Computers in Biology and Medicine 101 (2018) 70–81

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# Focus on an infrequently used quantity in the context of competing risks: The conditional probability function

Bastien Cabarrou[a], Florence Dalenc[a], Eve Leconte[b], Jean-Marie Boher[c,d], Thomas Filleron[a,*]

[a] Institut Claudius Regaud, IUCT-Oncopole, Toulouse, France
[b] TSE Université Toulouse 1 Capitole, Toulouse, France
[c] Institut Paoli Calmettes, Biostatistics Unit, Marseille, France
[d] Aix Marseille University, INSERM, IRD, SESSTIM, Marseille, France

ARTICLE INFO

*Keywords:*
Competing risks
Conditional probability
Cumulative incidence
Regression model
Metastatic breast cancer

ABSTRACT

In clinical studies of hematologic and oncologic diseases, the outcomes of interest are generally composite time to event endpoints which are usually defined by occurrence of different event types. Nonetheless, clinicians are interested in studying only one event type, which leads to a competing risks situation. In this context, Pepe and Mori presented a quantity directly derived from the cumulative incidence: the conditional probability. This function defines the probability that a given event occurs, conditionally on not having had a competing event by that time. The objective of this paper is to present this conditional cumulative incidence function and to compare its use to the cumulative incidence in different data sets. Different scenarios highlight the importance of the competing event on the interpretation of the conditional probability. Conditional probability needs to be interpreted jointly with the cumulative incidence. This quantity can be of interest especially when the risk of the competing event is large, strongly precludes the risk of the event of interest and provides useful additional information.

## 1. Introduction

In clinical studies of hematologic and oncologic diseases, the outcomes of interest are generally composite time to event endpoints: overall survival according to cause of death, disease free survival according to type of failure. These composite endpoints are usually defined by the occurrence of different event types. For example, overall survival is related to death from cancer and death from others causes. Nonetheless, most clinicians want to focus on a single event as for example relapse, non-relapse mortality (death before relapse), acute Graft-versus-Host Disease (aGvHD) or Chronic Graft-versus-Host Disease (cGvHD). In this competing risks situation, the occurrence of an event can preclude the appearance of other events, for example non-relapse mortality precludes the appearance of relapse. Different methods can be used to analyse competing risks data. Until recently, the most commonly used approach consisted in analysing separately each event type. Competing events were thus considered as censored at the time of occurrence of the first competing event and the "net probability" was estimated by one minus the Kaplan-Meier estimate [1]. This quantity corresponds to the probability of the occurrence of the event of interest before a given time in the hypothetical situation where this event would be the only cause of failure acting on the population. This approach, named in the literature the "censor method" [2], assumes the non-informative nature of the censoring. If this is not the case, the *censor* method provides an overestimate of the cumulative incidence function and can lead to misinterpretation of results [2]. A more appropriate method for the analysis of competing risks data, named the "include method" [3], has received considerable attention. This alternative approach takes into account the informative nature of the censoring, without the need for an unrealistic hypothesis of independence between failure types and the "crude probability" or cumulative incidence function was estimated according to Kalbfleisch and Prentice estimator. This quantity corresponds to the cumulative probability of occurrence by a given time for a particular type of event in the presence of competing events. To get a full understanding of competing risks, these quantities should be viewed simultaneously for all possible events. Differences between these two methods of estimation have been largely discussed in the medical literature [4–7] as in methodological papers [2,8]. Different guidelines and recommendations were also provided to analyse competing risks data [7,9–12]. In fact, the development and applications in the framework of competing risks are on the rise [13]. There has been much work published on methods for

* Corresponding author. Institut Claudius Regaud and IUCT-Oncopole, 1 avenue Irène Joliot-Curie, 31059, Toulouse Cedex 9, France.
*E-mail address:* filleron.thomas@iuct-oncopole.fr (T. Filleron).

https://doi.org/10.1016/j.compbiomed.2018.08.002
Received 7 May 2018; Received in revised form 2 August 2018; Accepted 2 August 2018

inference and modelling of competing risks data [14–16] and statistical routines written in R and SAS are available to apply this methodology [17–20]. In this context, Pepe and Mori presented a quantity directly derived from the cumulative incidence function and based on conditional probability [21]. This function defines the probability that a given event occurs, conditionally on not having had a competing event by that time. Different publications have used this function in the context of bone marrow transplantation (BMT) [22–24], but nevertheless it remains infrequently used. Recently, Allignol et al. developed a regression model for the conditional probability [25].

The goal of this paper is not to provide one more debate on the differences between crude and net probabilities, but to present the conditional cumulative incidence function and to compare its use to the cumulative incidence in different data sets. In section 2, notations and methodology associated to competing risks and estimation of these two functions will be briefly presented. In section 3, the methods are illustrated and compared on examples in Haematology, Metastatic Breast cancer and Head and Neck Mucosal Melanoma (HNMM). In section 4, different scenarios will be studied in order to illustrate the behavior of these two functions in different situations. Discussion takes place in section 5.

## 2. Methodology

This section presents a brief description of the methodology of competing risks. After presentation of the notations, three functions of interest are presented.

### 2.1. Notations - Definitions

In order to simplify the notations, we consider only two competing risks in the setting of BMT: Relapse denoted with subscript *R* and Transplantation Related Mortality with subscript *TRM*. Let $t_1 < t_2 < ... < t_J$ denote the ordered observed times for which at least one event occurs, in considering for each patient only the first event that occurs among all types of events. Let $n_j$ define the number of patients at risk just before $n_j$ and $d_j$ the number of all event types which occur at time $n_j$. $d_{Rj}$ and $d_{TRMj}$ represent respectively the number of R and TRM which occur at time $n_j$ ($d_j = d_{Rj} + d_{TRMj}$).

### 2.2. Functions of interest and estimation

In this section, we present the different functions of interest.

#### 2.2.1. Disease free survival

The disease free survival (DFS) is defined as the probability of being alive free of disease at a given point in time. Thus, death or disease relapse are treated as events, and patients alive and free of disease at their last follow-up are censored. Disease free survival is estimated by the Kaplan-Meier estimator:

$$\hat{S}(t) = \sum_{j:t_j \le t} \left[1 - \frac{d_j}{n_j}\right]$$

#### 2.2.2. Cumulative incidence of relapse: Relapse incidence

Relapse Incidence (RI), denoted by $\hat{I}_R(t)$, is defined as the probability of having had a relapse before time t, death without experiencing a relapse being considered as a competing event. As TRM and R are exclusive events, i.e. occurrence of R can preclude occurrence of TRM, the correct method of estimation of this function is based on the include method [26]. In a first step, the DFS rates were estimated using Kaplan-Meier method. In a second step, the cause-specific hazard of relapse was estimated by

**Table 1**
Summary of available macros/packages in standard softwares.

| Software | R | STATA (ado) | SAS |
|---|---|---|---|
| **Cumulative Incidence** | | | |
| Non Parametric | | | |
| Estimation | cuminc[a] survfit[b] prodlim[c] | stcompet[h] | %cif[k] |
| Comparison | | | |
| Gray | cuminc[a] | | %cif[k] |
| WP | compCIF[d] | stpepemor[i] | %cif[k] compcif[d] |
| Fine & Gray model | crr[a] | stcrreg[j] | %pshreg[l] |
| Test of proportionality | PSH.test[e] | | |
| Selection of covariates | crrstep[f] | | |
| **Conditional Probability** | | | |
| Non Parametric | | | |
| Estimation | cpf[g] | | cuminc[d] |
| Comparison WCP | cpf[g] compCP[d] | stpepemori[i] | compcp[d] |
| Regression models | cpfpo[g] pseudocpf[g] | | |

[a] Package cmprsk [18].
[b] Package survival [41].
[c] Package prodlim [42].
[d] http://www.uhnres.utoronto.ca/labs/hill/People_Pintilie.htm [43].
[e] Package crrSC [44].
[f] Package crrstep [45].
[g] Package Cprob [35].
[h] Ado [46].
[i] Ado [47].
[j] [48].
[k] http://support.sas.com/resources/papers/proceedings12/344-2012.pdf [49].
[l] http://cemsiis.meduniwien.ac.at/kb/wf/software/statistische-software/pshreg/ [50].

$$\hat{h}_R(t_j) = \frac{d_{Rj}}{n_j}$$

where $n_j$ is the number of patients who are alive without relapse just before $n_j$ and $d_{Rj}$ is the number of relapses at $n_j$.

Finally, the relapse incidence or "crude probability", i.e. the probability of a relapse occurring before time *t*, is estimated by:

$$\hat{I}_R(t) = \Pr[\text{Re}\,lapse\ \ by\ \ t] = \sum_{t_j | t_j \le t} \hat{S}(t_{j-1}) \times \hat{h}_R(t_j)$$

where $\hat{S}(t_{j-1})$ is the estimate of the DFS estimate just before time $t_j$ with $\hat{S}(t_0) = 1$.

Similar reasoning has to be applied to obtain the cumulative incidence of TRM. The complement of the DFS and the cumulative incidence of R and TRM are linked by the formula:

$$\hat{I}_R(t) + \hat{I}_{TRM}(t) = 1 - \hat{S}(t)$$

Different tests were proposed in the statistical literature to compare cumulative incidence between groups [21,27,28]; the most commonly used is the Gray test. Fine & Gray proposed to model the subdistribution hazard of the cumulative incidence [29]. The subdistribution hazard of event $k$, denoted by $\gamma_k$, can be interpreted as the hazard for an individual who either fails from cause $k$ or does not, and in the latter case has an infinite failure time for cause $k$ [10]. A semi-parametric proportional model was proposed for the subdistribution hazard of the cumulative incidence:

$$\gamma_k(t|x) = \gamma_{k0}(t)\exp(\beta_k{}'x)$$

with $\gamma_{k0}$ is the baseline subdistribution hazard of cause $k$, and is the vector of coefficients for the covariates $x$. The term $\exp(\beta_k{}'x)$ is named subHazard Ratio (*sHR*). This quantity can be interpreted as follows: