Survival Functions in the Presence of Several Events and Competing Risks: Estimation and Interpretation Beyond Kaplan-Meier

Patrizia Boracchi* and Annalisa Orenti

Department of Clinical Sciences and Community Health, Laboratory of Medical Statistics, Epidemiology and Biometry G. A. Maccacaro, University of Milan, Italy

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. A particular case is semi-competing risks when only fatal events can prevent the observation of the non fatal ones.

Kaplan-Meier is the most popular method to estimate overall or event free survival. On the other hand when a subset of events is considered and net survival is of concern, different estimators have been proposed. Kaplan-Meier method can be used only under the independence assumptions otherwise estimators based on multivariate distribution of times are needed. If causes of death are unknown, relative survival can approximate net survival only under specific assumptions on the mortality pattern.

Kaplan-Meier method cannot be used to estimate crude cumulative incidence of specific events.

The aim of this work is to present the survival functions used in competing risks framework, their non parametric estimators and semi parametric estimators for net survival based on Archimedean Copulas. This would be a help for the reader who is not experienced in competing risks analysis.

A simulation study is performed to evaluate performances of net survival estimators. To illustrate survival functions in presence of different causes of death and of different kind of events a numerical example is given, a literature dataset on prostate cancer and a case series of breast cancer patients have been analysed.

Keywords: Survival analysis, competing risks, crude cumulative incidence, net survival, relative survival, breast cancer.

1. INTRODUCTION

In several clinical studies, the evaluation of the effect of a therapy or the impact of prognostic factors is based on the time elapsed form the date of disease diagnosis or the beginning of treatment and the occurrence of events related to treatment failure. In the case of severe diseases, time to death (for all causes) is one of the main end-points and survival probability as a function of follow-up time is a clinically interpretable measure of prognostic impact.

It is well known that a peculiar characteristic of survival analysis is that time to death (or time to any event of interest) may be not available for all patients (censoring). As an example, some patients are still alive at the study ending (administrative censoring) or they are alive and lost to follow-up (right censoring). In such a case, a bivariate distribution of the random variables time to death (T) and censoring time (C) is then of concern. 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 to be observed and this condition is named ”competing risks”. The interest is on the marginal survival function of time to death. The information provided by such a kind of available data is sufficient to determine uniquely marginal survival only under the assumption of independence between T and C [1]. Under this assumption Kaplan-Meier method [2] is used to estimate marginal survival curves. Since its proposal, this method has been widely used in the analysis of time to event data derived from clinical studies also in the presence of different kinds of events, but, often, without an appropriate evaluation of the underlying assumptions and a correct interpretation of the estimated probabilities. Hereinafter, in the main probabilities useful in the presence of different kind of events during follow-up will be presented and commented. Moreover, the possible correct application of Kaplan-Meier method will be underlined.

In several studies information on the causes of deaths is also considered in order to evaluate their specific impact. 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 death for causes related to the disease and on its pertinent marginal survival function.

*Address correspondence to this author at the Department of Clinical Sciences and Community Health, University of Milan, Via Vanzetti 5, 20133 Milan, Italy; Tel: +39 02 23903204; Fax: 02 50320866; E-mail: patrizia.boracchi@unimi.it

E-ISSN: 1929-6029/15

In fact results are often reported in terms of “cause specific survival”. In the analysis of data, available information is usually based on times to death and corresponding causes or censoring time for alive patients. Times to death for causes not related to the disease censor the times to deaths for causes related to the disease. A multivariate distribution is then of concern and competing risks are acting. Under the assumption of independent censoring (administrative and lost to follow-up), if the further independence assumption between time to death for the cause of interest and time to death for other causes is also tenable, Kaplan-Meier method can be used to estimate marginal “cause specific” survival probabilities. This approach is widely used since it can be simply implemented after considering as censored the times to other causes of deaths. Example can be found in several papers e.g in the case of breast cancer see [3] 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, problems arise because observed data do not allow an unbiased non parametric estimate (non identifiability, [4]).

A proposed solution is based on the assumption of a particular structure of the multivariate distribution. Several multivariate survival distributions have been proposed, most of them are based on parametric distribution of marginal survival functions [5]. 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 [6,7]. An advantage of copulas is that the marginal distributions need not to be defined, thus they can be non parametric as well.

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 [8] 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 health status. Nevertheless adequate information on causes 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 [9].

Relative survival is based on the relative survival ratio (RSR) which is defined as the observed survival in the patient group divided by 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 disease in the study sample [10] 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, the death rate for other causes acts in the same way in the sample patients and in the reference population. Considering the same assumptions, a direct estimation of net survival which does not need the knowledge of the causes of death has been proposed [11].

The above considerations concerning the estimates of survival probability and cause specific survival probabilities in the framework of competing risks (except for relative survival) can be applied also when the efficacy of a therapy is evaluated in terms of the onset of adverse events, which are relevant for the study aims (e.g. in the case of breast cancer: local relapses, distant metastases, contralateral cancer, tumours in other sites, death without evidence of neoplastic progression). In the most comprehensive end-point all possible events should be considered directly or indirectly related to the failure of the therapy (as in the above case of death for all causes), otherwise a subsample of events can be considered aiming to a deep investigation of the reason of treatment failure (as in the above case of death for cause related to the disease). The latter estimate has to be interpreted in terms of marginal (net) survival function, i.e. a hypothetical situation where one of the events of interest considered in the subsample can be observed for all patients. An example of application of Kaplan-Meier method in such a context for breast cancer data can be found in [12].

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 to as “semi-competing risks”, since the occurrence of a fatal event preclude the occurrence of non fatal events but not vice-versa. 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 [13].

Although “survival” probability is the widespread measure of treatment effect, in the presence of competing risks the cumulative probability of events is also of interest. In the case of the most comprehensive end-point where all possible events are considered (death “for all causes” or treatment failure, as above defined), the cumulative probability can be obtained by complement to one of Kaplan-Meier estimates. The cumulative probability of death or failure for a specific cause (crude cumulative incidence) cannot be estimated by the complement to one of the “cause specific free survival”. Crude incidences, in fact, estimate the probability of dying or failing for each cause in the “real” situation where different causes are acting. The pertinent estimator proposed by Kalbfleish and Prentice needs to be considered [14]. This estimator is now frequently used, although in some applications a “naive” estimation by Kaplan-Meier method is still reported.

Following a “tutorial approach”, the aim of this work is to provide an initial support to a reader who knows “basic” method of survival analysis but she/he has a limited knowledge on competing risks methodological aspects. Using a standard statistical notation, the work should provide information which could be used to identify the most suitable function according to the study aim and to identify an adequate estimator for survival/incidence probability. Particular attention is given to net probability, including some recent developments related to semi-competing risks. These latter, according to the best knowledge of the Authors, are not included in classical text book for survival analysis to this day. Firstly the different survival functions are described starting from the multivariate distribution of latent failure times. The hazard functions related to the survival functions are defined and compared by some of their relationships. Kaplan-Meier and Kalbfleish and Prentice non parametric estimators for the crude cumulative incidence are shown and compared. Concerning the estimates of the net survival, the general definition of copulas function is provided. Although several copula functions can be used to estimate net survival in clinical applications [6], only the particular case of Archimedean Clayton Copulas is detailed. This is motivated by the availability of a simple closed form estimator which can be easily implemented by standard statistical software in the case of competing risks. Moreover, this copula is considered in the case of semi-competing risks as well. Some available literature is only cited for a deepened knowledge of Copula Functions [6,7], being a detailed Copulas dissertation out of the scope of the present work. For sake of simplicity only the case of two (semi) competing events is considered referring to bivariate copula functions. Given the availability of a strong consistent estimator of association parameter for copula function only in the presence of semi-competing risks, to exploit the performance of net survival estimate a simple simulation study will be presented only in this framework. In the case of causes of death, the interpretation of relative survival as net survival is also discussed.

To provide a numerical example on the difference among survival functions a small dataset was generated by simulation from a bivariate Clayton Copula.

For illustrative purposes a literature data set on prostate cancer is analyzed and discussed to allow readers to repeat and eventually extend the evaluation on the causes of death in relationship to the treatment in the competing risks framework.

Moreover an example on small breast carcinoma is used to illustrate measures of survival in the presence of several events during the follow-up and of different causes of death at the end of follow-up. This dataset allows to consider analyses in competing and semi-competing risks settings.

2. MATERIALS AND METHODS

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 ($t_1,\dots,t_K$), 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:

$$S(t) = S(t, \ldots, t, \ldots, t) = P\,(Y_1 > t, \ldots, Y_k > t, \ldots, Y_K > t) \quad (1)$$

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

$$S_k(t) = S(0, \ldots, t, \ldots, 0) = P(Y_1 > 0, \ldots, Y_k > t, \ldots, Y_K > 0) \quad (2)$$

This is the net survival function from event k [15].

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

On the other hand the crude survival function is based on the time to the first event for each subject, which is always observed: $T=\min(Y_1,\ldots,Y_k,\ldots,Y_K)$:

$$\begin{aligned} S_{(k)}(t) = P[&\min(Y_1, \ldots, Y_k, \ldots, Y_K) > t, \\ &\min(Y_1, \ldots, Y_k, \ldots, Y_K) = Y_k] \end{aligned} \quad (3)$$

The following relationship between overall and crude survival functions always holds: $S(t) = \sum_{k=1}^{K} S_{(k)}(t)$.

The crude cumulative incidence is the probability of k as first event:

$$I_k(t) = P(Y_1 > t, \ldots, Y_k \le t, \ldots, Y_K > t) \quad (4)$$

Obviously: $1 - S(t) = \sum_{k=1}^{K} I_k(t)$.

Survival/Incidence Probability Functions and Corresponding Hazards in the Presence of Competing Risks

Indicating as F(t) a "survival function", i.e. (1), (2) or (3) and as h(t) the corresponding hazard function, the following relationship holds:

$$F(t) = e^{-H(t)} \quad (5)$$

where H(t) is the cumulative hazard: $H(t) = \int_o^t h(u)du$.

In the case of *"overall" survival* (1):

h(t) is the *overall hazard* function, or instantaneous failure rate, which enables studying the dynamic process of the disease over time:

$$\lambda(t) = \lim_{\Delta t \to 0+} \frac{P(t \le T < t + \Delta t \,|\, T \ge t)}{\Delta t} \quad (6)$$

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.

In the case of net (or marginal) survival (2):

h(t) is the net (or marginal) hazard:

$$\phi_k(t) = \lim_{\Delta t \to 0+} \frac{P(t \le Y_k < t + \Delta t \,|\, Y_k \ge t)}{\Delta t} \quad (7)$$

where $\phi_k(t) \cdot \Delta t$ is the probability of dying for cause k in the infinitesimal interval between t and t+Δt, conditionally to the fact that event k had not occurred before time t, in the hypotetical situation where all patients experience event k.

In the case of crude cumulative incidence (4):

h(t) is the sub-distribution hazard:

$$\tilde{\lambda}_k(t) = \lim_{\Delta t \to 0+} \frac{P(t \le T < t + \Delta t;\, K = k \,|\, T \ge t \text{ or } (T < t;\, K \ne k))}{\Delta t}$$

where $\tilde{\lambda}_k(t) \cdot \Delta t$ is the probability that k occurs as first event in the infinitesimal interval between t and t+Δt, conditionally to the fact that no events have occurred before t or an event different from k have occurred before t [16]. How it can be argued from its definition, sub-distribution hazard is a measure which is not of direct clinical interpretation.

A hazard function which is not directly related to the above survival functions (1,2,3) is the cause specific hazard (or crude) hazard rate:

$$\lambda_k(t) = \lim_{\Delta t \to 0+} \frac{P(t \le T < t + \Delta t,\, K = k \,|\, T \ge t)}{\Delta t} \quad (9)$$

where $\lambda_k(t) \cdot \Delta t$ 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 time 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 $S(t) = \prod_k S_k^*(t)$ always 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.

Estimates of Survival/Incidence Probability 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 on the basis of overall hazard (6) and to obtain the corresponding overall incidence.

$$S(t) = \prod_{s=0}^{t} \left(1 - \lambda(s)\right) \tag{10}$$

$\lambda(s)$ is estimated only in correspondence to times in which events occurred by $\frac{d(s)}{n(s)}$, where d(s) is the number of events and n(s) the number of subjects exposed to risk at time s.

In order to estimate net survival, under the assumption of independence among events, given the relationship between cause specific hazard and net hazard, the formula (10) can be used by substituting $\lambda(s)$ with $\lambda_k(s)$, which is estimated by the ratio of the number of events of type k and the number of subjects at risk at time s:

$$\hat{\lambda}_k(s) = \frac{d_k(s)}{n(s)} \tag{11}$$

In this case, the estimation of net survival for the event k is obtained considering as censored times to occurrence of the events different from k and applying to this data Kaplan-Meier method.

If the incidences of different events are considered, crude cumulative incidence can be estimated by Kalbfleish and Prentice method [14-15]:

$$I_K(t) = P(T \le t, K = k) = \sum_{s=1}^{t} \lambda_k(s) \cdot S(s-1) \tag{12}$$

In this case, $\lambda_k(s)$ is the cause-specific hazard estimated by (11) and $S(s-1)$ is overall survival estimated by Kaplan-Meier method (10) considering the occurrence of any event. The estimate of sub-distribution hazard can be obtained as follows

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

The crude survival function is estimated by

$$S_{(k)}(t) = \sum_{s>t} \lambda_k(s) \cdot S(s-1) \tag{13}$$

and $I_k(t) + S_{(k)}(t) \le 1$

It is worth noting $S_k^*(t)$ obtained by Kaplan-Meier method considering as censored time to other events does not provide an estimate of crude survival and that $1 - S_k^*(t)$ 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 all n patients, 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 equation (12) with $1 - S_k^*(t) = \sum_{s=0}^{t} \lambda_k(s) \cdot S_k^*(s-1)$

it can be shown that the overall survival is always less than or equal to the cause-specific survival: $S(s-1) \le S_k^*(s-1)$.

Relative Survival and Net Survival

The cumulative relative survival is defined as:

$$\text{Relavive survival} = \frac{S_0(t)}{S_E(t)}$$

where the $S_0(t)$ is the overall survival in the sample under study and the $S_E(t)$ 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 of birth). Several methods have been proposed to calculate expected mortality from the population mortality tables. The population mortality tables give, 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 [10, 17-18].

In order to a correct estimation of relative survival, the effect of age on the mortality can be accounted for to reduce bias [19]. Different performances are reported for the proposed methods but there is no a general agreement on the more appropriate one.

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:

$$\frac{S_0(t)}{S_E(t)} = \frac{S_{01}^{*}(t) \cdot S_{02}^{*}(t)}{S_{E1}^{*}(t) \cdot S_{E2}^{*}(t)}.$$

In the presence of independence between the causes of death, the cause specific survival correspond to net survival:

$$\frac{S_0(t)}{S_E(t)} = \frac{S_{01}(t) \cdot S_{02}(t)}{S_{E1}(t) \cdot S_{E2}(t)}.$$

If the contributes of the cause of interest is negligible in the general population $S_{E1}(t) \approx 1$.

If the other cause acts in the same way in the sample under study and in the general population $S_{02}(t) \approx S_{E2}(t)$.

Then $\frac{S_0(t)}{S_E(t)} \approx S_{01}(t),$

thus relative survival can be interpreted as a net survival for the causes of interest.

Pohar Perme *et al.* [11] proposed an innovative method to estimate net survival. For the estimation of net survival, overall observed hazard λ_0 can be decomposed as follows: $\lambda_0 = \lambda_P + \lambda_e$ where λ_P is the hazard of death in the reference population and λ_e the excess hazard. Under the assumption that the hazard of death for the causes which are not of interest is given by the population mortality and, the observed hazard is larger than the population hazard, a survival function obtained by the excess hazard is named "net survival" [11]. Moreover, under the assumptions of non informative censoring and conditional independence between times to death for the cause of interest and times to death for the other causes (given sex, age and calendar year) an estimation of net survival for the cause of interest can be obtained starting from the cumulative weighted excess hazard [11]:

$$\Lambda_e(L) = \sum_{l=1}^{L} a_1 \frac{\sum_i w_{il} d_{il} - \sum_i w_{il} \lambda_{Pil} y_{il}}{\sum_i w_{il} y_{il}}$$

where l are interval times (which need to be small since the method was derived for continuous times), w_{il} is a weight for the subject i in the interval l, a_1 is the width of the interval, y_{il} is time at risk for the subject i in the interval l, d_{il} is the event indicator for the subject i in the interval l, λ_{Pil} is the population hazard corresponding to subject i in the interval l. The estimate of net survival is then obtained as $S_e(L) = e^{-\Lambda(L)}$

The Use of Kaplan-Meier and Kalbfleisch and Prentice Estimators for Net Survival Functions when Independence among Events cannot be Assumed

As the wrong estimation of net survival by Kaplan-Meier method on cause specific hazard is related to the dependence of censoring due to times of other events, some modification of Kaplan-Meier estimator accounting for the dependence have been proposed (see [20] among others).

Without assumption on independence of time to events, net survival is not estimable in a non parametric way on the basis of observed data in competing risks framework. Several works deal with this problem [21-24]. In particular Peterson [21] 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:

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

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).

To improve the above mentioned bounds a bivariate structure accounting for the dependence can be considered. In the case of two events (times y_1 and y_2)

with marginal survival functions $S_1(y_1)$ and $S_2(y_1)$. Klein and Moeschberger [23] proposed the following bivariate distribution:

$$S(y_1,y_2)=\left[S_1(y_1)^{1\text{-}\theta}\text{ - }S_2(y_2)^{1\text{-}\theta}-1\right]^{\frac{1}{1-\theta}} \qquad (14)$$

This distribution has the advantage that the association parameter θ is directly related to Kendall's tau: $\tau=\dfrac{\theta-1}{\theta+1}$ and it can be interpreted as the predictive hazard ratio:

$$\lim_{\Delta t\to 0}\frac{P(t\le Y_2<t+\Delta t\,|\,Y_2\ge t,Y_1=t)}{P(t\le Y_2<t+\Delta t\,|\,Y_2\ge t,Y_1>t)}$$

τ = -1 indicates a perfect negative association, i.e. subjects who experienced the event of interest have no chance to experience other competing events in the future; τ = 1 indicates a perfect positive association i.e. subjects who experienced the event of interest experience other competing events in the near future and τ = 0 implies a perfect independence among times to different events.

Given the relationship between net and crude survival, the estimation of net survival can be obtained solving a differential equation which, in the absence of ties and after fixing a value of θ, can be easily calculated on the bases of observed data [1].

$$S_k(t)=\left[1+(\theta-1)n^{\theta-1}\sum_{T(i)\le t,I(i)=k}\frac{1}{n-i+1}\theta\right]^{-\frac{1}{\theta-1}}$$

where $T_{(i)}$ are the ordered times for the occurrence of the all events (k=1 ,2) and times to the event which is not of interest are considered as censored. I(i)=k is the indicator function for the time i and the event k. It is worth of note that the function (14) pertains to a special multivariate distribution families called Copulas, which are 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. (14) is the Clayton Copula, a particular case of Archimedean Copulas.

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

$$C_\theta(S_1(y_1);S_2(y_2))=\Phi_\theta{}^{[-1]}\left(\Phi_\theta\left(S_1(y_1)+\Phi_\theta(S_2(y_2)\right)\right)$$

where $\Phi_\theta=[0,1]\times\odot\to[0,\infty)$ 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 is its pseudo-inverse defined by

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

Moreover, the above formula for C_θ yields a copula for $\Phi_\theta{}^{-1}$ if and only if $\Phi_\theta{}^{-1}$ is continuous and non-increasing on [0, ∞] and strictly decreasing on $\left[0,\Phi_\theta{}^{-1}(0)\right]$ [7].

The association parameter of Archimedean copulas has a direct relationship with Kendall's tau:

$$\tau=4\int_0^1\frac{\Phi_\theta(u)}{\Phi'_\theta(u)}\quad du+1$$

Several structured copula functions have been proposed in the literature and some detailed discussion of their properties and estimation procedures are also provided [6]. To estimate marginal function of Copulas, a graphic estimator has been proposed, that, in the special case of Archimedean Copulas, can be expressed in a closed form [25]. The original approach for copula graphic estimator is shown as a method to modify Kaplan-Meier estimator for the presence of dependent censoring by a Copula structure. It requires the observation of the variable Z=min(Y_1,Y_2) and the event observed at time Z. Times to event which are not of interest are then considered as censored.

Starting from the relationship:

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

Where $S(t)$ is the overall survival estimated by Kaplan-Meier method. The closed form for net survival estimator is:

$$S_k(t)=\Phi_\theta{}^{[-1]}\left(-\sum\nolimits_{ti\le t,\delta i=k}\left(\Phi_\theta\left(S(t_i)-\Phi_\theta(S\ \ (t_i)-1/n\right)\right)\right)$$

where n is the number of subject considered.

The approach does not allow for the presence of censored times in the case of individuals for which no events are observed. To overcome this limitation a recent improvement of the original estimator has been proposed [26].

Indicating with D the potential censoring time, δ the indicator for the event of interest (δ=1 if the event occurs and 0 otherwise) ρ the indicator for censoring (ρ=0 if the time is censored and 1 otherwise) and U=min (Z, D), the marginal distribution of $S_k(t)$ can be obtained from $J(t)=1-P(Z\le t)$ and $\tilde{J}(t)=P(Z\le t,\quad\delta=1)$ as follows:

$$S_k(t) = \Phi_\theta^{-1}\left[-\int_0^t \Phi'(J(s))d\tilde{J}(s)\right] \qquad (15)$$

Then it can be estimated plugging into (15) proper estimators of J(t) and $\tilde{J}(t)$ on the bases of the observed data U,ρ,ρδ. Under the assumption of independent censoring a suitable estimator for J(t) is the Kaplan-Meier (using U and ρ) and a suitable estimator for $\tilde{J}(t)$ the is the crude cumulative incidence $I_k(t)$.

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 of Kendall's tau which could be used as a first insight has been proposed [27], but it is not biased only in the case of independence between times to events.

In the case of semi-competing risks an alternative method based on Clayton copula has been proposed by Fine [13]. Two events are acting (k=1,2) and the event of type 1 is considered as the non terminal event and the event of type 2 the terminal one. The Clayton copula (14), is defined in the upper wedge $(1 \le y_1 \le y_2 \le \infty)$.

Fine proposed a strongly consistent estimator for the association parameter θ.

A closed-form estimator for net survival of the non fatal event k is obtained as:

$$\hat{S}_1(t) = \left[\hat{S}(t)^{1-\hat{\theta}} - \hat{S}_2(t)^{1-\hat{\theta}} + 1\right]^{\frac{1}{1-\hat{\theta}}}$$

where $\hat{\theta}$ is a consistent estimator for θ, $\hat{S}$ is the "overall" survival estimated by Kaplan-Meier method considering as events both non fatal and fatal events and $\hat{S}_2$ is survival from fatal event, estimated by Kaplan-Meier method considering as events only the fatal events, which are always observable.

3. SIMULATION STUDY

3.1. Monte Carlo Simulation for Net Survival Estimates

A Monte Carlo simulation was conducted to evaluate the performance of the net survival estimator proposed by Fine *et al*. [13] in a semi-competing risks context. The first aim is to evaluate the bias obtained by naive Kaplan-Meier avoiding to account for the dependence. The second aim is to evaluate the robustness of the estimator under the situation where the underlying structure is not a Clayton Copula. 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.

In order to generate multivariate survival data we refer to the simulation procedure based on copulas proposed by Rotolo *et al*. [28].

The simulations scheme is based on Clayton's and Frank's copula. The dependence parameters are those corresponding to unconditional Kendall's tau of τ = 0, 0.333, 0.5 and 0.75. Samples of sizes 200 are used. The random variable Y_1 has a unit exponential distribution; Y_2 has a unit exponential distribution as well, such that $P(Y_1 > Y_2) = 0.5$. The censoring variable C follows a uniform distribution on [0, a], where a is such that $P(Y_2 > C) = 20\%$. All simulations are based on 1000 replicates.

As regards estimators for $S_1(t)$, they are evaluated at $t_i=-\log(i/10)$, for i = 1, 3, 5, 7, 9, corresponding to the

Table 1: Simulation Results for the Survival Function of the non Terminal Event Estimated by Fine *et al*. Method and Kaplan-Meier Method. For every Simulation Scenario 1000 Sample of size 200 are Generate by Clayton 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

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

Table 2: Simulation results for the survival function of the non terminal event estimated by Fine *et al*. method and Kaplan-Meier method. For every simulation scenario 1000 sample of size 200 are generate by 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

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

10th, the 30th, the 50th, the 70th, and the 90th percentile of $S_1(t)$, the unit exponential survival function. Table **1** reports the simulation results of Fine and Kaplan-Maier estimator when data are generated from a Clayton copula models. Semi-parametric estimator proposed by Fine is unbiased. 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. Table **2** reports results of the simulation when data are generated from a Frank copula model. The bias for the estimate by Fine method is something biased and it increases at increasing association.

4. A SIMULATED DATA SET FOR THE COMPARISON OF THE ESTIMATED SURVIVAL/ INCIDENCE PROBABILITIES

The small dataset was generated by simulating data from the following bivariate distribution: $P\left(Y_1 > y_1, Y_2 < y_2 = C\left(\exp\left(-\lambda_1 y_1\right), \exp(-\lambda_2 y_2)\right)\right)$, where C is the Clayton Copula: $C(u_1,u_2) = \left[u_1^{-\theta} + u_2^{-\theta} - 1\right]^{-1/\theta}$.

The simulation algorithm suggested in [26] was used by the following steps:

20 values were generated from each of two independent random variables V_1 and V_2 distributed according to the exponential distribution with parameter=1.

20 values were generated from a random variable Z distributed according to a Gamma distribution with parameters 1/θ and 1. Then, for i=1,2 $U_i = (1 + V_i / Z)^{-1/\theta}$ were calculated and times $Y_i = -\ln + (U_i) / \lambda_i$ where finally obtained.

"observed" data in the competing risks framework were obtained considering the event corresponding to the minimum of (Y_1,Y_2).

The followings parameters values were used: $\lambda_1=\lambda_2=1$, θ=2.

For sake of simplicity in the interpretation of the estimated survival/incidences probabilities only non censored observation were generated (Table **3**).

Concerning deaths for all causes, survival probability S(t) was calculated by Kaplan-Meier method and the cumulative incidence I(t) as 1-S(t). It can be noted that, in absence of censoring, the estimates correspond to the proportion of subjects alive after time t and deceased before time t, respectively (Table **4**).

Table 3: The Dataset of 20 Times and Cause of Death, Simulated from Clayton Copula Model

times	Cause of death
0.02247599	1
0.03135967	1
0.04276071	1
0.11677077	2
0.15205448	1
0.16618929	2
0.24683757	2
0.28932287	1
0.35059856	2
0.39596928	1
0.53914335	1
0.68546373	1
0.69948798	1
0.96073401	2
1.08091976	1
1.58229144	1
2.03223993	1
2.91893249	1
3.21199164	1
3.69451010	2

Concerning the two causes of deaths (Tables **5-6**) net survival probability was calculated as:

$$S_i(t)=\left[1+\frac{\theta}{20}\Sigma_{k=1}^{t}\left(\frac{a.r(k)}{20}\right)^{-(\theta+1)}e(k)\right]^{-1/\theta}$$, for i=1,2.

where a.r(k) are subjects exposed at risk at time k and e(k) the number of death for the cause i at time k.

The cause specific survival functions ($S_1^*(t)$ and $S_2^*(t)$) were estimated by Kaplan-Meier method after considering as censored times to death for the competing cause. It can be verified that the property: $S(t)=S_1^*(t)\cdot S_2^*(t)$ holds. Crude survival probabilities ($S_{(1)}(t)$, $S_{(2)}(t)$) correspond to the proportions of subjects who deceased for the considered cause at a time greater than t. Crude survivals are always lesser than (or equal to) cause specific survivals. It can be verified that the property: $S(t)=S_{(1)}(t)+S_{(2)}(t)$ holds. Crude cumulative incidences ($I_1(t)$,$I_2(t)$) correspond to the proportion of patients whose occurrence of the death for the cause is lesser or equal to t. It can be verified that the property: $I(t)= I_1(t)+I_2(t)$ holds. 1-$I_1(t)$ and 1-$I_2(t)$ do not correspond neither to cause specific survivals nor to crude survivals.

5. EXAMPLE ON PROSTATE CANCER

The original study consists of 506 patients randomly allocated to one of four treatment regimes: placebo, 0.2 mg, 1.0 mg, and 5.0 mg DES daily. Further details regarding these data are given in [29]. As reported in the paper [30], placebo and 0.2 mg were designated as low-dose DES and 1.0 mg and 5.0 mg as high-dose. A dataset of this study was found at the following link http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets, where are reported records of 502 patients. Of these, 354 died (190 in low-dose and 164 in high dose) with 155 classified as cancer deaths (91 in low-dose and 64 in high dose), 127 classified as cardiovascular (59 in low-dose and 68 in high dose) and 72 classified as other causes (40 in low-dose and 32 in high dose). Survival time was recorded in months.

Aiming to compare treatment effect, the main end-point could be overall survival. Nevertheless, to deeply evaluate treatment effect, it is appropriate to consider, the causes of death.

Overall survival according to treatment is shown in Figure **1**. Patients treated with high dose, have a better

Table 4: Death for Each Cause. For each time t number of subject at risk (# a.r.), number of events (# e), Overall survival S(t) and overall incidence I(t) are reported

t	# a.r.	# e	S(t)	I(t)
0.02247599	20	1	0.95	0.05
0.03135967	19	1	0.90	0.10
0.04276071	18	1	0.85	0.15
0.11677077	17	1	0.80	0.20
0.15205448	16	1	0.75	0.25
0.16618929	15	1	0.70	0.30
0.24683757	14	1	0.65	0.35
0.28932287	13	1	0.60	0.40
0.35059856	12	1	0.55	0.45
0.39596928	11	1	0.50	0.50
0.53914335	10	1	0.45	0.55
0.68546373	9	1	0.40	0.60
0.69948798	8	1	0.35	0.65
0.96073401	7	1	0.30	0.70
1.08091976	6	1	0.25	0.75
1.58229144	5	1	0.20	0.80
2.03223993	4	1	0.15	0.85
2.91893249	3	1	0.10	0.90
3.21199164	2	1	0.05	0.95
3.69451010	1	1	0.00	1.00

Table 5: Death for cause 1. For each time t are reported: number of subject at risk (# a.r.), number of events (#e), cause specific survival $S_1^*(t)$, net survival $S_1(t)$, crude survival $S_{(1)}(t)$, 1- crude cumulative incidence ($1-I_1(t)$), cumulative incidence ($I_1(t)$)

t	# a.r	#e	$S^*_1(t)$	$S_1(t)$	$S_{(1)}(t)$	$1-I_1(t)$	$I_1(t)$
0.02247599	20	1	0.95000000	0.95346259	0.65	0.95	0.05
0.03135967	19	1	0.90000000	0.90660860	0.60	0.90	0.10
0.04276071	18	1	0.85000000	0.85945126	0.55	0.85	0.15
0.11677077	17	0	0.85000000	0.85945126	0.55	0.85	0.15
0.15205448	16	1	0.79687500	0.80344697	0.50	0.80	0.20
0.16618929	15	0	0.79687500	0.80344697	0.50	0.80	0.20
0.24683757	14	0	0.79687500	0.80344697	0.50	0.80	0.20
0.28932287	13	1	0.73557692	0.72295890	0.45	0.75	0.25
0.35059856	12	0	0.73557692	0.72295890	0.45	0.75	0.25
0.39596928	11	1	0.66870629	0.63065361	0.40	0.70	0.30
0.53914335	10	1	0.60183566	0.54929249	0.35	0.65	0.35
0.68546373	9	1	0.53496503	0.47609871	0.30	0.60	0.40
0.69948798	8	1	0.46809441	0.40912886	0.25	0.55	0.45
0.96073401	7	0	0.46809441	0.40912886	0.25	0.55	0.45
1.08091976	6	1	0.39007867	0.32144698	0.20	0.50	0.50
1.58229144	5	1	0.31206294	0.24939359	0.15	0.45	0.55
2.03223993	4	1	0.23404720	0.18706167	0.10	0.40	0.60
2.91893249	3	1	0.15603147	0.13107214	0.05	0.35	0.65
3.21199164	2	1	0.07801573	0.07950353	0.00	0.30	0.70
3.69451010	1	0	0.07801573	0.07950353	0.00	0.30	0.70

Table 6: Death for cause 2. For each time t are reported: number of subject at risk (# a.r.), number of events (#e), cause specific survival ($S_2^*(t)$), net survival ($S_2(t)$), crude survival ($S_{(2)}(t)$), 1- crude cumulative incidence ($1-I_2(t)$), cumulative incidence $I_2(t)$.

t	# a.r	#e	$S^*_1(t)$	$S_1(t)$	$S_{(1)}(t)$	$1-I_1(t)$	$I_1(t)$
0.02247599	20	0	1.0000000	1.00000000	0.30	1.00	0.00
0.03135967	19	0	1.0000000	1.00000000	0.30	1.00	0.00
0.04276071	18	0	1.0000000	1.00000000	0.30	1.00	0.00
0.11677077	17	1	0.9411765	0.92734486	0.25	0.95	0.05
0.15205448	16	0	0.9411765	0.92734486	0.25	0.95	0.05
0.16618929	15	1	0.8784314	0.84519340	0.20	0.90	0.10
0.24683757	14	1	0.8156863	0.76890882	0.15	0.85	0.15
0.28932287	13	0	0.8156863	0.76890882	0.15	0.85	0.15
0.35059856	12	1	0.7477124	0.68130096	0.10	0.80	0.20
0.39596928	11	0	0.7477124	0.68130096	0.10	0.80	0.20
0.53914335	10	0	0.7477124	0.68130096	0.10	0.80	0.20
0.68546373	9	0	0.7477124	0.68130096	0.10	0.80	0.20
0.69948798	8	0	0.7477124	0.68130096	0.10	0.80	0.20
0.96073401	7	1	0.6408964	0.47210060	0.05	0.75	0.25
1.08091976	6	0	0.6408964	0.47210060	0.05	0.75	0.25
1.58229144	5	0	0.6408964	0.47210060	0.05	0.75	0.25
2.03223993	4	0	0.6408964	0.47210060	0.05	0.75	0.25
2.91893249	3	0	0.6408964	0.47210060	0.05	0.75	0.25
3.21199164	2	0	0.6408964	0.47210060	0.05	0.75	0.25
3.69451010	1	1	0.0000000	0.03525661	0.00	0.70	0.30

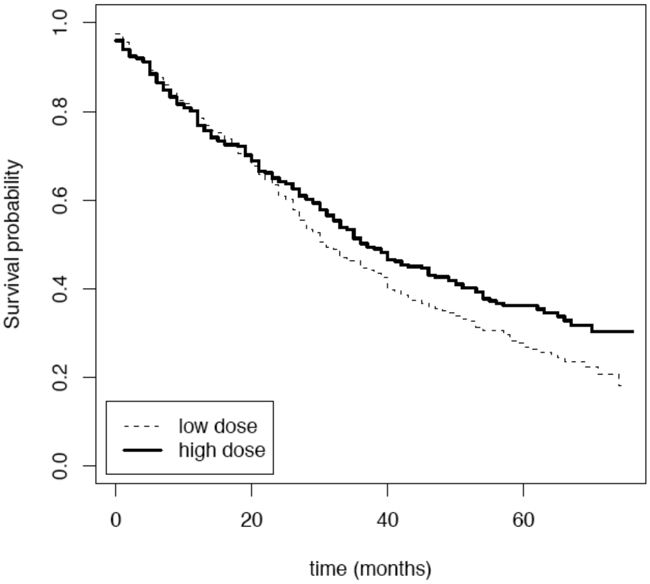

Figure 1: Overall survival estimated by Kaplan-Meier method according to treatment groupμ in prostate cancer dataset.

survival experience. Subdividing mortality according to the causes of dealth (Figure **2**), it can be observed that the greatest impact on mortality incidence is the mortality related to cancer (panel(a)) where the advantage of high dose treatment is evident. On the other hand, mortality incidence for cardiovascular disease is higher for patients with high dose treatment than that for patients with low-dose treatment. Incidences of mortality for other causes seem quite similar in the two treatment groups. The global advantage of high dose treatment is thus reduced because of the reverse effect of treatment in mortality for cardiovascular cause. If modification of treatment could allow to eliminate the mortality for cardiovascular death (and other causes of mortality), what would be treatment effect? The information can be obtained considering net survival for cancer. Before using a specific method for estimating net survival the association among times to death for different causes was evaluated. The estimated Kendall's tau for times to death for cardiovascular and cancer were: -0.0003 in patients treated with low dose and 0.0004 in patients treated with high dose. Thus having no evidence of sensible association, Kaplan-Meier method after censoring time to deaths for non cancer causes was adopted (Figure **3**). It is evident that the advantage of high dose treatment is greater than the corresponding one given by overall survival. It should be interesting to represent crude survival curves for the causes of deaths. The problem is the application of formula (13) in the presence of censored times when a death is not observed for all patients. At t=∞ all patients will die but we don't know the exact time and the cause of death.

6. EXAMPLE ON SMALL BREAST CANCER

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. Historical dataset regarding three clinical trials have been analyzed here.

Between 1973 and 1980, 701 women with breast cancers measuring no more than 2 cm in diameter were randomly assigned to undergo radical mastectomy (349 patients) or breast-conserving

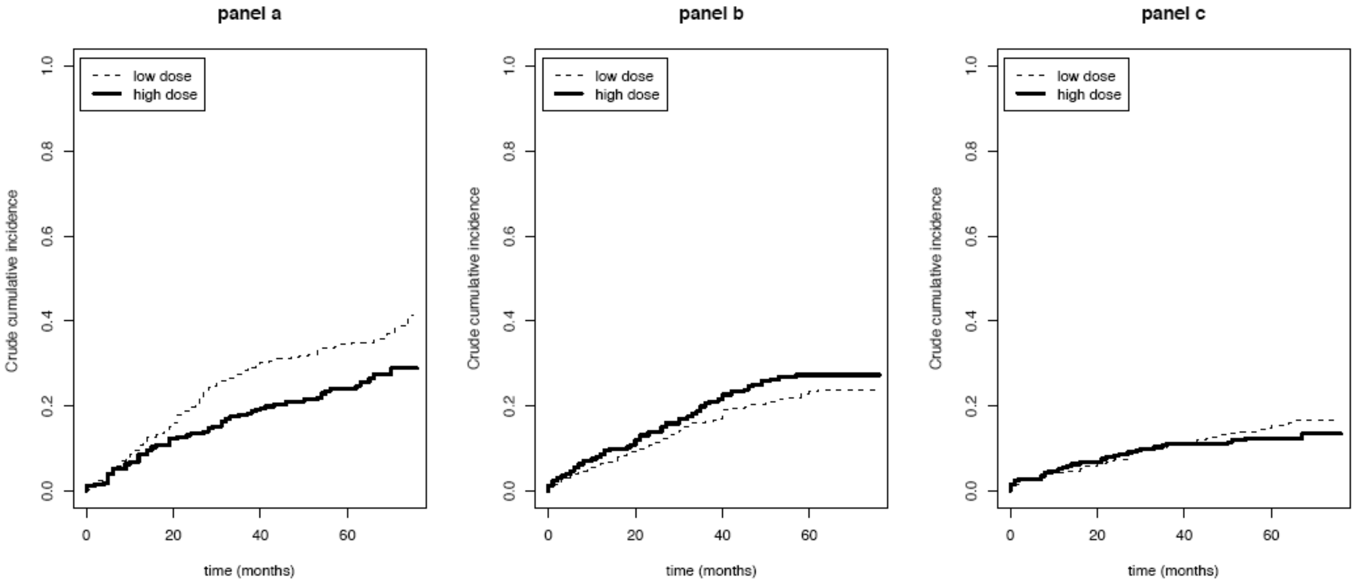

Figure 2: Estimates of crude cumulative incidences in prostate cancer dataset according to treatment group and cause of death: cancer (panel **a**), cardiovascular (panel **b**) and other (panel **c**).

surgery (quadrantectomy) followed by radiotherapy to the ipsilateral mammary tissue (QUART, 352 patients). After 1976, patients in both groups who had positive axillary nodes also received adjuvant chemotherapy with cyclophosphamide, methotrexate, and fluorouracil. Details on the results of this trial are reported in Veronesi *et al*. [31].

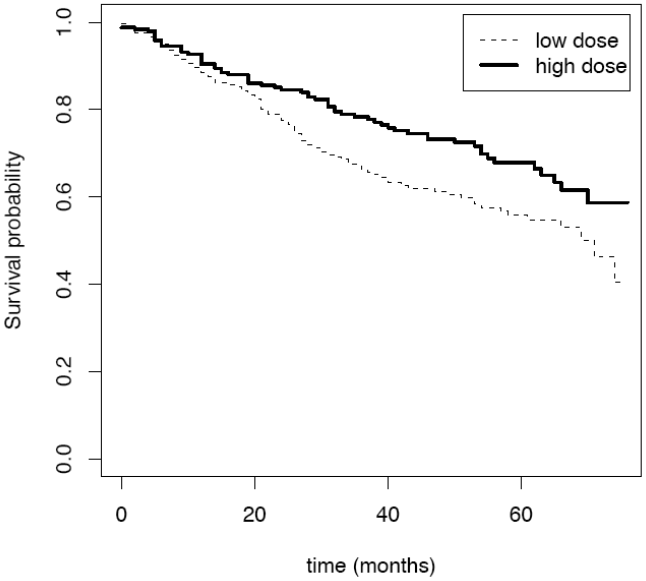

Figure 3: Estimates of net cancer survival in prostate cancer dataset according to treatment group.

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*. [32].

Between 1987 and 1989, 567 eligible women with carcinoma of the breast were randomly assigned to quadrantectomy, axillary dissection and radiotherapy (QUART, 294 women) and to quadrantectomy with axillary dissection without radiotherapy (QUAD, 273 women). Details on the results of this trial are reported in Veronesi *et al*. [33].

In both two last 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 estrogen receptors received tamoxifen.

Concerning overall survival and disease free survival, no statistically significant differences were found between surgical treatments. In any case, aiming to consider a homogeneous group of patients according to treatment, only quadrantectomy followed by radiotherapy (QUART) was considered. Thus a dataset formed by 1006 women is analysed (352 of the first trial, 360 of the second trial and 294 of the third trial). In this case series, the clinical well known prognostic factor is axillary lymph-node metastases, since all patients were classified in T1 tumour stage.

In the original papers [31-33] overall survival functions and crude cumulative incidences of first failure were presented according to treatment group, whereas our aim is the preliminary analysis on the whole case series in order to discuss the interpretation of the different survival curves and pertinent estimators, focusing attention on net breast cancer survival and net relapse free survival. A related purpose is the evaluation of association between different events occurring during the follow up, in particular of association between times to relapse and times to death. This quantity, although of clinical interest, was not evaluated in original articles [31-33].

Analysis of Death

As regards the clinical example we can observe that 272 of the 1006 women died within 15 years of follow-up. The overall survival function, given in Figure **4**, shows that the probability of surviving from death due to any cause 5, 10 or 15 years from surgery is respectively 0.91, 0.8 and 0.7.

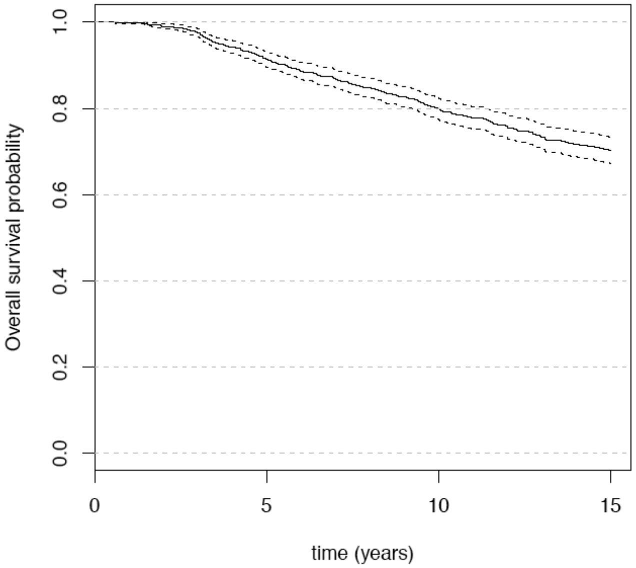

Figure 4: Overall survival estimated by Kaplan-Meier method (continuous line) and 95% confidence interval (dotted lines).

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 is 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, 200 deaths were classified as related to breast cancer and 72 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 other causes (Figure **5**, 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 [27] can be used, as first insight. The estimate is $\tau_K=15.5*10^{-5}$ thus the assumption of independence can be tenable. However Clayton copula graphical estimator can be used to compute net survival with association parameter $\alpha=3.1*10^{-4}$, corresponding to the Kendall's tau previously estimated. The estimated net survival probability is shown in Figure **5**, panel b. As expected, net survival estimates obtained by Kaplan-Meier method and copula graphic estimator, are practically overlapping. Concerning relative survival, the expected survival of the reference population was obtained by ISTAT mortality tables. The estimated net survival probability at 5, 10, 15 years is 0.93, 0.84 and 0.78 respectively, whereas relative survival at 5, 10, 15 year is 0.94, 0.85 and 0.79 respectively. It can be observed that the estimate is 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.

If the interest is on the incidence of death, the overall incidence is estimated by 1-Kapln-Meier overall survival estimates. Overall incidence can be decomposed in the incidence of death related to breast cancer and incidence of death related to other causes (Figure **6**). The greatest incidence of mortality is due to breast cancer. This incidence is the estimate of the probability of death as a consequence of breast cancer.

Analysis of Events

Within 15 years of follow-up the following events were observed: intra breast tumour recurrences (IBTR), distant metastases (DM), contralateral tumours (CT), other primary tumours, deaths without evidence of neoplastic events. The most comprehensive end-point

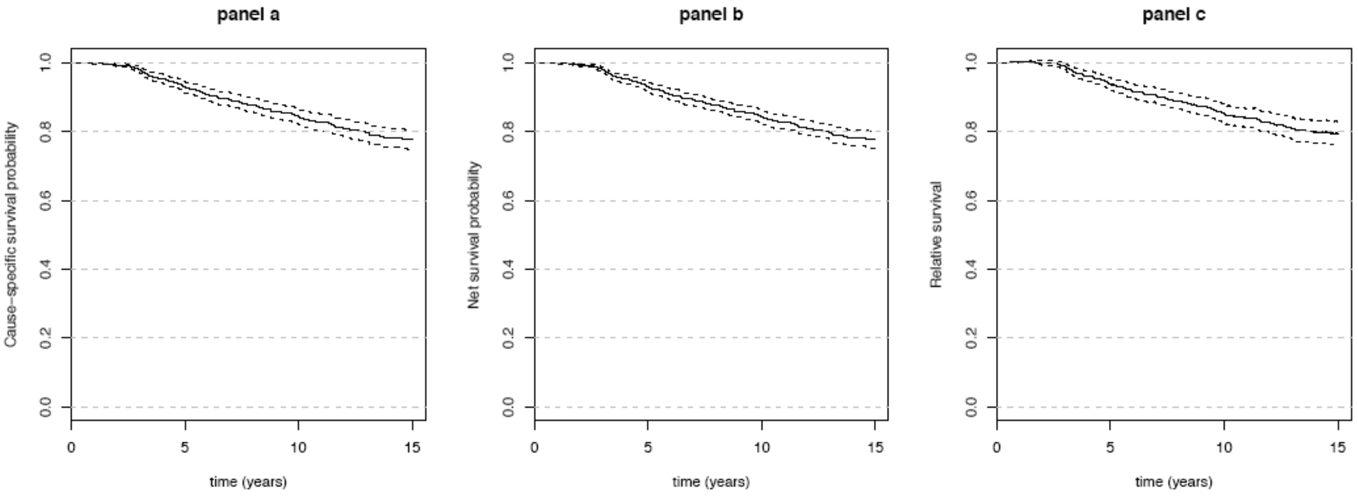

Figure 5: Estimates of (net) survival for breast cancer by Kaplan-Meier method (panel **a**), copula graphic estimator (panel **b**) and relative survival (panel **c**) (continuous line) and 95% confidence intervals (dotted lines).

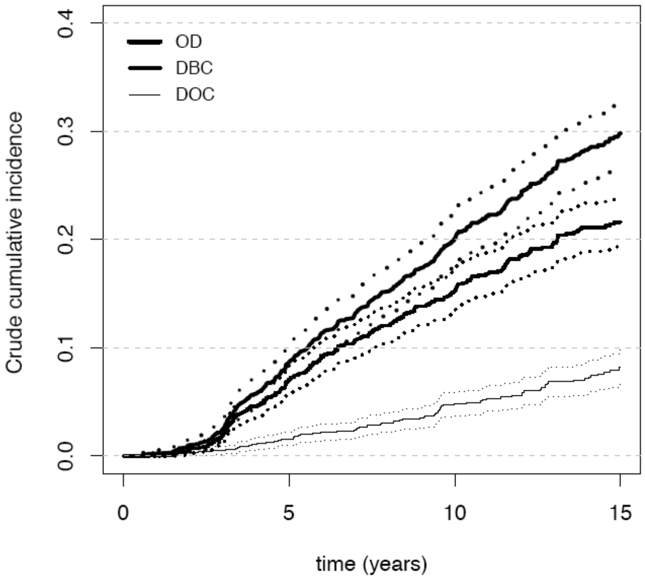

Figure 6: Estimates of overall incidence of mortality (OD) and of crude cumulative incidence of mortality for specific causes of death: death related to breast cancer (DBC) and death related to other causes (DOC) (continuous line) and 95% confidence intervals (dotted lines).

is composed by the first observed event, irrespective to the type of event (436 patients). In this framework, each event is considered directly or indirectly a failure to the therapeutic strategy and here named "first failure". First failure is then decomposed in events related to tumour progression (IBTR+DM+CT), named "relapse" (347 patients) and not related to tumour progression (89 patients). Crude cumulative incidences are of concerns. The crude cumulative incidence of relapse estimates the probability of observing a relapse as first event (Figure **7**).

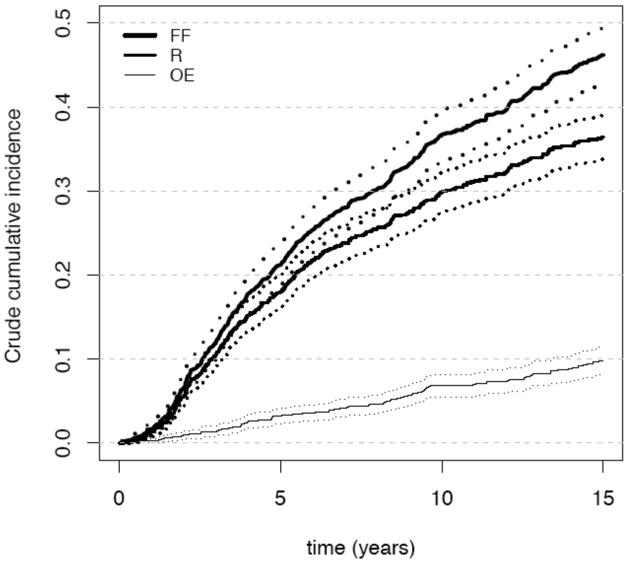

Figure 7: Estimates of first failure (FF) and of crude cumulative incidence of relapse (R) and of other events (OE) (continuous line) and 95% confidence intervals (dotted lines).

Relapse free survival probability is the measure most frequently reported in papers on breast cancer prognosis. Relapse free survival need to be interpreted as a "net survival" because it estimates the probability of surviving form relapse, in the hypothetical situation where a relapse is observable in all patients. As after the occurrence of other primary tumours, breast cancer tumour progression is no more considered as clinically interpretable and death as first event prevents the observation of the relapse, but no vice versa, thus the semi-competing risks framework is of concern. In this case the assumption of independence is not reliable because relapse can reasonably increase the risk of death. Thus a copula structure for the joint survival distribution is needed. The Fine estimate of the Clayton association parameter is 7.36, corresponding to a Kendall τ=0.79 and indicating a strong positive correlation between the two events. In particular, the hazard of death or other primary tumour after a relapse is 7.36 times bigger than the hazard of death or other primary tumour without previous relapse.

Relapse free survival estimated by Kaplan-Meier method is biased because of the lack of independence (Figure **8**, panel a) and it is higher than the copula estimation of Fine method (Figure **8**, panel b). Estimated relapse free survival at 5, 10, 15 years by Kaplan-Meier method is 0.82, 0.69 and 0.62 respectively and the corresponding estimates by Fine method is 0.82, 0.66 and 0.56 respectively.

The net relapse free survival must be included between the lower bound, representing survival to first failure and upper bound representing 1-crude cumulative incidence of relapse (Figure **9**). It can be noted that Fine estimates of relapse free survival is near to lower bound because of the strong positive association between times to events.

7. CONCLUDING REMARKS

One can say that at the beginning of the treatment each patient is potentially exposed to the risk of different kinds of events, each one at a different time period. According to the study aims usually composite end-points are defined considering the occurrence of at least one of the possible events or subsamples of events. Competing risk setting is usually cited if the occurrence of some events prevents the observation of the occurrence of other events of interest. Nevertheless it can be considered that also in the case of the most comprehensive composite end-point the occurrence of events is not always observable because of patients

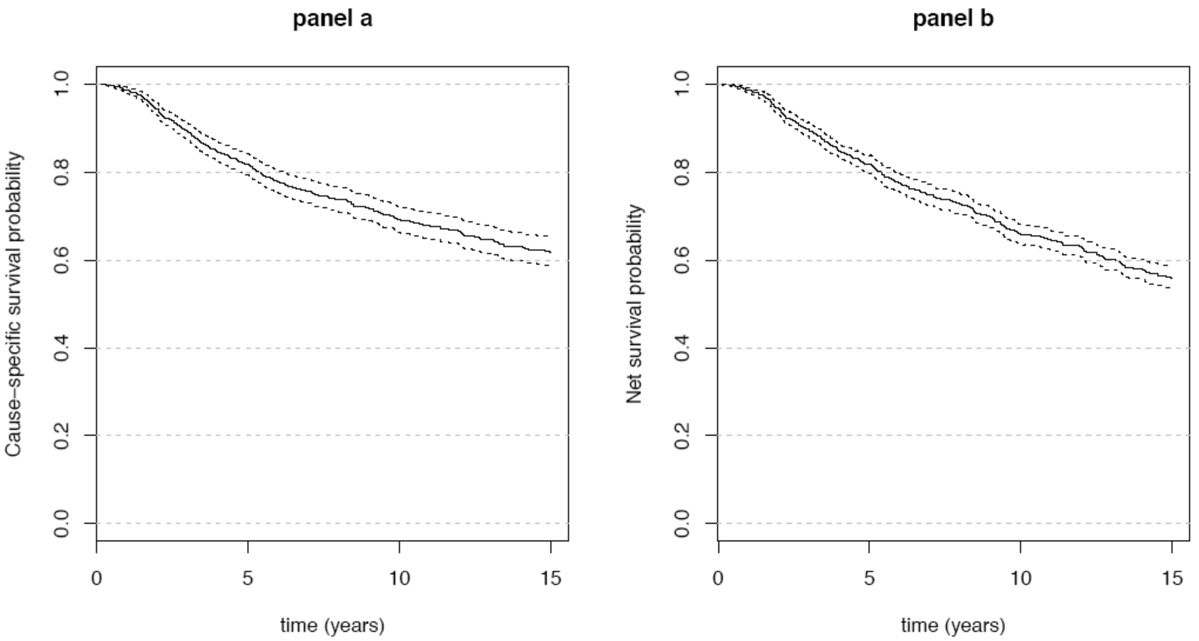

Figure 8: Estimates of (net) relapse free survival by Kaplan-Meier method (panel **a**) and Fine's method (panel **b**) (continuous line) and 95% confidence intervals (dotted lines).

who are lost to follow-up or because of administrative censoring, thus competing risks setting is always of concern. The independence assumption between the process of censoring and the occurrence of events should be carefully considered because biased estimates are obtained if dependence is ignored.

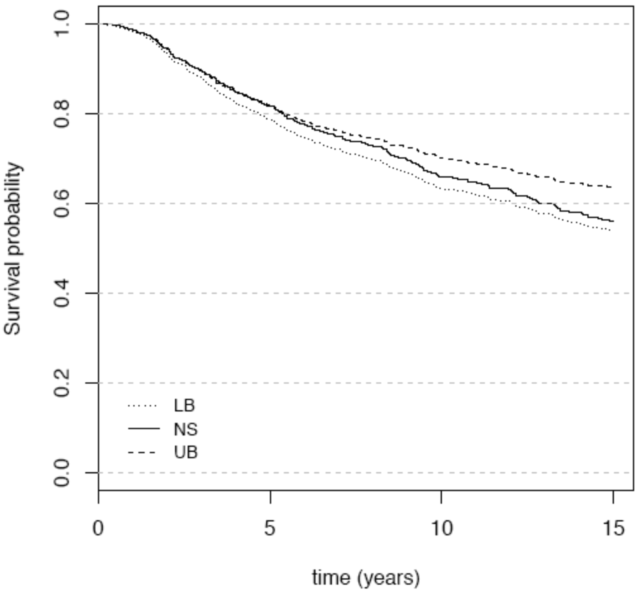

Figure 9: Bounding of net relapse free survival (NS) between lower bound (LB) represented by first failure free survival and upper bound (UB) represented by complement to 1 of the crude cumulative incidence of relapse.

When several kinds of events could occur during follow-up, partial event history is observed, independence among time to different events rarely can be assumed. This allows a simple non parametric estimate of crude survival or crude cumulative incidence while it avoids the possibility to estimate net survival functions without assumption on multivariate dependence structure. The attention is frequently on cumulative incidence of events (see [34] among others). We question whether this is always the most appropriate evaluation criteria or the choice is mainly related to the estimability of this measure and the availability of dedicated software. Since long time, in several clinical studies Kaplan-Meier method has been used in presence of competing risks to estimate the "event free survival curve" considering as (independent) censoring times to occurrence of other events. The interpretation of this estimate as "net" survival function is not explicitly provided, although it could be argued that "net" probability is of main interest for the Authors. On the other hand several methodological papers appeared on clinical journal to show that Kaplan-Meier method is not adequate and crude cumulative incidence needs to be used. In particular, some papers include consideration about the inappropriateness of the Kaplan-Meier methods to estimate crude cumulative incidence (see [34-35] among others). Again, the possibility of the interest in net survival is not taken into account. It is a matter of fact that the interpretation of net survival when more events are acting is more difficult than that of cumulative incidence, being related to an "hypothetical" situation of cause removal. Nevertheless marginal survival received attention since long time and an increasing number of studies are focussed on its estimation, mainly in the evaluation of mortality data on population registry (the most recent paper we found is [36].

From methodological point of view papers on the estimate of marginal survival can be found mainly since the beginning of 1980 and literature is increasing by proposing complex estimation methods (see [37] among others) and regression models [38-40]. Although the potential interest in clinical application and the availability of methodological issues, the use of net survival is till now not so diffuse as crude cumulative incidence. One of the reason may be that the non identifiability cited by Tsiatis's paper [4] is well known and the need of assumptions on multivariate distribution without the possibility to evaluate their tenability may arise doubts in "applied" statisticians. On the other hand, on classical survival books ([15, 41] among others) the theory of competing risk is mainly focused on estimable functions. Also considering a well known book specific for competing risk analysis [42], the survival/incidence functions are widely and deeply afforded, several multivariate distributional examples are given but the theory of Copulas is not shown. Copulas have been proposed as a flexible multivariate structures which can be used to estimate net survival probabilities. Multiple families of Copulas, having very different structures are described on books ([7] among others) and several evaluable papers on their properties, interpretations and sensitivity are available ([6] among others). This may render possible to chose a structure adequate to a specific clinical situation but the general estimation process of net survival, based on the relationship between net and estimable functions, requires the solution of a system of differential equations, which may be not easy to implement. An incentive to use a particular family of Copula function for clinical applications is the interpretability of its association parameter. The interpretation of the association parameter as predictive hazard ratio and the relationship with Kendall's tau, are some motivations for which Archimedean Copulas (in particular Clayton Copula) are suggested in clinical applications. Another incentive is the possibility to built a simple algorithms which can be adopted in a simple way on the basis of Kaplan-Meier and crude incidence probabilities. Archimedean Copulas give this possibility [1, 25-26]. A problem is that the above mentioned estimation algorithms are based on the knowledge of the value of the association parameter. To avoid this difficulty, some non parametric procedures available for Kendall's tau in the presence of bivariate censored data may be used ([27] among others). In the case of semi-competing risks, an unified approach has been proposed by Fine for a Clayton Copula [13]. The interpretation of the survival/probabilities functions available in competing risks framework allow to evaluate different aspects related to the prognosis and may give a more comprehensive knowledge of the disease progression and /or this impact on general population. In order to facilitate the comparison of the different survival/incidence estimates, a simple simulated dataset on 20 observation from Clayton Copula was proposed. The absence of censoring enables to use proportions of events which are easily interpreted.

Concerning the literature example on prostate cancer, the interpretation of treatment effect needs to decompose the overall incidence of death according to the different causes. This enable pinpointing the impact of the putative harmful effect of high dose treatment which is related to higher cardiovascular mortality. Although the overall survival experience is better for high dose treatment, the elimination of the harmful effect should improve the advantage of high dose treatment, as estimated by net survival.

Concerning the dataset on breast cancer, the estimation of the association parameter provides putative evidence that the independence assumption between causes of death is tenable then Kaplan-Meier approach can be used to estimate net breast cancer survival. On the contrary, in a semi-competing framework, a strong association between time to relapses and time to death was found. Thus the net relapse free survival estimate based on Clayton Copula is very different from the corresponding Kaplan-Meier one. The information is relevant as relapse free survival is one of the most considered end-point in cancer prognostic studies. It is worth of note that in the majority of published cancer prognostic studies it is estimated by Kaplan-Meier method censoring times to non considered events. The incidence of different events, estimated by Kalbfleish and Prentice method, gives a deep knowledge on the composition of treatment failure. In our example the impact of relapses is greater of that of death without evidence of previous neoplastic event. A deeper investigation of the contribution of the neoplastic events on treatment failure shows that metastases is the event of greatest impact (data not shown).

With this note we hope to stimulate applied statisticians to evaluate all possible measures for study end-points, including marginal survival and to develop simulation studies to investigate the robustness of multivariate distribution assumptions in several scenarios of clinical situations.

REFERENCES

[1] Zheng M, Klein JP. Estimates of marginal survival for dependent competing risks based on an assumed copula. Biometrika 1995; 82(1): 127-38. http://dx.doi.org/10.1093/biomet/82.1.127

[2] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Statist Assoc 1958; 53: 457-81. http://dx.doi.org/10.1080/01621459.1958.10501452

[3] Tai P, Joseph K, El-Gayed A, Yu E. Long-term outcome of breast cancer patients with one to two nodes involved - application of nodal ratio. Breast J 2012; 18(6): 542-8. http://dx.doi.org/10.1111/tbj.12010

[4] Tsiatis A. A nonidentifiability aspect of the problem of competing risks. Proc Nat Acad Sci USA 1975; 72(1): 20-2. http://dx.doi.org/10.1073/pnas.72.1.20

[5] Hougaard P. Modelling multivariate survival. Scand J Stat 1987: 291-304.

[6] Kaishev, VK, Dimitrova DS, Haberman S. Modelling the joint distribution of competing risks survival times using copula functions. Insurance: Mathematics and Economics 2007; 41(3): 339-361. http://dx.doi.org/10.1016/j.insmatheco.2006.11.006

[7] Nelsen RB. An Introduction to Copulas. New York: Springer, 1999. http://dx.doi.org/10.1007/978-1-4757-3076-0

[8] Martelli G, Boracchi P, Orenti A, *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 2014; 40(7): 805-12. http://dx.doi.org/10.1016/j.ejso.2014.03.029

[9] Rutherford MJ, Dickman PW, Lambert PC. Comparison of methods for calculating relative survival in population-based studies. Cancer Epidemiol. 2012; 36(1): 16-21. http://dx.doi.org/10.1016/j.canep.2011.05.010

[10] Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. Natl Cancer Inst Monograph 1961; 6: 101-21.

[11] Pohar Perme M, Stare J, Estève J. On estimation in relative survival. Biometrics 2012; 68(1): 113-120. http://dx.doi.org/10.1111/j.1541-0420.2011.01640.x

[12] Moliterni A, Ménard S, Valagussa P, *et al*. HER2 overexpression and doxorubicin in adjuvant chemotherapy for resectable breast cancer. J Clin Oncol 2003; 21(3): 458-62. http://dx.doi.org/10.1200/JCO.2003.04.021

[13] Fine JP, Jiang H, Chappell R. On semi-competing risks data. Biometrika 2001. 88(4): 907-919. http://dx.doi.org/10.1093/biomet/88.4.907

[14] Kalbfleish JD, Prentice RL. The Statistical Analysis of Failure Time Data. 2nd ed. Hoboken, New Jersery: John Wiley and Sons; 2002.

[15] Marubini E, Valsecchi MG. Analysing Survival Data from Clinical Trials and Observational Studies. Chichester: John Wiley and Sons; 1995.

[16] Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 1999; 94: 496-509. http://dx.doi.org/10.1080/01621459.1999.10474144

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

[18] Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. Biometrics 1982; 38: 933-42. http://dx.doi.org/10.2307/2529873

[19] Brenner H, Hakulinen T. On crude and age-adjusted relative survival rates. Journal of clinical epidemiology 2003; 56(12): 1185-1191. http://dx.doi.org/10.1016/S0895-4356(03)00209-9

[20] Slud EV, Rubinstein LV. Dependent competing risks and summary survival curves. Biometrika 1983; 70(3): 643-649. http://dx.doi.org/10.1093/biomet/70.3.643

[21] Peterson AV, Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. Proc Nat Acad Sci USA 1976; 73(1): 11-13. http://dx.doi.org/10.1073/pnas.73.1.11

[22] Peterson AV. Dependent competing risks: bounds for net survival functions with fixed crude survival functions. Environment International 1978; 1(6): 351-5. http://dx.doi.org/10.1016/0160-4120(78)90013-2

[23] Klein JP, Moeschberger ML. Bounds on net survival probabilities for dependent competing risks. Biometrics 1988: 529-38. http://dx.doi.org/10.2307/2531865

[24] Dignam JJ, Weissfeld LA, Anderson SJ. Methods for bounding the marginal survival distribution. Stat Med 1995; 14(18): 1985-98. http://dx.doi.org/10.1002/sim.4780141805

[25] Rivest LP, Wells MT. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. J Multiv Anal 2001; 79: 138-55. http://dx.doi.org/10.1006/jmva.2000.1959

[26] de Uña-Álvarez J, Veraverbeke N. Generalized copula-graphic estimator. Test 2013; 22(2): 343-360. http://dx.doi.org/10.1007/s11749-012-0314-2

[27] Brown BW, Hollander M, Korwar RM. Nonparametric tests of independence for censored data, with applications to heart transplant studies. Reliability and Biometry 1974: 327-54.

[28] Rotolo F, Legrand C, Van Keilegom I. A simulation procedure based on copulas to generate clustered multi-state survival data. Comput Meth Prog Bio 2013; 109: 305-12. http://dx.doi.org/10.1016/j.cmpb.2012.09.003

[29] Byar DP, Green SB. The choice of treatment for cancer patients based on covariate information. Bulletin du cancer 1979; 67(4): 477-490.

[30] Kay R. Treatment effects in competing-risks analysis of prostate cancer data. Biometrics 1986: 203-211. http://dx.doi.org/10.2307/2531258

[31] Veronesi U, Cascinelli N, Mariani L, *et al*. Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. N Engl J Med 2002; 347(16): 1227-32. http://dx.doi.org/10.1056/NEJMoa020989

[32] Mariani L, Salvadori B, Marubini E, *et al*. Ten Year Results of a Randomised Trial Comparing Two Conservative Treatment Strategies for Small Size Breast Cancer. Eur J Cancer 1998; 34(8): 1156-62. http://dx.doi.org/10.1016/S0959-8049(98)00137-3

[33] Veronesi U, Marubini E, Mariani L, *et al*. Radiotherapy after breast-conserving surgery in small breast carcinoma: long-term results of a randomized trial. Ann Oncol 2001; 12: 997-1003. http://dx.doi.org/10.1023/A:1011136326943

[34] Resche-Rigon M, Azoulay E, Chevret S. Evaluating mortality in intensive care units: contribution of competing risks analyses. Crit Care 2006; 10(1): R5. http://dx.doi.org/10.1186/cc3921

[35] Kim HT. Cumulative incidence in competing risks data and competing risks regression analysis. Clin Cancer Res 2007; 13(2): 559-65. http://dx.doi.org/10.1158/1078-0432.CCR-06-1210

[36] Allemani C, Weir HK, Carreira H, *et al*. Global surveillance of cancer survival 1995-2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2). Lancet 2014 Nov 26; Available from http://www.thelancet.com/journals/lancet/article/PIIS0140-6736%2814%2962038-9/abstract

[37] Yen Yen F, Ahmad N, Kassim S. A multistate approach to estimating the net survival function in the presence of competing risks. Malays J Math Sci 2011; 5(1): 125-41.

[38] Peng L, Fine JP. Regression modeling of semicompeting risks data. Biometrics 2007; 63(1): 96-108. http://dx.doi.org/10.1111/j.1541-0420.2006.00621.x

[39] Hsieh JJ, Huang YT. Regression analysis based on conditional likelihood approach under semi-competing risks data. Lifetime Data Anal 2012; 18: 302-20. http://dx.doi.org/10.1007/s10985-012-9219-3

[40] Lo SMS, Wilke RA. A regression model for the copula-graphic estimator. Journal of Econometric Methods 2014; 3(1): 21-46. http://dx.doi.org/10.1515/jem-2012-0016

[41] Elandt-Johnson RC, Johnson NL. Survival models and data analysis. New York: John Wiley & Sons, 1980.

[42] Crowder MJ. Classical competing risks. Boca Raton, FL: Chapman & Hall/CRC, 2001.

Received on 12-12-2014 Accepted on 30-01-2015 Published on 16-02-2015

http://dx.doi.org/10.6000/1929-6029.2015.04.01.14