Crude Cumulative Incidence in the form of a Horvitz-Thompson like and Kaplan-Meier like Estimator

Laura Antolini[*] Elia Mario Biganzoli[†]
Patrizia Boracchi[‡]

[*] Unità Operativa di Statistica Medica e Biometria Istituto Nazionale Tumori di Milano, laura.antolini@unimib.it

[†] Unità Operativa di Statistica Medica e Biometria Istituto Nazionale Tumori di Milano, elia.biganzoli@istitutotumori.mi.it

[‡] Istituto di Statistica Medica e Biometria Università di Milano, patrizia.boracchi@unimi.it

Crude Cumulative Incidence in the form of a Horvitz-Thompson like and Kaplan-Meier like Estimator

Laura Antolini, Elia Mario Biganzoli, and Patrizia Boracchi

Abstract

The link between the nonparametric estimator of the crude cumulative incidence of a competing risk and the Kaplan-Meier estimator is exploited. The equivalence of the nonparametric crude cumulative incidence to an inverse-probability-of-censoring weighted average of the sub-distribution function is proved. The link between the estimation of crude cumulative incidence curves and Gray's family of nonparametric tests is considered. The crude cumulative incidence is proved to be a Kaplan-Meier like estimator based on the sub-distribution hazard, i.e. the quantity on which Gray's family of tests is based. A standard probabilistic formalism is adopted to have a note accessible to applied statisticians.

Crude Cumulative Incidence in the form of a Horvitz-Thompson like and Kaplan-Meier like Estimator

L. Antolini[1] [1], E. Biganzoli[1] and P. Boracchi[2]

[1] Unità di Statistica Medica e Biometria, Istituto Nazionale per lo Studio e la Cura dei Tumori di Milano, Via Venzian 1, 20133 Milano, Italy.

[2] Istituto di Statistica Medica e Biometria, Università degli Studi di Milano, Via Venezian 1, 20133 Milano, Italy.

ABSTRACT

The link between the nonparametric estimator of the crude cumulative incidence of a competing risk and the Kaplan-Meier estimator is exploited. The equivalence of the nonparametric crude cumulative incidence to an inverse-probability-of-censoring weighted average of the sub-distribution function is proved. The link between the estimation of crude cumulative incidence curves and Gray's family of nonparametric tests is considered. The crude cumulative incidence is proved to be a Kaplan-Meier like estimator based on the sub-distribution hazard, i.e. the quantity on which Gray's family of tests is based. A standard probabilistic formalism is adopted to have a note accessible to applied statisticians.

Key words: Survival analysis, Competing risks, Nonparametric estimation, Gray's test, Sub-distribution hazard.

1 INTRODUCTION

In the competing risks setting several events may originate the occurrence of failure and are thought as competing causes of failure. The crude cumulative incidence of a given event (CCI), i.e. the cumulative probability of observing the event as first, is a quantity is of theoretical interest and practical application. The decomposition of the overall cumulative incidence of failure (CI) in sum of CCIs of each event, enable to analyse the events contributions in originating the failure.

[1] *Correspondence: L. Antolini, Unità di Statistica Medica e Biometria, Istituto Nazionale per lo Studio e la Cura dei Tumori di Milano, Via Venzian 1, 20133 Milano, Italy; E-mail: laura.antolini@unimib.it

The nonparametric maximum likelihood estimator of the CCI (Kalbfleish and Prentice, 1980) is the sum of unconditional probabilities of failure (due to the event of interest), obtained multiplying the cause-specific hazard by the overall survival probability. The Gray's family of nonparametric tests for comparing CCIs considers weighted averages of the sub-distribution hazards (Gray, 1988). This family of tests is akin to the Harrington and Fleming's one (Harrington and Fleming, 1982) for comparing overall survivals (or overall CIs, i.e. the complement to one of the overall survivals), which in turn, are usually estimated by the Kaplan and Meier method (Kaplan and Meier, 1958). The estimator of the overall survival obtained by this method is the product of conditional survival probabilities derived from the overall hazard of failure, which in turn, is the key quantity on which Harrington and Fleming's family of tests is based. The estimator of the overall CI can also be expressed as an inverse-probability-of-censoring weighted average (Satten and Datta, 2001), having the form of a Horvitz-Thompson estimator (Horvitz and Thompson, 1952) of the distribution function of the time to failure.

It is worth noticing as although there are analogies between the structure of the estimator of the CCI and that of the overall CI, the literature shows as if the interest is focused on the CCI, the use of the Kaplan-Meier method treating as censored events different from the one of interest, leads to an overestimate of the underlying CCI. The applied literature shows several examples where this incorrect use of the Kaplan-Meier methodology is present. From a theoretical viewpoint this is discussed in (Gooley et. al, 1999; Satagopan et. al, 2004). Nonetheless, to facilitate the applied statistician in understanding the link between nonparametric estimation and testing of the CCI, there is a need to explicitate whether the CCI's estimator can be expressed as a Kaplan-Meier like estimator based on the sub-distribution hazard, which, in turn is the key quantity on which Gray's family of tests is based.

The aim of the present note is to show as the CCI's estimator can be written in two equivalent forms: as an inverse-probability-of-censoring weighted average of the sub-distribution function of the time to failure, and as a Kaplan-Meier like estimator based on the sub-distribution hazard.

In the first section, setting the notation, the estimators of the CCI and of the overall CI (with its equivalent forms) are reviewed. In the second section, the CCI is expressed in the form of an inverse-probability-of-censoring weighted average. In the third section, the interpretation of the CCI as the cumulative incidence of an artificial failure time random variable whose hazard is the sub-distribution hazard (Elandt-Johnson and Johnson, 1980), is reviewed. Then, the estimator is proved to be obtainable as a Kaplan-Meier like estimator of the cumulative incidence of the artificial failure time random variable.

2 NOTATION AND BACKGROUND

Let T_f denote the possibly right censored failure time random variable in the competing risks setting, where there are $R \geq 2$ mutually exclusive causes of failure ($r = 1, ..., R$). Let ε denote the actual cause of failure, $T = min\left\{T_f; T_c\right\}$ the observed time, where T_c is the right censoring time, and δ the status indicator ($\delta = 1$ if $T = T_f$ and $\delta = 0$ otherwise). T_f and T_c are assumed independent (i.e. random censoring). Let $(t_i, d_i, d_i{\cdot}e_i)$ $(i = 1, ..., N)$ a sample of observations independent and identically distributed (IID) to $(T,\ \delta,\ \delta \cdot \varepsilon)$. The goal is to estimate the CCI of a cause r, defined as the sub-distribution function

$$F_r(t) = \Pr\left\{T_f \leq t; \varepsilon = r\right\} \tag{1}$$

where the semicolon is the intersection operator. Let us consider the following notation: $\tau_0 = 0$, $\tau_1 < \tau_2$ $, ..., < \tau_J$ $(J \leq N)$ are the distinct observed times; $n_{jr} = \sum_{i=1}^{N} \mathrm{I}\{t_i = \tau_j; e_i = r\}$ is the number failures at τ_j due to the cause r; $n_j = \sum_{r=1}^{R} n_{jr}$ is the number of failures (due to any cause) at τ_j; $m_j = \sum_{i=1}^{N} \mathrm{I}\{t_i = \tau_j; e_i = 0\}$ is the number of censorings at τ_j. If $m_j{\cdot}n_{jr} > 0$ for some j and r, i.e. there are ties among failures and censorings at τ_j, the m_j censorings are assumed to occur right after the n_{jr} failures. The observed number of subjects at risk of failure at τ_j is

$$Y(\tau_j) = \sum_{i=1}^{N} I\left\{t_i \geq \tau_j\right\} = \sum_{k=j}^{J} (n_k + m_k) \tag{2}$$

and the observed number of subjects at risk of being censored is $Y(\tau_j) - n_j$. The nonparametric maximum likelihood estimator of F_r (1) proposed by Kalbfleish and Prentice (1980) (KP) is

$$\widehat{F}_r(t) = \sum_{j:\tau_j \leq t} \widehat{h}_r(\tau_j) \cdot \widehat{S}_{T_f}(\tau_{j-1}) \tag{3}$$

The term $\widehat{h}_r(\tau_j) = n_{jr}/Y(\tau_j)$ estimates

$$\lim_{\Delta t \to 0^+} \Pr\left\{t < T \leq t + \Delta t; \delta \cdot \varepsilon = r | T > t\right\} / \Delta t$$

in $t = \tau_j$, which under random censoring, equals the cause-specific hazard function (Kalbfleish and Prentice, 1980)

$$h_r(t) = \lim_{\Delta t \to 0^+} \Pr\left\{t < T_f \le t + \Delta t; \varepsilon = r | T_f > t\right\} / \Delta t$$

The term $\widehat{S}_{T_f}(\tau_{j-1})$ is the Kaplan-Meier estimator (Kaplan and Meier, 1958; KM) of the overall survival $S_{T_f}(t) = \Pr\{T_f > t\}$

$$\widehat{S}_{T_f}(t) = \prod_{j:\tau_j \le t} \left(1 - \widehat{h}_{T_f}(\tau_j)\right) \tag{4}$$

where $\widehat{h}_{T_f}(\tau_j) = n_j / Y(\tau_j)$ is the estimate of

$$\lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \cdot \Pr\left\{t < T \le t + \Delta t; \delta = 1 | T > t\right\} \tag{5}$$

in $t = \tau_j$, which, under random censoring, is equal to the overall hazard of failure

$$h_{T_f}(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \cdot \Pr\left\{t < T_f \le t + \Delta t | T_f > t\right\}$$

The complement to one of (4), $\widehat{F}_{T_f}(t) = 1 - \widehat{S}_{T_f}(t)$, is the KM estimator of the overall CI ($F_{T_f}(t) = \Pr\{T_f \le t\}$). It is worth of note as $\widehat{F}_{T_f}$ can be equivalently written as

$$\widehat{F}_{T_f}(t) = \sum_{j:\tau_j \le t} \widehat{h}_{T_f}(\tau_j) \cdot \widehat{S}_{T_f}(\tau_{j-1}) \tag{6}$$

In fact, for t such as $\tau_k \le t < \tau_{k+1}$, the complement to one of (4) can be written as

$$\begin{aligned}
\widehat{F}_{T_f}(t) &= 1 - \left[1 - \widehat{h}_{T_f}(\tau_k)\right] \cdot \widehat{S}_{T_f}(\tau_{k-1}) \\
&= 1 - \widehat{S}_{T_f}(\tau_{k-1}) + \widehat{h}_{T_f}(\tau_k) \cdot \widehat{S}_{T_f}(\tau_{k-1}) \\
&= \widehat{F}_{T_f}(\tau_{k-1}) + \widehat{h}_{T_f}(\tau_k) \cdot \widehat{S}_{T_f}(\tau_{k-1})
\end{aligned}$$

and, reiterating this argument for $\widehat{F}_{T_f}(\tau_{k-1}), ..., \widehat{F}_{T_f}(\tau_2)$, follows the (6).

The KM estimator of the overall CI has also been expressed by Satten and Datta (2001) in the form of a Horvitz-Thompson (Horvitz and Thompson, 1952) estimator as the inverse-probability-of-censoring weighted average

$$\widehat{F}_{T_f}(t) = \frac{1}{N} \sum_{i=1}^{N} \frac{I\{t_i \le t; e_i \cdot d_i = 1\}}{\widehat{S}_{T_c}(t_{i-})} = \frac{1}{N} \sum_{j:\tau_j \le t} \frac{n_j}{\widehat{S}_{T_c}(\tau_{j-1})} \tag{7}$$

where: $t_{i-} = \max\{\tau_k : \tau_k < t_i\}$, and $\widehat{S}_{T_c}$ is the KM estimator of the censoring free survival function $S_{T_c}(t) = \Pr\{T_c > t\}$

$$\widehat{S}_{T_c}(t) = \prod_{j:\tau_j \le t} \left[1 - \frac{m_j}{Y(\tau_j) - n_j}\right]$$

It is worth of note as, unlikely the overall CI (6), which involves the hazard and the survival function of the same random variable (T_f), the CCI (3) involves again $\widehat{S}_{T_f}$, but now the hazard is $\widehat{h}_r$, which is defined in terms of (T_f, δ) and does not match the definition of hazard function of any random variable (Boracchi et al., 2005). Thus, $\widehat{F}_r$ cannot be thought as the generalisation of the (6), and equivalent expressions with a structure of the (7) and of the complement to one of (4) cannot be directly derived.

3 $\widehat{F}_r$ AS A HORVITZ-THOMPSON LIKE ESTIMATOR

Let us consider first the case of absence of censoring. The observed data would be (t_{fi}, ε_i) (for $i = 1, ..., N$). An estimate of F_r (1) equivalent to the (3), is the empirical sub-distribution function

$$\widehat{F}_r(t) = \frac{1}{N}\sum_{i=1}^{N} I\{t_{fi} \le t; \varepsilon_i = r\} = \frac{1}{N}\sum_{j:\tau_j \le t} n_{jr} \tag{8}$$

Considered as random variable for each t, $\widehat{F}_r(t)$ is an average of N terms IID to the random variable I$\{T_f \le t; \varepsilon = r\}$, which follows a Bernoulli distribution of parameter $F_r(t)$. The equality between (8) and (3) follows observing as from (2) we can write $Y(\tau_k) - n_k = \sum_{l=k+1}^{J} n_l = Y(\tau_{k+1})$. Thus, for t such as $\tau_{j-1} \le t < \tau_j$, $\widehat{S}_{T_f}$ (4) becomes

$$\widehat{S}_{T_f}(t) = \prod_{k:\tau_k \le t} \frac{Y(\tau_{k+1})}{Y(\tau_k)} = \frac{Y(\tau_2)}{N} \cdot \frac{Y(\tau_3)}{Y(\tau_2)} \cdots \frac{Y(\tau_j)}{Y(\tau_{j-1})} = \frac{Y(\tau_j)}{N}$$

Finally, substituting this result in (3), follows the (8). In the presence of right censoring, similarly to (Jewell et al, 2006) we can write

$$\widehat{F}_r(t) = \frac{1}{N}\sum_{i=1}^{N} \frac{I\{t_i \le t; d_i \cdot \varepsilon_i = r\}}{\widehat{S}_{T_c}(t_{i-})} = \frac{1}{N}\sum_{j:\tau_j \le t} \frac{n_{jr}}{\widehat{S}_{T_c}(\tau_{j-1})} \tag{9}$$

which is a weighted average of N terms IID to the random variable $\mathrm{I}\{T_f \leq t; \varepsilon = r\}$ and weighted inversely by $\widehat{S}_{T_c}(t_{i-})$. The equality between (9) and (3) can be proved observing as both are right continuous step functions with possible jumps at τ_j. At τ_j, the jump of $\widehat{F}_r$ (3) is $\widehat{F}_r(\tau_j) - \widehat{F}_r(\tau_{j-1}) = n_{jr} \cdot \widehat{S}_{T_f}(\tau_{j-1})/Y(\tau_j)$. The corresponding jump of $\widehat{F}_r$ (9) is $\widehat{F}_r(\tau_j) - \widehat{F}_r(\tau_{j-1}) = n_{jr}/[N \cdot \widehat{S}_{T_c}(\tau_{j-1})]$. Thus, the equality between the functions holds if and only if

$$\widehat{S}_{T_f}(\tau_{j-1}) \cdot \widehat{S}_{T_c}(\tau_{j-1}) = \frac{Y(\tau_j)}{N} \tag{10}$$

for any j. Now, starting from the left hand of (10), we can write

$$\begin{aligned} \widehat{S}_{T_f}(\tau_{j-1}) \cdot \widehat{S}_{T_c}(\tau_{j-1}) &= \prod_{k=1}^{j-1} \left[1 - \frac{n_k}{Y(\tau_k)}\right] \cdot \left[1 - \frac{m_k}{Y(\tau_k) - n_k}\right] \\ &= \prod_{k=1}^{j-1} \frac{Y(\tau_k) - (n_k + m_k)}{Y(\tau_k)} \end{aligned} \tag{11}$$

Finally, from (2) $Y(\tau_k) - (n_k + m_k) = \sum_{l=k+1}^{J} (m_l + n_l) = Y(\tau_{k+1})$ and substituting this result in (11), it follows

$$\widehat{S}_{T_f}(\tau_{j-1}) \cdot \widehat{S}_{T_c}(\tau_{j-1}) = \prod_{k=1}^{j-1} \frac{Y(\tau_{k+1})}{Y(\tau_k)} = \frac{Y(\tau_2)}{Y(\tau_1)} \cdot \frac{Y(\tau_3)}{Y(\tau_2)} \cdots \frac{Y(\tau_j)}{Y(\tau_{j-1})} = \frac{Y(\tau_j)}{N}$$

4 $\widehat{F_r}(t)$ AS A KAPLAN-MEIER LIKE ESTIMATOR

The CCI (1) can be thought as the cumulative incidence of the artificial random variable T^* (where the subscript r was omitted for sake of writing) having support $R^+ \cup \{+\infty\}$ (Gray, 1988), defined from $(T_f; \varepsilon)$

$$T^* = T_f \cdot I\{\varepsilon = r\} + \infty \cdot I\{\varepsilon \neq r\} \tag{12}$$

For $t \in R^+$, $F_{T^*}(t) = \Pr\{T^* \leq t\} = \Pr\{T_f \leq t; \varepsilon = r\} = F_r(t)$. The hazard function of T^* (sub-distribution hazard), for $t \in R^+$, is

$$h_{T^*}(t) = \lim_{\Delta t \to 0^+} \Pr\{t < T^* \leq t + \Delta t | T^* > t\} / \Delta t$$

which, under random censoring, can be equivalently written as

$$h_{T^*}(t) = \lim_{\Delta t \to 0^+} \cdot \Pr\{t < T^* \leq t + \Delta t | T^* > t; T_c > t\} / \Delta t$$

Using the definition (12), h_{T^*} can also be written in terms of the observable random variables (T, δ, $\delta \cdot \varepsilon$) and of T_c (which in general is observable only for actually censored observations) as

$$= \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \cdot \Pr \{ t < T \leq t; \delta \cdot \varepsilon = r | T > t \cup (T \leq t; \delta \cdot \varepsilon \neq r; T_c > t) \} \quad (13)$$

Let us consider now a sample (t_i, d_i, $e_i \cdot d_i$) $(i = 1, ..., N)$. Before deriving a suitable estimator of $h_{T^*}(\tau_j)$, let us observe as the number of subjects at risk of failure due to any cause at τ_j (2) can be also written as

$$Y(\tau_j) = N \cdot \widehat{S}_{T_f}(\tau_{j-1}) \cdot \widehat{S}_{T_c}(\tau_{j-1}) \quad (14)$$

where $\widehat{S}_{T_f}(\tau_{j-1}) \cdot \widehat{S}_{T_c}(\tau_{j-1})$ is the estimate of the probability of the conditional event in (5) (the equality between (2) and (14) follows from (10). Now, we can argue similarly to derive the number of subjects at risk according to the conditional event in (5). The probability of the latter is

$$S_{T_f}(t) \cdot S_{T_c}(t) + \sum_{s \neq r} F_s(t) \cdot S_{T_c}(t)$$

thus, number of subjects to at risk at τ_j is

$$Y^*(\tau_j) = N \cdot \left[\widehat{S}_{T_f}(\tau_{j-1}) \cdot \widehat{S}_{T_c}(\tau_{j-1}) + \sum_{s \neq r} \widehat{F}_s(\tau_{j-1}) \cdot \widehat{S}_{T_c}(\tau_{j-1}) \right] \quad (15)$$

From (14) and (9), $Y^*(\tau_j)$ becomes

$$Y^*(\tau_j) = Y(\tau_j) + \sum_{i=1}^{N} I \{ t_i \leq \tau_j; d_i \cdot e_i \neq r, 0 \} \cdot w_{t_i}(\tau_j) \quad (16)$$

where $w_{t_i}(\tau_j) = \widehat{S}_{T_c}(\tau_{j-1}) / \widehat{S}_{T_c}(\tau_{i-})$. Thus, in addition to $Y(\tau_j)$, $Y^*(\tau_j)$ includes also the subjects who failed before τ_j for an event $v \neq r$, weighted by $w_{t_i}(\tau_j)$. The latter is an estimate of $\Pr \{ T_{ci} \geq \tau_j | T_{ci} \geq t_i \} = S_{T_c}(\tau_{j-1}) / S_{T_c}(t_{i-})$, which for $t_i \leq \tau_j$ is less or equal than one and decreases the less is t_i. Let us observe as (16) is equal to the number of subjects considered at risk in the estimating equation of the Fine and Gray's (1999) regression model.

Moreover, $Y^*(\tau_j)$ can be also written expanding $Y(\tau_j)$ as

$$Y^*(\tau_j) = Y(\tau_j) \cdot \left[\frac{1 - \widehat{F}_r(\tau_{j-1})}{\widehat{S}_{T_f}(\tau_{j-1})} \right] \quad (17)$$

The multiplier of $Y(\tau_j)$ is greater or equal than one, being from (7) and (9)

$$1 - \widehat{F}_r(\tau_{j-1}) = \widehat{S}_{T_f}(\tau_{j-1}) + \sum_{s \neq r} \widehat{F}_s(\tau_{j-1}) > \widehat{S}_{T_f}(\tau_{j-1}) \tag{18}$$

The greater is the competing action of events $v \neq r$ (within τ_{j-1}), i.e. the greater is $\sum_{s \neq r} \widehat{F}_s(\tau_{j-1})$ in (18), the greater is the multiplier of $Y(\tau_j)$. To prove the equality between (16) and (17), starting from (15), and observing as from (10) $\widehat{S}_{T_c}(\tau_{j-1}) = Y(\tau_j)/[N \cdot \widehat{S}_{T_f}(\tau_{j-1})]$, we can write

$$Y^*(\tau_j) = \left[\widehat{S}_{T_f}(\tau_{j-1}) + \sum_{s \neq r} \widehat{F}_s(\tau_{j-1})\right] \cdot \frac{Y(\tau_j)}{\widehat{S}_{T_f}(\tau_{j-1})}$$

and from (18) the (17) follows. Let us observe as (17) is equal to the number of subjects considered at risk in the statistic for the comparison of crude cumulative incidence curves (Gray, 1988).

An estimate of F_r (1) , equivalent to (3), having the structure of the KM estimator can be obtained generalising the (4) to T^* , by

$$\widehat{F}_r(t) = 1 - \prod_{j:\tau_j \leq t} \left[1 - \widehat{h}_{T^*}(\tau_j)\right] \tag{19}$$

where $\widehat{h}_{T^*}(\tau_j) = n_{jr}/Y^*(\tau_j)$ is the estimate of the sub-distribution hazard function (13) in $t = \tau_j$. The equality between the step functions (19) and (3) holds true if

$$\sum_{k=1}^{j} \widehat{h}_r(\tau_k) \cdot \widehat{S}_{T_f}(\tau_{k-1}) = 1 - \prod_{k=1}^{j} \left[1 - \widehat{h}_{T^*}(\tau_k)\right] \tag{20}$$

for any j, which can be proved by induction. For $j = 1$, $\widehat{h}_{T^*}(\tau_1) = \widehat{h}_{T_f}(\tau_1)$ and $\widehat{S}_{T_f}(\tau_0) = 1$, thus (20) is verified. Now, assuming that (20) holds true for a given j, this implies (20) holds true for $j+1$, in fact

$$\sum_{k=1}^{j+1} \widehat{h}_r(\tau_k) \cdot \widehat{S}_{T_f}(\tau_{k-1}) = \sum_{k=1}^{j} \widehat{h}_r(\tau_k) \cdot \widehat{S}_{T_f}(\tau_{k-1}) + \widehat{h}_r(\tau_{j+1}) \cdot \widehat{S}_{T_f}(\tau_j)$$

and using the hypothesis

$$= 1 - \prod_{k=1}^{j} \left[1 - \widehat{h}_{T^*}(\tau_k)\right] + \widehat{h}_r(\tau_{j+1}) \cdot \widehat{S}_{T_f}(\tau_j)$$

Now, observing as $\widehat{h}_r(\tau_{j+1}) = \widehat{h}_{T^*}(\tau_{j+1}) \cdot Y^*(\tau_j)/Y(\tau_j)$ and using (17), it follows

$$= 1 - \prod_{k=1}^{j} \Big[1 - \widehat{h}_{T^*}(\tau_k)\Big] + \widehat{h}_{T^*}(\tau_{j+1}) \cdot \Big[1 - \widehat{F}_r(\tau_j)\Big]$$

finally, using again the hypothesis

$$= 1-\prod_{k=1}^{j} \Big[1 - \widehat{h}_{T^*}(\tau_k)\Big]+\widehat{h}_{T^*}(\tau_{j+1})\cdot\prod_{k=1}^{j} \Big[1 - \widehat{h}_{T^*}(\tau_k)\Big] = 1-\prod_{k=1}^{j+1} \Big[1 - \widehat{h}_{T^*}(\tau_k)\Big]$$

5 DISCUSSION

The KP estimator of the CCI is a fundamental tool when dealing with survival data in the presence of competing risks. As a part of the estimation process, it invlolves the KM estimator of the overall survival/CI. Both estimators are usually introduced as nonparametric maximum likelihood estimators.

The KM estimator is the product of conditional survival probabilities, which are directly obtained from the nonparametric estimate of the underlying overall hazard. KM curves, derived from two or more groups of subjects, are usually accompanied by the result of a nonparametric hypothesis testing on their equality. The log rank test, and more generally the Harrington and Fleming's family of tests (Harrington and Fleming, 1982), considering weighted averages of hazards, are based on the equality of the underlying hazards, and indeed rely on the one-to-one correspondence holding between hazard functions and survival/cumulative incidence functions. A natural link between estimation and the hypothesis testing procedures is that the overall hazard common to the groups under the null hypothesis, is estimated employing the same nonparametric estimator of the hazard used to derive the curves. The KM estimator can be alternatively expressed as an inverse-probability-of-censoring weighted average having the form of a Horvitz-Thompson estimator of the distribution function of the time to failure, leading to a convenient form for the asymptotic theory (Satten and Datta, 2001). The estimators of the CCI and of the overall CI are intrinsecaly linked as the summation of the CCI estimates equals the estimate of the overall CI, in a coherent way with the corresponding population quantities. Moreover, both CI and CCI estimators can be written as sum of estimated unconditional probabilities. In the first case, they are the product of the overall survival by the overall hazard. This leads, indeed, to a third equivalent form for the KM estimator of the overall CI, which is derived eventually in terms of the overall hazard. In the case of the CCI, the unconditional probability involved is the product of the overall survival by the cause-specific hazard. However, in this case, the CCI estimator cannot be written only in terms of the cause-specific hazard as the overall survival involves the overall hazard, which in turn, depends on the cause specific hazards of all causes. As a consequence, the KM procedure cannot be employed to derive CCI estimates substituing the cause-specific hazard in place of the overall hazard (Gooley et al., 1999; Satagopan et. al, 2004). Nonetheless, an alternative expression of the CCI estimator as an inverse-probability-of-censoring weighted average having the form of a Horvitz-Thompson estimator can still be derived as showed in section 3.

Concerning the comparison among CCI curves, it has to be pointed out as the equality among CCI curves do not necessarily imply the equality between the corresponding overall survivals and cause specific hazards (Gray, 1988). Gray's family of tests, in fact, relies on the equality among the groups of the sub-distribution hazard, which being the hazard of a fictitious improper random variable having the CCI as cumulative incidence guarantees a one-to-one correspondence between the CCIs and sub-distribution hazards. To facilitate the applied statistician in understanding the link between non-parametric estimation and testing of the CCI, we proved in section 4 as the KP estimator of the CCI can be expressed in terms of the sub-distribution hazard present in Gray's tests obtaining a KM like estimator. The sub-distribution hazard was estimated observing as the number of subjects at risk when estimating the overall hazard can be thought as the expected number of sampling subjects free from failure and from censoring. This concept was generalised to the case of the sub-distribution hazard, considering the expected number of subjects free from any event or who developed an event different from the one interest, and free from censoring by the time point considered. This leads to an estimate of the sub-distribution hazard equal the one employed in Gray's tests to compute the sub-distribution hazard common to the groups under the null hypothesis. The Horvitz-Thompson form of the CCI estimator was employed to prove the alternative KM form in section 4. Further work is needed to generalise the asymptotic results obtained using the Horvitz-Thompson form in the absence of competing risk to the case of the presence competing risks.

References

[1] Boracchi, P., Antolini, L., Biganzoli, E., Marubini, E., (in press) Competing

Risks: Modeling Crude Cumulative Incidence Functions, *Italian Journal of Applied Statistics*

Elandt-Johnson, R., Johnson, N. (1980), *Survival Models and Data Analysis.* New York: Wiley.

Fine, J. P., and Gray, R. J. (1999), A Proportional Hazards Model for the Subdistribution of a Competing Risk, *Journal of the American Statistical Association*, 94, 496-509.

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

Gray, R. J. (1988), A class of K-sample tests for comparing the cumulative incidence of a competing risk, *The Annals of Statistics*, 16, 1140-1154.

Harrington, D. P., Fleming, T. R. (1982), A class of rank test procedures for censored survival data, *Biometrika*, 69, 553-566.

Horvitz, D. G., Thompson, D. J. (1952), A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, 47, 663-685.

Jewell, N P., Lei, X., Ghani, A. C., Donnelly, C. A., Leung, G. M, Ho, L. M., Cowling, B., Hedley, A. J. (2006), Estimation of the Case Fatality Ratio with Competing Risks Data: An Application to Severe Acute Respiratory Syndome (SARS), *to appear on Statistics in Medicine.*

Kalbfleisch, J. D., Prentice, R. L. (1980), *The statistical analysis of failure time data*, New York: Wiley.

Kaplan, E. L., Meier, P. (1958), Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, 53, 457-481.

Satagopan, J. M., Ben-Porat, L., Berwick, M., Robson, M., Kutler, D., Auerbach, A. D. (2004), A note on competing risks in survival data analysis, *Breast Journal Cancer*, 91, 1229-35.

Satten, G. A., Datta, S. (2001), The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average, *The American Statistician*, 55, 207-210.