Description of survival data extended to the case of competing risks: a teaching approach based on frequency tables

DAVIDE PAOLO BERNASCONI[1], LAURA ANTOLINI[2]

Abstract

Survival analysis is a powerful statistical tool to study failure-time data. In introductory courses students learn how to describe right-censored survival time data using the product-limit estimator of the survival function on a given *end-point* relying on a product of conditional survival probabilities. In the case of a composite *end-point*, the next step is to account for the presence of competing risks. The complement to one of the survival function is decomposed into the sum of cause-specific incidences, which are obtained as sum of unconditional probabilities due to the single competing risk. However, this algebraic decomposition is not straightforward, given the difference between the structures of the involved estimators. In addition, one is tempted to use the Kaplan-Meier estimator, leading to an erroneous decomposition of the overall incidence. Here we discuss a simple reinterpretation of the Kaplan-Meier formula in terms of sum of non-conditional probabilities of developing the *end-point* in time, adjusted for the presence of censoring. This approach could be used for describing survival data through simple frequency tables which are directly generalized to the case of competing risks. In addition, it makes clear how the estimation of the single cause-specific incidence through the Kaplan-Meier estimator, simply considering the occurrence of competing events as censored data, leads to an overestimation of the cause-specific incidence. Two examples are provided to support the explanation: the first one could help to clarify the procedure described by the formulas; the second one simulates real data in order to present graphically the results.

Key words: Crude incidence; Non-parametric estimator; Cause of failure

---

1 INTRODUCTION

Survival analysis deals with the behavior of variables representing time T elapsed between an initial time and the occurrence of an event of interest (*end-point*). Examples in medicine include the time elapsed between the diagnosis of a disease and the patient death, the beginning of exposure to a pollutant and the onset of a disease, the start of a treatment and the treatment

---
[1] *Correspondig Author*, Department of Health Sciences, University of Milano-Bicocca, Via Cadore 48, 20900 Monza, Italy. *e-mail*: davide.bernasconi@unimib.it. *Tel*: +39 02 64488098
[2] Department of Health Sciences, University of Milano-Bicocca, Via Cadore 48, 20900 Monza, Italy. *e-mail*: laura.antolini@unimib.it

failure [1], [2]. Survival analysis is helpful even in the industrial context, where it is referred to as "reliability analysis": for example, considering a particular equipment, could be of interests to study the time elapsed from its first use to the occurrence of the first failure. In practical applications it often happens that the *end-point* is composite, because it is defined as the occurrence of the first between two or more different types of event. In this case an analysis that accounts for the presence of competing risks is usually recommended [3]. A typical example in the clinical context is the occurrence of the first event between relapse of disease and death due to severe toxicity in the absence of relapse. Consequences will be different, both for the patient and for clinical interpretation, depending on whether the achievement of the composite *end-point* was due to relapse or death.

Basics on competing risks are required even in introductory courses in survival analysis since with the improvement of care, applications in medicine move from hard *end-points* (such as death) to *end-points* including non-fatal events. In our experience, one of the difficulties a teacher encounters in handling non parametric estimation tools for competing risks description is that the Aalen-Johansen estimator of cause-specific incidences differs from the Kaplan-Meier estimator of the overall incidence in terms of algebraic form [4], [5]. This makes more difficult to understand that the sum of the cause-specific probabilities leads exactly to the overall incidence and encourages a typical mistake, also seen in the application literature, that arise when using the Kaplan-Meier formula: to treat competing risks as cause of censoring. Indeed this leads to an overestimation of the crude incidence function.

Here we discuss a simple reinterpretation of the Kaplan-Meier formula in terms of sum of non-conditional probabilities of developing the *end-point* in time, adjusted for the presence of censoring. This approach is aimed to help students to describe survival data through simple frequency tables and is directly generalized to the case of competing risks.

We consider, for the sake of simplicity, the case of a discrete failure time random variable. The paper is organized as follows: in Section 2, we explain, with the help of a tutorial example, the general approach to analyze survival data through the Kaplan-Meier method. We review an alternative formulation of the Kaplan-Meier estimator defined in an iterative way as sum of non-conditional probabilities [6]. This helps in understanding the analysis of competing risks data using the Aalen-Johansen estimator of the cause-specific incidence presented in Section 3 where also the improper use of the Kaplan-Meier method is discussed. In Section 4 we give a graphical presentation of our results, using a simulated example that mimics real data. We conclude with a discussion in Section 5.

2   NON PARAMETRIC ESTIMATION OF THE INCIDENCE FUNCTION: THE KAPLAN-MEIER METHOD

The key quantities to describe the random variable survival time T are the functions of time t:

- incidence: $F(t) = P(T \leq t)$, proportion of subjects for whom the *end-point* has occurred within time t;

- survival: $S(t) = 1 - F(t)$, proportion of subjects event free at time t;

- hazard: $h(t) = P(T = t \mid T \geq t)$, instantaneous probability that the *end-point* occurs at time t conditional on being event free right before t.

The observed data are realizations of the pair $Y = min(T, C)$ and $\Delta = I(T \leq C)$ where C is the censoring time and Δ the *end-point* indicator. The random variables T and C are assumed

independent. On a sample of N subjects the observed data are $\{y_i; \delta_i\}$ with $i = 1, ..., N$. With respect to the observed discrete times $t_1 < t_2 < ... < t_j < ...$, where the *end-point* was observed on at least one subject, data can be organized by the counts of subjects at risk $n_j = \sum_{i=1}^{N} I(y_i \geq t_j)$ and of subjects reaching the *end-point* at time t_j, $d_j = \sum_{i=1}^{N} I(y_i = t_j)\delta_i$.

The hazard function, which can be thought as the "velocity" of occurrence of the event of interest, increases in time the proportion of subjects $F(t)$ reaching the *end-point* and, conversely, reduces the proportion of event free subjects $S(t)$. This relationship is described by the recursive equation

$$F(0) = 0 \quad F(t) = F(t-) + P(T = t) = F(t-) + S(t-)h(t) \tag{1}$$

where $t-$ is the discrete time right before t. Let us note that moving in time from $t-$ to t, the incidence (1) increases of a quantity equal to $S(t-)h(t)$ because the hazard $h(t)$ acts on the proportion of subjects $S(t-)$ event free at $t-$.

Let us consider a simple example about the time (measured in months) elapsed between the start of an oncologic treatment and the achievement of the *end-point*, defined as relapse of disease or death for severe toxicity in the absence of relapse. In a sample of $N = 4$ subjects with observed survival times equal to: 10, 20, 30, 40, the estimated incidence function is equal to the ratio between the number of subjects who have reached the *end-point* by time t and the total of 4 subjects considered: $\hat{F}(10) = \frac{1}{4}, \hat{F}(20) = \frac{1}{2}, \hat{F}(30) = \frac{3}{4}, \hat{F}(40) = 1$. In the absence of censoring, as in this example, it is easy to notice that the incidence function at time t_j can be thought as the cumulative relative frequency of the *end-point* at time t_j, arguing as follows. The frequency of the subjects reaching the *end-point* at $t_{j'}$ is $\frac{d_{j'}}{N}$, thus the cumulative frequency at t_j is $F(t_j) = \sum_{j' \leq j} \frac{d_{j'}}{N}$. The incidence function can also be estimated, obtaining the same values, using the recursive formula (1), as it is shown in Table 1. It can be noticed that:

1. the estimates of the risk function $\hat{h}(t_j)$ are obtained as a ratio between the number of subjects d_j reaching the *end-point* at time t_j and the number n_j of event free subjects at t_j;

2. the relative frequency of the subjects reaching the *end-point* at t_j, $\hat{P}(T = t_j)$, is calculated by $\hat{S}(t_{j-1})\hat{h}(t_j)$ which turns to be equal to $\frac{d_j}{N}$;

3. the estimate of the incidence function $\hat{F}(t)$ is the cumulative of $\hat{S}(t-)\hat{h}(t)$.

Table 1. Incidence function estimated by the iterative method (1)

j	t_j	n_j	d_j	$\hat{h}(t_j)$	$\hat{S}(t_{j-1})\hat{h}(t_j)$	$\hat{F}(t_j)$
1	10	4	1	1/4	(1-0)*1/4=1/4	0+1/4=1/4
2	20	3	1	1/3	(1-1/4)*1/3=1/4	1/4+1/4=1/2
3	30	2	1	1/2	(1-1/2)*1/2=1/4	1/2+1/4=3/4
4	40	1	1	1/1	(1-3/4)*1/1=1/4	3/4+1/4=1/4

The approach we used in this simple example with no censoring is known as Kaplan-Meier method [5]. Although in (1) the Kaplan-Meier estimator is defined as a sum of non-conditional

probabilities, it is usually presented by the formula based on product of conditional probabilities:

$$F(t) = 1 - \prod_{u \leq t}(1 - h(u)) \qquad (2)$$

where u are discrete times, as in [7]. Formulas (1) and (2) are equivalent. This can be proved by showing that the difference between the value of the function at two contiguous time points, $F(t) - F(t-)$, according to both (1) and (2), is equal to $S(t-)h(t)$.

The Kaplan-Meier method allows to include also the contribution of censored data. Let us assume we have another subject, whose survival time is known to be greater than 30 (for example the last control visit was performed at 30 months after the beginning of the treatment administration). This subject will give a contribution to the hazard estimation $h(t)$ only at times less or equal to 30. Specifically he/she will give a unitary contribution to the denominator and no contribution to the numerator. The estimating process for this case is shown in Table 2.

Table 2. Incidence function estimated by the Kaplan Meier method, using formulation (1), in presence of censored data. The superscript ^{C} indicates the presence of a censored data

j	t_j	n_j	d_j	$\hat{h}(t_j)$	$\hat{S}(t_{j-1})\hat{h}(t_j)$	$\hat{F}(t_j)$
1	10	5	1	1/5	(1-0)*1/5=1/5	0+1/5=1/5
2	20	4	1	1/4	(1-1/5)*1/4=1/5	1/5+1/5=2/5
3	30^C	3	1	1/3	(1-2/5)*1/3=1/5	2/5+1/5=3/5
4	40	1	1	1/1	(1-3/5)*1/1=2/5	3/5+2/5=1

At time $t_3 = 30$ there are $n_3 = 3$ subjects at risk, 1 subject reaching the *end-point* and 1 censored subject, thus at time $t_4 = 40$

1. the estimate of the risk function $\hat{h}(40)$ is obtained as a ratio between the number of subjects $d_4 = 1$ reaching the *end-point* and the number of event free subjects $n_4 = 1$ which is equal to $n_3 = 3$ eroded by both the subject reaching the *end-point* and the subject censored at $t_3 = 30$;

2. the relative frequency of the subjects reaching the *end-point* at t_4, $\hat{P}(T = t_4)$, is calculated by $\hat{S}(t_3)\hat{h}(t_4) = \frac{2}{5}$ which now is different from $\frac{d_4}{N} = \frac{1}{5}$ due to the presence of censoring.

The estimate of the incidence function $\hat{F}(t)$ is again the cumulative of $\hat{S}(t-)\hat{h}(t)$.

3 NON PARAMETRIC ESTIMATION OF THE CAUSE-SPECIFIC INCIDENCE OF A COMPETING RISK

3.1 The Aalen–Johansen estimator

When we need to handle a composite *end-point*, defined as the occurrence of the first between one or more events, we could be interested in splitting the *end-point* into two (or more) components by determining the cause-specific quantities. It becomes useful to distinguish between the causes of the *end-point* development, especially in order to interpret the impact of treatment. To this

purpose we introduce, in addition to the survival time T, another indicator ε, where $\varepsilon = 1$ if T is the time to occurrence of event 1 (e.g. relapse of the disease), and $\varepsilon = 2$ if T is the time to occurrence of event 2 (e.g. toxic death). The incidence and hazard functions are defined depending on the event (cause) occurred, as follows:

- cause-specific incidence due to cause 1 (relapse): $F_1(t) = P(T \leq t; \varepsilon = 1)$, proportion of subjects for reached the composite *end-point* before time t due to cause 1;

- cause-specific hazard: $h_1(t) = P(T = t; \varepsilon = 1 \mid T \geq t)$, instantaneous probability to develop the composite *end-point* due to cause 1 at time t in subjects free from the composite *end-point* (i.e. free from both relapse and death) right before t.

Similarly for death we can define

- $F_2(t) = P(T \leq t; \varepsilon = 2)$ and $h_2(t) = P(T = t; \varepsilon = 2 \mid T \geq t)$.

Notice that $F_1(t) + F_2(t) = F(t)$ and $h_1(t) + h_2(t) = h(t)$, so the cause-specific quantities are the decomposition of the incidence and the hazard of the composite *end-point* in the two components due to relapse and death.

The relationship between the cause-specific incidence (also known as crude incidence) and the cause-specific hazard functions is described by the recursive equation [1]

$$F_1(0) = 0 \quad F_1(t) = F_1(t-) + P(T = t; \varepsilon = 1) = F_1(t-) + S(t-)h_1(t) \tag{3}$$

where $t-$ is the discrete time right before t. It is worth to note that (3) involves both the cause-specific hazards due to relapse $h_1(t)$ and death $h_2(t)$. In fact $S(t-)$ is derived from (1) and depends on the overall hazard $h(t) = h_1(t) + h_2(t)$. Moving in time from $t-$ to t the cause-specific incidence increases by the quantity $S(t-)h_1(t)$ due to the action of the hazard of relapse $h_1(t)$ on the proportion of subjects $S(t-)$ free from the composite *end-point* at time $t-$. Let us note that experiencing death as the first event implies having no relapse as first event and thus an apparent protection against relapse. This can be seen considering that, by (3), an increase in "velocity" of the onset of death $h_2(t)$ will lead to an increased hazard of composite *end-point* $h(t)$ and a consequent decline in the proportion of subjects $S(t-)$ on which $h_1(t)$ can act. Let us consider again the example on the $N = 4$ subjects with no censoring, assuming that the first and the third time are due to relapse while the second and the fourth are due to death. We can still estimate the cause-specific incidence function as the ratio between the number of subjects reaching the composite *end-point* due to the event considered by time t_j and the total of 4 subjects: $\hat{F}_1(10) = \hat{F}_1(20) = \frac{1}{4}, \hat{F}_1(30) = \hat{F}_1(40) = \frac{1}{2}, \hat{F}_2(10) =, \hat{F}_2(20) = \hat{F}_2(30) = \frac{1}{4}, \hat{F}_2(40) = \frac{1}{2}$. The cause-specific incidence function can also be estimated by the iterative formula (3), as shown in Table 3.

The example describes the situation where the *end-point* is composite and no censoring occurs. Thus, Table 3 can be thought as the extension of Table 1 to the case of competing risks, and analogous considerations can be done:

1. the estimates of the cause-specific risk function $\hat{h}_1(t_j)$ are obtained as a ratio between the number of subjects d_{1j} reaching the *end-point* 1 (relapse) at time t_j and the number n_j of subjects free from both events at time t_j; the same applies to $\hat{h}_2(t_j)$;

2. the relative frequency of the subjects reaching the *end-point* due to cause 1 at t_j, $\hat{P}(T = t_j; \varepsilon = 1)$, is calculated by $\hat{S}(t_{j-1})\hat{h}_1(t_j)$ which turns to be equal to $\frac{d_{1j}}{N}$; analogous calculations are valid for *endpoint* 2;

3. again, the estimates of the two cause-specific incidence functions $F_1(t)$ and $F_2(t)$ are the cumulative of $\hat{S}(t-)\hat{h}_1(t)$ and $\hat{S}(t-)\hat{h}_2(t)$ respectively.

Table 3. Cause-specific estimated incidence functions by the iterative method (3).

j	t_j	n_j	d_{1j}	d_{2j}	$\hat{h}_1(t_j)$	$\hat{h}_2(t_j)$	$\hat{S}(t_{j-1})\hat{h}_1(t_j)$	$\hat{S}(t_{j-1})\hat{h}_2(t_j)$	$\hat{F}_1(t_j)$	$\hat{F}_2(t_j)$	$\hat{F}(t_j)$
1	10	4	1	0	1/4	0	(1-0)*1/4=1/4	(1-0)*0=0	1/4	0	1/4
2	20	3	0	1	0	1/3	(1-1/4)*0=0	(1-1/4)*1/3=1/4	1/4	1/4	1/2
3	30	2	1	0	1/2	0	(1-1/2)*1/2=1/4	(1-1/2)*0=0	1/2	1/4	3/4
4	40	1	0	1	0	1/1	(1-3/4)*0=0	(1-3/4)*1/1=1/4	1/2	1/2	1

When censored data are present, the Aalen-Johansen estimator (3) enables to take this into account [1]. The method relies on this principle: if we consider a new subject with censored survival time equal to 30, he/she will give a contribution to the hazard estimation only for times less or equal to 30. Specifically, he/she will give a unitary contribution to the denominator and no contribution to the numerator. The estimating process for this case is shown in Table 4.

Table 4. Cause-specific incidence estimated function by the iterative method (3) in presence of censored data (Aalen-Johansen method). The superscript ^C indicates the presence of a censored data.

j	t_j	n_j	d_{1j}	d_{2j}	$\hat{h}_1(t_j)$	$\hat{h}_2(t_j)$	$\hat{S}(t_{j-1})\hat{h}_1(t_j)$	$\hat{S}(t_{j-1})\hat{h}_2(t_j)$	$\hat{F}_1(t_j)$	$\hat{F}_2(t_j)$	$\hat{F}(t_j)$
1	10	5	1	0	1/5	0	(1-0)*1/5=1/5	(1-0)*0=0	1/5	0	1/5
2	20	4	0	1	0	1/4	(1-1/5)*0=0	(1-1/5)*1/4=1/5	1/5	1/5	2/5
3	30^C	3	1	0	1/3	0	(1-2/5)*1/3=1/5	(1-2/5)*0=0	2/5	1/5	3/5
4	40	1	0	1	0	1/1	(1-3/5)*0=0	(1-3/5)*1/1=2/5	2/5	3/5	1

The approach here described is the correct way to deal with survival time data when censoring and competing risks are present. Thus, Table 4 represents the extension of Table 2 to the case of competing risks.
At time $t_3 = 30$ there are $n_3 = 3$ subjects at risk, 1 subject reaching the *end-point* due to cause 1 and 1 censored subject, thus at time $t_4 = 40$

1. the estimate of the cause-specific risk function $\hat{h}_2(40)$ is obtained as a ratio between the number of subjects $d_{2;4} = 1$ reaching the *end-point* due to cause 2 and the number of subjects $n_4 = 1$ free from both events, which is equal to $n_3 = 3$ eroded by both the subject reaching the composite *end-point* and the subject censored at $t_3 = 30$;

2. the relative frequency of the subjects reaching the *end-point* due to cause 2 at t_4, $\hat{P}(T = t_4; \varepsilon = 2)$, is calculated by $\hat{S}(t_3)\hat{h}_2(t_4) = \frac{2}{5}$ which now is different from $\frac{d_{2;4}}{N} = \frac{1}{5}$ due to the presence of censoring.

The estimate of the two cause-specific incidence functions $\hat{F}_1(t)$ and $\hat{F}_2(t)$ are the cumulative of $\hat{S}(t-)\hat{h}_1(t)$ and $\hat{S}(t-)\hat{h}_2(t)$, respectively.

3.2 The improper use of the Kaplan-Meier estimator

It is important to point out that the cause-specific estimates of the incidence described in section 3.1 cannot be obtained by the Kaplan-Meier method (2) simply considering the occurrence of the event different from that of interest as censored. This approach would lead to an overestimation of the cause-specific incidence as shown in Table 5 [8].

Table 5. Overestimation of the cause-specific incidence functions by the Kaplan-Meier method (4) considering times relative to the competing event as censored.

j	t_j	n_j	d_{1j}	d_{2j}	$\hat{h}_1(t_j)$	$\hat{h}_2(t_j)$	$\hat{F}_{1KM}(t_j)$	$\hat{F}_{2KM}(t_j)$
1	10	4	1	0	1/4	0	0+(1-0)*1/4=1/4	0+(1-0)*0=0
2	20	3	0	1	0	1/3	1/4+(1-1/4)*0=1/4	0+(1-0)*1/3=1/3
3	30	2	1	0	1/2	0	1/4+(1-1/4)*1/2=5/8	1/3+(1-1/3)*0=1/3
4	40	1	0	1	0	1/1	5/8+(1-5/8)*0=5/8	1/3+(1-1/3)*1/1=1

Let us note that for every time t the estimates $\hat{F}_{1KM}(t_j)$ and $\hat{F}_{2KM}(t_j)$ in Table 5 are greater or equal to the corresponding estimates $\hat{F}_1(t_j)$ and $\hat{F}_2(t_j)$ shown in Table 4.
An intuitive explanation of the cause of this overestimation can be given as follows. Let us suppose we want to calculate the cause-specific incidence function at time t for relapse, $\hat{F}_1(t_j)$, using equation (2). Since (2) is equivalent to (1) (Section 2), the cause-specific incidence function, according to the Kaplan-Meier method, would be:

$$F_1(0) = 0 \quad F_1(t) = F_1(t-) + S_1(t-)h_1(t) \qquad (4)$$

where $S_1(t-) = 1 - F_1(t-)$. Let us observe that the velocity of occurrence of event 1, represented by the cause-specific hazard $h_1(t)$, would cause a proportion of event-free people $S_1(t)$ greater than the true $S(t)$, which is eroded not only by $h_1(t)$ but also by $h_2(t)$ since $h(t) = h_1(t) + h_2(t)$. As a consequence, the cause-specific incidence function at time next to t, $F_1(t+)$, will be overestimated. This, in other words, means that $S_1(t-)h_1(t) \geq S(t-)h_1(t) = P(T = t; \varepsilon = 1)$.

4 GRAPHICAL EXAMPLE

Here we consider a sample of simulated data that mimics a randomized study comparing a "standard" against an "experimental" treatment, the latter associated with a lower incidence of failure, defined as occurrence of relapse or toxic death. Figure 1 shows the estimated incidence function, according to the treatment group, obtained by the cumulative relative frequency (1). Let us note that, for example, after 50 weeks from the start of treatment, the cumulative relative frequency of subjects reaching the *end-point* is about 40% in the experimental group and about 60% in the standard one.

The incidence of the composite *end-point* shown in Figure 1 can be split into two cause-specific components using (3): one due to relapse and the other due to death (Figure 2).

Figure 1. Example of incidence function in a simulated randomized study.
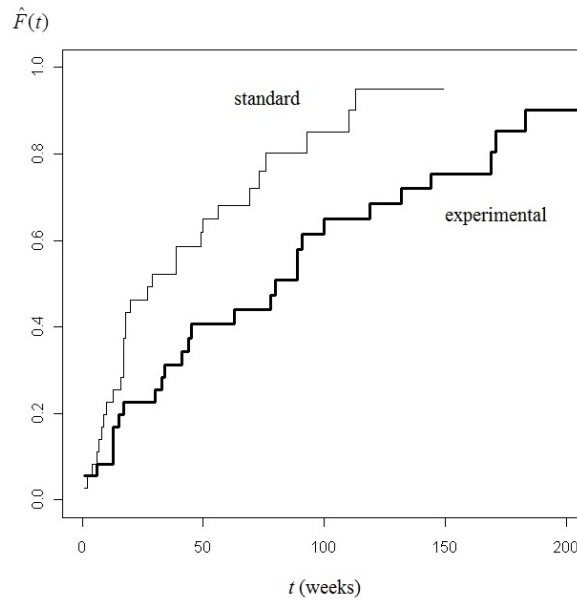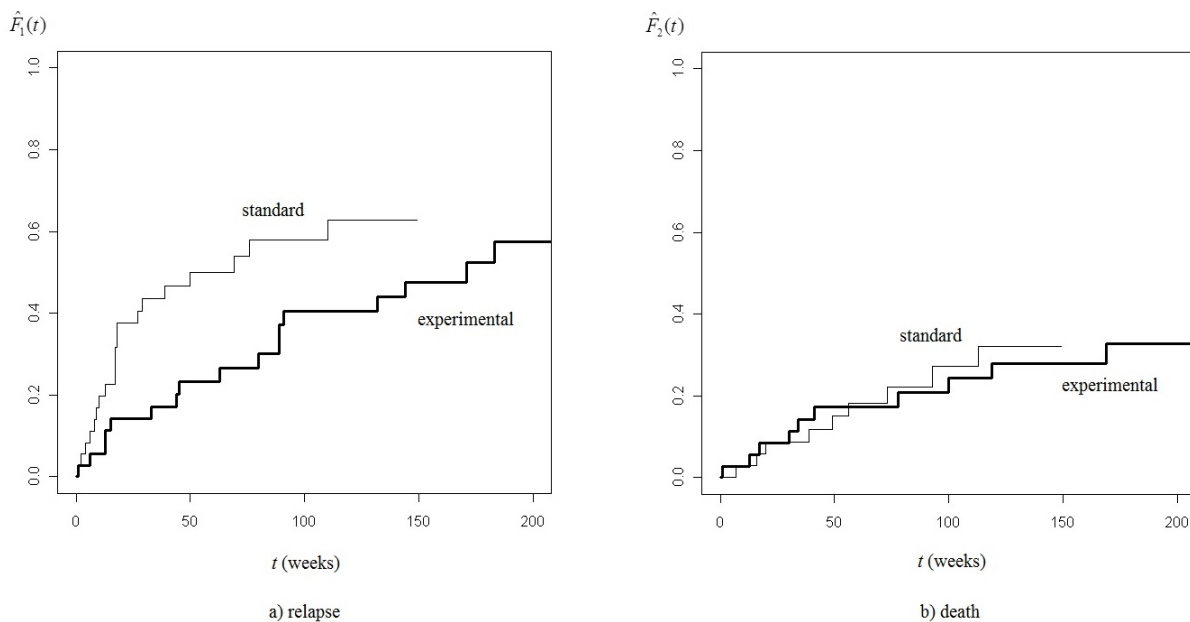

Figure 2. Cause-specific incidences in the two groups, standard and experimental treatment.


a) relapse

b) death

In each treatment group, at each t the sum of the values laying on the y-axis in figures 2a and 2b equals the y-axis values of the corresponding group in Figure 1. This is consistent with the relationship between the incidence of the composite *end-point* and the cause-specific incidences $F_1(t) + F_2(t) = F(t)$ in the population quantities.

The decomposition of the total incidence (Figure 1) into the cause-specific incidences (Figure 2) shows that the gain provided by the experimental treatment reducing the occurrence of the com-

posite *end-point* compared with standard treatment is mainly due to a reduction of occurrence of relapse (Figure 2a) and not of death (Figure 2b). Figure 3 shows the overestimation of $F_1(t)$ obtained by the Kaplan-Meier method, censoring the competing event, in the two treatment groups.

Figure 3. Cause-specific incidence of relapse obtained by the Aalen-Johansen estimator (solid line) and by the Kaplan-Meier estimator (dashed line).
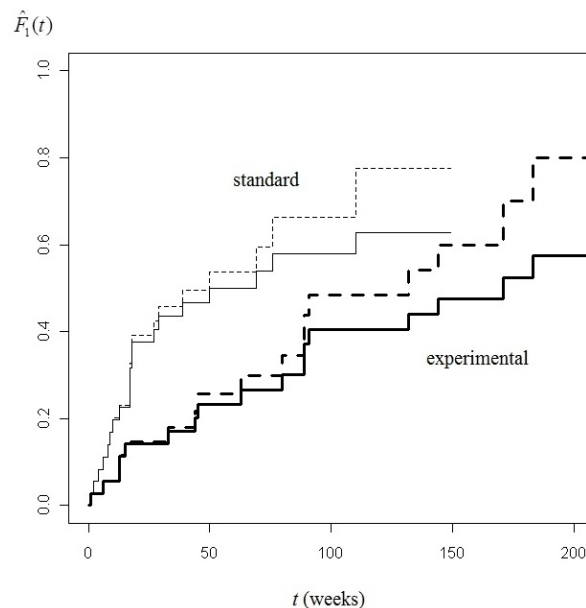
5   CONCLUSION

We have provided a reinterpretation of the Kaplan-Meier incidence as the cumulative relative frequency of event in time, where the frequency of event at each time point is adjusted for the presence of censoring. This can be helpful for students to move from simple analysis of frequency tables to the description of survival data.

This approach can be of interest from a pedagogical perspective also in the more complex scenario involving competing risks where the Kaplan-Meier incidence, usually presented as a product of conditional probabilities, is decomposed into the cause-specific components due to the causes of failure using the Aalen-Johansen estimator. The latter, which is a sum of unconditional probabilities, is coherent with the expression of the Kaplan-Meier incidence written as a cumulative relative frequency.

In this context, the estimator of the overall incidence of the composite *end-point* (1) and the cause-specific incidence (3) are written in the same fashion. The only difference is the type of hazard (overall vs cause-specific) applied to the proportion of subjects $S(t-)$ free from the composite *end-point*. From this similarity it follows directly, i.e. with no need of algebraic proof, that the equality between the incidence of the composite end-point and the sum of the cause-specific incidences holds also for the estimates. A further advantage of this approach is that it makes clear how the Kaplan-Meier incidence (written as product of conditional probabilities) using the cause-specific hazard instead of the overall hazard leads to an overestimation of the cause-specific incidence.

References

[1] Kalbfleisch JD and Prentice RL. The statistical analysis of failure time data 2nd edition. New York: Wiley, 2002.

[2] Lau B, Cole SR, and Gange SJ. Competing risk regression models for epidemiologic data. *American Journal of Epidemiology* 2009;170:244-256.

[3] Putter H, Fiocco M and Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* 2007;26(11):2389-2430.

[4] Aalen O and Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* 1978;5:141-150.

[5] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958;53:457-481.

[6] Antolini L. Kaplan-Meier estimation of the cumulative incidence of a competing risk. *Biomedical Statistics and Clinical Epidemiology* 2009;3:129-134.

[7] Marubini E and Valsecchi MG. Analysing survival data from clinical trials and observational studies. John Wiley & Sons: Chichester, 1995.

[8] Gooley TA, Leisenring W, Crowley J and Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine* 1999;18:695-706.