Hacettepe Journal of Mathematics and Statistics
Volume 45 (4) (2016), 1161 – 1173

Kaplan-Meier estimator in competing risk contexts

Pablo Martínez-Camblor* †‡

Abstract

Survival analysis has become in a common procedure in biomedical researches. Conventionally, the well-known nonparametric Kaplan-Meier (KM) estimator is used in order to approximate the real survivor curve. However, in competing risk contexts where more than one failure cause compete to occur and only one of them is of interest, the direct use of the Kaplan-Meier statistic does not perform correctly and, in order to obtain a good estimation, it must be adapted. In this work, via Monte Carlo simulations, the author explores the behavior of the Kaplan-Meier estimator in a competing risk context. In addition, differences between KM and multiple decrement methods are pointed out. Finally, a real-data problem is used in order to illustrate the situation.

Keywords: Competing risks, Kaplan-Meier estimator, Multiple decrement, Survival Analysis.

2000 AMS Classification: 62N01

Received : 11.02.2015 *Accepted* : 13.06.2015 *Doi* : 10.15672/HJMS.20157111086

1. Introduction

Conventionally, survival analysis is devoted to the study of data where the response of interest is the time required for certain (studied) event, which inevitably happens, to occur. Main particularities of these studies are: *i)* on one hand, the distribution of time is often strongly asymmetric and usual parametric models based on the normal law do not perform adequately and, *ii)* the researcher frequently does not have a complete knowledge on the time to event for each subject included in the study; he/she knows that the event does not occur in a period of time but he/she does not know how long the event is needing to occur. These situations are frequently repeated in the nature; perhaps the bio-sanitary (the study of time to death in patients with some particular disease) is one of the most known fields. Of course, there exists a vast literature about

*Hospital Universitario de Asturias (HUCA), Email: pablomc@ficyt.es

†Universidad Autonoma de Chile, Chile

‡Pablo Martínez-Camblor. Dpto. Estadística e IO y DM, Facultad de Ciencias, c/ Calvo Sotelo s/n, 33005 Oviedo, Asturies, Spain.

statistical survival methods, among all, we want to remark the indispensable monograph of Kalbfleish and Prentice (2002).

Let T be the non-negative random variable representing the failure time of interest, as is well-known, in survival analysis, there are mainly three different ways to specify its distribution (see, for instance, Kalbfleish and Prentice (2002)): the survivor function, the probability density function, and the hazard function. The survivor function stands for the probability that the event occurs after a fixed value of time, t, that is,

$$S(t) = \mathcal{P}\{T > t\}, \qquad 0 \leq t < \infty. \tag{1.1}$$

Note that if F denotes the standard cumulative distribution function (CDF) for the random variable T, $S(t) = 1 - F(t)$ $(0 \leq t < \infty)$. Directly, when T is an absolutely continuous variable, the probability density function (PDF) is defined in the standard form,

$$f(t) = d[1 - S(t)]/dt = dF(t)/dt, \qquad 0 \leq t < \infty. \tag{1.2}$$

Obviously, it holds $S(t) = \int_t^\infty f(u)du$. Finally, the hazard function stands for the rate of that the event occurs instantaneously after the time t when it is known that it does not happen before t; that is,

$$\begin{aligned} \lambda(t) &= \lim_{h \to 0^+} \mathcal{P}\{T < t + h | T \geq t\}/h \\ &= f(t)/S(t) = -d \log(S(t))/dt, \qquad 0 \leq t < \infty. \end{aligned} \tag{1.3}$$

Integrating with respect to t and taking into account that $S(0) = 1$, it holds the equality

$$S(t) = \exp\left\{-\int_0^t \lambda(u)du\right\} = \exp\{-\Lambda(t)\}, \qquad 0 \leq t < \infty. \tag{1.4}$$

where $\Lambda(t) = \int_0^t \lambda(u)du$ is known as the *cumulative hazard* function. Standard analysis of survival data usually includes the non-parametric Kaplan-Meier (KM) estimator (Kaplan and Meier (1958) for the survivor curve estimation and the semi-parametric proportional hazard Cox regression (Cox (1972)) in order to explore possible covariate effects.

Under the usual assumption of independence between time to event and censoring time, the KM estimator has really good properties (in the Section 2, some properties of the KM estimator are pointed out); in addition, it has a direct and simple probabilistic interpretation. However, when the studied event not necessarily happens; i.e., there exists one (or more) event which is incompatible with the studied one, the KM estimator overestimated the probability that the event happens. In practice, these situations are really frequent; for instance, when the studied variable is the time to recurrence of some disease; obviously, death without recurrence makes not possible the disease relapses or, when the researcher is interested in the time to death by a particular cause; the death for other cause is, logically, not compatible with the considered event. In this work, the author explores the survival curve estimation in the *competing risk* setting. Particularly, the advantages of using the multiple decrement (MD) estimator (Aalen (1978)) are investigated via Monte Carlo simulations (Section 4). From a real problem dataset, in Section 5, the differences between the KM and the MD estimators are pointed out; particularly, the distribution of the time-free of leukemia in patients with myelodysplasia is analyzed. Finally, in Section 6, the author presents his conclusions.

2. The Kaplan-Meier estimator

The well-known Kaplan-Meier or product-limit estimator was proposed in 1958 in one of the most (or the most, depending on the consulted source) cited and popular statistical paper (Kaplan and Meier (1958)). In that work, the authors proposed a

non-parametric method for the estimation of the cumulative distribution function from incomplete observations. The standard mathematical formulation is as follows: let $\boldsymbol{T} = \{T_1, \ldots, T_N\}$ be the times to event and let $\boldsymbol{C} = \{C_1, \ldots, C_N\}$ be the censor times, let F and G be the CDFs for the time to event and the censor time, respectively. The observed times are $\boldsymbol{Z} = \{Z_1, \ldots, Z_N\}$ where $Z_j = \min\{T_j, C_j\}$ $(1 \leq i \leq N)$. In addition, it is also known what time is really observed; i.e., the final available information are the pairs $\{(Z_1, \delta_1), \ldots, (Z_N, \delta_N)\}$, where $\delta_j = I_{T_j}(Z_j)$ (takes the value 1 if the time to event is observed and 0 otherwise). Then, the KM estimator for the survivor function is defined by

$$\hat{S}_N(t) = \prod_{j=1}^{N} \left\{ 1 - \frac{\delta_{(j)} \cdot I_{(-\infty,t]}(Z_{(j)})}{N - j + 1} \right\}, \tag{2.1}$$

where for $j \in 1 \ldots, N$, the pairs $(Z_{(j)}, \delta_{(j)})$ satisfy that $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(N)}$. In this context, the Kaplan-Meier is the maximum-likelihood estimator. In addition, their properties have been deeply studied; asymptotic normality can be derived from the work of Csörgő (1996) in which, under usual and mild assumptions, the so-called *Hungarian embeddings* (Komlós, Major and Tusnády (1975)) and the law of the iterated logarithm are generalized to the random right censorship case. Although some alternative methods have been proposed (see, for instance Peto et al. (1975) or Simon and Lee (1982)), the variance of the KM estimator is usually approximated from the Greenwood's formula (Greenwood (1926)),

$$\mathbb{V}[\hat{S}_N(t)] = \sum_{j=1}^{N} \frac{\delta_{(j)} \cdot I_{(-\infty,t]}(Z_{(j)})}{N - \sum_{j=1}^{N} I_{(-\infty,t]}(Z_{(j)})}.$$

On the other hand, Bitouzé et al. (1999) provided a Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator; in particular, they established that there exists an absolute constant K such that,

$$\mathcal{P}\left\{ \sup_{t \in \mathbb{R}} |(1 - G(t)) \cdot (\hat{S}_N(t) - S(t))| > \lambda/\sqrt{n} \right\} \leq 2.5 \cdot e^{-2\lambda^2 + K\lambda},$$

for any positive value λ. Figure 1 depicts the Kaplan-Meier estimation joint with a 95% confidence band (computed using the Greenwood's formula), for the time to death (at left) and the time-free of leukemia (right) for the *Myelodysplastic* dataset. This dataset is from a retrospective study that included high-risk patients reported to the Spanish Group of Myelodysplastic Syndromes Registry (RESMD) between years 2000 and 2013. This data will be used in order to illustrate the considered problem (see Section 6). Anyway, interested readers are referred to Bernal et al. (2015) for additional information about this study. The dataset includes a total of 968 patients (1,273.7 persons-year), 616 of them died during the follow-up. Two-hundred sixty eight patients (27.7%) developed leukemia during the follow-up and 403 died without leukemia. In spite that, of course, these 403 patients are not going to develop leukemia anymore, they are considered as censored for the KM estimator; i.e., their weights are spread among the subjects who are still at risk.

The Kaplan-Meier estimator, like the traditional empirical estimator for the CDF, initially assigns to each sample point a weight of $1/N$ (N stands for the sample size). The main particularity is that, at the time that one subject is censored, KM assumes that its (future) behavior will be similar to the behavior of subjects who are still at risk; therefore, these subjects inherit the weight of the censored subject. Suppose that the minimum time $Z_{(1)}$ corresponds with an event, at this time KM produces a jump of $1/N$, the second time $Z_{(2)}$ is a censored subject; then, subjects who are still in the study

Figure 1. For the Myelodysplastic dataset: at left, Kaplan-Meier estimation for the time to death, at right, Kaplan-Meier estimation for the leukemia-free time. In both panels, 95% confidence bands are included (in gray).

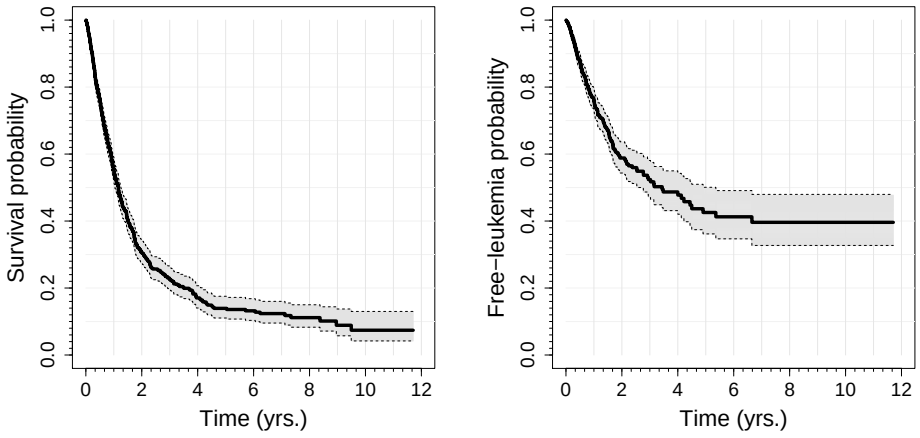


Table 1. Kaplan-Meier contruction for the case described in the manuscript: considered sorted sample is: $\{(Z_{(1)},1),(Z_{(2)},0),(Z_{(3)},1),\cdots,(Z_{(N)},\delta_{(N)})\}$.

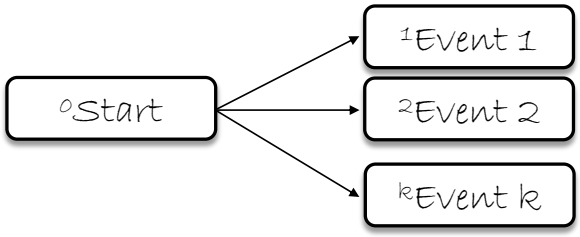
Time	**at risk**	$\boldsymbol{\delta}$	**survival**
0	N		1
$Z_{(1)}$	N	1	$1-[1/N]$
$Z_{(2)}$	$N-1$	0	$1-[1/N]$
$Z_{(3)}$	$N-2$	1	$1-[1/N]-[(1/N)+1/N(N-2)]$
		.....	

$(N-2)$ inherit its weight $(1/N)$; therefore the new weight of these $N-2$ subjects will be $1/N+1/N(N-2)$. Hence, if the third observed time, $Z_{(3)}$, is again an event, KM will produce, at time $Z_{(3)}$, a jump of $1/N+1/N(N-2)$. Table 1 depicts schematically the KM construction.

3. The competing risk context

There are many real situations in which the event of interest does not always occur; i.e., there exist other events, incompatible with the studied one, which can happen before. The study of the time to death for some particular cause; death for other causes makes not possible the studied event (see, for instance, Verduijn et al. (2011)), the study of the time-free of one particular disease; death for other causes makes impossible the relapse of the considered disease (Boo et al. (2015)), or the study of the transplanted organ

Figure 2. Usual competing risk schema. Transitions from the state ^{0}Start to k the different events are the quantities of interest.



survivor; death of patient does not permit the study of the organ failure (Martínez-Camblor et al. (2015)) are just a few examples of the so-called *competing risk* context. Of course, there exists a vast literature on this topic; see, for instance, Tiatsis (1998) and references therein and Andersen et al. (2002) for the multi-state models approach to competing risk, but our purpose is not to make a revision. Rather, we discuss the problem of the Kaplan-Meier estimator on this context. Figure 2 depicts the standard schema for the competing risk setting; $P_{0,i} = \mathcal{P}_{0,i}(t) = \mathcal{P}\{T_i \leq t\}$ where T_i is the time required to achieve the jth event, with $j \in 1, \ldots, k$ are the main quantities of interest.

In the competing risk contests, the sample must provide information about the observed time and on what event has been really observed. Therefore,
$Z_j = \min\{C_j, T_{1,j}, \ldots, T_{k,j}\}$ ($T_{i,j}$ $(1 \leq j \leq N)$ is the time that the subject j would need to achieve the event (state) i) and $\delta_j = i$ with $i \in 0, 1, \ldots, k$ stands for the observed event for the subject j (0 when no event has still happened, i.e., at the final of the study, the subject is still at risk; censored subjects). In order to study the distribution of the time to one particular event (for instance, the ith one, with i taking any value in $1, \ldots, k$), a frequent -and wrong- practice is to consider the rest of the events as censored and then, to estimate the distribution of interest from the Kaplan-Meier estimator. The main issues of this procedure are:

i) Although the independence assumption between the times to event and the time to censoring is plausible, usually, the times to the different events involved in a competing risk setting are strongly dependent. Notice that a patient died before having a relapse, is not going to relapse anymore; the censorship provides information about the considered even. This effect is known as informative censorship.

ii) Due to patients which experiencing a competing event, different to the studied one, are not going to achieve, directly, the event of interest anymore (they are not going to do the transition from ^{0}Start to the studied event), subjects which are still at risk; i.e., those which can still experimenting the event of interest, must not inherit their weights.

iii) In the standard survival analysis, the probability of survival and the probability of event are equivalent quantities ($1 = \mathcal{P}\{T > t\} + \mathcal{P}\{T \leq t\}$). In the competing risk context, there are more involved events and the fact that a subject does not suffer the studied event does not imply that this subject is free of events.

In this context, it holds the equality,

$$\begin{aligned} 1 =& \mathcal{P}\{T > t\} + \mathcal{P}\{T \leq t\} \\ =& \mathcal{P}\{T > t\} + \mathcal{P}\{T \leq t \wedge \delta = 1\} + \cdots + \mathcal{P}\{T \leq t \wedge \delta = k\}. \end{aligned} \tag{3.1}$$

The quantities $\mathcal{P}\{T \leq t \wedge \delta = i\}$ $(1 \leq i \leq k)$ are the *cumulative incidence functions*. However, Andersen, Abildstrom and Rosthoj (2002) claimed that: 'this is, in fact, a rather unfortunate name for this quantity as it may give the incorrect impression that it is a cumulative intensity'. Alternative proposed names are marginal or crude failure probabilities.

4. Multiple decrements method

The nonparametric Kaplan-Meier estimator can be adapted for the competing risk setting in the so-called multiple decrement (MD) method. The considered estimator for the *general transition probabilities* was proposed by Aalen (1978). However, and in spite that different papers have tried to popularize this procedure (see, for instance, Martínez-Camblor et al. (2009) and references therein) it is still little used by practitioners and it is unknown by the physicians. The MD procedure assumes that the probability that two different events occur simultaneously is zero (i.e., $\mathcal{P}\{T_i = T_l\} = 0$ for $1 \leq i \neq l \leq k$). From this proviso, $P_{0,l} = \mathcal{P}\{T \leq t, \delta = l\}$ (transition probability between the states 0 and l, $1 \leq l \leq k$) is equivalent to the probability that all the involved times were greater or equal to t and the studied one was exactly t, that is

$$P_{0,l} = \int_0^t S(u)\lambda_l(u)dt = \int_0^t S(u)d\Lambda_l(u), \tag{4.1}$$

where $\lambda_l(u)$ is the hazard function referred to event l. A direct plug-in method using the KM estimator for estimate $S(u)$, and the Nelson-Aalen estimator to estimate the cumulative incidence function, let us to obtain the MD estimator by

$$\mathrm{MD}_l(t) = \sum_{t_j \leq t} \frac{r_{l,j}}{N_j} \prod_{t_i \leq t} \left(1 - \frac{\sum_{l=1}^k r_{l,i}}{N_i}\right), \tag{4.2}$$

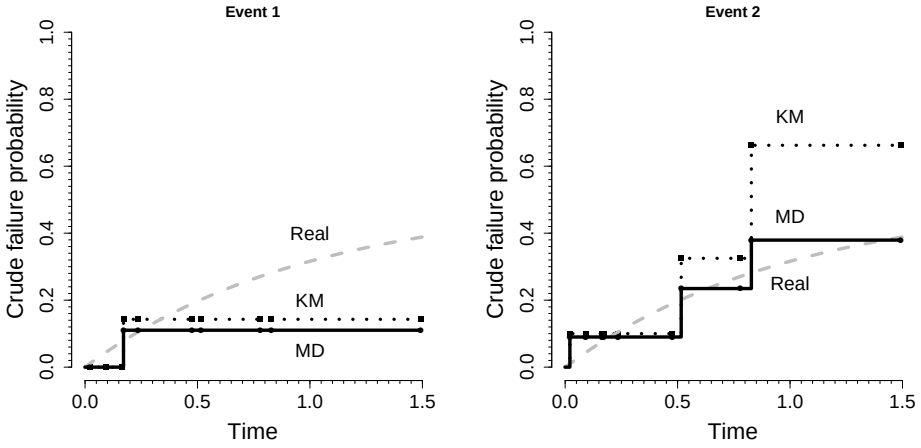
where $r_{l,j}$ $(1 \leq l \leq N)$ and N_j are the number of subjects which have suffered the event l and which were at risk just before of moment t_j $(1 \leq j \leq N)$, respectively. Of course, theoretical properties of the $\mathrm{MD}_l(\cdot)$ estimator have been deeply studied. In Aalen (1978) is proved its uniform consistency (with rate $\log(N) \cdot N^{-1/2}$) and its weak convergence to an adequate Gaussian process (with the usual rate $N^{-1/2}$). Recently, Njamen-Njomen and Ngatchou-Wandji (2014) developed adapted stochastic processes to the Nelson-Aalen and Kaplan-Meier estimators.

In order to illustrate the problem we simulate a three independent times from an exponential law (with mean 1): $T_{1,j}$, $T_{2,j}$ and C_j, in ten subjects $(1 \leq j \leq 10)$. We compute $Z_j = \min\{T_{1,j}, T_{2,j}, C_j\}$ and define $\delta_j = i$, where $i = 1$ if $Z_j = T_{1,j}$, $i = 2$ if $Z_j = T_{2,j}$ and $\delta_j = 0$ if $Z_j = C_j$ $(1 \leq j \leq 10)$. Table 2 depicts the computed estimations by using the KM and the MD methods for the events 1 and 2. Real values (for both events, they are the CDF of an exponential distribution with mean 1) are also reported. Note that in this case, all involved subdistributions are the same. Figure 3 depicts the curves. Since the KM considers censored all events different to the studied one, its 'jumps' are frequently bigger than the MD ones. Obviously, for a fixed point of time t, the MD estimator considers at risk only those subjects which at this time, have not suffer any event.

Table 2. Results for one simulation example of competing risk setting. Sample size was 10 and two different events were simulated ($\delta = 1, 2$; $\delta = 0$ stands for censored data). Direct Kaplan-Meier (KM) and its modification for the multiple decrement (MD). Real values are the same for both considered events.

	Subjects			KM		MD	
Time	at risk	δ	Real	1	2	1	2
0.021	10	2	0.010	0.000	0.100	0.000	0.090
0.091	9	0	0.043	0.000	0.100	0.000	0.090
0.164	8	0	0.076	0.000	0.100	0.000	0.090
0.171	7	1	0.079	0.143	0.100	0.110	0.090
0.235	6	0	0.105	0.143	0.100	0.110	0.090
0.476	5	0	0.189	0.143	0.100	0.110	0.090
0.516	4	2	0.202	0.143	0.325	0.110	0.234
0.779	3	0	0.271	0.143	0.325	0.110	0.234
0.828	2	2	0.281	0.143	0.662	0.110	0.379
1.492	1	0	0.388	0.143	0.662	0.110	0.379

Figure 3. Referred to the data shown in Table 2. At left, real (gray), KM and MD estimations for the event 1. At right, real (gray), KM and MD estimations for the event 2.



5. Monte Carlo simulation study

In order to study the behavior of the direct Kaplan-Meier (KM) and the Multiple decrement (MD) estimators on the competing risk setting, a Monte Carlo simulation study was carried out. The time of studied event, $T_1 = \exp\{D_1\}$, where D_1 was drawn from a normal distribution with mean μ (values of $-1/2$ and $1/2$ were considered) and variance one; the time to the competing risk event, $T_2 = \exp\{D_2\}$, with D_2 generated

Table 3. Mean ± standard deviation of the 1,000 Monte Carlo iterations for the quantity $100 \cdot \tau^{-1} \cdot \int_0^\tau |\hat{S}(t) - S(t)|dt$ where $S(t)$ is the real subdistribution function and $\hat{S}(t)$ its estimation based on KM and on MD estimators and τ is the maximum observed time for $\mu = -1/2$.

		$\rho = 0.0$		$\rho = 0.25$	
N	c	**KM**	**MD**	**KM**	**MD**
50	10	7.05 ± 2.39	2.10 ± 1.39	5.90 ± 2.26	1.79 ± 1.19
	-1/4	7.88 ± 3.72	3.85 ± 2.20	6.85 ± 3.40	3.55 ± 1.99
	-1/2	7.99 ± 3.87	4.19 ± 2.37	7.15 ± 3.67	4.11 ± 2.28
250	10	5.55 ± 1.09	0.62 ± 0.41	4.35 ± 0.96	0.48 ± 0.33
	-1/4	6.67 ± 1.52	1.23 ± 0.74	5.67 ± 1.42	1.15 ± 0.65
	-1/2	6.86 ± 1.85	1.46 ± 0.90	5.76 ± 1.73	1.37 ± 0.79
1000	10	4.60 ± 0.81	0.23 ± 0.17	3.47 ± 0.72	0.16 ± 0.11
	-1/4	6.12 ± 0.91	0.48 ± 0.29	5.01 ± 0.81	0.45 ± 0.26
	-1/2	6.27 ± 0.97	0.57 ± 0.33	5.24 ± 0.90	0.54 ± 0.31
		$\rho = 0.75$		$\rho = -0.50$	
50	10	3.17 ± 1.44	1.17 ± 0.81	9.84 ± 2.90	3.12 ± 1.85
	-1/4	4.69 ± 2.67	3.17 ± 1.81	10.03 ± 3.99	4.43 ± 2.48
	-1/2	5.44 ± 3.17	3.72 ± 1.98	10.16 ± 4.47	5.03 ± 2.71
250	10	2.11 ± 0.62	0.29 ± 0.21	8.79 ± 1.14	1.05 ± 0.68
	-1/4	3.24 ± 1.08	1.00 ± 0.59	9.23 ± 1.70	1.53 ± 0.91
	-1/2	3.52 ± 1.31	1.24 ± 0.66	9.36 ± 2.04	1.82 ± 1.03
1000	10	1.55 ± 0.38	0.10 ± 0.07	8.12 ± 0.78	0.42 ± 0.27
	-1/4	2.91 ± 0.59	0.39 ± 0.22	8.99 ± 0.94	0.66 ± 0.38
	-1/2	2.96 ± 0.69	0.49 ± 0.26	8.98 ± 0.99	0.76 ± 0.44

from a standard normal distribution and $\mathbb{E}[D_1 \cdot D_2] = \rho$ (values of 0, 1/4, 1/2 and 3/4 were considered). Finally, the censoring time, $C = exp\{N\}$, where N was drawn, independently, from a normal distribution with mean c (values of 10, -1/4 and -1/2 were considered) and variance one. The (simulated) observed data were the pairs (Z, δ) where $Z = \min\{C, T_1, T_2\}$ and $\delta = i$ ($i = 0$ if $Z = C$, $i = 1$ if $Z = T_1$, and $i = 2$ if $Z = T_2$). Mean ± standard deviation of the average error, $100 \cdot \tau^{-1} \int_0^\tau |\hat{S}(t) - S(t)|dt$ with $\tau = \max_{1 \leq j \leq N} Z_j$ based on 1,000 Monte Carlo iterations are reported (N stands for the sample size, $S(t)$ denotes the real subdistribution function and $\hat{S}(t)$ its estimation).

Table 3 depicts the observed results when $\mu = -1/2$. In this case, the probability that the considered event happens is: $\mathcal{P}\{T_1 < T_2\} = 0.638$, 0.658, 0.761 and 0.611 for $\rho = 0$, 1/4, 3/4 and $-1/2$, respectively. The expected censorship percentages were 0% ($c = 10$); 32.6%, 34.5%, 40.1% and 28.3% ($c = -1/4$) for $\rho = 0$, 1/4, 3/4 and $-1/2$, respectively; and 39.7%, 41.9%, 47.2% and 37.5% ($c = -1/2$) for $\rho = 0$, 1/4, 3/4 and $-1/2$, respectively. The MD method clearly obtained better results than KM.

Table 4 shows the coverage percentages and mean±sd (standard deviations below 0.00 were denoted by 0.01) of the length of the 95% symmetric confidence intervals (computed by using the naive bootstrap method) for the subdistribution function at times $t = 1/2$ and $t = 1$ using the KM and MD estimators. Observed results endorses the previous obstained ones: KM is not an estimator for the subdistribution function, especially, for larger censorship percentages. The DM estimator works adequately although it shows

Table 4. Coverage percentages and mean±sd (standard deviations below 0.00 were denoted by 0.01) of the length of the 95% symmetric confidence intervals (computed by using the naive bootstrap method with 200 iterations) for the subdistribution function at times $t = 1/2$ and $t = 1$ using the KM and MD estimators when $\mu = -1/2$.

$\rho = 0.0$		$t = 1/2$				$t = 1$			
N	c	KM		MD		KM		MD	
50	10	87.3%	0.289 ± 0.02	91.6%	0.237 ± 0.02	55.0%	0.312 ± 0.04	87.3%	0.234 ± 0.02
	-1/2	90.8%	0.348 ± 0.04	91.4%	0.265 ± 0.02	75.4%	0.452 ± 0.10	85.1%	0.266 ± 0.02
1000	10	20.0%	0.065 ± 0.01	92.9%	0.059 ± 0.01	5.1%	0.070 ± 0.01	93.6%	0.061 ± 0.01
	-1/2	37.1%	0.077 ± 0.01	96.0%	0.069 ± 0.01	3.2%	0.105 ± 0.01	94.9%	0.080 ± 0.01
$\rho = 0.75$		$t = 1/2$				$t = 1$			
50	10	92.4%	0.275 ± 0.02	92.8%	0.238 ± 0.02	77.4%	0.285 ± 0.03	91.2%	0.233 ± 0.02
	-1/2	91.9%	0.325 ± 0.03	92.3%	0.266 ± 0.02	83.3%	0.426 ± 0.08	89.5%	0.869 ± 0.03
1000	10	71.2%	0.062 ± 0.01	93.9%	0.059 ± 0.01	1.4%	0.065 ± 0.01	93.0%	0.961 ± 0.01
	-1/2	76.6%	0.073 ± 0.01	95.1%	0.069 ± 0.01	1.7%	0.095 ± 0.01	93.5%	0.083 ± 0.01

itself a little bit unconservative for the largest censorship percentage ($c = -1/2$ and $t = 1$).

Table 5 is similar to Table 3 for $\mu = 1/2$. In this case, the probability that the considered event happens is: $\mathcal{P}\{T_1 < T_2\} = 0.361$, 0.341, 0.239 and 0.387 for $\rho = 0$, 1/4, 3/4 and $-1/2$, respectively. The expected censorship percentages were 0% (for $c = 10$); for $c = -1/4$, approximately 47.2%, 49.3%, 54.3% and 43.5% for $\rho = 0$, 1/4, 3/4 and $-1/2$, respectively; and for $c = -1/2$, 54.8%, 56.7%, 61.2% and 51.5% for $\rho = 0$, 1/4, 3/4 and $-1/2$, respectively. The observed results were similar to the ones observed in the Table 3. Notice that, due to, in this case, the effect of the competing event was higher ($\mathcal{P}\{T_1 < T_2\} < 1/2$), the difference between the MD and the KM methods was bigger.

Finally, Table 6 is similar to Table 4 when $\mu = 1/2$. Although the KM estimator obtained better results, observed results are in the same way to the previous one and endorse the conclusions.

6. Real-world problem: the Myelodysplastic data

As has been claimed above, competing risk appears frequently in biomedicine researches, in fact, it is more a rule than an exception. The study of a specific cause of death and the time-free of disease are, probably, the most repeated examples. The main objectives of this section are the estimation of the time-free of leukemia and the time to death without leukemia in a cohort of patients with Myelodysplastic syndromes. The *Myelodysplastic* data was used with this goal. This dataset has been previously introduced in the Section 2 and were collected by the Spanish Group of Myelodysplastic Syndromes Registry (RESMD). Remember that a total of 968 patients (1,273.7 persons-year) were finally included in the study. There were 603 males (62.3%) and 365 females (37.7%); the median age at diagnosis was of 72.8 (ranged between 63.5 and 79.1) years. Two-hundred sixty eight patients (27.7%) developed leukemia during the follow-up and 403 died without leukemia. Figure 4 depicts a flowchart for the Myelodysplastic data. Interested readers are referred to Bernal et al. (2015) for complete information about the cohort and the problem.

By using the KM estimator and assuming as censored those events different to the studied one, the median time for developing leukemia was 3.37 years and, during the

Table 5. Mean ± standard deviation of the 1,000 Monte Carlo iterations for the quantity $100 \cdot \tau^{-1} \cdot \int_0^\tau |\hat{S}(t) - S(t)|dt$ where $S(t)$ is the real subdistribution function and $\hat{S}(t)$ its estimation based on KM and on MD estimators and τ is the maximum observed time for $\mu = 1/2$.

		$\rho = 0.0$		$\rho = 0.25$	
N	c	KM	DM	KM	MD
50	10	6.03 ± 1.96	1.16 ± 0.81	5.12 ± 1.82	0.96 ± 0.70
	-1/4	6.30 ± 3.57	2.67 ± 1.72	5.93 ± 3.18	2.21 ± 1.34
	-1/2	6.38 ± 3.68	2.78 ± 1.75	5.80 ± 3.48	2.52 ± 1.41
250	10	5.29 ± 0.99	0.34 ± 0.24	4.34 ± 0.88	0.28 ± 0.20
	-1/4	5.99 ± 1.77	0.92 ± 0.55	5.06 ± 1.67	0.83 ± 0.48
	-1/2	5.86 ± 2.08	1.06 ± 0.61	5.01 ± 2.01	1.02 ± 0.61
1000	10	4.58 ± 0.71	0.13 ± 0.09	3.60 ± 0.66	0.09 ± 0.07
	-1/4	5.92 ± 0.99	0.38 ± 0.23	5.04 ± 0.94	0.35 ± 0.21
	-1/2	5.77 ± 1.14	0.48 ± 0.27	5.07 ± 1.11	0.42 ± 0.25
		$\rho = 0.75$		$\rho = -0.50$	
50	10	3.41 ± 1.49	0.56 ± 0.43	8.38 ± 2.19	1.68 ± 1.08
	-1/4	4.05 ± 2.59	1.71 ± 1.00	7.76 ± 3.55	2.80 ± 1.59
	-1/2	4.20 ± 2.90	1.99 ± 1.23	7.78 ± 3.79	3.08 ± 1.70
250	10	2.76 ± 0.72	0.16 ± 0.12	6.43 ± 1.01	0.45 ± 0.32
	-1/4	3.35 ± 1.59	0.63 ± 0.38	7.93 ± 1.03	0.59 ± 0.40
	-1/2	3.30 ± 1.80	0.80 ± 0.47	7.82 ± 2.13	1.25 ± 0.71
1000	10	2.25 ± 0.51	0.05 ± 0.04	7.43 ± 0.73	0.23 ± 0.15
	-1/4	3.27 ± 0.94	0.26 ± 0.15	8.07 ± 0.92	0.48 ± 0.29
	-1/2	3.26 ± 1.07	0.34 ± 0.20	7.98 ± 1.04	0.54 ± 0.30

Table 6. Coverage percentages and mean±sd (standard deviations below 0.00 were denoted by 0.01) of the length of the 95% symmetric confidence intervals (computed by using the naive bootstrap method with 200 iterations) for the subdistribution function at times $t = 1/2$ and $t = 1$ using the KM and MD estimators when $\mu = 1/2$.

$\rho = 0.0$		$t = 1/2$				$t = 1$			
		KM		MD		KM		MD	
N	c								
50	10	92.0%	0.188 ± 0.04	92.3%	0.150 ± 0.03	76.2%	0.315 ± 0.04	91.8%	0.203 ± 0.02
	-1/2	91.2%	0.223 ± 0.07	90.6%	0.168 ± 0.05	87.3%	0.490 ± 0.14	89.4%	0.243 ± 0.04
1000	10	64.7%	0.043 ± 0.01	94.2%	0.037 ± 0.01	0.2%	0.070 ± 0.01	94.0%	0.051 ± 0.01
	-1/2	79.1%	0.051 ± 0.01	95.2%	0.044 ± 0.01	5.5%	0.107 ± 0.01	95.1%	0.071 ± 0.01
$\rho = 0.75$		$t = 1/2$				$t = 1$			
50	10	92.7%	0.137 ± 0.05	92.6%	0.110 ± 0.04	88.9%	0.240 ± 0.05	91.6%	0.159 ± 0.03
	-1/2	87.1%	0.149 ± 0.08	89.6%	0.116 ± 0.06	91.2%	0.344 ± 0.16	90.1%	0.189 ± 0.07
1000	10	81.9%	0.031 ± 0.01	92.6%	0.027 ± 0.01	11.4%	0.053 ± 0.01	95.1%	0.039 ± 0.01
	-1/2	85.5%	0.037 ± 0.01	94.9%	0.032 ± 0.01	39.6%	0.080 ± 2.26	94.0%	0.055 ± 0.01

follow-up, the estimated percentage of leukemia was 60.4%, while this estimation was only the 34.9% with the MD method (because this percentage does not lead the 50%, it is not possible to estimate the median time). In the same way, the median time to direct death (without developing leukemia) was 1.67 years when it was estimated by using the

Figure 4. Flowchart for the Myelodysplastic data.

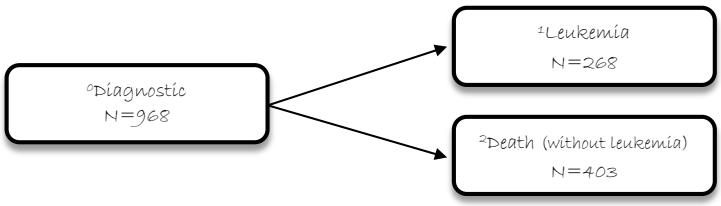
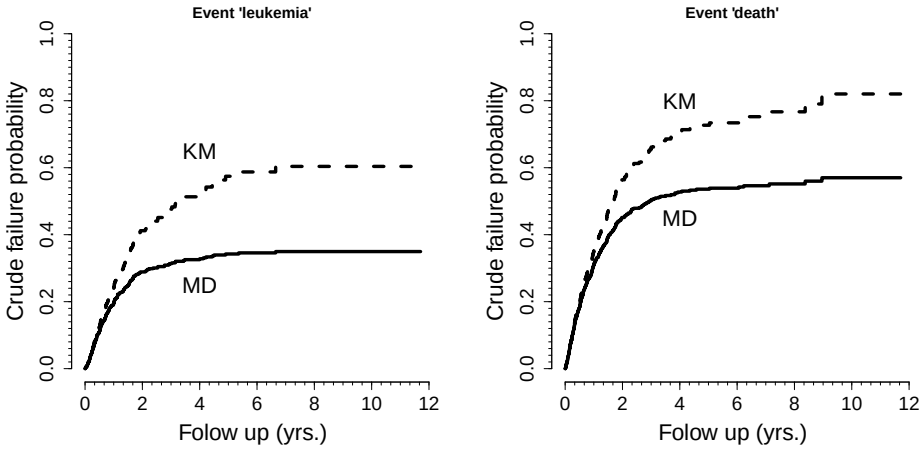


Figure 5. Crude failure probabilities computed by the KM and MD estimators for the time-free of leukemia, at left, and the time to direct death (without a previous leukemia), at right for the Myelodysplastic data.



KM estimator and 2.93 years when the MD method is employed. Figure 5 depicts the crude failure probabilities computed by the KM and MD estimators for the time-free of leukemia, at left, and the time to direct death (without a previous leukemia), at right; also called transition probabilities from the state 0 to 1 and 0 to 2, respectively.

It is worth to make note that the sum of the two KM estimations can take values larger than 1. In the considered problem, for $t \geq 3$, it does.

7. Main conclusions

Even when there exists a number of papers (see, for instance, the works of Putter et al. (2007) or Martínez-Camblor et al. (2015) among many others) trying to avoid the existing gap between theoretical and practical backgrounds, the advances in the statistical

methodology are still far from the methods commonly used by practitioners. In addition physicians and basic investigators are usually reluctant to apply in their studies new statistical techniques even when they may be more appropriate to deal with the problem at hand. Multi-state and, particularity, competing risk methods are examples of this situation; in spite of these techniques are the appropriate ones in order to study complex survival schemes, direct Kaplan-Meier and Cox regression are still the used methodologies even when some of the necessary assumptions are violated.

This paper considered the Kaplan-Meier estimator behavior in the competing risk setting. Monte Carlo simulations show that the direct use of the KM estimator produces serious mistakes in those scenarios where the probability of the competing event is high. However, in this context, the MD procedure works fine. In particular, under usual and mild conditions, it is an asymptotically unbiased estimator for the subdistribution functions (see, Kalb fleisch and Prentice (2002)). In addition, and in spite of MD is not include in most popular software, this procedure is easy to implement from the KM outcomes. In addition, several specific and friendly `R packages` [18] which are freely available in the CRAN (`http://cran.r-project.org/web/package`) have been developed with this goal; for example, Meira-Machado and Roca-Pardiñas (2011) describe the `p3state.msm` package and give a complete revision about previously existing software.

Finally, it is worth to remark that friendly statistical packages make easy the data analysis process. Particularly, most of the commercial software includes routines which perform Kaplan-Meier estimations and proportional hazard Cox models. However, using these techniques without checking (and, of course, knowing) conditions required for their correct performing, can produce erroneous conclusions. Remark that, in the practical problem considered, differences between the estimations provided by the KM and the MD methods were beyond ten percent.

Acknowledgments

The author is really grateful to the Associated Editor and to the three anonymous referees for their comments and suggestions which can improve the present work. In addition, he is grateful with the Spanish Group of Myelodysplastic Syndromes Registry (RESMD) and, in special, to Teresa Bernal for the permission to use the Myelodysplastic data.

References

[1] Aalen, O. *Nonparametric estimation of partial transition probabilities in multiple decrement models*, Annals of Statistics **6** (3), 534-545, 1978.

[2] Andersen, P.K., Abildstrom, S.Z. and Rosthoj, S. *Competing risks as a multi-state model*, Statistical Methods in Medical Research **11**, 203-215, 2002.

[3] Bitouzé, D., Laurent, B. and Massart, C. *A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator*, Ann. Inst. Henri Poincaré **35** (6), 735-763, 1999.

[4] Bernal, T., Martínez-Camblor, P., Sánchez-García, J., de Paz, R., Luño, E., Nomdedeu, B., Ardanaz, M.T., Pedro, C., Amigo, M.L., Xicoy, B., del Cañizo, C., Tormo, M., Bargay, J., Valcárcel, D., Brunet, S., Benlloch, L. and Sanz, G. *Effectiveness of azacitidine in unselected high-risk myelodysplastic syndromes: Results from the Spanish Registry*, Leukemia, `doi: 10.1038/leu`, 2015.

[5] de Boo, L., Pintilie, M., Yip, P., Baniel, J., Fleshner, N. and Margel, D. *Time from first detectable PSA following radical prostatectomy to biochemical recurrence: A competing risk analysis*, Canadian Urology Association Journal **9** (1-2), 14-21, 2015.

[6] Cox, D.R., *Regression models and life table (with discussion)*, Journal of the Royal Statitical Society, Ser. B, **34**, 187-220, 1972.

[7] Csörgő, S. *Universal Gaussian approximations under random censorship*, Annals of Statistics **24** (6), 2744-2778, 1996.

[8] Greenwood, M. *The errors of sampling of the survivorship tables*, Reports on Public Health and Statistical Subjects 33, 1-26, 1926.

[9] Kalbfleisch, R. and Prentice, R.L. *The statistical analysis of failure time data*, Wiley, New Jersey, 2002.

[10] Kaplan, E.L. and Meier, P. *Nonparametric estimation from imcomplete observations*, Journal of the American Statistical Association **53**, 457-481, 1958.

[11] Komlós, J., Major, P. and Tusnády, G. *An approximation of partial sums of independent rvs and the sample df. I*, Wahrsch verw Gebiete 32, 111-131, 1975.

[12] Martínez-Camblor, P., Larrañaga, N., Sarasqueta, C., Mitxelena, M.J. and Basterretxea, M. *That deadly and pink corporeity: Analysis of disease-free survival analysis breast cancer in Gipuzkoa (Spain) in the presence of competing risks*, Gaceta Sanitaria **23** (6), 554-557, 2009.

[13] Martínez-Camblor, P., de Uña-Álvarez, J. and Díaz Corte, C. *Expanded renal transplantation: A multi-state model approach*, Journal of Applied Statistics, DOI:10.1080/02664763.2015.1043866, 2015.

[14] Meira-Machado, L. and Roca-Pardiñas, J. *p3state.msm: Analyzing survival data from an illness-death mode*, Journal of Statistical Software **38** (3), 1-18, 2011.

[15] Njamen-Njomen, D. and Ngatchou-Wandji, J. *Nelson-Aalen and Kaplan-Meier estimators in competing risks*, Applied Mathematics **5**, 465-47, 2014.

[16] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. and Smith, P.G. *Design and analysis of randomised clinical trials requiring prolonged observation of each patient, II*, British Jorunal of Cancer, 35, 1-39, 1977.

[17] Putter, H., Fiocco, M. and Geskus, R. *Tutorial in biostatistics: Competing risks and multistate models*, Statistics in Medicine 26, 2389-2430, 2007.

[18] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2007.

[19] Simon, R. and Lee, Y.J. *Nonparametric confidence limits for survival probabilities and median survival time*, Cancer Treatment Reports 66, 37-42, 1985.

[20] Tiatsis, AA. Competing risks, *Encyclopedia of Biostatistics*, 1, 824-834, 1988.

[21] Verduijn, M., Grootendorst, D.C., Dekker, F.W., Jager, K.J. and le Cessie, S. *The analysis of competing events like cause-specific mortalitybeware of the KaplanMeier method*, Nephrology Dialisys Transplantation 26, 5661, 2011.