parameter has a direct relationship with Kendall's τ:

$$\tau = 4\int_0^1 \frac{\phi_\theta(u)}{\phi'_\theta(u)}du + 1$$

where ϕ_θ is the copula generator function.

A Kendall's τ is a well known association measure whose possible values range from -1 to 1. -1 indicates a perfect negative association, i.e. subjects who have experienced the event of interest have no chance to experience other competing events in the future. 1 indicates a perfect positive association i.e. subjects who have experienced the event of interest will experience other competing events in the near future. A 0 value implies a perfect independence among times to different events.

Since in the presence of competing or semi-competing risks it is not possible to identify the joint times distribution, in statistical practice the choice of the Archimedean copula is based on clinical consideration on the putative adequacy of copula's properties.

In the case of competing risks if the copula for Y_1 and Y_2 is known, the marginal distributions are uniquely determined and a graphic estimator based on estimable quantities has been proposed (Zheng and Klein, 1995). The simplest form of this estimator is available for Archimedean copulas (Rivest and Wells, 2001).

A copula C is called Archimedean if it admits the representation:

$$C_\theta\big(S_1(y_1); S_2(y_2)\big) = \phi_\theta^{[-1]}\Big(\phi_\theta\big(S_1(y_1)\big) + \phi_\theta\big(S_2(y_2)\big)\Big)$$

Where $\phi_\theta: [0,1] \times \Theta \to [0,\infty)$ is a continuous, strictly decreasing and convex function such that $\phi_\theta(1) = 0$. θ is a parameter within some parameter space Θ. ϕ_θ is the so-called generator function and $\phi_\theta^{[-1]}$ is its pseudo-inverse defined by

$$\phi_\theta^{[-1]}(t) = \begin{cases} \phi_\theta^{-1}(t) & \text{if } 0 \le t \le \phi_\theta(0) \\ 0 & \text{if } \phi_\theta(0) \le t \le \infty \end{cases}$$

Moreover, the above formula for C yields a copula for ϕ_θ^{-1} if and only if ϕ_θ^{-1} is continuous and non-increasing on [0, ∞] and strictly decreasing on $[0, \phi_\theta^{-1}(0)]$ (Nelsen, 1999).

The copula graphic estimator is shown as a method to afford the presence of dependent censoring, being times to event which are not of interest considered as censored. Indicating with δ=1 the event of interest and with δ=2 other events, the copula graphic estimator of the marginal $S_1(y_1)$ is a right continuous decreasing step function, with jumps at the points t_i, where the event 1 occurs. Starting from the relationship:

$$\phi_\theta^{[-1]}\left(\phi_\theta\big(S_1(t)\big) + \phi_\theta\big(S_2(t)\big)\right) = S(t),$$

Where $S(t)$ is the overall survival estimated by Kaplan-Meier method.

The closed form for estimate $S_1(t) = \phi_\theta^{[-1]}\left(\sum_{t_i \le t, \delta_i = 1}\left(\phi_\theta\big(S(t_i)\big) - \phi_\theta(S(t_i) - 1/n)\right)\right)$

Where n is the number of subject considered. It is worth of note that the Copula function depends on the association parameter, which is not estimated by the above mentioned method. An empirical estimator ok Kendall's τ which could be used as a first insight has been proposed (Brown, 1974), but it is unbiased only in the case of independence between times to events.

2.5 Estimates of the association between two censored times

With semi-competing risks data, the dependence structure between the non terminal (X) and terminal (Y) event is often of biological interest, and not only a nuisance parameter in the problem specification, because it can be of interest to know the extent to which the occurrence of an intermediate non terminal event hastens the occurrence of a more severe terminal event.

Different methods for estimating the association of two variables under censoring have been proposed in the literature (Brown, 1974; Wang and Wells, 2000; Beaudoin, Duchesne and Genest, 2007; Lakhal, Rivest and Beaudoin, 2009, Hsieh, 2010). However in the presence of semi-competing risks the estimate of the association parameter is even more complex, as the

two event times are both independently censored by C, moreover the non terminal event X can be dependently censored by the terminal event Y. Thus more specific approaches have been proposed and have to be adopted in a semi-competing risks analysis, by specifying the form of the bivariate distribution of times to events (Fine et al., 2001; Lakhal et al., 2008; Xu et al., 2010).

I propose here the methods of estimation proposed by Brown (1974) and Fine et al. (2001) and later I will compare them by means of a Monte Carlo simulation.

To characterize the semi-competing risks data, assume C is the censoring time independent of (X, Y) and has continuous distribution function. The random variable Y can be right censored by C, while X can be independently censored by C if Y > C and X > C or dependently censored by Y if Y < C and X > Y.

Let:

$Y' = \min(Y, C)$ the time to death or censoring,

$U = \min(X, Y)$ the time to relapse or death,

$X' = \min(U, C) = \min(X, Y')$ the time to relapse, death or censoring,

$\delta_{Y'} = I(Y \leq C)$ the status indicating if the subject died or was censored,

$\delta_{X'} = I(U \leq C)$ the status indicating if the subject relapsed/died or was censored,

$\delta_X = I(X \leq Y')$ the status indicating if the subject relapsed or died/was censored,

where I(·) is the indicator function.

The observable data are $[(Y'_i, \delta_{Y'_i}, X'_i, \delta_{X'_i}, \delta_{X_i}),\ i = 1, \ldots n]$, n independent and identically distributed realizations of $(Y', \delta_{Y'}, X', \delta_{X'}, \delta_X)$. They are used to estimate $\hat{S}_X(x)$.

Method proposed by Brown (1974)

Let (X, Y) be possibly correlated random variables, and let (X_i, Y_i) and (X_j, Y_j) $(i \neq j)$ be independent realizations from (X, Y). The (i, j)th pair is called concordant if $(X_i - X_j)(Y_i - Y_j) > 0$ and discordant if $(X_i - X_j)(Y_i - Y_j) < 0$. The popular version of Kendall's τ is defined as the difference of concordance and discordance probabilities between the (i, j)th pair. If X and Y are continuous, $\tau = pr\{(X_i - X_j)(Y_i - Y_j) > 0\} - pr\{(X_i - X_j)(Y_i - Y_j) < 0\}$.

It is easy to see that $-1 \leq \tau \leq 1$ and if (X, Y) are independent, $\tau = 0$. In the absence of censoring one observes i.i.d replications of (X, Y). Then τ can be easily estimated by taking the difference of sample concordance and discordance proportions.

This is equivalent to applying the formula,

$$\hat{\tau} = \binom{n}{2} \sum_{1 \le i < j \le n} a_{ij} b_{ij}$$

where $a_{ij} = 1$ if $X_i < X_j$, $a_{ij} = -1$ if $Y_i > Y_j$ and b_{ij} is similarly defined. Notice that the "score", $a_{ij}b_{ij}$, is 1 if the (i, j) pair is concordant and is −1 if discordant. In the presence of ties modified scores are used, that is $a_{ij}=0$ if $X_i=X_j$ and $b_{ij}=0$ if $Y_i=Y_j$. Alternatively, a second formula, which excludes tied pairs in computing the total number of combinations, is given by:

$$\Gamma = \frac{\sum_{i,j=1}^{n} a_{ij} b_{ij}}{\sqrt{\sum_{i,j=1}^{n} a_{ij}^2 \sum_{i,j=1}^{n} b_{ij}^2}}$$

Brown et al. (1974) proposed an estimator of Kendall's τ which utilized the marginal Kaplan-Meier estimates to modify the scores for those pairs whose concordance/discordance relationships are not clear.

Except for ties, he assigned $a_{ij} = 2\mathrm{pr}\{X_i > X_j | (X_i', X_j', \delta_{1i}, \hat{S}_X)\} - 1$ and $b_{ij} = 2\mathrm{pr}\{Y_i > Y_j | (Y_i', Y_j', \delta_{1i}, \hat{S}_Y)\} - 1$, where $\hat{S}_X$ and $\hat{S}_Y$ are the marginal Kaplan-Meier estimators of S_X and S_Y respectively. Table 1 lists the values of a_{ij} given in Brown et al. (1974). The values of b_{ij} are similarly defined.

$(\delta_{xi}, \delta_{xj})$	$X_i' > X_j'$	$X_i' = X_j'$	$X_i' < X_j'$
(1, 1)	1	0	-1
(0, 1)	1	1	$2\{\hat{\mathbb{F}}_X(x_j')/\hat{\mathbb{F}}_X(x_i'\} - 1$
(1, 0)	$1 - 2\{\hat{\mathbb{F}}_X(x_i')/\hat{\mathbb{F}}_X(x_j'\}$	-1	-1
(0, 0)	$1 - \{\hat{\mathbb{F}}_X(x_i')/\hat{\mathbb{F}}_X(x_j'\}$	0	$\{\hat{\mathbb{F}}_X(x_j')/\hat{\mathbb{F}}_X(x_i'\} - 1$

Table 1. Values of a_{ij} of Brown et al.'s estimator.

To normalize the measure to lie between [-1, 1], Brown et al. (1974) adopted Γ as their estimate of τ , denoted as $\hat{\tau}_B$. Note that this method takes partial information provided by the Kaplan-Meier estimates into account. For singly censored observations, this approach seems quite intuitive for determining the unknown relationship. However for pairs with doubly censored observations, the modifications may not be sufficient because joint information is ignored.

Method proposed by Fine et al. (2008)

When the non terminal and terminal events are positively correlated, it is natural to posit the gamma frailty model (Clayton, 1978). This model is quite easy, because it has a closed form and the association parameter is clearly interpretable as the predictive hazard ratio, i. e. the ratio between the hazard of dying at y conditionally to the fact that at x a relapse has already occurred and the hazard of dying at y conditionally to the fact that at x a relapse has not occurred yet:

$$\lim_{\Delta t \to 0} \frac{P(t \leq Y < t + \Delta t | Y \geq t, X = t)}{P(t \leq Y < t + \Delta t | Y \geq t, X > t)}$$

Since S(x, y), the joint survival function of events X and Y, is only identified when X < Y, Fine et al. (2001) define the model on the upper wedge (Day et al., 1997):

$$S(x, y) = \left[S_X(x)^{1-\theta} - S_Y(y)^{1-\theta} - 1\right]^{\frac{1}{1-\theta}} \qquad 1 \leq \theta \leq x \leq y \leq \infty \quad [1]$$

where $S_X(x)$ and $S_Y(y)$ are continuous marginal survival functions.

The authors developed a concordance estimator for θ by generalizing Oakes (1982, 1986), which is valid only when the model is on the upper wedge.

Let $\Delta_{ij} = I\left[(X_i - X_j)(Y_i - Y_j) > 0\right]$ denote the concordance status for each independent pair (X_i, Y_i) and (X_j, Y_j), representing two different subjects. With semi-competing risks data, Δ_{ij} is computable only when $\widetilde{X}_{ij} < \widetilde{Y}_{ij} < \tilde{C}_{ij}$, where $\widetilde{X}_{ij} = \min(X_i, X_j)$, $\widetilde{Y}_{ij} = \min(Y_i, Y_j)$, $\tilde{C}_{ij} = \min(C_i, C_j)$. Let $D_{ij} = I(\widetilde{X}_{ij} < \widetilde{Y}_{ij} < \tilde{C}_{ij})$ and define $\widetilde{X}'_{ij} = \min(\widetilde{X}_{ij}, \widetilde{Y}_{ij}, \tilde{C}_{ij})$ and $\widetilde{Y}'_{ij} = \min(\widetilde{Y}_{ij}, \tilde{C}_{ij})$.

The estimating equation $U(\theta) = \sum_{i<j} W(\widetilde{X}'_{ij}, \widetilde{Y}'_{ij}) D_{ij} \left[\Delta_{ij} - \frac{\theta}{1+\theta}\right] = 0$ can be used to obtain a closed-form estimator for θ:

$$\hat{\theta} = \frac{\sum_{i<j} W(\widetilde{X}'_{ij}, \widetilde{Y}'_{ij}) D_{ij} \Delta_{ij}}{\sum_{i<j} W(\widetilde{X}'_{ij}, \widetilde{Y}'_{ij}) D_{ij} (1 - \Delta_{ij})}$$

where $W(\widetilde{X}'_{ij}, \widetilde{Y}'_{ij}) = \frac{n}{\sum_i I\left[X'_{ij} \geq \widetilde{X}'_{ij}, Y'_{ij} \geq \widetilde{Y}'_{ij}\right]}$ is a useful weight function, analogous to the weighted estimator in Oakes (1986).

This parameter θ can be interpreted as the odds of concordance, being a weighting ratio between concordant and discordant pairs of Y and X times.

Positing the model [1] for X < Y, Fine et al. (2001) demonstrated that $\hat{\theta}$ is almost surely consistent for θ and that $n^{1/2}(\hat{\theta}-\theta)$ is asymptotically normal with mean zero and variance Σ which is consistently estimated by $\hat{\Sigma}=\hat{I}^{-2}\hat{J}$, where

$$\hat{I}=n^{-2}\sum_{i<j}W(\tilde{X}'_{ij},\tilde{Y}'_{ij})D_{ij}(1+\theta)^{-2},$$

$$\hat{J}=2n^{-3}\sum_{k<l<m}(\hat{Q}_{kl}\hat{Q}_{km}+\hat{Q}_{kl}\hat{Q}_{lm}+\hat{Q}_{km}\hat{Q}_{lm}),$$

$$\hat{Q}_{kl}=W(\tilde{X}'_{kl},\tilde{Y}'_{kl})D_{kl}\left[\Delta_{kl}-\frac{\hat{\theta}}{1+\hat{\theta}}\right].$$

2.6 Estimates of net survival function in the presence of semi-competing risks

In a semi-competing risks framework the interest is often focused in estimating the net survival of the intermediate non terminal event. When non terminal and terminal events are independent, Kaplan Meier method censoring times to terminal events can be used. Otherwise the association between the non terminal and terminal events has to be evaluated and specific statistical methods have to be adopted.

As in the case of competing risks, copula graphic estimator can be adopted in a semi-competing risks setting as discussed in Lakhal et al. (2008). Furthermore in the case of semi-competing risks an alternative method based on Clayton copula has been proposed by Fine (Fine et al., 2001). I describe here these two methods and later I will compare them by means of a Monte Carlo simulation.

Method proposed by Fine et al. (2001)

Fine et al. (2001) solved the problem of dependent censoring between terminal and non terminal events, by modelling the joint distribution of (X, Y) in the observable region, avoiding extrapolations in the lower wedge of (X, Y) where X>Y. In this way their model captures the identifiable features of the dependence structure of the random variables, by leaving the marginal survival distribution unspecified (Genest & MacKay, 1986), i.e. it does not focus on a particular form of the marginal survival distributions of the non terminal and terminal event and estimates them by means of non parametric Kaplan Meier method.

The model has the following functional form defined through a copula model:

$$S(x,y) = P(X > x; Y > y) = C_\theta[S_X(x), S_Y(y)] \qquad 0 \le x \le y,$$

where $C_\theta(\cdot,\cdot)$ belongs to a one-dimensional parametric family of copulas indexed by θ and models the dependency between X and Y and $S_X(\cdot)$, $S_Y(\cdot)$ are the marginal survival (or distribution) functions of the non terminal and terminal event, respectively.

When the events are positively correlated, the well-known gamma frailty copula (Clayton, 1978) is used and we resort to model [1]. Moreover when x=y=t, we obtain:

$$P(X > t; Y > t) = P(U > t) = S_U(t) = g[S_X(t), S_Y(t), \theta] = \left[S_X(t)^{1-\theta} + S_Y(t)^{1-\theta} - 1\right]^{\frac{1}{1-\theta}}$$

A closed-form estimator for $S_X(t)$ is obtained as:

$$\hat{S}_X(t) = \left[\hat{S}_U(t)^{1-\hat{\theta}} - \hat{S}_Y(t)^{1-\hat{\theta}} + 1\right]^{\frac{1}{1-\hat{\theta}}}$$

where $\hat{\theta}$ is a strongly consistent estimator for θ, $\hat{S}_U$ and $\hat{S}_Y$ are the Kaplan-Meier estimators for S_U and S_Y, using $\{(Y_i', \delta_{Y_i'}),\ i = 1, \dots n\}$ and $\{(X_i', \delta_{X_i'}),\ i = 1, \dots n\}$ respectively. When X and Y are assumed independent, $\hat{S}_X(t)$ reduces to the Kaplan-Meier estimator based on $\{(X_i', \delta_{X_i}),\ i = 1, \dots n\}$.

Recall that $\hat{\theta}$ is strongly consistent for θ. Since U and Y are subject only to independent censoring by C, $\hat{S}_U$ and $\hat{S}_Y$ are strongly consistent estimator for S_U and S_Y (Fleming & Harrington, 1991, Ch. 6). Since g has bounded derivatives, it can be shown that $\hat{S}_X(t)$ strongly converge to $S_X(t)$. This implies that, if the independence assumption holds (θ=1) for x<y, then the usual Kaplan-Meier estimator is consistent for S_X.

The covariance function estimator is:

$$\hat{\sigma}(s,t) = n^{-3} \sum_{k<l<m} \left[\hat{V}_{kl}(s)\hat{V}_{km}(t) + \hat{V}_{lm}(s)\hat{V}_{km}(t) + \hat{V}_{kl}(s)\hat{V}_{lm}(t) + \hat{V}_{lm}(s)\hat{V}_{kl}(t) + \hat{V}_{km}(s)\hat{V}_{kl}(t) + \hat{V}_{km}(s)\hat{V}_{lm}(t)\right] + n^{-3} \sum_{i<j} \hat{V}_{ij}(s)\,\hat{V}_{ij}(t)$$

where $\hat{V}_{ij}(t) = -g_1\left\{\hat{S}_U(t), \hat{S}_Y(t), \hat{\theta}\right\}\hat{S}_U(t) \int_0^t \hat{\pi}_U(u)^{-1}\left\{d\hat{M}_{Ui}(u) + d\hat{M}_{Uj}(u)\right\}$

$$-g_2\left\{\hat{S}_U(t), \hat{S}_Y(t), \hat{\theta}\right\}\hat{S}_Y(t) \int_0^t \hat{\pi}_Y(u)^{-1}\left\{d\hat{M}_{Yi}(u) + d\hat{M}_{Yj}(u)\right\} + g_3\left\{\hat{S}_U(t), \hat{S}_Y(t), \hat{\theta}\right\}\hat{I}^{-1}\hat{Q}_{ij}$$

$$g_1(a,b,c) = a^{-c}(a^{1-c} - b^{1-c} + 1)^{\frac{c}{1-c}}$$

$$g_2(a,b,c) = -b^{-c}(a^{1-c} - b^{1-c} + 1)^{\frac{c}{1-c}}$$

$$g_3(a,b,c) = (a^{1-c} - b^{1-c} + 1)^{\frac{1}{1-c}} \left[\frac{\log(a^{1-c} - b^{1-c} + 1)}{(1-c)^2} + \frac{-a^{1-c}\log(a) + b^{1-c}\log(b)}{(a^{1-c} - b^{1-c} + 1)(1-c)} \right]$$

$$\hat{\pi}_U(t) = n^{-1} \sum_{i=1}^{n} I(X_i' \geq t)$$

$$\hat{\pi}_Y(t) = n^{-1} \sum_{i=1}^{n} I(Y_i' \geq t)$$

$$\hat{M}_{Ui}(t) = I(X_i' \leq t, \delta_{X'} = 1) - \int_0^t I(X_i' \geq u) d\hat{\Lambda}_U(u)$$

$$\hat{M}_{Yi}(t) = I(Y_i' \leq t, \delta_{Y'} = 1) - \int_0^t I(Y_i' \geq u) d\hat{\Lambda}_Y(u)$$

$\hat{\Lambda}_U(u)$ and $\hat{\Lambda}_Y(u)$ are Nelson-Aalen estimators for cumulative hazards at time u, corresponding to -log(S_U) and -log(S_Y).

To construct confidence intervals for S_X, let consider that $n^{\frac{1}{2}}\{m[\hat{S}_X(t)] - m[S_X(t)]\}$ is asymptotically normal, where m is an invertible and differentiable function. For example $m(x) = \log\left[\frac{x}{1-x}\right]$. If the δ-method is applied, a (1-2α) interval for $S_X(t)$ has endpoints: $m^{-1}\left\{m[\hat{S}_X(t)] \pm n^{-\frac{1}{2}} \dot{m}[\hat{S}_X(t)] \hat{\sigma}(t,t)^{\frac{1}{2}} \phi_{1-\alpha}\right\}$, where $\dot{m}(x) = \frac{\partial[m(x)]}{\partial x}$

The estimator $\hat{S}_X$ is a step-function. The changes are at the observed values of X and Y at which $\hat{S}_U(t)^{1-\hat{\theta}} - \hat{S}_Y(t)^{1-\hat{\theta}}$ jumps. In finite samples, $\hat{S}_U(t)$ may be greater than $\hat{S}_Y(t)$, although $S_U(t) \leq S_Y(t)$, for all t. Also, $\hat{\theta}$ may be less than one. This means that $\hat{S}_X(t)$ may not be monotone or may not be well defined. In contrast, the Kaplan-Meier estimator decreases at each X' with $\delta_X = 1$. The difficulties arise in estimating probabilities in the tail of S_X with heavy censoring of X by Y.

To address the instability, we restrict inferences to the interval [0, t*], where $t* \leq \max[s: \hat{S}_U(u)^{1-\hat{\theta}} - \hat{S}_Y(u)^{1-\hat{\theta}} > -1, 0 \leq \hat{S}_X(u) \leq 1, u \leq s\}]$. For t≤t*, define the monotone estimator $\hat{S}_X^*(t) = \min_{s \leq t}[\hat{S}_X(s)]$. This estimator accepts $\hat{S}_X(t)$ if it satisfies the monotonicity constraint; if not, it carries forward the smallest value of $\hat{S}_X(s)$ for s≤t. Since $\hat{S}_X$ is uniformly consistent, so too is $\hat{S}_X^*$. We conjecture that $n^{\frac{1}{2}}[\hat{S}_X^*(t) - S_X(t)]$ and $n^{\frac{1}{2}}[\hat{S}_X(t) - S_X(t)]$ have the same limiting distribution.

Method proposed by Lakhal et al. (2008)

To estimate survival from a non terminal event, Lakhal et al. (2008) proposed to apply the copula graphic estimator. This estimator was first introduced by Zheng and Klein (1995) to estimate the survival function under a dependent censoring. They assumed that the joint distribution of the failure and censoring times follow a known copula and derived estimating equations for the marginal survival functions. When this copula is Archimedean, Rivest and Wells (2001) found a closed-form expression for the copula graphic estimator and investigated its asymptotic behaviour using martingale theory. Lakhal et al. (2008) demonstrated that in this context, the copula-graphic estimating function for $S_X(x)$ satisfies: $\hat{S}_U(t) = \phi_{\hat{\theta}}^{-1}[\phi_{\hat{\theta}}\{S_X(t)\} + \phi_{\hat{\theta}}\{S_Y(t)\}]$. Let t be an observed failure time for X, then

$$\phi_{\hat{\theta}}[\hat{S}_U(t)] = \phi_{\hat{\theta}}[\hat{S}_X^{CG}(t)] + \phi_{\hat{\theta}}[\hat{S}_Y(t)] \quad [2]$$

where $\hat{S}_X^{CG}(t)$ is the copula-graphic estimator of S_X, for example, a non-increasing step-function with jumps at the observed values of X. Because X is not censored by Y at t and U=min(X, Y), this point is also an observed failure time for U and thus is a discontinuity point for $\hat{S}_U(\cdot)$. On the other hand, t cannot be a failure time for Y by continuity and hence, $\hat{S}_Y(\cdot)$ does not jump at t. Writing [2] at t and t^- and subtracting the resulting equations yields

$$\phi_{\hat{\theta}}[\hat{S}_U(t)] - \phi_{\hat{\theta}}[\hat{S}_U(t^-)] = \phi_{\hat{\theta}}[\hat{S}_X^{CG}(t)] - \phi_{\hat{\theta}}[\hat{S}_X^{CG}(t^-)].$$

Summing these terms over all observed failure times of X prior to t gives

$$\phi_{\hat{\theta}}[\hat{S}_X^{CG}(t)] = \sum_{t_i \le t; \delta_{xui}=1} \phi_{\hat{\theta}}[\hat{S}_U(t_i)] - \phi_{\hat{\theta}}[\hat{S}_U(t_i^-)],$$

and the copula-graphic estimator for $S_X(t)$ is

$$\hat{S}_X^{CG}(t) = \phi_{\hat{\theta}}^{-1}\left\{\sum_{t_i \le t; \delta_{xui}=1} \phi_{\hat{\theta}}[\hat{S}_U(t_i)] - \phi_{\hat{\theta}}[\hat{S}_U(t_i^-)]\right\}.$$

2.7 Regression models in survival analysis

Let S(t|Z) be the survival function of T given Z. A general regression model can be expressed as:

$$g[S(t|Z)] = h(t) + Z^T\beta$$

Where g is a known decreasing function and h(t) is a completely unspecified strictly increasing function and β is a px1 vector of unknown regression coefficients.

When g[S(t|Z)]=log(-log(S(t|Z))=Λ(t|Z), Cox proportional hazard model is obtained;

when g[S(t|Z)]=logit(S(t|Z)), proportional odds model is obtained (Cheng et al., 1995).

For two individuals with covariate vectors Z_1 and Z_2, the model satisfy a vectical shift after transformation:

$$g[S(t|Z_1)] - g[S(t|Z_2)] = (Z_1 - Z_2)\beta$$

If the model is on "overall" survival, the model is commonly expressed in terms of $\lambda(t|Z)$, i.e. the instantaneous hazard of event. If g is complementary log log and the proportional hazard holds this is equivalent to model $\Lambda(t|Z)$, otherwise, in the case of covariate effects which varies in times, the covariate effects on $\lambda(t|Z)$ could differ from the effects on $\Lambda(t|Z)$. Because of the relationship between hazard and survival if regression coefficients are different from 0, this implies a difference between survival functions.

If h(t) is unspecified these correspond to semi-parametric regression models.

In the case of competing risks models can be expressed in terms of $\lambda_k(t|Z)$ or $\Lambda_k(t|Z)$. This is a widely used approach since time to occurrence of other events is simply censored. In the case of independence among events this is equivalent to modelling net hazard.

If crude cumulative incidence is of concern, the regression model on $\lambda_k(t|Z)$ cannot be adopted to evaluate differences in the crude cumulative incidences, because the lack of direct relationship between $\lambda_k(t|Z)$ and $I_k(t|Z)$. Regression models must be expressed in terms of the subdistribution hazard $\tilde{\lambda}_k(t)$. The general regression model is represented as g(I(t|Z)). In the case of complementary log-log function, Fine and Gray regression model is obtained (Fine and Gray, 1999).

Model estimates are based on specific likelihood functions because of the semi-parametric model definition. In the case of Cox model a partial likelihood is of concern. Likelihood ratio test or Wald test can be used for inference. In the case of Fine and Gray model the likelihood has been modified regarding the risks set and introducing IPCW (inverse probability of censoring weighted) for censoring. Because of the IPCW the likelihood is not proper, thus likelihood ratio test cannot be used.

The regression model previously given is equivalent to the linear transormation model:

$$h(T) = -Z^T\beta + \varepsilon$$

where ε is a random error with distribution $F=1-g^{-1}$.

In the case of specified h(T) or ε, parametric regression models are obtained, as in the case of Weibull which is an Accelerated Failure Times model, after defining ε as extreme values distribution.

A bridge from semi-parametric to parametric regression models can be obtained by piecewise models, where time is partitioned and for each partition a parametric distribution, which may vary in the different partitions, is considered. The most popular is the piecewise exponential model, in which the parameter of the exponential distribution depends on the partition. The relationship between the likelihood function of some parametric survival models and generalized linear models is well known (Aitkin et al., 1989). This allows to implement regression model for survival analysis by software for generalized linear models.

In particular for piecewise exponential regression models, the dataset needs to be organized in such a way that a subject is replicated for each time he/she is at risk, including a status variable for each replication. The dependent variable is now status, the error function is poisson and the link function is log. The advantage is the possibility to model the shape of the hazard function during time including into linear predictor splines for time intervals. This approach renders easy to model non proportional covariate effects by interaction between covariate values and splines for time intervals.

The approach can be easily extended to cause specific hazards by modifying the pertinent status. Thus GLM approach can be used for net hazard in the case of independence among events.

If relative survival need to be considered, GLM approach can be used. Several model structure are available (Estève et al., 1990; Hakulinen and Tenkanen, 1987; Dickman et al., 2003; Andersen et al. 1985). In particular, Dickman proposed the following one:

$$\lambda_O(t|Z) = \lambda_E(t|Z) + \exp(Z\beta)$$

This is estimated by a generalized regression model where the number of death is the response, the error sistributrion is Poisson, the offset is the logarithm of the person time at risk and the link function is the logarithm of the difference between the mean of observed and expected number of deaths (Dickman et al., 2003).

In the case of crude cumulative incidence, no direct relationship can be found between the likelihood of models on subdistribution hazard and the likelihood of generalized linear models. This has been solved by considering regression models on “pseudo-values” on crude cumulative incidence and using GEE (Generalized Estimating Equations) for estimates and inference. It is worth of note that pseudo-values can be used in every survival regression models.

Pseudo-observations in survival analysis

One way of setting up regression models for any function f(X) and check such models with censored survival data (and with more general incomplete event history data) is to replace f(X) by the "*pseudo-observations*" (Andersen and Pohar-Perme, 2010).

The basic idea is simple. If the data were complete, $f(X_i)$ would be observed for each individual i and the expected value E(f(X)) could be estimated by $\frac{1}{n}\sum_i f(X_i)$. Conversely, suppose that the data are incomplete (e.g. some observations are censored and therefore not all $f(X_i)$ are observed), but a well-behaved estimator for the expectation ϑ=E(f(X)) is available anyway, e.g. the Kaplan–Meier estimator for S(t)=E(I(X>t)). The pseudo-observation for f(X) for individual i, i = 1, . . . , n, is then defined as

$$\hat{\vartheta}_i = n \cdot \hat{\vartheta} - (n-1) \cdot \hat{\vartheta}^{-i}$$

where $\hat{\vartheta}^{-i}$ is the estimator applied to the sample of size n−1 obtained by eliminating the i-th individual from the data set. Intuitively the i-th pseudo-observation can be viewed as the contribution of the individual i to the E(f(X)) estimate on the sample of size n.

In the absence of censoring, at each time t pseudo values can assumed only two values: 1 if the subject is still alive at the time t and 0 if the subject has died before t. In the presence of censored times, values can be lower than 0 or greater than 1.

The pseudo-observations are computed for every individual at predefined time points, usually corresponding to specified quantiles of the survival function.

If a sample of n subjects and k time points are considered, a dataset composed by n*k not independent pseudo-observations is obtained.

The idea is now to replace the incompletely observed $f(X_i)$ by $\hat{\vartheta}_i$, that is

(1) $\hat{\vartheta}_i$ may be used as an outcome variable in a generalised linear regression model with some link function g:

$$g[E(f(X)|Z)] = \beta_0 + \sum_j \beta_j Z_j \qquad [9],$$

or

(2) $\hat{\vartheta}_i$ may be used to compute residuals or in a scatterplot when assessing model assumptions.

Regardless of the application, the pseudo-observations $\hat{\vartheta}_i$ will always be used for all n subjects and not only for those where $f(X_i)$ was unobserved.

The pseudo-observations at a fixed time-point t exhibit the following two properties:

(P1) the $\hat{\vartheta}_i(t)$'s are approximately i.i.d. and

(P2) the $\hat{\vartheta}_i(t)$'s are conditionally unbiased given the covariates, $E\left[\hat{\vartheta}_i(t)|Z_i\right] = F_1(t|Z_i) + o_p(1)$

provided that (i) the censoring time is independent of covariates, failure time, and cause of failure, and (ii) $t < t^*$, where t^* is such that the survival function of the censoring time, G, satisfies $G(t^*) > \upsilon$ for a fixed $\upsilon > 0$ (Graw et al., 2009, Lemma 2). These two properties make pseudo-observations suitable to use as alternative outcomes for regression purposes when there is censoring. Indeed, if pseudo-observations are set as the outcome variables for both censored and uncensored individuals, then generalized estimating equations (GEE) can be used to fit model [9] because properties (P1) and (P2) guarantee the consistency and asymptotic normality of the estimates obtained in this way.

In the absence of censoring canonical link function for proportion would be used, as complementary log log or logit with binomial error distribution. Nevertheless in the presence of censoring pseudo-values can be lower than 0 or greater than 1, thus it is usually adopted the same link function, but a Gaussian error distribution.

Pseudo-observations provide a common approach to various kinds of models by replacing the incompletely observed outcome and then fitting using generalised estimating equations. For example they are currently used in survival analysis in the presence of competing risks and their good performance in terms of consistency and asymptotic normality in estimating the effect of covariate on the crude cumulative incidence function has already been demonstrated (Graw et al., 2009).

2.8 Regression models on net survival when independence assumption is not tenable

In the presence of competing or semi-competing risks, the main problem of interest is to estimate the covariate effect on the survival function of the terminal event X at time t: $\beta_X(t)$. Because X is subject to dependent censoring by the terminal event Y, the inference of $\beta_X(t)$ becomes complicated and difficult.

Specific regression models on net survival have been proposed. Taking into account that the marginal distribution of latent variables can be identify for a given dependence structure, Lo and Wilke (2011) suggest a plug-in regression framework for the copula-graphic estimator. Their model is an attractive empirical approach, as it does not require parametric knowledge

of the marginal distribution, but it expresses the marginal distributions ($S_k(t|Z)$) in terms of estimable quantities (crude cumulative incidences $I_k(t|Z)$, for k=1, ..., K), given a specified copula structure. After estimating the marginal distribution ($S_k(t|Z)$), the effect of covariate Z is obtained as $\frac{\partial S_k(t|Z)}{\partial Z}$.

As regards regression models in a semi-competing risks setting two models are available from the literature (Peng and Fine, 2007; Hsieh and Huang, 2012). They both employ a copula function and enable varying the effect of covariate and association parameter on time, but Peng and Fine use an approach based on method of moment, while Hsieh and Huang proposes a conditional likelihood approach.

I propose here to adopt the methodology based on pseudo-observations and pertinent regression methods based on generalized linear models.

Method proposed by Peng and Fine (2007)

To formulate covariate effects on time to the non terminal event X, it is tempting to employ the popular proportional hazards model, that is:

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta_0^T Z),$$

where $\lambda(t|Z)$ denotes the hazard function of X conditional on Z, $\lambda_0(t)$ is an unspecified baseline hazard function, and β_0 is a $p \times 1$ coefficient vector.

In practice, restricting the hazard functions associated with two sets of covariates to be proportional over time may be unrealistic. The proportional hazards model can be equivalently represented as

$$S_X(t|Z) = \exp\{-\exp[\log\Lambda_0(t) + \beta_0^T Z]\}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is the cumulative hazard. Peng and Fine (2007) proposed accommodating time-varying covariate effects on the survival function of X via a generalized functional linear model:

$$S_X(t|Z) = g[\theta_0(t)^T \tilde{Z}] \qquad [4]$$

where g(·) is a known monotone function, $\tilde{Z} = (1, Z^T)^T$, and $\theta_0(t)$ is a $(p+1) \times 1$ vector of unknown time-dependent coefficients and is completely unspecified in t but is assumed to be a right-continuous function with left-hand limits. This model defines a rich family of varying-coefficient regression models. Choosing $g = \{\exp[-\exp(\cdot)]\}$ and $g = \exp/(1+\exp)$, the model [4] accommodates respectively the standard proportional hazards model and the

proportional odds model. The survival based functional regression modelling facilitates estimation without involving smoothing. It also renders straightforward interpretations of the time-varying parameter $\theta_0(t)$ via the generalized linear model representation, namely, $g^{-1}[S_X(t|Z)] = \theta_0(t)^T Z$.

With semi–competing risks, estimation of θ_0 requires a model for the dependence structure of (X, Y), since Y may dependently censor X. Peng and Fine (2007) proposed linking the joint distribution of (X, Y) to its marginal distributions through a known time-independent copula function. It is assumed that in the observable region of the data

$$P(X > s, Y > t|Z) = C\{S_X(s|Z), S_Y(t|Z), \alpha_0(s,t)\} \qquad 0 \le s \le t \qquad [5]$$

where $\alpha_0(s,t)$ is an unknown time-varying parameter, which is also a right-continuous function with left-hand limits. In general, it can be interpreted as the standard odds ratio based on the binary random variables I(X>s) and I(Y>t). Depending on the parameterization, larger values of $\alpha_0(s,t)$ generally correspond to either increasing positive or negative association defined by $\frac{P(X>s,Y>t)}{P(X>s)P(Y>t)}$ >1 or <1, respectively (Nelsen, 1999). Unlike parameterizations based on hazard association measures, the copula parameterization in [5] yields an explicit form for the joint distribution.

Since Y is subject to censoring only by C, the regression model for Y can be chosen among existing models for standard independently right-censored data. To simplify the developments, the model for $S_Y(t|Z)$ is assumed to take the form

$$S_Y(t|Z) = h\left[\eta_0(t)^T \tilde{Z}\right] \qquad [8]$$

where h is a known link function and $\eta_0(t)$ is estimable with existing methods.

Under models [6]-[8], the covariate effects on T_1 and the dependence parameter can be estimated simultaneously on the basis of a set of nonlinear estimating equations, which adopts a "working independence" assumption across time. Let $\alpha(t) = \alpha(t,t)$. The estimator $\left[\hat{\alpha}(t), \hat{\theta}(t)\right]$ is obtained as the solution of

$U[\alpha(t), \theta(t), \hat{\eta}(t), t] = n^{-\frac{1}{2}} \sum_{i=1}^{n} A_i[\alpha(t), \theta(t), \hat{\eta}(t), t]$, where

$A_i[\alpha(t), \theta(t), \eta(t), t] =$

$V_i[\alpha(t), \theta(t), t] D_i[\alpha(t), \theta(t), \eta(t)]\{I(X_i > t) - I(Y_i > t)\Psi[\alpha(t), \theta(t)^T \tilde{Z}_i, \eta(t)^T \tilde{Z}_i]\}$,

$$D_i\{\alpha(t), \theta(t), \eta(t)\} = \frac{\partial \Psi[\alpha(t), \theta(t)^T \tilde{Z}_i, \eta(t)^T \tilde{Z}_i]}{\partial \begin{pmatrix} \alpha(t) \\ \theta(t) \end{pmatrix}}$$

and V_i is a scalar weight function, i=1, ... , n. One can show that $\hat{\alpha}(t)$ and $\hat{\theta}(t)$ are step functions that jump only at observed failure and censoring times. The estimating equation needs to be solved only at finitely many timepoints.

Under certain regularity conditions including restrictions on $\hat{\eta}(t)$, as n approaches infinity, there exists a unique solution to $U\{\alpha(t), \theta(t), \eta(t), t\} = 0$ in a neigborhood of (α_0, θ_0) that converges to ($\alpha_0(t)$, $\theta_0(t)$) in probability, uniformly in $t \in [l, u]$. It is further shown that $n^{\frac{1}{2}}\left\{\left[\hat{\alpha}(t)^T, \hat{\theta}(t)^T\right]^T - [\alpha(t)^T, \theta(t)^T]^T\right\}$ converges weakly to a tight Gaussian process. The conditions on $\hat{\eta}(t)$ for validity of $\hat{\alpha}(t)$ and $\hat{\theta}(t)$ are verified under proportional hazards models.

Method proposed by Hsieh and Huang (2012)

Hsieh and Huang (2012) proposes a conditional likelihood approach to estimate ($\alpha_0(t)$, $\theta_0(t)$). Consider the complete likelihood function based on observable indicators.

Let $I_X(t) = I(X > t)$, $I_Y(t) = I(Y > t)$. Because $X = T_1 \wedge T_2 \wedge C$ and $Y = T_2 \wedge C$, the possible values of $(I_X(t), I_Y(t))$ are

(i) $I_X(t) = 1$, $I_Y(t) = 1$,

(ii) $I_X(t) = 0$, $I_Y(t) = 1$,

(iii) $I_X(t) = 0$, $I_Y(t) = 0$.

The conditional probabilities of $I_X(t)$ given $I_Y(t)$ are:

$$\text{(i)}\quad P(I_X(t) = 1 | I_Y(t) = 1) = \frac{P(T_1 > t, T_2 > t|Z)P(C > t|Z)}{P(T_2 > t|Z)P(C > t|Z)} = \frac{P(T_1 > t, T_2 > t|Z)}{P(T_2 > t|Z)}$$

$$\text{(ii)}\quad P(I_X(t) = 0 | I_Y(t) = 1) = \frac{P(T_2 > t|Z)P(C > t|Z) - P(T_1 > t, T_2 > t|Z)P(C > t|Z)}{P(T_2 > t|Z)P(C > t|Z)} =$$

$$= \frac{P(T_2 > t|Z) - P(T_1 > t, T_2 > t|Z)}{P(T_2 > t|Z)}$$

$$\text{(iii)}\ P(I_X(t) = 0 | I_Y(t) = 0) = 1.$$

Based on the observed indicators $\{(I_{X_i}(t), I_{Y_i}(t)) : i = 1, \ldots, n\}$, consider the likelihood function:

$$L(\alpha(t), \theta(t), \eta(t) = P\left(I_{X_1}(t) = x_1, I_{Y_1}(t) = y_1, \ldots, I_{X_n}(t) = x_n, I_{Y_n}(t) = y_n\right) =$$

$$= P\left(I_{X_1}(t) = x_1, \ldots, I_{X_n}(t) = x_n \middle| I_{Y_1}(t) = y_1, \ldots, I_{Y_n}(t) = y_n\right)$$

$$\times P(I_{Y_1}(t) = y_1, I_{Y_2}(t) = y_2, \ldots, I_{Y_n}(t) = y_n) = L_c(\alpha(t), \theta(t), \eta(t)) \times L_m(\eta(t)).$$

Note that $L_m(\eta(t)) = P(I_{Y_1}(t) = y_1, I_{Y_2}(t) = y_2, \dots, I_{Y_n}(t) = y_n)$ only contains the information of $\eta_0(t)$, and all the information of $(\alpha_0(t), \theta_0(t))$ is contained in

$$L_c\big(\alpha(t), \theta(t), \eta(t)\big) = P\big(I_{X_1}(t) = x_1, \dots, I_{X_n}(t) = x_n \big| I_{Y_1}(t) = y_1, \dots, I_{Y_n}(t) = y_n\big) =$$

$$= P(I_{X_1}(t) = x_1 | I_{Y_1}(t) = y_1) \times P(I_{X_2}(t) = x_2 | I_{Y_2}(t) = y_2) \times \cdots \times P(I_{X_n}(t) = x_n | I_{Y_n}(t) = y_n).$$

Note that $L_c\big(\alpha(t), \theta(t), \eta(t)\big)$ does not involve the distribution of C, which is unknown, because $P(C > t)$ in $L_c\big(\alpha(t), \theta(t), \eta(t)\big)$ can be canceled. Thus, fixed $\eta(t) = \hat{\eta}(t)$, to maximize $L\big(\alpha(t), \theta(t), \hat{\eta}(t)\big)$ respective to $\alpha(t)$ and $\theta(t)$ is equivalent to maximize $L_C(\alpha(t), \theta(t), \hat{\eta}(t))$ respective to $\alpha(t)$ and $\theta(t)$. Therefore, it is possible to obtain the maximized conditional likelihood estimator of $(\alpha_0(t), \theta_0(t))$ by maximizing $\log[L_c\{\alpha(t), \theta(t), \hat{\eta}(t)\}]$ where

$$\log[L_c\{\alpha(t), \theta(t), \eta(t)\} =$$

$$= \sum_{i=1}^{n} \Big\{ I_{X_i}(t) I_{Y_i}(t) \log\{\Psi[\alpha(t), \theta(t)^T Z_i, \eta_0(t)^T Z_i]\}$$

$$+ \big[1 - I_{X_i}(t)\big] I_{Y_i}(t) \log\{1 - \Psi[\alpha(t), \theta(t)^T Z_i, \eta(t)^T Z_i]\}\Big\},$$

where $\Psi\{\alpha(t), \theta(t)^T Z_i, \eta(t)^T Z_i\} = \dfrac{P\big(T_{1_i} > t, T_{2_i} > t \big| Z_i\big)}{P\big(T_{2_i} > t \big| Z_i\big)} = \dfrac{C\{g[\theta_0^T(t) Z_i], h[\eta_0^T(t) Z_i]\}}{h[\eta_0^T(t) Z_i]}$.

Therefore, the estimator of $(\alpha_0(t), \theta_0(t))$, denoted as $(\hat{\alpha}(t), \hat{\theta}(t))$, is the solution of $U\big(\alpha(t), \theta(t), \hat{\eta}(t)\big) = 0$, where

$$U\big(\alpha(t), \theta(t), \eta(t)\big) = \frac{\partial \log[L_c\{\alpha(t), \theta(t), \eta(t)\}]}{\partial \begin{pmatrix} \alpha(t) \\ \theta(t) \end{pmatrix}} =$$

$$= \sum_{i=1}^{n} \left\{ D_i\{\alpha(t), \theta(t), \eta(t)\} \frac{I_{X_i}(t) I_{Y_i}(t) 1}{\Psi\{\alpha(t), \theta(t)^T Z_i, \eta(t)^T Z_i\}} - \frac{\big(1 - I_{X_i}(t)\big) I_{Y_i}(t) 1}{1 - \Psi\{\alpha(t), \theta(t)^T Z_i, \eta(t)^T Z_i\}} \right\},$$

where $D_i\{\alpha(t), \theta(t), \eta(t)\} = \dfrac{\partial \Psi\{\alpha(t), \theta(t)^T Z_i, \eta(t)^T Z_i\}}{\partial \begin{pmatrix} \alpha(t) \\ \theta(t) \end{pmatrix}}$.

Estimate $\alpha_0(t)$ and $\theta_0(t)$ separately at each t because the estimating function, $U(\alpha(t), \theta(t), \hat{\eta}(t))$, jointly estimates $(\alpha_0(t), \theta_0(t))$ adopting the "work-independence" assumption across time. Note that $\hat{\alpha}(t)$ and $\hat{\theta}(t)$ are step functions because they only jump at observed failure times and censoring times. Hence, the estimating function only needs to

be solved at finite time points. Moreover, the standard statistical software may facilitate this optimization.

This study also presents the large sample properties of $\left(\hat{\alpha}(t), \hat{\theta}(t)\right)$. Replacing $\eta_0(t)$ with $\hat{\eta}(t)$ in the estimating procedures complicates the proofs of the large sample properties.

This study extends the technique proposed by Peng and Fine (2007) to prove the large sample properties of $\left(\hat{\alpha}(t), \hat{\theta}(t)\right)$. Two theorems provide uniform consistency of $\left(\hat{\alpha}(t), \hat{\theta}(t)\right)$ and the Gaussian process of $\sqrt{n}\begin{pmatrix}\alpha(t) - \alpha_0(t)\\ \theta(t) - \theta_0(t)\end{pmatrix}$.

Method based on pseudo-observations

I propose here to extend the use of pseudo-observations in survival analysis in the presence of semi-competing risks, by estimating survival probability with a semi-parametric method, as that adopted by Fine at al. (2001), which is a consistent estimator and then computing pseudo-observations and analysing them in a generalized linear model to evaluate the effect of specific covariates on survival to intermediate events.

Pseudo-values observations could be used also in a semi-competing risks setting for fitting a regression model on survival for the non terminal event. In this context the pseudo-observations are computed as follows:

$$\hat{\vartheta}_i = n \cdot \hat{\vartheta} - (n-1) \cdot \hat{\vartheta}^{-i}$$

where $\hat{\vartheta}$ is the survival estimator for the non-terminal event $S_x(t)$ computed on the whole dataset and $\hat{\vartheta}^{-i}$ is the survival estimator for the non-terminal event applied to the sample of size n−1 obtained by eliminating the i-th individual from the data set.

After computing the pseudo-observation for every individual at predefined time points, usually corresponding to specified quantiles of the survival function, a generalized linear model with a chosen link function is applied. They enables estimating the effect of covariate on the occurrence of the non-terminal event.

3. Results

3.1 Monte Carlo simulation

Monte Carlo simulations were conducted to investigate the performance of methods outlined in section 2 for the estimate of the copula association parameter in the presence of semi-competing risks, to compare the copula-graphic and Fine's estimator for survival function of non terminal event and to investigate the performance of regression method based on pseudo-observations theory in a context of semi-competing risks.

In order to generate multivariate survival data I refer to the simulation procedure based on copulas proposed by Rotolo et al. (2013). The conditional survival function of the k-th time, given the k-1 previous ones, is:

$$S_{(k)|(1),\dots,(k-1)}\left(t_{(k)}\middle|t_{(1)},\dots,t_{(k-1)}\right)=\frac{\frac{\partial^{k-1}}{\partial t_{(1)}\dots\partial t_{(k-1)}}S_{J_0}(t_{(1)},\dots,t_{(k)},0,\dots,0)}{\frac{\partial^{k-1}}{\partial t_{(1)}\dots\partial t_{(k-1)}}S_{J_0}(t_{(1)},\dots,t_{(k-1)},0,\dots,0)}$$

For ease of notation and of presentation, consider as an example simulation from a bivariate Clayton copula function, after fixing the association between times to different events and the marginal function of the two events of interest. Given $U_i \sim U(0,1), i.i.d.$, the simulation algorithm is illustrated below:

1. Generate a value for X from its marginal survival function:
   $X = S_x^{-1}(U_1)$
2. Conditionally on X=x, generate Y:

$$Y|x = S_{y|x}^{-1}(U_2|x) = S_y^{-1}\left(\left\{\left[U_2^{-\frac{\theta}{1+\theta}}-1\right]S_x(x)^{-\theta}+1\right\}^{-\frac{1}{\theta}}\right)$$

Simulation on the association parameter and survival function

Clayton Copula is the widest used function to model the joint distribution of multivariate time to events. It is worth of note that, as the bivariate distribution is unknown, there is no guarantee that Clayton copula structure is adequate to estimate net survival function of non terminal event. Nevertheless, in the presence of competing or semi-competing risk it is not possible to verify the underlying complete bivariate structure.

In the case Clayton Copula is the true structure, the aim of the simulation is to investigate the correctness and coverage of the estimates.

In the case of miss-specified copula structure, the aim of simulation is to investigate the robustness of the estimator. To this issue we generate data from a Frank copula and use the Clayton copula structure in order to estimate both association parameter and survival function.

The simulations scheme is based on Clayton's copula, with $\phi_\theta(t) = \frac{1}{\theta}(t^{-\theta} - 1)$ and Frank's copula, with $\phi_\theta(t) = \log\left(\frac{1-e^{-\theta}}{1-e^{-\theta t}}\right)$. The dependence parameters are those corresponding to unconditional Kendall's τ of τ = 0, 0.333, 0.5 and 0.75. Samples of sizes 200 are used. The random variable X has a unit exponential distribution; Y has a unit exponential distribution as well, such that P(X > Y) = 0.5. The censoring variable C follows a uniform distribution on [0, a], where a is such that P(Y > C) = 20%. All simulations are based on 1000 replicates.

Results on the association parameter

In Table 1, I report the means and standard deviations of the Kendall's τ computed by methods proposed by Brown (1974) and those obtained using the relationship with the copula association parameter estimated by methods proposed by Fine et al. (2001). Data generated by Clayton's and Frank's copula are considered with different degrees of association. Table 1 shows that the method proposed by Brown for the association of doubly censored survival times is unbiased only in the presence of independence between times to different events (τ=0), whereas it is not appropriate in the presence of dependent censoring, as arises in semi-competing risks setting with non null association between times to non terminal and terminal events. These considerations are valid both when data are generated by Clayton or Frank copula models. On the other hand method proposed by Fine has a better performance. In particular this method has a very good performance when the data distribution actually follows a Clayton copula model, in fact the association estimates are approximately unbiased and with a small standard deviation; whereas it has a lower performance when data are actually generated from a Frank copula models, even though the biases are quite small.

Data generated by Clayton copula							
τ=0		τ=0.333		τ=0.5		τ=0.75	
Brown	Fine	Brown	Fine	Brown	Fine	Brown	Fine
0.005 (0.002)	0.000 (0.004)	0.208 (0.002)	0.335 (0.004)	0.312 (0.002)	0.501 (0.002)	0.475 (0.001)	0.750 (0.001)

Data generated by Frank copula							
τ=0		τ=0.333		τ=0.5		τ=0.75	
Brown	Fine	Brown	Fine	Brown	Fine	Brown	Fine
0.005 (0.045)	0.000 (0.061)	0.216 (0.043)	0.258 (0.062)	0.327 (0.041)	0.415 (0.057)	0.489 (0.037)	0.673 (0.042)

Table 1. Simulation results for the Kendall's τ (true values in the second row) generate by Clayton and Frank Copulas and estimate by methods of Brown (1974) and Fine et al. (2001). Mean and standard errors ( ) are shown on 1000 sample of size 200.

Results on the estimating methods of net survival in the presence of semi-competing risks

As regards estimators for $S_X(t)$, they are evaluated at $t_i=-\log(i/10)$, for i = 1, 3, 5, 7, 9, corresponding to the 10th, the 30th, the 50th, the 70th, and the 90th percentile of $S_X(t)$, the unit exponential survival function. Three estimators for $S_X(t)$ are compared: that proposed by Fine et al. (2001), the copula graphic estimator proposed by Lakhal et al. (2008) and the naive Kaplan Meier estimator. When data are generated from a Clayton copula models, both semi-parametric estimator proposed by Fine and copula graphic estimator are almost unbiased, whereas when data are generated from a Frank copula model they are something biased and their performance decreases at increasing association. The efficiency of the two estimators is quite similar. On the contrary the Kaplan Meier estimator is accurate only under independence, whereas it may severely overestimate the survival probabilities when there is positive association between times to different events.

Data generated by Clayton copula												
	τ=0			τ=0.333			τ=0.5			τ=0.75		
	Fine	CG	KM	Fine	CG	KM	Fine	CG	KM	Fine	CG	KM
s10	0.077 (0.004)	0.108 (0.006)	0.106 (0.005)	0.095 (0.002)	0.105 (0.002)	0.228 (0.004)	0.099 (0.002)	0.105 (0.001)	0.267 (0.004)	0.100 (0.001)	0.105 (0.001)	0.299 (0.003)
s30	0.291 (0.004)	0.300 (0.003)	0.299 (0.003)	0.297 (0.003)	0.303 (0.002)	0.420 (0.002)	0.298 (0.002)	0.303 (0.002)	0.466 (0.002)	0.301 (0.002)	0.303 (0.001)	0.519 (0.002)
s50	0.494 (0.003)	0.498 (0.002)	0.499 (0.002)	0.499 (0.003)	0.502 (0.002)	0.579 (0.002)	0.495 (0.003)	0.501 (0.002)	0.614 (0.002)	0.500 (0.002)	0.502 (0.002)	0.669 (0.001)
s70	0.698 (0.002)	0.701 (0.001)	0.701 (0.001)	0.697 (0.002)	0.700 (0.001)	0.734 (0.001)	0.696 (0.002)	0.699 (0.002)	0.754 (0.001)	0.698 (0.002)	0.701 (0.002)	0.794 (0.001)
s90	0.899 (0.000)	0.900 (0.000)	0.900 (0.000)	0.898 (0.001)	0.899 (0.000)	0.904 (0.000)	0.898 (0.001)	0.899 (0.001)	0.908 (0.000)	0.899 (0.001)	0.901 (0.001)	0.920 (0.000)

Data generated by Frank copula												
	τ=0			τ=0.333			τ=0.5			τ=0.75		
	Fine	CG	KM	Fine	CG	KM	Fine	CG	KM	Fine	CG	KM
s10	0.077 (0.067)	0.108 (0.075)	0.106 (0.067)	0.045 (0.030)	0.071 (0.043)	0.159 (0.066)	0.046 (0.024)	0.068 (0.033)	0.194 (0.066)	0.063 (0.023)	0.068 (0.033)	0.245 (0.059)
s30	0.291 (0.063)	0.300 (0.059)	0.299 (0.053)	0.305 (0.062)	0.307 (0.053)	0.401 (0.050)	0.300 (0.060)	0.303 (0.049)	0.444 (0.049)	0.299 (0.050)	0.303 (0.049)	0.503 (0.044)
s50	0.494 (0.052)	0.498 (0.048)	0.499 (0.045)	0.563 (0.057)	0.534 (0.048)	0.587 (0.042)	0.578 (0.064)	0.539 (0.050)	0.624 (0.041)	0.568 (0.069)	0.539 (0.050)	0.672 (0.037)
s70	0.698 (0.040)	0.701 (0.037)	0.701 (0.037)	0.750 (0.041)	0.726 (0.037)	0.748 (0.033)	0.773 (0.044)	0.737 (0.038)	0.772 (0.032)	0.789 (0.051)	0.737 (0.038)	0.807 (0.030)
s90	0.899 (0.021)	0.900 (0.021)	0.900 (0.021)	0.910 (0.021)	0.905 (0.022)	0.908 (0.021)	0.917 (0.022)	0.909 (0.022)	0.914 (0.020)	0.930 (0.021)	0.909 (0.022)	0.928 (0.018)

Table 2. Simulation results for the survival function of the non terminal event estimated by Fine et al. method, copula graphic estimator (CG) of Lakhal et al. and Kaplan-Meier method (KM). For every simulation scenario 1000 sample of size 200 are generate by Clayton or Frank Copulas. Survival functions are evaluated at $t_i=-\log(i/10)$, for i=1, 3, 5, 7, 9, corresponding to the 10th, the 30th, the 50th, the 70th, and the 90th percentile of the true marginal survival function. Mean and standard errors () of survival functions are reported here.

Simulation on the regression model

I proposed here a simulation procedures to evaluate and compare the performances of two regression models in the presence of semi-competing risks: the one proposed by Hsieh and Huang (2012) based on a conditional likelihood approach and the regression method based on pseudo observations.

The simulation scheme mimics that already adopted by Hsieh and Huang.

Consider the models: $\log(X_i/3) = -\beta_X Z_i + e_{Xi}$ and $\log(Y_i/3) = -\beta_Y Z_i + e_{Yi}$, for i = 1, ... , n, where Z_i is a normal random variable with mean 1 and variance 0.5 constrained in [0,2], $Pr(e_{Xi} > x)$ and $Pr(e_{Yi} > x)$ both follow e^{-e^x}. This gives proportional hazard models for X_i and Y_i, and the dependence structure of (e_{Xi} , e_{Yi}) follows the Clayton model as:

$$\Pr(e_{Xi} > x, e_{Yi} > y) = \left[\Pr(e_{Xi} > x)^{1-\theta} + \Pr(e_{Yi} > y)^{1-\theta} - 1\right]^{\frac{1}{1-\theta}}.$$

The parameter settings in this study include θ = 1.5 (corresponding to a Kendall's τ=0.2), $\beta_x = 1$ (meaning that the treatment has a small effect on the non terminal event X), $\beta_y = 0$ (meaning that the treatment has no effect on the terminal event Y) or $\beta_y = 0.2$ (meaning that the treatment has a small effect on the terminal event Y); and sample sizes n of 200 and 500. For $\beta_y = 0$, the independently censoring time C_i is generated from U(1,10), in which the censoring percentages for X and Y are 27% and 23%, respectively. For $\beta_y = 0.2$, the independent censoring time C_i is generated from U(0,1) if γ = 1 and from U(1,1.2) if γ = 0, where γ is from Bernulli (0.2). In this case, the censoring percentages of X and Y are 52% and 67%, respectively. All simulations are based on 1000 replicates. Although Hsieh and Huang method estimates simultaneously the association parameter and the regression coefficients for the non terminal event, my attention is focused here on the performance in estimating the regression coefficient β_X. As regard the regression model based on pseudo-observations a generalized linear model with Gaussian error and complementary log-log link function. Two alternative proportional hazard regression models on pseudo-observations are considered: the first one is a regression model which assume a constant hazard over time thus the model includes the treatment covariate only. The second one is a regression model which does not assume a constant hazard over time thus the model includes the treatment covariate and time point used to generate pseudo values by dummy variables.

Table 3 presents the mean, bias, empirical standard deviation (EmpSD), average modified standard deviation (ModSD) and the coverage probability (CP) of the nominal 95%

confidence interval of β_X estimated by pseudo-observations method and mean, bias, empirical standard deviation (EmpSD) of β_X estimated by Hsieh and Huang's method. Note that our pseudo-observation method estimates a single coefficients for the treatment, while Hsieh and Huang's regression method estimates the regression coefficient at different time points previously determined. Hsieh and Huang regression model gives almost unbiased estimates of the regression coefficient and small empirical standard deviation. Model based on pseudo-observation with treatment covariate only has the lower performance in terms of bias and small empirical standard deviations, particularly when the covariate acts only on the non terminal event, and not non the terminal event (β_Y = 0). On the contrary regression model based on pseudo observations with covariate and time effect has a better performance in terms of bias but a bigger empirical standard deviation; however its empirical standard deviation is similar to the average modified standard deviation and the coverage probability of the 95% confidence interval is near 0.95.

n=200

β_Y=0 β_Y=0.2

Pseudo values method with constant hazard					Pseudo values method with constant hazard				
mean	bias	EmpSD	ModSD	CP	mean	bias	EmpSD	ModSD	CP
0.838	-0.162	0.162	0.161	0.790	0.951	-0.049	0.225	0.231	0.945
Pseudo values method with non constant hazard					Pseudo values method with non constant hazard				
mean	bias	EmpSD	ModSD	CP	mean	bias	EmpSD	ModSD	CP
1.022	0.022	0.208	0.209	0.959	1.013	0.013	0.238	0.244	0.959
Hsieh and Huang regression method					Hsieh and Huang regression method				
t	mean	bias	EmpSD		t	mean	bias	EmpSD	
0.22	0.996	-0.004	0.333		*0.2*	1.011	0.011	0.363	
0.47	0.986	-0.014	0.248		*0.4*	0.983	-0.017	0.276	
0.72	0.976	-0.024	0.216		*0.6*	0.971	-0.029	0.249	
0.97	0.982	-0.018	0.223		*0.8*	0.964	-0.036	0.249	
1.22	0.959	-0.041	0.246		*1.0*	0.946	-0.054	0.266	

Table 3A. Simulation results for the treatment effect on the non terminal event $\hat{\beta}_X$ estimated with pseudo-observations and with Hsieh and Huang regression model. Simulation scheme adopted by Hsieh and Huang is used. Data are generated by Clayton copula.

n=500

β_Y=0 β_Y=0.2

Pseudo values method with constant hazard					Pseudo values method with constant hazard				
mean	bias	EmpSD	ModSD	CP	mean	bias	EmpSD	ModSD	CP
0.838	-0.162	0.101	0.101	0.626	0.956	-0.044	0.144	0.146	0.938
Pseudo values method with non constant hazard					Pseudo values method with non constant hazard				
mean	bias	EmpSD	ModSD	CP	mean	bias	EmpSD	ModSD	CP
1.018	0.018	0.126	0.131	0.954	1.014	0.014	0.153	0.154	0.956
Hsieh and Huang regression method					Hsieh and Huang regression method				
t	mean	bias	EmpSD		t	mean	bias	EmpSD	
0.22	1.001	0.001	0.198		*0.2*	0.98	-0.02	0.211	
0.47	0.987	-0.013	0.151		*0.4*	0.967	-0.033	0.168	
0.72	0.971	-0.029	0.162		*0.6*	0.95	-0.05	0.169	
0.97	0.907	-0.093	0.159		*0.8*	0.891	-0.109	0.174	
1.22	0.942	-0.058	0.148		*1.0*	0.9	-0.1	0.159	

Table 3B. Simulation results for the treatment effect on the non terminal event $\hat{\beta}_X$ estimated with pseudo-observations and with Hsieh and Huang regression model. Simulation scheme adopted by Hsieh and Huang is used. Data are generated by Clayton copula.

In order to better analyze the performance of regression models for net survival in the presence of semi-competing risks, a further simulation is carried out. The simulation scheme is equal to that proposed by Hsieh and Huang and used above, except for the fact that the error is extended so that the dependence structure of (e_{Xi} , e_{Yi}) follows the Frank model as:

$$\Pr(e_{Xi} > x, e_{Yi} > y) = -\frac{1}{\theta}\ln\left[1 + \frac{\left(e^{-\theta \Pr(e_{Xi} > x)} - 1\right)\left(e^{-\theta \Pr(e_{Yi} > y)} - 1\right)}{e^{-\theta} - 1}\right]$$

Every simulation scenario is studied again with Hsieh and Huang regression model and with two different pseudo values regression models considering a constant and a not constant hazard over time. Table 4 reports the results in terms of mean of the estimated coefficients, bias, empirics standard deviation (EmpSD) and, where available, mean of the standard deviations estimated in the regression model (ModSD) and coverage probability (CP) of the 95% confidence interval of β_x.

In this context as well the performance of Hsieh and Huang regression model is quite good, giving almost unbiased regression coefficients estimates. The pseudo-observation regression

model performs equally good. In particular it gives more efficient coefficient regression estimates, in fact the standard deviation of pseudo-values regression model are always smaller than the standard deviation of Hsieh and Huang regression model. The coverage probability of pseudo-values model with constant hazard over time is something smaller than 0.95, whereas the coverage probability of the regression model with non constant hazard over time is near 0.95.

To conclude it can be said that in the presence of low association between times to different events the choice of copula is not limiting in the analysis of regression data in a semi-competing risks context with pseudo-values observations.

n=200

β_y=0

Pseudo values method with constant hazard				
mean	bias	EmpSD	ModSD	CP
0.885	-0.115	0.170	0.163	0.849
Pseudo values method with non constant hazard				
mean	bias	EmpSD	ModSD	CP
1.098	0.098	0.220	0.215	0.949

Hsieh and Huang regression method			
t	mean	bias	EmpSD
0.22	1.027	0.027	0.333
0.47	1.030	0.030	0.243
0.72	1.039	0.039	0.222
0.97	1.050	0.050	0.227
1.22	1.045	0.045	0.250

β_y=0.2

Pseudo values method with constant hazard				
mean	bias	EmpSD	ModSD	CP
0.986	-0.014	0.230	0.236	0.953
Pseudo values method with non constant hazard				
mean	bias	EmpSD	ModSD	CP
1.055	0.055	0.246	0.251	0.960

Hsieh and Huang regression method			
t	mean	bias	EmpSD
0.2	0.994	-0.006	0.349
0.4	0.986	-0.014	0.271
0.6	0.981	-0.019	0.241
0.8	0.978	-0.022	0.253
1.0	0.968	-0.032	0.272

Table 4A. Simulation results for the treatment effect on the non terminal event $\hat{\beta}_X$ estimated with pseudo-observations and with Hsieh and Huang regression model. Simulation scheme adopted by Hsieh and Huang is used. Data are generated by Frank copulas.

n=500

β_Y=0 β_Y=0.2

Pseudo values method with constant hazard					Pseudo values method with constant hazard				
mean	bias	EmpSD	ModSD	CP	mean	bias	EmpSD	ModSD	CP
0.889	-0.111	0.104	0.103	0.797	0.974	-0.026	0.151	0.147	0.940
Pseudo values method with non constant hazard					Pseudo values method with non constant hazard				
mean	bias	EmpSD	ModSD	CP	mean	bias	EmpSD	ModSD	CP
1.096	0.096	0.136	0.135	0.912	1.040	0.040	0.160	0.156	0.953
Hsieh and Huang regression method					Hsieh and Huang regression method				
t	mean	bias	EmpSD		t	mean	bias	EmpSD	
0.22	1.016	0.016	0.201		*0.2*	0.984	-0.016	0.216	
0.47	1.014	0.014	0.155		*0.4*	0.983	-0.017	0.175	
0.72	1.032	0.032	0.155		*0.6*	0.982	-0.018	0.170	
0.97	0.972	-0.028	0.172		*0.8*	0.931	-0.069	0.184	
1.22	1.006	0.006	0.158		*1.0*	0.920	-0.080	0.173	

Table 4B. Simulation results for the treatment effect on the non terminal event $\hat{\beta}_X$ estimated with pseudo-observations and with Hsieh and Huang regression model. Simulation scheme adopted by Hsieh and Huang is used. Data are generated by Frank copulas.

3.2 A clinical example

From 1973 to 1989 at the National Cancer Institute in Milan a series of clinical trials was done to compare different therapeutic strategy in women with small, non-metastatic primary breast cancer. I analyze here data regarding two clinical trials.

Between 1985 and 1987, 705 patients were accrued in a randomised clinical trial comparing two conservation treatment strategies: quadrantectomy, axillary dissection and radiotherapy (QUART, 360 women) versus tumorectomy and axillary dissection followed by external radiotherapy and a boost with Ir implantation (TART, 345 women). No second surgery was given to women with affected surgical margins. Details on the results of this trial are reported in Mariani et al. (1998).

Between 1987 and 1989, 579 women with carcinoma of the breast were randomly assigned to quadrantectomy, axillary dissection and radiotherapy (QUART, 299 women) and to quadrantectomy with axillary dissection without radiotherapy (QUAD, 280 women). Details on the results of this trial are reported in Veronesi et al. (2001).

In both randomized trials axillary node positive women received adjuvant medical therapy: premenopausal and postmenopausal patients negative for estrogen receptors received chemotherapy, while postmenopausal patients positive for estrogens receptors received tamoxifen. Form the published data, no significant difference in survival were found in the trials according to treatment and the analysis on events were focussed mainly on IBTR where QUART showed an advantage. In the present analysis, IBTR was not the end-point of interest, since it does not prevent the observation of subsequent severe events (distant metastases, contralateral tumours and new primary tumours) which can be able to influence survival probability. Difference related to treatment for the above mentioned events were not evidenced. For exemplificative aims, data were jointly considered without taking into account for surgery and radiotherapy.

Aiming to illustrate the above mentioned approaches on survival probabilities the following analyses were performed.

Causes of Deaths: non parametric estimates for overall survival, survival for deaths related to breast cancer and relative survival. Here competing risks were accounted for.

Severe events: non parametric estimates of survival free of severe events. Here, semi competing risks were accounted for estimating net severe event free survival.

Regression models were performed including as covariates ER, PGR status, tumour size, and axillary node involvement

For the analysis of death semi parametric models were used.

For the analysis of severe events pseudo values regression models were used and compared with regression models proposed by Hsieh and Huang.

Table 6 reports the main characteristic of patients, divided by trial and treatment. It can be noted that the trials are randomized and the characteristics of patients have similar distributions in different treatment group.

	Trial 1		Trial 2		
	QUAD (n=360)	TART (n=345)	QUAD (n=273)	QUART (n=294)	TOT (n=1272)
Median age	50	49.5	52.9	52.1	51
Estrogens receptor status					
- Negative (≤10)	76 (21.1%)	64 (18.6%)	48 (17.6%)	63 (21.4%)	251 (19.7%)
- Positive (>10)	249 (69.2%)	250 (72.5%)	184 (67.4%)	190 (64.6%)	873 (68.6%)
- NA	35 (9.7%)	31 (9%)	41 (15%)	41 (13.9%)	148 (11.6%)
Progesterone receptor status					
- Negative (≤25)	103 (28.6%)	98 (28.4%)	95 (34.8%)	97 (33%)	393 (30.9%)
- Positive (>25)	220 (61.1%)	216 (62.6%)	137 (50.2%)	156 (53.1%)	729 (57.3%)
- NA	37 (10.3%)	31 (9%)	41 (15%)	41 (13.9%)	150 (11.8%)
Tumour dimension					
- T1 (≤ 2 cm)	304 (84.4%)	300 (87%)	230 (84.2%)	242 (82.3%)	1076 (84.6%)
- T2 (> 2cm, but ≤ 5cm)	55 (15.3%)	41 (11.9%)	42 (15.4%)	47 (16%)	185 (14.5%)
- NA	1 (0.3%)	4 (1.2%)	1 (0.4%)	5 (1.7%)	11 (0.9%)
Axillary node involvement					
- N0 (0 positive nodes)	240 (66.7%)	225 (65.2%)	182 (66.7%)	211 (71.8%)	858 (67.5%)
- N1 (≥1 positive nodes)	120 (33.3%)	120 (34.8%)	91 (33.3%)	83 (28.2%)	414 (32.5%)

Table 6. Characteristics of patients in the two trials divided by treatment group.

Analysis of death

As regards the clinical example we can observe that 328 of the 1272 women die within 15 years of follow-up. The overall survival function, given in Figure 1, shows that the probability of surviving from death due to any cause 5, 10 or 15 years from surgery is respectively 0.91, 0.79 and 0.71.

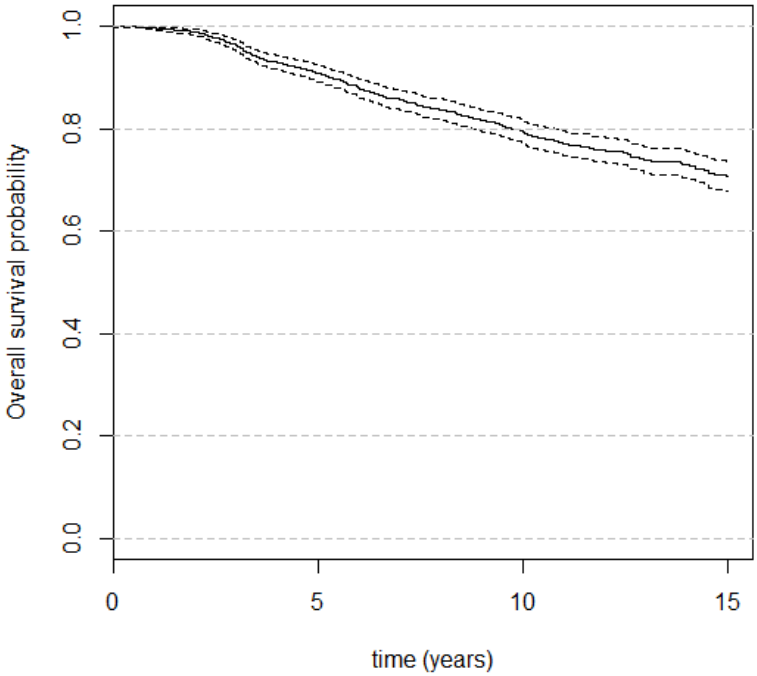

Figure 1. Overall survival estimated by Kaplan-Meier method (continuous line) and 95% confidence interval (dotted lines).

The causes of deaths were classified as related to breast cancer or related to other causes. Since data are taken from clinical trials, accurate follow-up is available and the classification of causes of death has been retained reliable by clinician. In this clinical example it can be useful to focus attention only on death due to breast cancer, thus to estimate survival from breast cancer death.

Death due to breast cancer

As regards the whole dataset, 244 deaths were classified as related to breast cancer and 84 as related to other causes. The net survival probability is thus of concern, i.e, the probability of surviving to breast cancer in the case this is the only acting cause of death in the population.

If independence between the two causes of death is assumed, Kaplan-Meier method can be used, considering as censored times to death for all causes (Figure 2, panel a). The independence, although in this case could be clinically reasonable, cannot be a priori assumed. To investigate this issue, Kendall's tau coefficient of concordance for bivariate

censored data (Brown et al, 1974) can be used, as first insight. The estimate is τ_K=0.0001, thus the assumption of independence can be tenable. However Clayton copula graphical estimator can be used to compute net survival with association parameter θ=0.0002, corresponding to the Kendall's tau previously estimated. The net survival probability is estimated is shown in Figure 2, panel b. As expected, net survival estimates obtained by Kaplan-Meier method and copula graphic estimator, are practically overlapping. The estimated net survival probability at 5, 10, 15 years is 0.92, 0.83 and 0.78 respectively. Finally relative survival is also computed, after obtaining the expected survival of the reference population by ISTAT mortality tables. Relative survival at 5, 10, 15 year is 0.93, 0.84 and 0.79 respectively. It can be observed that these estimates are slightly higher than those obtained by the two above mentioned methods. Although the assumption of independence between causes of deaths and the low contribution of the mortality related to breast cancer in the reference population can be considered as tenable, the study sample is conditioned by the protocol's inclusion criteria (absence of comorbidities which avoid the application of surgery or chemotherapy) thus other causes of deaths may not acting as in the reference population. This condition induce to caution in interpreting of relative survival as net survival.

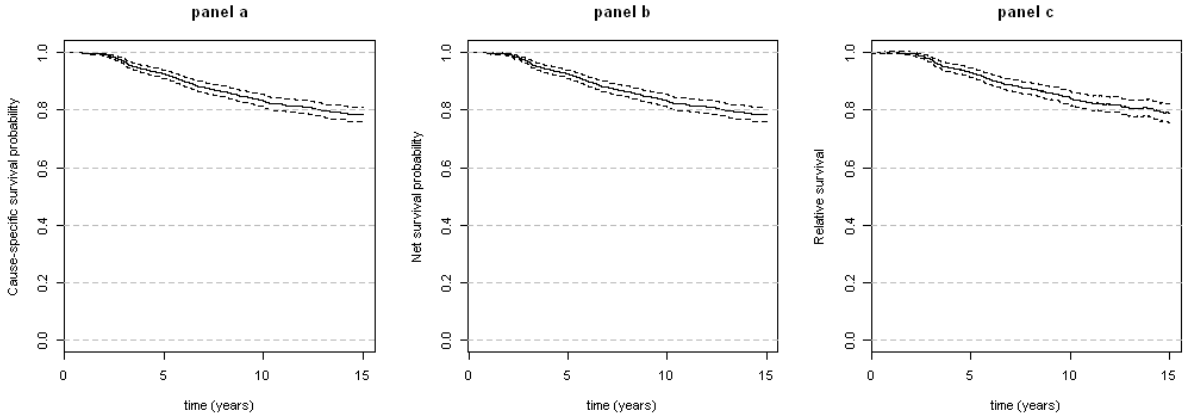

Figure 2. Estimates of (net) survival for breast cancer by Kaplan-Meier method (panel a), copula graphic estimator (panel b) and relative survival (panel c) (continuous line) and 95% confidence intervals (dotted lines).

Regression models

Cox regression models were performed to evaluate the effect of covariate on overall survival and breast cancer specific survival. ER has a time dependent effect on survival, in particular the protective effect tends to reduce in time. Regression estimates on overall and cause specific survival are slightly different as the overall survival include the effect on death due to other causes. Additive Poisson regression on relative survival should provide very similar estimate to model on cause-specific hazard, since the independence assumption is tenable. Although the estimates are similar, it can be observed a difference in regression coefficients estimates due to selection of patients in the clinical trials.

A

	Estimate	Standard error	z-value	p-value
ER status	-0.784	0.298	-2.630	0.008
PGR status	-0.515	0.131	-3.930	<0.001
Tumour dimension	0.304	0.150	2.030	0.042
Axillary nodes involvement	0.682	0.117	5.820	<0.001
ER status* time	0.0004	0.0001	3.590	<0.001

B

	Estimate	Standard error	z-value	p-value
ER status	-1.129	0.334	-3.380	0.001
PGR status	-0.408	0.153	-2.670	0.008
Tumour dimension	0.322	0.173	1.860	0.063
Axillary nodes involvement	0.803	0.135	5.960	<0.001
ER status* time	0.001	0.0002	3.850	<0.001

C

	Estimate	Standard error	z-value	p-value
ER status	-1.419	0.456	-3.110	0.002
PGR status	-0.480	0.177	-2.710	0.007
Tumour dimension	0.349	0.189	1.850	0.064
Axillary nodes involvement	0.944	0.158	5.960	<0.001
ER status* time	0.240	0.076	3.150	0.002

Table 7. Estimates of regression coefficients with pertinent standard error, wald statistics and p-value for Cox regression model on overall hazard (A), Cox regression model on cause specific hazard for breast cancer (B) and Poisson additive regression model for relative survival (C).

Analysis of s*evere events*

In this analysis I will focus attention on survival from severe events, such as regional or distant metastases, contralateral breast carcinoma and other primaries, as it was supposed that breast cancer is not a fatal disease by itself, but before dying all patients should experience a severe event which leads to death. Intra breast tumour recurrence (IBTR) is considered the first evidence of surgery failure, nevertheless this is not considered a "fatal event" unless the subsequent occurrence of distant metastasis, thus distant metastases are always recorded after IBTR. The occurrence of metastasis or contralateral carcinoma or other primary is a non terminal event which is connected to subsequent terminal event (death) and studying the distribution of time to severe events gives information on the progression of the disease and it is of concern in order to choice the best treatment strategy. However it is not always possible to observe time to severe events, as some patients die without experiencing them. This is a typical setting of semi-competing risks where a terminal event (death) can censor a non terminal event (metastasis, contralateral carcinoma) but not vice-versa and the censoring effect of death on times to severe events cannot be considered independent, as there is a clinical evidence of strong association between severe events and death. The last analysis of the above mentioned trials do not evidence a significant impact of the kind of treatment on the occurrence of distant metastases and death (Mariani et al., 1998 and Veronesi et al., 2001) thus in the analysis on semi competing risks treatment was not considered.

In the subsequent analysis follow up was stopped at 15 years for both trials. Table 6 summarizes the number of patients experiencing different events within 15 years of follow-up. 784 women of the 1984 recruited for the two trials experienced neither severe event neither death, 52 women died before having a severe event, 276 women had a severe event and hereafter died and 160 women had a severe event and hereafter were censored due to administrative censoring or lost to follow-up.

		Severe event	
		no	yes
Death	no	784	160
	yes	52	276

Table 6. Contingency table for the number of patients experiencing severe event or death.

Association between severe event and death

A piecewise exponential regression model with spline functions on time was applied as preliminary investigation on the shape of hazard of death. In patients with small breast carcinoma the hazard of death shows a typical shape with two peaks, indicating that the hazard of death reaches two local maximum at almost 3-4 years and 7-8 years of follow-up (Figure 5).

However to better understand the course of the disease from surgery to death it is useful to evaluate how the occurrence of severe events alters the risk of dying. To this a time dependent covariate indicating the occurrence of severe events was added to the above mentioned piecewise exponential model. Figure 6 shows how the hazard function before and after a severe event and it can be noted that there is a big change in the shape of hazard. After a severe events patients are very much more exposed to the risk of dying.

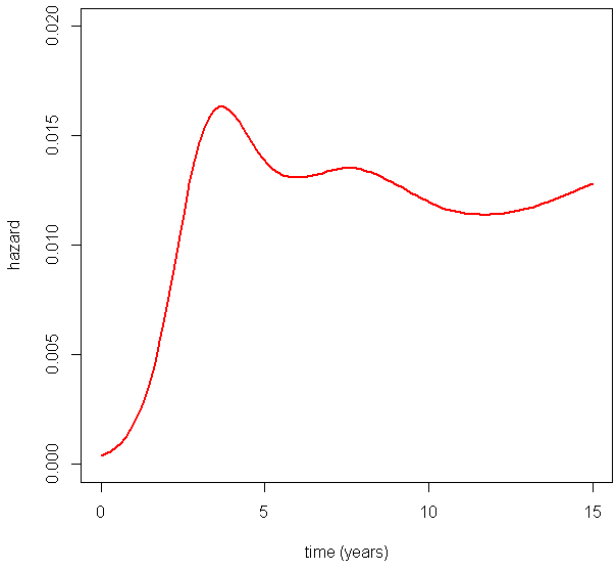

Figure 5. Hazard of death in the whole population under study

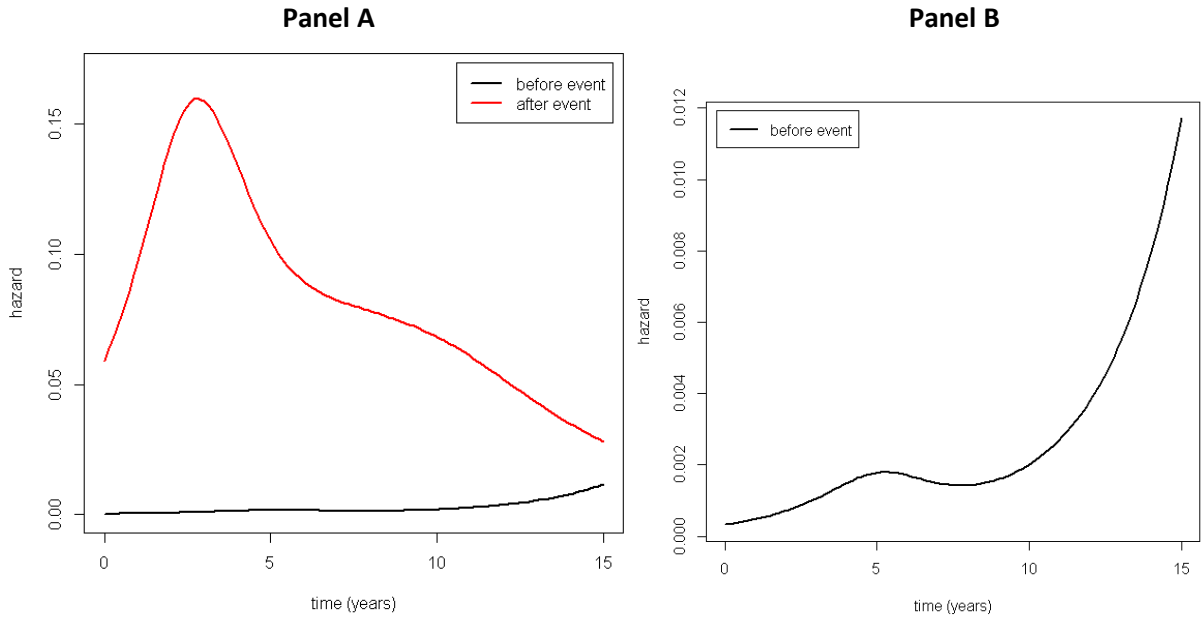

Figure 6. Hazard of death before and after a severe event (Panel A) and hazard of death before a severe events in a more accurate scale in order to better understand the shape of hazard function (Panel B).

The time elapsed from severe event to death is empirically analyzed and the corresponding survival function is reported in Figure 7. It is evident that it is quite short, in fact only 30% of patients survive 5 years after a severe event.

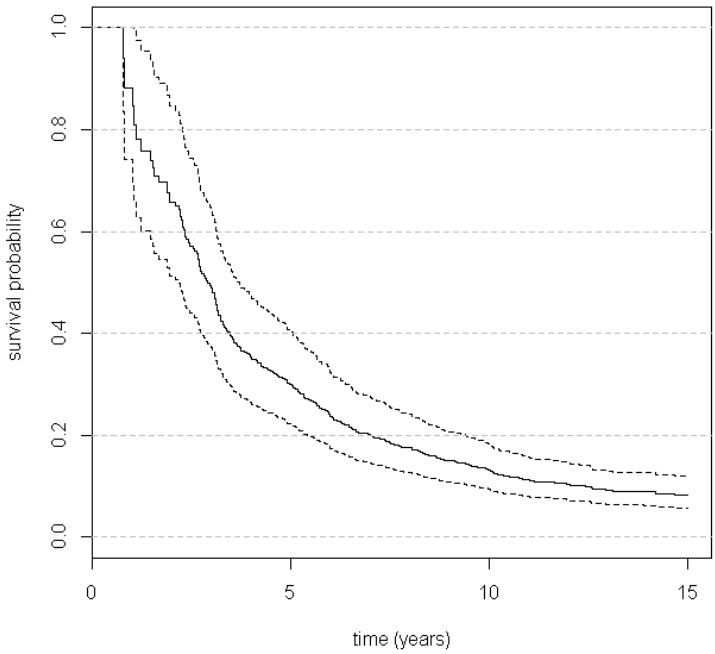

Figure 7. Survival probability from death after a severe event.

From the previous results it is evident that the occurrence of severe event alters the progression of the disease and increases the risk of dying. As described in section 2 there are several statistical methods to estimate more accurately the association between times to severe events (non terminal events) and times to death (terminal event). For example, in this sample Brown estimate of Kendall's τ in the presence of bivariate censoring is 0.651, confirming the strong association between time to severe event and time to death. However in the presence of semi-competing risks, as severe event (non terminal) and death (terminal), it was demonstrated in simulation section that more specific methods which define the bivariate distribution of times to different events have to be adopted to estimate association. For example Fine method uses Clayton copula to describe the bivariate distribution of time to non terminal and terminal event and its association parameter has a direct relationship with Kendall's τ and can be interpreted as a predicted hazard ratio too. The concordance parameter of the Clayton's copula is estimated in this sample by Fine's method as 10.92, corresponding to a Kendall's τ of 0.832 and meaning that the risk of dying for women who had experienced a severe event is 10.92 times bigger that the risk of dying for women who had not experienced a severe event yet. This means that after the occurrence of a severe event the subject have a significantly higher risks of dying.

Survival from severe events

Severe events are non terminal events whose times of occurrence can be censored by the occurrence of death and the censoring of death cannot be considered independent because of the positive strong association between severe events and death. Figure 8 report the survival curves of severe disease obtained by considering death independent censoring (cause-specific method) and by considering the semi-competing risks framework and estimating net survival curve by the method proposed by Fine using Clayton copula function. Moreover to make a comparison also the complement to 1 of the crude cumulative incidence of death, which for brevity in this text is named "crude survival". Actually the main differences arise after 10 years of follow-up between net survival and cause specific survival and crude survival which are instead almost overlapped. A possible explanation is that the majority of death without evidence of disease occurs later in the follow-up. From panel B of Figure 8 it can be noted that net survival function estimated in a semi-competing risks setting by Fine method, is almost always included between the lower bound given by the first failure

survival and the upper bound given by 1-crude cumulative incidence (Peterson et al.). Furthermore as the estimated association between severe events and death is strongly positive, net survival is nearer first failure survival, that is survival from the minimum of time between severe event and death. However at 3-4 years of follow up the net event free survival is slightly over the upper bound. This is a warning on the appropriateness of the joint distribution of times adopted. Maybe Clayton copula does not adapts perfectly to this dataset. In Figure 10 it is shown through ad hoc simulated data that when data are actually distributed like a Clayton copula model, the net event free survival estimated with Fine's method is always included in the bounding, whereas when data are distributed like a non Clayton copula, for example a Frank copula, the bounding property is not always respected but sometimes the net event free survival function can be inferior to the lower bound or superior to the upper bound.

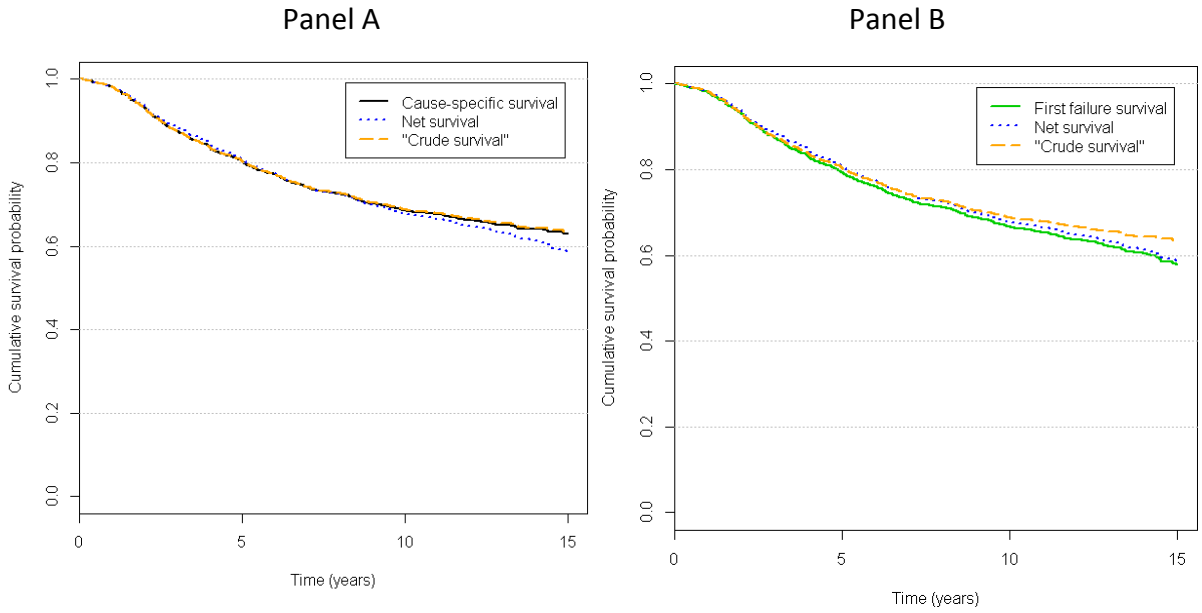

Figure 8. Panel A: Survival curve of severe disease obtained with cause-specific method, 1-crude cumulative incidence and Fine's semi-competing risks approach. Panel B: Fine's survival function is included between the lower bound given by the first failure survival and the upper bound given by 1-crude cumulative incidence.

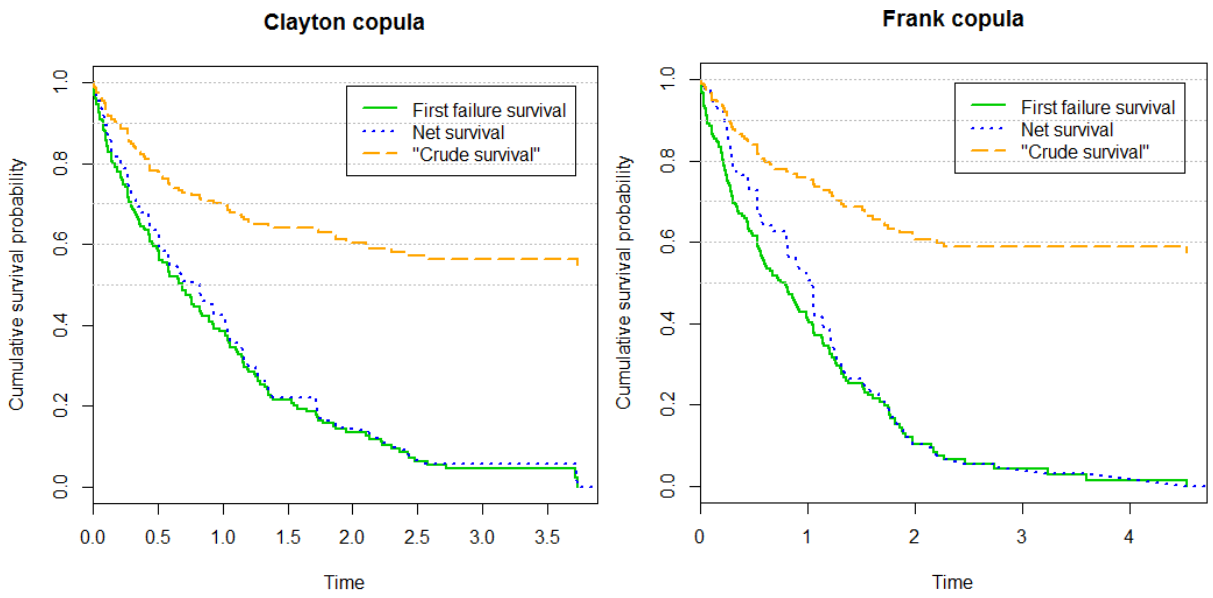

Figure 9. Check of the bounding property of net event free survival function between first failure survival and 1-crude cumulative incidence, for Clayton and Frank copula function.

The difficulties in taking into account the presence of semi-competing risks and properly estimating the survival function in a semi-competing risks setting, for examples by means of methods outlined in section 2.4, rely in the absence of appropriate statistical software function. This is why I had to implement ad hoc R functions for both Fine's non-parametrical method and Lakhal's copula graphic estimator for the estimate of survival function of non terminal events.

From Figure 10 it can be noted that the two previously cited methods for computing survival from non terminal events give similar results when applied to the whole dataset of the clinical example.

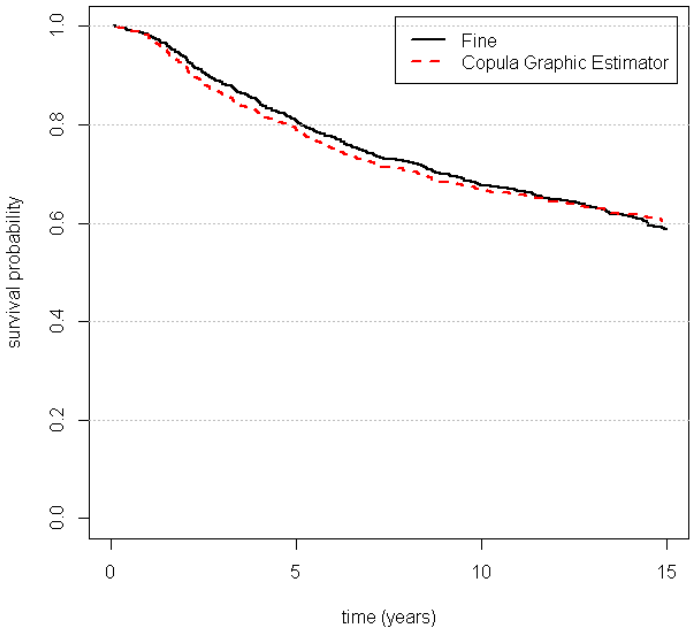

Figure 10. Survival from severe events, obtained with two different methods in a semi-competing risks setting.

The effect of covariate on survival from severe events

To have a deeper understanding of the treatment effect it is useful to evaluate its survival in patients with different characteristics. In fact population in the study is not completely homogeneous, as can be seen from table 5, but patients present different values of covariates like levels of hormones receptors (estrogen status, progesterone status) and tumour features (dimension, involvement of axillary nodes) which are potentially connected with a different survival probability from severe events.

As a preliminary analysis I computed non parametrical survival curves of severe events at different levels of the covariates of interest. The results are reported in Figure 11, together with the association parameter of times to severe events and times to death, estimated within every level of the covariates. It can be observed that patients with bigger tumour dimension, involvement of axillary nodes and negative progesterone receptors status have a lower survival from severe events. Furthermore the association estimates within different levels of these covariates is are similar and coherent with the estimates obtained in the whole population (θ=10.92).

The effect of estrogen status on survival from severe events merits comments as it is not straightforward: women with negative receptor status have a lower survival at the beginning of the follow-up, and a higher survival probability after 6 years of follow-up, indicating that the effect of estrogen status changes over time. Moreover the association between times to severe events and death in patients with negative estrogen status is very high θ=20.32 (95% IC: 14.729 - 26.792), corresponding to a Kendall's τ=0.906, and quite different from the association of severe events and death in patients with positive estrogen receptor status θ=9.146 (95% IC: 7.919 - 10.875), corresponding to a Kendall's τ=0.803.

ESTROGEN RECEPTOR STATUS

ER≤10 (neg) θ=20.32 (τ=0.906)

ER>10 (pos) θ=9.146 (τ=0.803)

PROGESTERONE RECEPTOR STATUS

PGR≤25 (neg) θ=12.734 (τ=0.854)

PGR>25 (pos) θ=9.73 (τ=0.814)

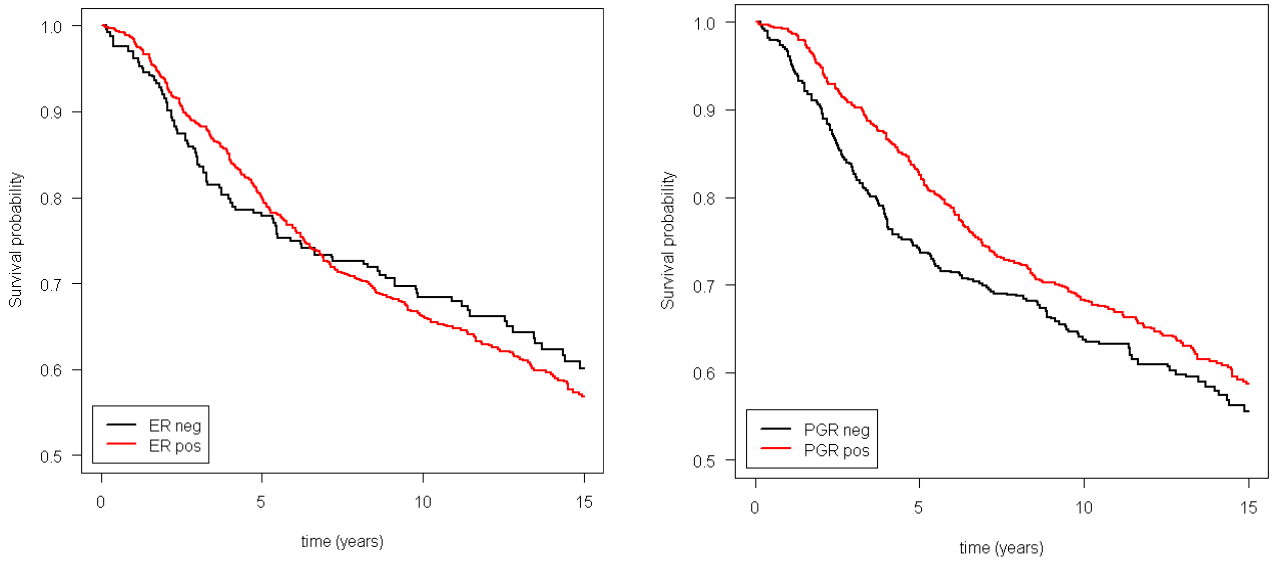

Figure 11A. Survival curves of severe events for different levels of the covariates estrogen and progesterone receptor status.

TUMOUR DIMENSION

T1 (≤ 1 cm)	θ= 10.421	(τ=0.825)
T2 (> 1 cm)	θ= 13.569	(τ=0.863)

AXILLARY NODE INVOLVEMENT

N0 (0)	θ= 11.342	(τ=0.838)
N1 (≥1)	θ= 9.675	(τ=0.813)

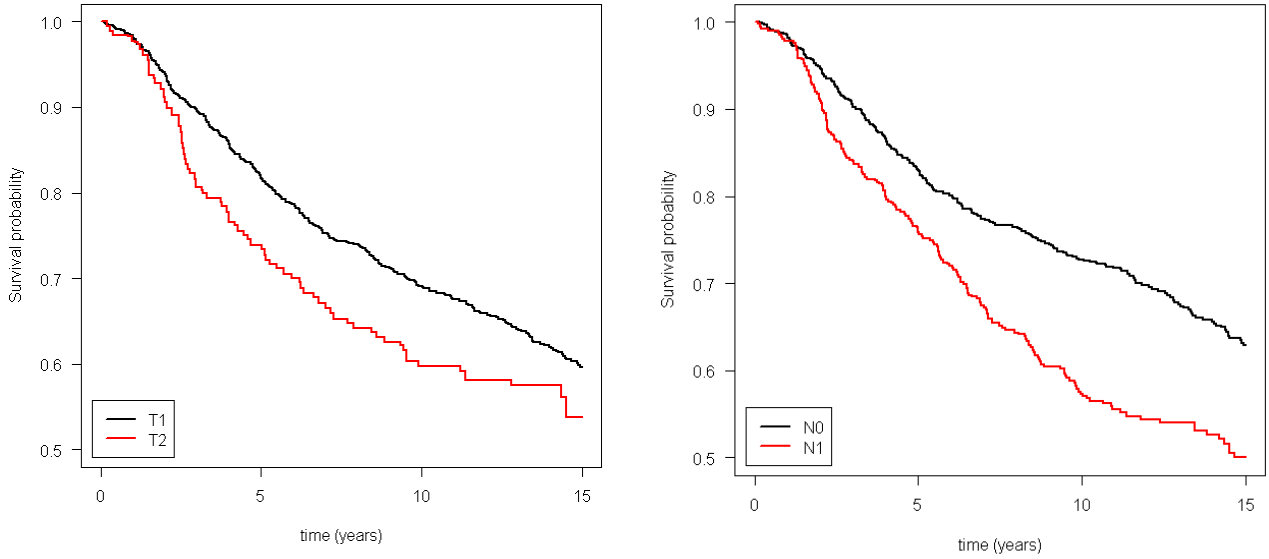

Figure 11B. Survival curves of severe events for different levels of the covariates tumour dimension and axillary node involvement.

A more complete regression analysis can be carried out in order to jointly evaluate the effect of different covariates on survival from severe events. Thus a regression analysis based on pseudo values is performed to evaluate the effect of dimension, axillary nodes involvement and hormones receptor status on survival from severe events. The association parameter used to compute pseudo-values is that of the whole population θ=10.92, as it was shown that the parameter estimated within different levels of the covariate of interest are not significantly different. A generalized linear model with complementary log log link function and Gaussian error distribution is adopted. As known from the literature and confirmed by Figure 11 the effect of estrogen receptor status on survival from severe events changes over time, thus a time dependent effect for estrogen status is considered. A b-spline for time with 3 degrees of freedom is considered for time. The results of the complete regression model are reported in Table 7. It can be noted that after adjusting for the effect of all covariates of interest, the only significant effects are those of progesterone status and axillary nodes involvement, indicating that women with positive PGR

status have significantly lower risk of severe events than women with negative PGR status and women with positive axillary nodes involvement have significantly higher risk of severe events than women without axillary nodes involvement.

	Estimate	Standard error	p-value
(Intercept)	-3.091	0.362	<0.001
spline time 1	2.359	0.669	<0.001
spline time 2	1.638	0.312	<0.001
spline time 3	2.319	0.401	<0.001
ER status	-0.645	0.450	0.152
PGR status	-0.245	0.127	**0.054**
Tumour dimension	0.179	0.147	0.223
Axillary nodes involvement	0.490	0.112	**<0.001**
spline time 1 * ER status	1.034	0.813	0.203
spline time 2 * ER status	0.921	0.378	0.015
spline time 3 * ER status	0.864	0.489	0.078

Table 7. Results of regression analysis using pseudo-observation: coefficients estimates together with the corresponding standard errors and p-values

The effect of ER status changing over time can be easily evaluated in Figure 12. Until 5 years of follow up women with positive ER status have a lower risk of severe events, whereas after 5 years of follow-up women with positive ER status have a higher risk of severe events. This is confirmed by Hsieh and Huang regression method on ER status, whose results are reported in Figure 12 as well. Pseudo-values regression model has the advantages of enabling adjusting ER status effect by all other covariate of interest and enabling flexibly modelling the effect of time on survival from severe events and interaction between time and ER status.

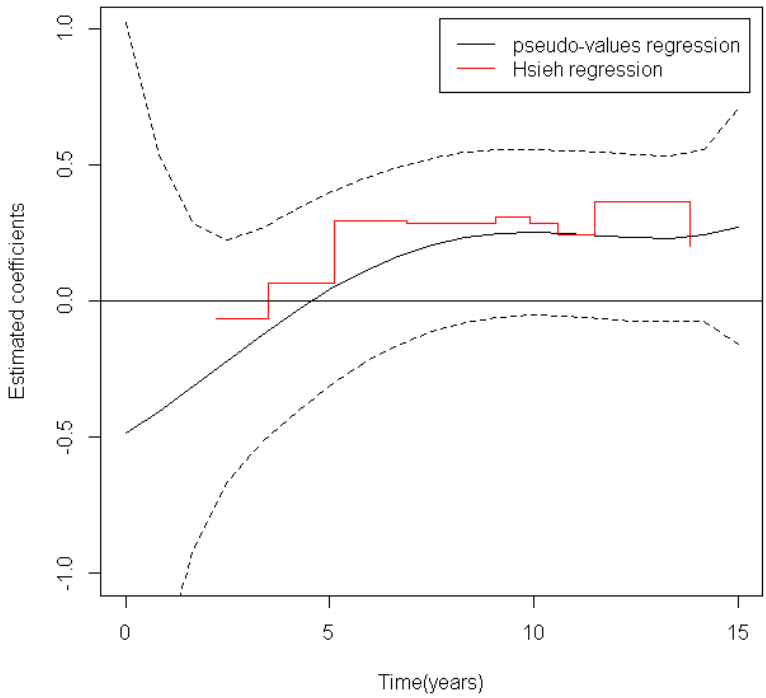

Figure 12. Estimated regression coefficient for ER status in regression model based on pseudo-values and Hsieh and Huang regression model based on conditional likelihood.

4. Discussion

In biomedical studies, it is often of interest to evaluate the efficacy of treatment or the effect of specific covariates such as stage of cancer. To this aim time to death for any cause or death due to a specific cause is an important endpoint. In the analysis of the causes of death, since long time crude cumulative incidence was the main considered estimates. It is worth of note that the "concept" of net survival is not straightforward being based on "ideal situation where the only cause of interest is acting". Nevertheless an increasing interest in net survival is shown. In fact, several papers concerning relative survival are focussed in interpreting results in terms of net survival. A reason is that relative survival enables the comparison of the cause of interest in different countries, "removing" the effect of the remaining causes of death. When relapse is of interest the investigation of time to intermediate non terminal events, such as local relapse or distant metastasis is also essential, as it provides additional information pertaining to the disease progression process. A situation where individuals are exposed to the risk of non terminal event and terminal events is usually called "semi-competing risks" setting, as the occurrence of the terminal event can prevent the observation of the non terminal event, but the occurrence of a non terminal event does not prevent the occurrence of the terminal event.

Also in this framework, crude cumulative incidence of the non terminal event is a wide used estimate, being this quantity estimable. Crude cumulative incidence estimates the probability of the non terminal event before death.

Thus crude cumulative incidence does not provide the estimate of the probability of the occurrence of the intermediate non terminal event in the situation where this event can be observed for all patients before dying. This scenario can be based on the consideration that nobody is “cured” from cancer and a disease progression is ever observed. Death can preclude this observation if time to death is shorter than the hypothetical time to relapse. Nevertheless this estimate is not straightforward as terminal events occurring before the non terminal events cannot be considered independent censoring, since non terminal and terminal event are usually associated. Thus it is very important to assess the degree of association between the two events before apply any specific survival method.

Since the limited available literature and lack of software functions on semi-competing risks data, crude cumulative incidence is used avoiding “net” quantities. Kaplan Meier method is often used to estimate net survival from intermediate events without correctly take into account the association between intermediate and fatal events. In any case, an initial estimate of the association can be performed to evaluate the degree of association and then to realize if standard software can be used without bias.

After the estimates of survival/incidence probability during follow-up, statistical analysis are often performed to evaluate prognostic covariates effect by regression models.

Cox model on cause specific hazards has been adopted since long time because it simple implementation. The difficulty was on the interpretation of results in terms of effect on a specific survival function. In the case of independence between time to different events, the effect is on net survival given the relationship between net hazard and net survival. Unfortunately, before the proposal of sub-distribution hazard regression model, the wrong interpretation of cause specific hazard in terms of effect on crude cumulative incidences was rather diffuse. Now, because of the software availability, Fine and Gray regression model is widely applied. On the contrary, if net survival is of concern and the independence assumption is not tenable, software for regression models is not available. Although on the papers showing models, models system of equations are reported, their implementation is not straightforward. AS an example, for the model proposed by Hsieh only a single covariate can be used and the need to estimate covariate effect for a single follow-up time makes the estimation procedure cumbersome.

In the previous sections I summarized some analytical methods for estimating survival function and regression model, which properly account for competing and semi-competing risks data.

An example on breast cancer data was focussed on the estimation of net survival in the case of typical classification of causes of death (related or not related to cancer) and severe events.

The choice on analysis limited to net survival was suggested by the consideration that analysis on crude cumulative incidence can be found on the same datasets but net survival, although of possible clinical interest ,was not considered. In the case of causes of death, independence can be assumed and the analysis can be done with standard methods. In the case of severe events, a strong association between time to events and time to death avoid the use of standard software. Copula models were of concern. In this framework I propose a new regression method which can be implemented by GLM software. As demonstrated through a Monte Carlo simulation Kaplan Meier method is not suitable to estimate survival of intermediate event in the presence of association between intermediate and fatal events, but specific survival function as those proposed by Fine et al. (2001) or Lakhal et al. (2008) are required. As regards the evaluation of covariate effect, the method proposed here based on pseudo-observation. When data are generate form a Clayton copula and net survival estimates according to Fine et al were used in pseudo-value regression model, pseudo value estimation is almost unbiased and performs almost as well as the method based on conditional likelihood proposed by Hsieh and Huang. Furthermore pseudo-values regression model has the advantages of enabling adjusting covariate effect by all other covariate of interest and enabling flexibly modeling the effect of time on net event free survival and time dependent effect of covariates in a flexible way as well.

However further work is needed in order to accurately evaluate pseudo-values properties for the estimate of net event free survival in the presence of semi-competing risks. For example consistency or asymptotic normality have to be studied.

5. References

M. Aitkin, D. Anderson, B. Francis, J. Hinde. Statistical Modelling in GLIM. Oxford: Oxford University Press, 1989.

F. Ambrogi, E. Biganzoli, P. Boracchi. Estimates of useful measures in competing risks survival analysis. Statistics in medicine 27: 6407-6425, 2008.

P.K. Andersen, K. Borch-Johnsen, T. Deckert, A. Green, P. Hougaard, N. Keiding, S. Kreiner. A Cox regression model for relative mortality and its application to diabetes mellitus survival data. Biometrics 41: 921–932, 1985.

P.K. Andersen, M. Pohar Perme. Pseudo-observations in survival analysis. Statistical Methods in Medical Research 19: 71–99, 2010.

B.W. Brown, M. Hollander, R.M. Korwar, Nonparametric tests of independence for censored data, with applications to heart transplant studies. Reliability and Biometry: 327-354, 1974.

S.C. Cheng, L.J. Wei, Z. Ying. Analysis of transformation models with censored data. Biometrika 82(4): 835-845, 1995.

D.G. Clayton. A model for association in bivariate life tables and its application to epidemiological studies of familial tendency in chronic disease epidemiology. Biometrika 65: 141-151, 1978.

D.R. Cox. Regression models and life tables. Journal of the Royal Statistical Society, Series B 34(2): 187-220, 1972.

R. Day, J. Bryant, M. Lefkopoulou. Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic factors. Biometrika 84: 45-56, 1997.

P.W. Dickman, A. Sloggett, M. Hills, T. Hakulinen. Regression models for relative survival. Statistics in medicine 23: 51–64, 2003.

J.J. Dignam, L.A. Weissfeld, S.J. Anderson. Methods for bounding the marginal survival distribution. Stat Med 14(18): 1985-1998, 1995.

F. Ederer, L.M. Axtell, S.J. Cutler. The relative survival rate: A statistical methodology. National Cancer Institute Monograph 6: 101–121, 1961.

F. Ederer, H. Heise. Instructions to IBM 650 programmers in processing survival computations. Methodological note No. 10, End Results Evaluation Section, National Cancer Institute, Bethesda MD, 1959.

J. Estève, E. Benhamou, M. Croasdale, M. Raymond. Relative survival and the estimation of net survival: elements for further discussion. Statistics in medicine 9: 529–538, 1990.

J.P. Fine, R.J. Gray. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association 94: 496-509, 1999.

J.P. Fine, H. Jiang, R. Chappell. On semi-competing risks data. Biometrika 88 (4): 907-919, 2001.

F. Graw, T.A. Gerds, M. Schumacher. On pseudo-values for regression analysis in competing risks models. Lifetime Data Analysis 15: 241–255, 2009.

H. Jiang, J.P. Fine, M.R. Kosorok, R. Chappell. Pseudo Self-Consistent Estimation of a Copula Model with Informative Censoring. Scandinavian Journal of Statistics 32 (1): 1-20, 2005.

T. Hakulinen. Cancer survival corrected for heterogeneity in patient withdrawal. Biometrics 38: 933-942, 1982.

T. Hakulinen, L. Tenkanen. Regression analysis of relative survival rates. Journal of the Royal Statistical Society, Series C 36: 309–317, 1987.

P. Hougaard. Modelling multivariate survival. Scand J Stat 14: 291-304, 1987.

J.J. Hsieh, Y.T. Huang. Regression analysis based on conditional likelihood approach under semi-competing risks data. Lifetime Data Analysis 18: 302-320, 2012.

J.D. Kalbfleish, R.L. Prentice. The Statistical Analysis of Failure Time Data. 2nd ed. Hoboken, New Jersery: John Wiley and Sons, 2002.

E.L. Kaplan, P. Meier. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53: 457–481, 1958.

J.P. Klein, P.K. Andersen. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. Biometrics 61: 223–229, 2005.

J.P. Klein, M.L. Moeschberger. Bounds on net survival probabilities for dependent competing risks. Biometrics: 529-538, 1988.

TA. Kpanzou. Copulas in statistics. African Institute for Mathematical Sciences (AIMS) 2007.

L. Lakhal, L.P. Rivest, B. Abdous. Estimating Survival and Association in a Semicompeting Risks Model. Biometrics 64: 180–188, 2008.

L. Lakhal, L.P. Rivest, D. Beaudoin. IPCW Estimator for Kendall's Tau under Bivariate Censoring. The International Journal of Biostatistics 5 (1): Article 8, 2009.

S.M.S. Lo, R.A. Wilke. A regression model for the copula-graphic estimator. Nottingham School of Economics, Discussion Paper No. 11-04, 2011.

S.M.S. Lo, R.A. Wilke. A regression model for the copula-graphic estimator. Journal of Econometric Methods 3(1): 21–46, 2014.

L. Mariani, B. Salvadori, E. Marubini, et al. Ten Year Results of a Randomised Trial Comparing Two Conservative Treatment Strategies for Small Size Breast Cancer. European Journal of Cancer, 34 (8): 1156-1162, 1998.

G. Martelli, P. Boracchi, A. Orenti, et al. Axillary dissection versus no axillary dissection in older T1N0 breast cancer patients: 15-year results of trial and out-trial patients. Eur J Surg Oncol 40(7): 805-812, 2014.

E. Marubini, M.G. Valsecchi. Analysing Survival Data from Clinical Trials and Observational Studies. Chichester: John Wiley and Sons, 1995.

A. Moliterni, S. Ménard, P. Valagussa, et al. HER2 overexpression and doxorubicin in adjuvant chemotherapy for resectable breast cancer. J Clin Oncol 21(3): 458-462, 2003.

R.B. Nelsen. An Introduction to Copulas. New York: Springer, 1999.

D. Oakes. A model for association in bivariate survival data. J. R. Statist. Soc. B 44: 414-422, 1982.

D. Oakes. Semiparametric inference in bivariate survival data. Biometrika 73: 353-361, 1986.

L. Peng, J.P. Fine. Regression Modeling of Semicompeting Risks Data. Biometrics 63 (1): 96-108, 2007.

A.V. Peterson. Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. Proc Nat Acad Sci USA 73 (1): 11-13, 1976.

A.V. Peterson. Dependent competing risks: bounds for net survival functions with fixed crude survival functions. Environment International 1(6): 351-355, 1978

R.L. Prentice, J.D. Kalbfleisch, A.V. Peterson, Jr, N. Flournoy, V.T. Farewell, N.E. Breslow: The Analysis of Failure Times in the Presence of Competing Risks. Biometrics, 34 (4): 541-554, 1978.

L.P. Rivest, M.T. Wells. A Martingale Approach to the Copula-Graphic Estimator for the Survival Function under Dependent Censoring. Journal of Multivariate Analysis 79: 138-155, 2001.

F. Rotolo, C. Legrand, I. Van Keilegom. A simulation procedure based on copulas to generate clustered multi-state survival data. Computer methods and programs in biomedicine 109: 305–312, 2013.

M.J. Rutherford, P.W. Dickman, P.C. Lambert. Comparison of methods for calculating relative survival in population-based studies. Cancer Epidemiol 36(1): 16-21, 2012.

P. Tai, K. Joseph, A. El-Gayed, E. Yu. Long-term outcome of breast cancer patients with one to two nodes involved - application of nodal ratio. Breast J 18(6): 542-548, 2012

A. Tsiatis. A nonidentifiability aspect of the problem of competing risks. Proc Nat Acad Sci USA 72(1): 20-2, 1975

U. Veronesi, E. Marubini, L. Mariani, et al. Radiotherapy after breast-conserving surgery in small breast carcinoma: Long-term results of a randomized trial. Annals of Oncology 12: 997-1003, 2001.

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