International Journal of Statistical Analysis

Research Article Open Access

Estimation of the Incidence for Non-Terminal Events in Presence of a Terminal Event and Evaluation of Covariate Effects: Sub-Distribution and Marginal Distributions Based on Copulas. An Application to Disease Progression on a Breast Cancer Trial Dataset

Marano G[1] and Boracchi P[1,2]

[1]Department of Clinical Sciences and Community Health, Laboratory of Medical Statistics, University of Milan, Milan, Italy.

[2]DSRC Data Science Research Center, University of Milan, Milan, Italy.

***Corresponding Author: :** Giuseppe Marano, Department of Clinical Sciences and Community Health, Laboratory of Medical Statistics, Biometry and Epidemiology "G.A. Maccacaro", Via Vanzetti 5-20133 Milan, Italy, Tel: +39 02 50320855, E-mail: giuseppe.marano@unimi.it

Citation: Marano G (2019) Estimation of the Incidence for Non-Terminal Events in Presence of a Terminal Event and Evaluation of Covariate Effects: Sub-Distribution and Marginal Distributions Based on Copulas. An Application to Disease Progression on a Breast Cancer Trial Dataset. International Journal of Statistical Analysis. V1(2): 1-3.

Received Date: Dec 21, 2019 **Accepted Date**: Jan 31, 2020 **Published Date**: Feb 01, 2020

1. Abstract

In clinical studies, during follow-up several kinds of events related to disease progression may be observed. In the semi-competing risks setting, some events, such as death, may prevent the observation of disease progression, thus acting as competitors for the event of interest. Methods of analysis specific for semi-competing risks data referring to marginal distribution of the non-competing events constitute a recent area of methodological research which has received a great impulse in latest years. However, in clinical applications the analysis is traditionally based on crude cumulative incidences, and inference on marginal distributions is seldom considered, even when the principal aim concerns the probability of observing disease progression and death occurred without progression is a "nuisance". Aim of this work is making a comparative review of semi-parametric marginal and sub-distribution methods of analysis, with particular reference to marginal regression models based on copulas. More specifically, two structures were considered for marginal models: in the first one all parameters are time-dependent, while in the second one parameters vary with covariates but does not depend on time. Applications to breast cancer clinical trial data and to a simulated dataset are reported, to show the differences and the similarities among marginal and sub-distribution approaches. Results highlight that, when the competing event acts during the whole follow-up, the marginal approach became essential for the correct estimation of marginal incidences and covariate effects. Regression methods based on copulas are promising, however there is a need of refinements concerning model building strategies, and, of standardized software routine for the practical application of these methods.

2. Keywords: Semi-competing risks; Dependent censoring; Copula models; Time-indexed dependence structure; Net incidence; Crude cumulative incidence; Progression-free survival

3. Introduction

In several clinical settings the course of a disease is characterized by the occurrence of different unfavorable events during follow-up which can be ended with a terminal one (death). For each patient the event history is then recorded by the time to occurrence of each event starting from the beginning of the treatment or the disease diagnosis until the occurrence of the terminal event or the date of the end of the study.

The evaluation of treatment efficacy or the identification of prognostic factors are based on events which can be considered related to the disease and frequently grouped as a composite end-point (named usually "disease progression"). The time to disease progression is the time elapsed from the beginning of follow-up to the occurrence of the first evidence of progression, which is clinically interpretable as the first evidence of treatment failure. In the case of severe diseases, disease progression is expected to be observed for all patients before death, whose cause can be likely attributable to disease itself. On the contrary, such a condition does not occur for non-severe diseases, where long-follow-up is available and a patient can die without disease progression being

observed. The impact of mortality without disease progression is particularly relevant when the age at disease onset is greater than 50 years: in fact severe comorbidities which are not related to the investigated disease become more common with increasing age reducing patient's life expectancy. As an example, this is the case of early breast cancer where death without tumour recurrence is cited in several studies [6, 40] and in some studies it is analyzed as a specific end-point [42]. It is a matter of fact that in most studies disease progression is the primary end-point and the occurrence of death before disease progression is not an event of interest. As its occurrence prevents the observation of the primary end-point, death is considered a competing risk and it is not further investigated (several examples of this attitude can be found in breast cancer prognostic studies, see [21]) In such cases the statistical analysis is focused on the probability of occurrence of disease progression before death during follow-up (sub-distribution, indicated usually as crude cumulative incidence (CCI): [28]).

In a different perspective the marginal distribution of time to progression is of interest; thus a bivariate distribution of time to death and time to progression is of concern. The clinical interpretation of marginal distribution is the probability of disease progression in a condition where death cannot censor time to progression (i.e. disease progression would be observed for all patients). Concerning time to progression and time to death only the latter may compete with the first one to be the observed event, but no vice versa, and this setting is called "semi-competing risks" [14]. In this setting, occurrence of death without disease progression causes incomplete data on time to progression, therefore preventing the estimation of the bivariate distribution. As a consequence, the marginal distribution of time to disease cannot be estimated without making assumptions on the bivariate distribution.

Although, historically, the main aim of competing risks analysis was seen as the estimation of the marginal distributions [8] the sub-distribution approach is the most applied one. From a statistical perspective this approach is advantageous because crude cumulative incidence can be directly estimated from incomplete data and does not require assumptions about the dependence structure between time to events. On the contrary, the estimation of marginal distribution requires the knowledge of the bivariate distribution, which cannot be estimated from incomplete data. A wide literature about methodological research on crude incidence functions is available. The most diffuse methods in clinical applications are: extensions of log-rank test to compare crude incidence distributions [18] and semi-parametric regression models, which can be considered as extension of Cox model for competing risks [13]. These methods are implemented in statistical software and tutorials on clinical journals suggest the appropriate application and interpretation of results [29]. The clinical application of analysis based on marginal distributions is more controversial because the interpretation of marginal probability seems to be based on strong clinical a priori considerations on disease course which cannot be tested. Nevertheless, in some clinical papers the probability of being free form disease progression is roughly estimated by Kaplan-Meier method, considering time to death without progression as censored, under the assumption of independence between time to progression and time to death [36]. This attitude suggests a potential clinical interest in the marginal probability. Since in several clinical settings independence between time to disease progression and time to death cannot be assumed, the correct procedures of estimation of marginal probability are more complex than the use of Kaplan-Meier curves, as above described.

In the semi-competing risks setting fully non parametric methods cannot be applied, and a simple estimation approach is based on assumed bivariate structures (copulas) with non-parametric marginal distributions. Starting from year 2000, several methodological papers deal with the use of copula functions in estimating the marginal distribution of a non-terminal event censored by a terminal one [14,26,30,49]. The structure of copulas involves a parameter which is related to the dependence between the time to terminal event and the time to non-terminal event. This is a further advantage of the model because of the clinical interest on this feature. At the best knowledge of the authors, inference methods based on copulas for assessing covariates effects on marginal distributions have been proposed by regression models [4,22,23,33]; some of them by quantile regression [24,25]; and one taking into account of clustered data [34]. For sake of simplicity we consider proposals of semi-parametric regression models by [4,22,23,33]. Although the models we considered are based on the Clayton copula, parameter constraints and estimation methods differ, thus results depend on the specific method used.

The aim of the present work is to compare and to discuss data analysis of crude cumulative incidence and marginal incidence when semi-competing risks are of concern. For each model, parameters interpretation will be pointed out allowing the reader to have a correct interpretation of results. Moreover, pro and cons will be pointed out for the different approaches. To the above

ends, a practical application will be shown using data on a breast cancer clinical trial where information about 15 years follow up are available. Moreover, to have a deeper insight, simulated data were used where the times to non-terminal and terminal events compete along the whole follow-up.

4. Methods

This section is organized as follows: in the first paragraph, the functions of interest - i.e. survival functions, cumulative incidences and instantaneous rates - for each of the sub-distribution and the marginal approaches, will be discussed. In the second paragraph the basic properties of copulas will be sketched, in order to provide the necessary basis for interpretation of applied regression models. In the third one, semi-parametric methods of analysis will be reviewed, including the recent proposals cited in the introduction. The focus will be on the bivariate context, in which only two events are considered. This choice will allow to discuss the topics of interest for the purposes of this paper without substantial losses of generality. For more details on the statistical methods discussed here, readers are referred to the following textbooks: [8,28,38] and to the other references provided in the followings.

4.1. Probabilistic Functions of Interest

The non-terminal and the terminal event will be indicated by the letters X and Y, respectively. Let T_X and T_Y be the corresponding occurrence times. Let $S(t_x,t_y) = P\{T_x > t_x, T_y > t_y\}$ be the bivariate survival function of (T_X , T_Y), with respective marginal: $S_x(t_x) = P\{T_x > t_x\} = S(t_x,0)$ and $S_y(t_y) = P\{T_y > t_y\} = S(0,t_y)$. The corresponding hazard functions will be indicated by: $h_x(t_x)$ and $h_y(t_y)$. Let also C be the censoring time, which will be assumed to be independent of both T_X and T_Y. The information about the first observed event among X and Y is "stored" within the variable: $T = min\{T_x,T_y\}$; and the functions: $\delta_x = I(T = T_x)$ and $\delta_y = I(T = T_y)$; where I(.) represents the indicator function. Covariates are indicated by the matrix Z with N rows (subjects) and K columns (covariates).

The sub-distribution approach is mainly concerned with assessing the probability of occurrence of each event as the first one. In this context cumulative probabilities, frequently called Crude Cumulative Incidences (CCIs), are expressed through the sub-distribution functions:

$$F_x^-(t) = P\{T \le t, \delta_x = 1\} \quad ,t \ge 0$$
$$F_y^-(t) = P\{T \le t, \delta_y = 1\} \quad ,t \ge 0 \qquad (1)$$

A distinctive feature of $F_x^-(t)$ and $F_y^-(t)$ is that they do not represent proper distributions. This property is shown by using the relationships with the distribution of the first event's time T: $F_T(t) = F_x^-(t) + F_y^-(t)$ for each t ≥ 0: as a consequence, for t tending to infinity the two functions cannot reach the upper limit of 1. Hence the name "sub-distribution".

Closely related to the sub-distribution functions, sub-hazard functions represent the instantaneous rates of the events of interest. The definition is given below for the non-terminal event (an analogous formula, with term Y in place of X, holds for the terminal event) :

$$h_x^-(t) = \lim_{\Delta t \to 0} \frac{P[t \le T < t+\Delta t,\ \delta_x = 1 \mid T > t \cup (T \le t, \delta_x = 0)]}{\Delta t}, \ t \ge 0. \qquad (2)$$

This expression represents the instantaneous rate (i.e. the "hazard function") of T_X based upon a risk set which includes (see the expression on the right side of the conditioning symbol: |) both the subjects withouth no event observed until t (in symbols: $T > t$) and the subjects who have "failed" before t for events other than X (in symbols: $T \le t, \delta_X = 0$). It must be underlined that this interpretation could be realistic in contexts where events other than X does not terminate the event-hystory of the subject. However, this is not the case, in general, in competing and semi-competing risks settings. In fact, in the present case expression (2) implies that among subjects who did not experience the non-terminal event X before time t, those who experienced a terminal event, e.g. death, would be still considered at risk for X at t. From this considerations it follows that the definition of sub-hazard function is not realistic when a terminal event (at least one) can modify the "history" of subjects within the target population.

The relationship between sub-distribution and sub-hazard functions is given by:

$$h_x^-(t) = -d\log\left(1 - F_x^-(t)\right) / \ dt; \qquad (3)$$

which is equivalent to:

$$1 - F_x^-(t) = \exp\left(-H_x^-(t)\right); \quad \text{with: } \mathbf{H}_x^-(t) = \int_0^t \mathbf{h}_X^-(u)du. \qquad (4)$$

Expressions (3) and (4) are fundamental for the parameters interpretation of CCIs regression models: see the following paragraphs.

In contrast to the sub-distribution hazard, the cause-specific (CS) hazard represent the instantaneous rate of event times in subjects who are event-free:

$$h_x^{CS}(t) = \lim_{\Delta t \to 0} \frac{P[t \le T < t+\Delta t, \delta_X = 1 | T > t]}{\Delta t}.$$

This definition implies the lack of a direct relationship with the sub-distribution function, which may hinder the interpretation of covariate effects in CCIs regression models see, e.g., Di Serio (1997).

Concerning the marginal approach, the functions of interest, also called net functions, are the marginal survival and the marginal hazard function of the non-terminal event: $S_x(t_x)$ and $h_x(t_x)$. In fact, for the terminal event standard univariate methods (e.g.: Kaplan-Meyer curves, Cox regression) may be used, because all pertinent observations are available. In this framework standard relationships between cumulative probabilities and hazard functions hold:

$$S_x(t) = 1 - F_x(t) = exp\left(\int_0^t h_x(u)du\right) \quad , t \ge 0.$$

We are confident that such expressions are well known to the potential readers, so we will avoid to discuss them further. Within the semi-competing risks setting $S_x(t)$ and $h_x(t_x)$ represent the marginal survival and the marginal instantaneous rate of T_X: therefore, no "external influence" - that is, no event that could influence the occurrence of X or modify the risk set - is accounted for. This interpretation intrinsically refers to a hypothetical context in which the non-terminal event would not be censored by the terminal one, and thus it would be observed for all subjects. This could be of interest for researchers who want to investigate, for example, the progression of a disease in a population, by removing any possible source of nuisance. The clinical meaning of net and sub-distribution functions has been widely discussed, for example see: [7,19,38].

Theoretical relationships between sub-distribution functions and net functions have been thoroughly investigated. The key point is that under general conditions net functions are not directly estimable from incomplete data, while sub-distribution functions are [44]. Exact relationships that allow to bypass this issue are available under the assumption of independence between T_x and T_y: for example, it may be shown that $h_x^-(t) = h_x^{CS}(t)$. Such relationships, however, are not a distinctive feature, because they may also hold true for dependent event-times. More in general, sub-distribution functions may be used to derive lower and upper bounds for the net functions. The most general result has been obtained with Peterson's bounds [35]; however, it is recognized that often such bounds are too broad for being useful in applications. Several refinements of Peterson's bounds have been proposed in literature. A detailed discussion of this topic is beyond the purposes of this paper, but it is worth mentioning that, as a general rule, narrower bounds require more restrictive assumptions. In conclusion, relationships between net functions and sub-distribution functions may not always be useful for making inference on the former ones. To quote Crowder: << it's possible to have a pretty good picture of the >> [sub-distribution functions] << without this picture's telling you much about the >> [joint or the marginal distributions] [7].

4.2. Distributional Models Specified Through Copulas

The theory of copulas provides peculiar models for the multivariate distribution of event times, where the joint survival function (n is expressed as functions of the respective marginals. Then, a general expression for the bivariate survival function is: $S(t_x, t_y) = C(S_X(t_x), S_Y(t_y); \boldsymbol{\theta})$; where $C(\cdot, \cdot; \boldsymbol{\theta})$ is a "grounded" non-increasing function (a survival copula) with parameter $\boldsymbol{\theta}$. For more theoretical details, see [15]. The most frequently adopted model is the Clayton copula [5], whose expression is:

$$C(S_X(t_x), S_Y(t_y), \theta) = \left[S_X(t_x)^{1-\theta} + S_Y(t_y)^{1-\theta} - 1\right]^{\frac{1}{1-\theta}} ; t_x \ge 0, t_y \ge 0, \theta \ge 1. \quad (5)$$

Like copulas in general, $S_x(t_x)$ and $S_y(t_y)$ can be specified by either parametric or non-parametric functions. In each case, for given $S_x(t_x)$ and $S_y(t_y)$, the value of the parameter $\boldsymbol{\theta}$ determines the particular expression of the bivariate survival function: thus $\boldsymbol{\theta}$ "tunes" the dependence between event times. For example, for the Clayton copula it may be shown that, whichever the expressions of the marginals, the "independence distribution" $S(t_x, t_y) = S_x(t_x) * S_y(t_y)$ is obtained for $\boldsymbol{\theta} \to 1$. Furthermore, as a general result $\boldsymbol{\theta}$ is linked to correlation coefficients, even though explicit formulas are available only for particular copula models. For example, for the Clayton copula, the following relationship holds with Kendall's correlation coefficient:

$$\tau = E\left[sign(T_x - ET_x) * sign(T_y - ET_y)\right] = \frac{\theta - 1}{\theta + 1} \quad (6)$$

thus showing a direct link with the degree of correlation between T_X and T_Y. Note that in the expression above, incorrelation (that is, $\tau = 0$) is easily obtained for $\boldsymbol{\theta} = 1$.

In the context of multivariate survival analysis, also including the competing and semi-competing risks settings, a large variety of copulas provides the means for flexible specification of distributional models. In order to obtain additional flexibility, the

models previously shown may be extended, by letting the copula parameter θ depend on time or on covariates and specifying regression models for the marginal distributions. These extended regression models will be illustrated in the following paragraph.

4.3. Statistical Models for Evaluating Covariate Effects

For the sub-distribution approach, non-parametric estimates of CCIs may be obtained by estimators akin to the Kaplan-Meyer estimates in univariate survival analysis [3,18] is usually performed for comparing cumulative incidences among two or more sub-groups. The regression method proposed by [13] is the most widespread in applications. The Fine and Gray model is formulated in terms of proportional sub-distribution hazards:

$$h_x^-(t|Z) = h_0^-(t)\exp(Z^T\beta) \quad , t \geq 0 \quad ; \qquad (7)$$

where $h_0^-(t)$ is an unspecified baseline sub-hazard, depending only on t, and the effect of Z consists in a multiplication of $h_0^-(t)$ for a proportionality constant. For each covariate, the magnitude of effect is expressed in terms of sub-hazard ratios: for example, for a categorical covariate, say Z_k, with modalities: z_0, z_1, …, z_{M-1}; and codified with M-1 dummy variables, then:

$$HR_{X(m)} = \frac{h_x^-(t|Z_k=z_m)}{h_x^-(t|Z_k=z_0)} = \exp(\beta_m) \quad , m = 1, \dots, M-1; \qquad (8)$$

where β_m is the regression coefficient associated to the m-th dummy variable. These hazard ratios express the relative variations of the sub-hazard between subjects with $Z_k=z_m$ and $Z_k=z_0$ (reference category), all the remaining covariates remaining constant. By similar arguments, when Z_k is numeric the hazard ratio $HR_X = \exp(\beta)$ quantifies the variation of the sub-distribution hazard due to a one unit increase of Z_k, all other covariates remaining constant.

By using expressions (2) and (3) the model (7) may be equivalently expressed in terms of sub-distribution functions:

$$F_x^-(t|Z) = 1 - (1 - F_0^-(t))^{\exp(Z^T\beta)};$$

with: $1 - F_0^-(t) = exp\left(H_0^-(t)\right)$. This expression shows that any increase (decrease) of the sub-distribution hazard due to covariates corresponds to an increase (decrease) in CCIs. However, as pointed out, amongst others, by Austin and Fine (Austin and Fine 2017), the amount of increase (decrease) is not the same in the two contexts: thus, the HR in expression (8) cannot be interpreted as a measure of relative incidence.

In the semi-competing risks setting, observable event times are expressed through the variables: $\tilde{T}_Y = \min\{T_y, C\}$ and $\tilde{T}_X = \min\{T_x, T_y, C\}$. In a geometrical perspective, the couples $(\tilde{T}_X, \tilde{T}_Y)$ lie in the region $W \equiv \{(t_x, t_y) | t_y \geq 0, 0 \leq t_x \leq t_y\}$. Furthermore, by recalling that $S_x(t_x) = S(t_x, 0)$, it may be noticed that the net survival $S_x(t_x)$ lies outside the "observable region" W : this is a distinctive feature that makes estimation of net functions a non-standard problem.

In Fine et al's paper [14], an estimation procedure for marginal incidence of non-terminal event was proposed, which can be considered the base for the subsequent regression models. The estimation procedure does not require assumptions about the shape of $S(t_x, t_y)$ outside the observable region. The distributional model can be defined as follows:

$$S(t_x, t_y) = \begin{cases} \left[S_X(t_x)^{1-\theta} + S_Y(t_y)^{1-\theta} - 1\right]^{\frac{1}{1-\theta}} & , (t_x, t_y) \in W \\ S'(t_x, t_y) & , otherwise \end{cases} \qquad (9)$$

The model corresponds to the Clayton copula (5) in the observable region (W), whereas outside such region it corresponds to an unspecified survival function: $S'(t_x, t_y)$. For estimating the net survival $S_x(t)$, a useful formula is derived by inverting the expression of the Clayton copula for $t_x=t_y=t$:

$$S_x(t) = \left[S(t,t)^{1-\theta} - S_y(t)^{1-\theta} + 1\right]^{\frac{1}{1-\theta}} \quad , t \geq 0 \qquad (10)$$

The function $S(t, t)$ above is related to the distribution of the first event time T: $S(t, t) = P\{T_x > t, T_y > t\} = P\{T > t\} = 1 - S_T(t)$. [14] propose a consistent estimator for $S_x(t)$, obtained by plugging in consistent estimators of $S_y(t)$, $S_T(t)$ and θ in then above expression. Of note, consistent estimators of $S_T(t)$ and $S_y(t)$ may be obtained, for example, by the Kaplan-Meyer method. For estimates of θ fulfilling the above requirements, several alternatives are available: e.g. those in [14,30].

In subsequent works (Peng and Fine 2007, Hsieh and Huang 2012, Chen 2012) semi-parametric regression methods have been proposed, assuming more general dependence structures than in expression (9). In the first two proposals [23,33] the general model allows dependence of the copula parameter on event times:

$$S(t_x, t_y|Z) = C(S_X(t_x|Z), S_Y(t_y|Z), \alpha(t_x, t_y)), \quad (t_x, t_y) \in W. \qquad (11)$$

For example, model (9) may be obtained as particular case of this expression by letting C(. , . , .) be the Clayton copula (expression (5)), and $\alpha(t_x, t_y)$ = θ. Marginal survival functions are represented through transformation models [9]:

$$S_x(t|Z) = g(Z^T\beta(t)) \ , t \geq 0$$
$$S_y(t|Z) = h(Z^T\gamma(t)) \ , t \geq 0 \quad . \qquad (12)$$

where g(.) and h(.) are monotone, differentiable and invertible functions, and $\beta(t)$ and $\gamma(t)$ indicating time-dependent effects of covariates. For example, by using the complementary log-log (cloglog) function, the net survival function for the non-terminal event has the following explicit form:

$$S_x(t|Z) = \exp\left[-\exp(\log H_0(t) + Z^T\beta(t))\right]. \qquad (13)$$

According to this expression, the parameters $\beta(t)$ are interpreted in terms of log-hazard ratios for the net cumulated hazards of the non-terminal event:

$$HR_{X(m)}(t) = \frac{H_x(t|Z_k = z_m)}{H_x(t|Z_k = z_0)} = \exp(\beta_m(t)) \ , m = 1, \dots, M-1; \qquad (14)$$

with $H_x(t|Z) = -\log S_x(t|Z)$. Of note, in this approach the baseline cumulative hazard ($H_0(t)$ in expression (13)) is estimated at each t along with $\beta(t)$ and $\alpha(t_x, t_y)$ (the latter one is estimated in a limited sub-region: t_x=t_y=t). Concerning estimation methods, Fine et al proposed a procedure based on generalized estimating equations, while Hsieh and Huang used a conditional likelihood approach. In both proposals, model parameters must be estimated separately for each observed event time. This requires the extra "working" assumption of independence among estimates across time, as in the context of longitudinal data analysis [47]. Furthermore, in both the proposals a consistent estimate of regression parameter $\gamma(t)$ (expression 12), obtained e.g. with a Cox model, must be "plugged in" in the estimating equation or the conditional likelihood. This feature is analogous to the method by [14] which requires consistent estimates of $S_T(t)$, $S_y(t)$ and θ in the copula model (expression 10).

In [4] the dependence structure is determined by choosing a specific copula, e.g. Clayton, but the copula parameter (here indicated by: $\alpha(Z)$) depends on covariates:

$$S(t_x, t_y|Z) = C(S_x(t_x|Z), S_y(t_y|Z), \alpha(Z)) \qquad (15)$$

This model could be useful, for example, when the target population may be divided in two or more strata with heterogeneous characteristics. Marginal distributions are specified through transformation models for counting process theory [48]:

$$H_X(t|\beta, R_x) = G_X\left[\int_0^t I(T_x \geq u)\exp(\beta_X^T Z_1(u))dR_x(u)\right]$$
$$H_Y(t|\beta, R_x) = G_Y\left[\int_0^t I(T_Y \geq u)\exp(\beta_Y^T Z_2(u))dR_y(u)\right] \quad . \qquad (16)$$

where R_X and R_Y represent the event-history before time t, and G_x, G_Y are non-negative, strictly increasing and continuously differentiable. Here regression parameters – i.e.: β_X and β_Y – do not vary with time. As previously discussed, transformation models allow to specify several distinct models for the survival function. Proportional hazard models are included as a particular case, allowing the interpretation of estimates in terms of net hazard ratios, as in expression (14). The method of estimation is Non-Parametric Maximum Likelihood (NPML), thus estimates are step functions, like in the methods previously discussed. However, in the NPML method a full Likelihood function is defined, which includes all model parameters, thus allowing to estimate them jointly.

5. Application

5.1. Plan of Analysis for Breast Cancer Data

Data were collected from 567 women with small, non-metastatic primary breast cancer who were recruited in a randomized controlled clinical trial at the National Cancer Institute in Milan between 1985 and 1989, submitted to surgery and subsequently followed for a period of 15 years [46]. Surgery was either quadrantectomy (QUAD) or quadrantectomy plus radiotherapy (QUART); complete information about this trial can be found in the reference above. For illustrative purposes, we report the application of the methods so far discussed, with the aim of evaluating and comparing the incidence of the first among several types of cancer related event of women in QUAD and QUART groups. The cancer-related events were: recurrence of breast tumour, omolateral or contralateral breast carcinoma, primary tumours in other site and regional or distant metastases. Overall, such events can be considered as manifestations of tumour progression: therefore, for the purposes of the present work they were grouped in a composite event. i.e., tumour progression. The terminal event was death.

Statistical methods described in the previous chapter have been applied. Differences of cancer progression incidence between the two experimental arms (QUAD, QUART) were assessed mainly by regression models. To such end, type of surgery was considered as categorical covariate and included in regression models by a dummy variable, with QUAD as reference category. Furthermore, for specifying transformation models, we used the cloglog link

(expression (12)) and the identity link (expression (16)), so that to have proportional hazard type expressions for the marginals. Results of regression models were reported in terms of estimates of regression parameters and net hazard ratios (expression (14)), estimates of copula parameters, and estimates of net cumulative incidences of cancer progression: $1 - S_1(t|Z)$; where $S_1(t|Z)$ is the marginal survival included in the model in expressions (11) and (12), or in the model in expressions (15) and (16). The goodness of fit of the models in expression (9) and in expressions (11) and (12) was evaluated by a graphical procedure specific for these models [33]. In this procedure, the estimates of the survival function for the first event ($S_T(t)$) derived from the fitted model are plotted against other estimates obtained by semi-parametric methods. In the figure, points eventually distant from the diagonal highlight disagreement between the two estimates, and, therefore, potential lack of fit of the copula model. An empirical check of the proportional marginal hazards was performed by plotting the estimates of log(-log($S_1(t)$)) obtained for model (9) versus log(t), separately for QUAD and QUART groups. The analyses were performed using R release 3.5.0 (R core group 2018) and Knime Analytic Platform version 3.6.0 [2], except for estimation of the Chen model (expressions (15) and (16)), which was performed using Matlab release 2019a [43].

5.2. Simulated Data

Data were generated for a sample size of 2000 observations, divided in two subgroups, say, group A and B. The distributional model was a bivariate Clayton copula with θ=3 (corresponding to Kendall's τ equal to 0.5). For the non-terminal event (say, cancer progression) we assumed an exponential marginal distribution with parameter 1/3, and a regression parameter β=2 in order to specify a higher instantaneous rate of events in group B than in group A. For the terminal event (say, death not related to cancer) we assumed an exponential marginal distribution with parameter 1/3 and a regression parameter β=1.5 (as before, specifying higher incidence in group B). This model represents a scenario in which the competing effect of terminal events acts within the whole follow-up. CCIs were estimated by the Fine and Gray model. Marginal distributions were estimated by the method of Hsieh and Huang.

5.3. Results for Breast Cancer Data

During a follow-up period of 15 years, cancer related events occurred to 123 women over 273 in the QUAD experimental arm, and to 99 over 294 in QUART. In QUAD, 67 women deceased, 57 of which after having experienced cancer related events, and 10 without previous cancer events. In QUART 60 women deceased, 53 with and 7 without previous cancer events, respectively. The first "conventional" approach was to consider cancer progression as first observed event. Non-parametric estimates of the respective CCIs are reported in (Figure 1). It may be seen that the CCIs for the QUART group (surgery plus radiotherapy) are lower at each time with respect to the QUAD group. The result of the Gray test: p=0.0018; indicates a significant difference of CCIs between the two groups. Furthermore, both the curves increase less steadily after about seven years from surgery, and become flat at about ten and eleven years, respectively for QUAD and QUART, as very few (n=4) cancer-related events were observed in the later follow-up period. The estimated sub-hazard ratio of cancer progression from the Fine-Gray model is 0.66 (95% C.I.: 0.51-0.86), indicating a lower hazard in QUART as compared to QUAD (reference category in the regression model).

Concerning the net incidences, we first considered the Clayton copula model (9). The model was fitted separately for QUAD and QUART groups. Estimates of the copula parameter θ are 6.2 and 13.0 for QUAD and QUART respectively. Since this parameter is unbounded on the right, it seems more convenient to consider the corresponding values of Kendall's tau (expression (7)), which are 0.72 and 0.86 respectively. Such values suggest a high correlation between time to progression and time to death, in particular for the QUART group. It must be recalled that the Clayton copula is assumed only in the "observable region" (see expression 9): thus

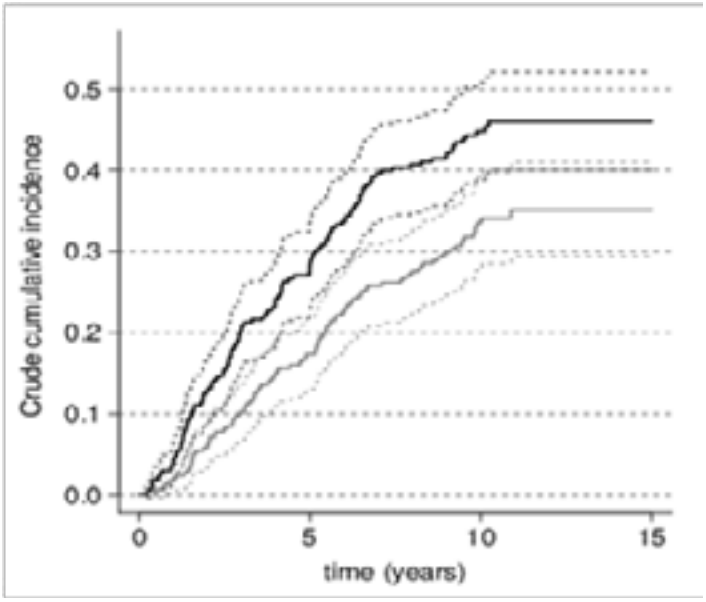

Figure 1: estimated CCIs of cancer related events with respective 95% Confidence Intervals
Solid lines: estimates, dashed lines: lower and upper 95% C.I.s. Black: QUAD; grey: QUART.

the estimates above cannot be interpreted as global correlations between event times. However, such estimates suggest strong dependence, and, consequently, inappropriate use of Kaplan-Meyer methods for estimation of net incidences.

In (Figure 2A) there are reported the estimates of the net cumulative incidence of cancer progression, together with the estimated CCIs previously shown. Net incidences are higher in the QUAD group. Within the figure evident differences between net and CCI estimates emerge only at later follow-up times, that is, after about ten years. In fact, before ten years' times to death are preceded by times to progression in the majority of cases, while after this period few cancer progressions are observed and causes of mortality are those expected to act in the same cohort population. Thus, CCI curves become flat, while, on the contrary, net incidence estimates increase. In the empirical check of proportional hazards for the two treatment groups (Figure 2B) no evidence emerged against the proportional hazards assumption. Finally, it is worth of note that no evidence of lack of fit of model (9) was shown in the diagnostic plot (Figure 2C).

The model specified by expressions (11) ad (12) was fitted by the Hsieh-Huang estimation method. We recall that for this model both regression and copula parameters vary with time, and estimates are calculated separately for each time. Estimates of regression and copula parameters and baseline cumulative hazard are reported in (Figure 3). First, we note that estimates of the cumulative hazard in the baseline group (Figure 3C) show a substantial decrease in later follow-up times. This is not admissible for a cumulated hazard function, which by definition is non-decreasing. In this case this may be happened because parameters are estimated separately without monotonicity constraints. However, in the current case only 4 cancer progressions were observed in the later follow-up (against 17 deaths), thus the above issue cannot be attributed to the performances of the estimation procedure, but, instead, to a low number of observed events. For this reason, in the subsequent analysis will consider only results related to times earlier than eleven years.

Considering times up to eleven years from surgery, the estimates of the copula parameter $\alpha(t, t)$ vary from to 8.28 and 8.41 (Figure 3A), suggesting no practical relevance against a standard model with constant θ, for the data under examination. Estimates of the regression parameter β(t) range from -1.69 to -0.12 (Figure 3B) corresponding to net HRs from 0.18 to 0.89. These results suggest lower incidence of cancer progression in the QUART group as compared to QUAD. However, at times earlier than about two years some estimates appear to be less stable than those obtained at subsequent times. This could be attributed to the low number of observed events: specifically, 9 deaths occurred before two years (1 before one year). Finally, the Peng and Fine diagnostic plot (Figure 3D) suggests a lack for the data in the QUAD group for the predictions in the range 0.4 to 0.6, corresponding to times close to 11 years (maximum time considered).

Finally, we consider results from the fitting of the model in expressions (15) and (16), estimated with the Chen method. In particular, this model has distinct copula parameters $\theta(Z)$ for QUAD and QUART groups, but no time varying parameters: $\alpha(t_x, t_y)$ and β(t). For avoiding potential problems in the

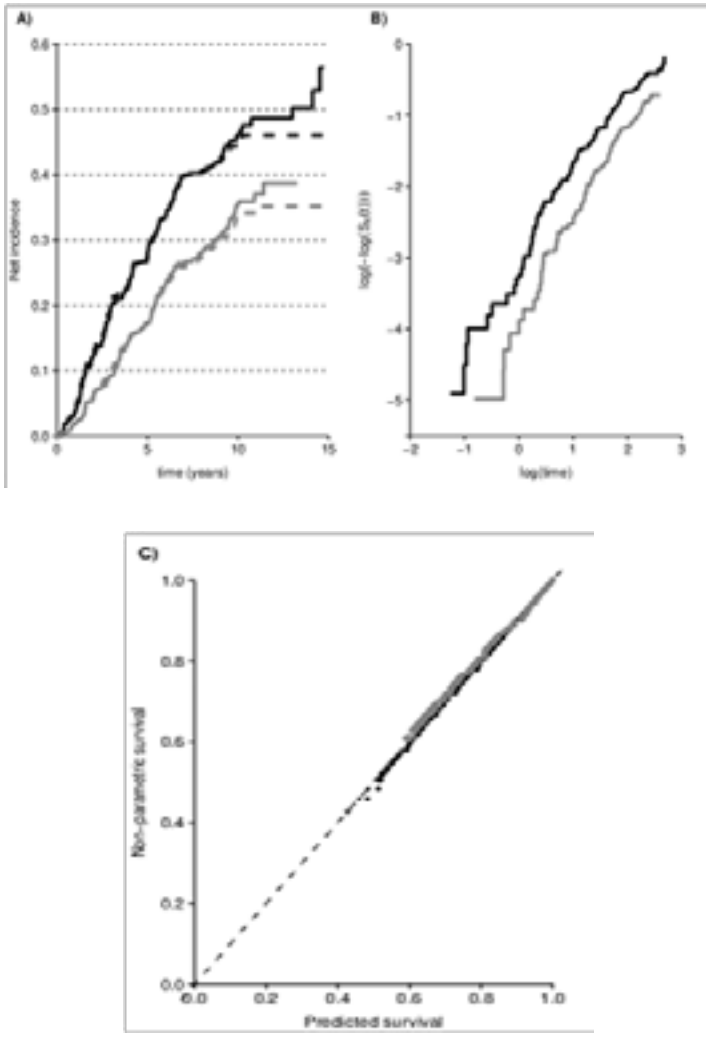

Figure 2: estimation of the semi-parametric model with a standard Clayton copula (expression (9))
Panel A: estimated net cumulated incidences for QUAD and QUART groups (solid lines). Dashed lines: estimates of CCIs for the two groups. Black: QUAD; grey: QUART.
Panel B: check of the proportional hazards assumption;
Panel C: goodness of fit for QUAD (black dots) and QUART (grey dots). The diagonal dashed line represents "perfect fit".

convergence of the estimation algorithm, we considered a restricted follow-up period, up to eleven years from surgery. Estimates of $\theta(Z)$ are 5.3 (95% C.I.: 3.9-6.8) and 10.4 (95% C.I.: 8.8-12.0) for QUAD and QUART respectively, corresponding to estimates of Kendall's τ of 0.73 (95% C.I.: 0.66-0.79) and 0.84 (95% C.I.: 0.66-0.79), respectively. These estimates are close to the ones shown in the first part of this paragraph, by applying model (9) separately for QUAD and QUART groups. The estimate of the regression parameter β comparing the incidence of cancer progression between QUART and QUAD groups is -0.36 (95% C.I.: -0.18 to -0.54), corresponding to a net HR of 0.70 (95% C.I.: 0.58 to 0.84). In Figure 4 the estimates of the net incidences are reported along with the estimates obtained for the previously considered models. It may be seen that such estimates are higher than the estimates obtained under model (9): that is, the model with a Clayton copula with constant θ. A better agreement between the two estimates could be expected, because the current model is rather similar to model (9). However, it should be considered that, in practice, estimation of the current model is cumbersome, because of the high number of parameters included in the likelihood function: 312 parameters in this application. In fact, the likelihood function

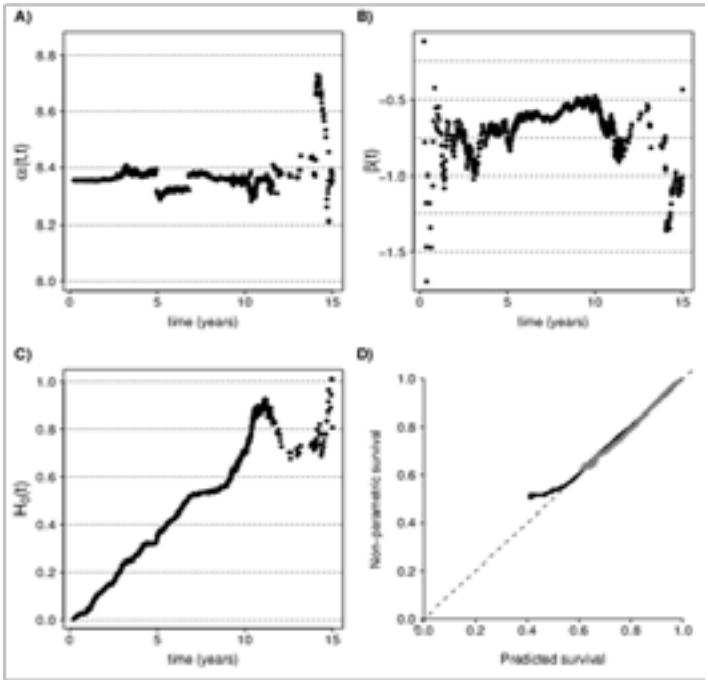

Figure 3: estimation of the semi-parametric model with extended time-varying copula (expressions (11) and (12)),
Panel A: estimates of the copula parameter $\alpha(t,t)$; B: estimated regression parameters β*(t)*; C: estimated baseline cumulative hazard $H_0(t)$; D: goodness of fit for QUAD (black dots) and QUART (gray dots) groups. The diagonal dashed line represents "perfect fit".

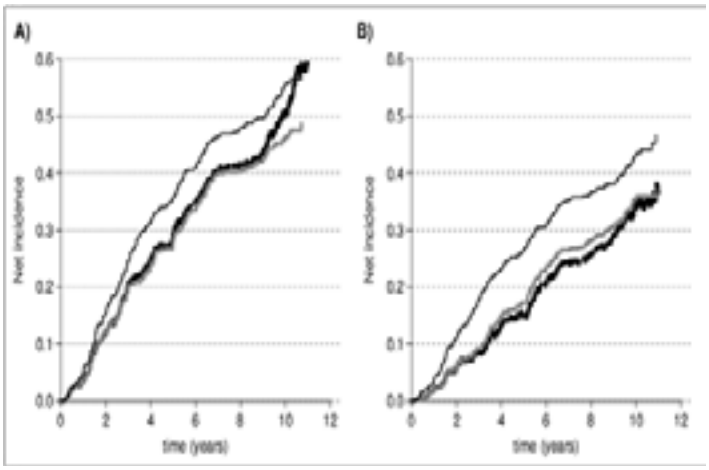

Figure 4. estimated cumulative incidences from the semi-parametric model with extended copula (expressions (15) and (16)),
Panel A: estimates for QUAD group; B: estimates for QUART group. Black solid lines: estimates for the extended copula model; grey solid lines: estimates for model (9); black dots: estimates for the extended time-varying copula model (expressions (15) and (16)).

includes, along with the parameters: $\theta(Z)$ and β, one parameter for each observed time to event. In conclusion, we are not able to guarantee that the reported estimates correspond to a global maximum of the likelihood function. A deeper examination of local and global maxima of a high-dimensional likelihood function could even more cumbersome, and however this task is beyond the purposes of this paper. Furthermore, a procedure for goodness of fit evaluation, similar to the one previously applied, is not defined for the current model.

5.4. Results for Simulated Data

In these "artificial" data, groups A and B include 998 and 1002 observations respectively. The number of cancer progressions was 189 and 713 respectively for groups A and B. In group A 286 deaths were recorded, 193 of which without previous cancer related events. Group B includes 785 recorded deaths, 203 of which without previous cancer related events. These frequencies are higher than those reported in the previous analysis, thus showing a higher "competitiveness" of the terminal event.

For what concerns CCIs, the estimate of regression coefficient β was 1.83, corresponding to a HR of 6.21. For the extended copula model (expressions (11) and (12)) estimates of the copula parameter $\theta(t)$ range from 1.12 to 4.37, while estimates of the regression parameter β(t) range from 1.12 to 4.37, corresponding to net HRs from 5.46 to 17.62. In figure 5 there are reported estimates of CCIs and of the marginal incidences of cancer progression, for groups A and B. It may be seen that in each group CCI estimates are lower

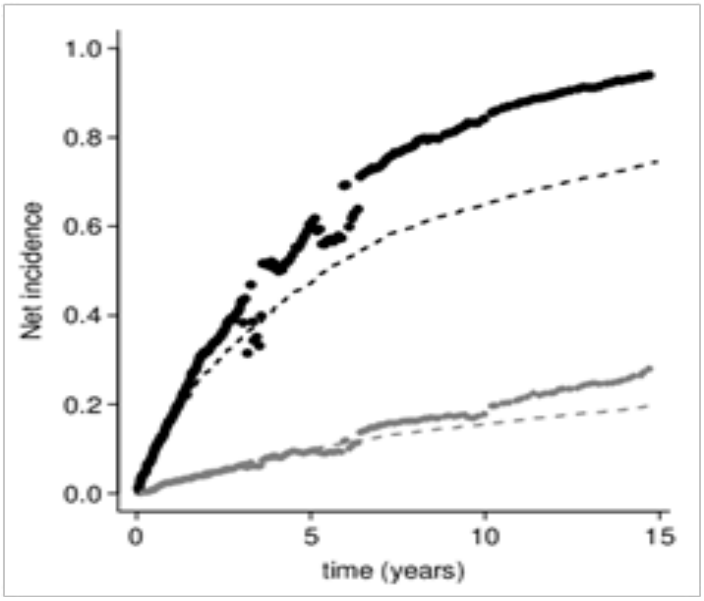

Figure 5. Analysis of simulated data
Solid lines: net incidences; dots: CCIs. Black: group B; grey: group A

with respect to estimates of the net incidences. This pattern was not observed in the previous analysis, where a differences between CCI and net estimates emerged only in the later follow-up.

6. Discussion and Conclusions

The presence of competing risks is a common issue in several clinical studies in which treatment failure can be due to different causes. Researchers are interested in the kind of event occurring as the first one, because of the possibility to investigate disease dynamics and to plan therapeutic strategies potentially preventing treatment failures. In this context, crude cumulative incidence of the different events is the best choice for the study aims. When the end-point is disease progression, the incidence of this event before death (for any cause) may be not of primary interest. This is the case of non-severe diseases with long follow-up available, or of diseases that can never be considered "cured": in such cases it could be preferable to estimate the incidence of disease progression under "removing" the effect of death without progression, that is, the marginal incidence. In this context, the approaches based on the bivariate distribution of time to death and time to progression are intuitive, and copulas with non-parametric-marginal distributions are the most flexible choice. Among several copula structures, the Clayton one is the most frequently considered because of its simplicity and its useful properties, among which, the direct link with gamma frailty models [32]. Frailty models have been proposed for semi-competing risks [12,17] nevertheless, despite the link between frailties and copulas, the two models are not interchangeable, as the parameters of the two models have a different interpretation [16]. In particular, regression models based on copulas are expressed in terms of net hazard, thus inference results can be extended to the marginal distribution but this is not possible for the parameters of frailty regression models. Frequently adopted methods for semi-competing risks data are those based on multi-state models [27,31,39]. These models have the advantage to be verifiable from incomplete data. Furthermore, information on transition intensities (from the beginning of treatment to disease progression or to death and from disease progression to death) are useful to investigate disease dynamics. However multi-state model coefficients are not directly linked to the marginal distribution. The variety of the above mentioned approaches has guaranteed the presence of a rich array of possibilities for methodological research, which has been "translated" into efforts in the development of methods of statistical analysis of semi-competing risks data. This work focuses only on specific topics: therefore, the references provided here are far from being a comprehensive review of literature.

Regression models based on Clayton copulas have been considered in the present study: an extended time-varying copula model [23,33] and an extended model with copula parameter constant with time [4]: the first includes a time dependent association parameter and time dependent regression coefficients for covariates, and the second has a simpler model structure (no time dependent effects). Issues were encountered with the algorithms for estimating model parameters. For the first model there are two estimation approaches; we found more easy to implement the approach described by Hsieh and Huang, which requires to find the maximum of a conditional likelihood function, separately for each time to event. For the second model, the author kindly shared a Matlab script, but, because of the large number of parameters included in the likelihood function, we actually cannot guarantee the optimal performance of the estimation procedure on our data. Moreover, the first model was more flexible, but in cases of lack of evidence of time dependent effect it is not possible to estimate a model with simplified structure using the same estimation approach.

For the breast cancer dataset, estimates of the crude cumulative incidence of disease progression until ten years from surgery were similar to those of marginal incidence. This results are explained by the fact that in this period the majority of deaths occurred after disease progression: thus, being disease progression occurred as the first event, crude cumulative incidence is in practice not distinct from marginal incidence. After the eleventh year, with an increased

occurrence of death as first event, differences emerged between the two incidences. However, estimates were unstable because of the low number of events in the late follow-up. In order to prevent those readers who have not a deep knowledge of competing risks topics, from concluding that the two procedures always provide the same results (anyway, marginal and sub-distributions refer to distinct concepts) we showed the analysis of a simulated dataset where both time to progression and time to death occur in the whole follow-up. In this case the difference between estimates of net and crude incidences can be appreciated.

Nevertheless, data analysis presented in clinical journals are usually based on crude cumulative incidences and/or multi-state models, even when the marginal probability of disease progression is of concern. The application of methods for the analysis of marginal distributions on clinical data are performed only by the respective authors or by "statistical followers" who appreciate the method. To our knowledge this is not attributable to lack of interest by clinical researchers in marginal distribution but, instead, to other causes, mainly to the novelty of proposed methods and, partly related to the previous one, to the lack of optimised software routines. Again, we underline that semi-competing risks constitute an appealing novel research area: thus, at the present time a standard, consolidated and widespread strategy of statistical analysis is far beyond being available to researchers. Nonetheless, the relevance of specific methods targeted on marginal distributions has been discussed by other authors, recently by [20,45] in the field of ageing research. In conclusion, we hope that this paper may be useful to stimulate biostatisticians to consider approaches based on marginal regression models when disease progression is of concern, supporting the relevant information for the clinical aims in spite of simplest solutions based on sub-distributions. These solutions, do not target the study outcomes when marginal incidences are of concern, and, as shown in this paper, could be misleading. At the same time, we hope to stimulate software developers to implement procedures for marginal regression in the most widespread statistical software, so that to provide easily available tools to researchers who are interested to apply and/or assess the performances of these approaches.

7. Acknowledgements

We are grateful to professor Yi-Hau Chen (Institute of Statistical Science, Academia Sinica, Taipei, Taiwan) for having shared the code for fitting the extended copula models: this has given us a precious opportunity that enriched our work. We acknowledge also professor Federico Ambrogi (Department of Clinical Sciences and Community Health, Laboratory of Medical Statistics, Biometry and Epidemiology "G.A. Maccacaro", University of Milan, Italy) and Annalisa Orenti (Department of Clinical Sciences and Community Health, Laboratory of Medical Statistics, Biometry and Epidemiology "G.A. Maccacaro", University of Milan, Italy) for the substantial support to research work.

References

1. Austin PC, Fine JP. Practical recommendations for reporting Fine Gray model analyses for competing risk data. Statistics in medicine. 2017; 36(27): 4391-400

2. Berthold MR, Cebron N, Dill F, Gabriel TR, Kotter T, Meinl T, et al. KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter. 2009; 11(1): 26-31.

3. Boracchi P, Orenti A. Survival Functions in the Presence of Several Events and Competing Risks: Estimation and Interpretation Beyond Kaplan-Meier. International Journal of Statistics in Medical Research. 2015; 4(1): 121-39.

4. Chen YH. Maximum likelihood analysis of semi-competing risks data with semiparametric regression models. Lifetime data analysis. 2012; 18(1): 36-57.

5. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika. 1978; 65(1): 141-51.

6. Colleoni M, Sun Z, Price KN, Karlsson P, Forbes JF, Thürlimann B, et al. Annual hazard rates of recurrence for breast cancer during 24 years of follow-up: results from the international breast cancer study group trials I to V. Journal of Clinical Oncology. 2016; 34(9): 927.

7. Crowder MJ. Classical competing risks. Chapman and Hall/CRC. 2001.

8. Crowder MJ. Multivariate survival analysis and competing risks. Chapman and Hall/CRC. 2012.

9. Dabrowska DM, Doksum KA. Partial likelihood in transformation models with censored data. Scandinavian journal of statistics. 1988; 1-23.

10. Day R, Bryant J, Lefkopoulou M. Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic indicators. Biometrika. 1997; 84(1): 45-56.

11. Di Serio C. The protective impact of a covariate on competing failures with an example from a bone marrow transplantation study. Lifetime data

analysis. 1997; 3(2): 99-122.

12. Do Ha I, Xiang L, Peng M, Jeong JH, Le, Y. Frailty modelling approaches for semi-competing risks data. Lifetime data analysis. 2019; 1-25.

13. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. JASA. 1999; 94: 496-509.

14. Fine JP, Jiang H, Chappell R. On semi-competing risks data. Biometrika. 2001; 88(4): 907-19.

15. Georges P, Lamy AG, Nicolas E, Quibel G, Roncalli T. Multivariate survival modelling: a unified approach with copulas. Available at SSRN 1032559. 2001; 72.

16. Goethals K, Janssen P, Duchateau L. Frailty models and copulas: similarities and differences. Journal of Applied Statistics. 2008; 35(9): 1071-9.

17. Ghosh D. Semiparametric inferences for association with semi-competing risks data. Statistics in medicine. 2006; 25(12): 2059-70.

18. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. The Annals of statistics. 1988; 16(3): 1141-54.

19. Grunkemeier G L, Jin R, Eijkemans MJ, Takkenberg JJ. Actual and actuarial probabilities of competing risks: apples and lemons. The Annals of thoracic surgery. 2007; 83(5): 1586-92.

20. Haneuse S, Lee KH. Semi-competing risks data analysis: accounting for death as a competing risk when the outcome of interest is nonterminal. Circulation: Cardiovascular Quality and Outcomes. 2016; 9(3): 322-31.

21. Hess KR, Esteva FJ. Effect of HER2 status on distant recurrence in early stage breast cancer. Breast cancer research and treatment. 2013; 137(2): 449-55.

22. Hsieh JJ, Wang W, Adam Ding A. Regression analysis based on semi-competing risks data. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008; 70(1): 3-20.

23. Hsieh JJ, Huang YT. Regression analysis based on conditional likelihood approach under semi-competing risks data. Lifetime data analysis. 2012; 18(3): 302-20.

24. Hsieh JJ, Ding AA, Wang W, Chi YL. Quantile regression based on semi-competing risks data. Open Journal of Statistics. 2013; 3(1): 12.

25. Hsieh JJ, Hsiao MF. Quantile regression based on a weighted approach under semi-competing risks data. Journal of Statistical Computation and Simulation. 2015; 85(14): 2793-807.

26. Jiang H, Fine JP, Kosorok MR, Chappell R. Pseudo self-consistent estimation of a copula model with informative censoring. Scandinavian Journal of Statistics. 2005; 32(1): 1-20.

27. Jiang F, Haneuse S. A semi-parametric transformation frailty model for semi-competing risks survival data. Scandinavian Journal of Statistics. 2017; 44(1): 112-29.

28. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. John Wiley & Sons. 1980; 360.

29. Kim HT. Cumulative incidence in competing risks data and competing risks regression analysis. Clinical Cancer Research. 2007; 13(2): 559-65.

30. Lakhal L, Rivest LP, Abdous B. Estimating survival and association in a semicompeting risks model. Biometrics. 2008; 64(1): 180-8.

31. Lin Y. Parametric estimation in competing risks and multi-state models. 2011.

32. Oakes D. Bivariate survival models induced by frailties. Journal of the American Statistical Association. 1989; 84(406): 487-93.

33. Peng L, Fine JP. Regression modeling of semicompeting risks data. Biometrics. 2007; 63(1): 96-108.

34. Peng M, Xiang L, Wang S. Semiparametric regression analysis of clustered survival data with semi-competing risks. Computational Statistics & Data Analysis. 2018; 124: 53-70.

35. Peterson AV. Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. Proceedings of the National Academy of Sciences. 1976; 73(1): 11-3.

36. Pickles MD, Manton DJ, Lowry M, Turnbull LW. Prognostic value of pre-treatment DCE-MRI parameters in predicting disease free and overall survival for breast cancer patients undergoing neoadjuvant chemotherapy. European journal of radiology. 2009; 71(3): 498-505.

37. Pintilie M. Competing risks: a practical perspective John Wiley & Sons. 2006; 58.

38. Pintilie M. Analysing and interpreting competing risk data. Statistics in medicine. 2007; 26(6): 1360-67.

39. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in medicine. 2007; 26(11): 2389-430.

40. Puttisri A, Pamarapa A, Jayanton Patumanond MD, Apichat Tantraworasin MD, Charoentum C. Recurrence and death from breast cancer after complete treatments: An experience from hospitals in Northern Thailand. J Med Assoc Thai. 2014; 97(9): 932-8.

41. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018. URL https://www.R-project.org/

42. Ring A, Sestak I, Baum M, Howell A, Buzdar A, Dowsett M, et al. Influence of comorbidities and age on risk of death without recurrence: a retrospective analysis of the Arimidex, Tamoxifen Alone or in Combination trial. J Clin Oncol. 2011; 29(32): 4266-72.

43. The MathWorks, MATLAB and Statistics Toolbox Release 2019a. URL https://www.mathworks.com/

44. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. Proceedings of the National Academy of Sciences. 1975; 72(1): 20-2.

45. Varadhan R, Xue QL, Bandeen-Roche K. Semicompeting risks in aging research: methods, issues and needs. Lifetime data analysis. 2014; 20(4): 538-62.

46. Veronesi U, Marubini E, Mariani L, Galimberti V, Luini A, Veronesi P, et al. Radiotherapy after breast-conserving surgery in small breast carcinoma: long-term results of a randomized trial. Ann Oncol. 2001; 12: 997-1003.

47. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986; 121-30.

48. Zeng D, Lin DY. Maximum likelihood estimation in semiparametric regression models with censored data. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2007; 69(4): 507-64.

49. Zhou R, Zhu H, Bondy M, Ning J. Analysing semi-competing risks data with missing cause of informative terminal event. Statistics in medicine. 2017; 36(5): 738-53.