SiDOR
Statistical Inference
Decision & Operations Research Group
Universidade Vigo

XI XORNADA DE USUARIOS DE R EN GALICIA

Comparison of cumulative incidence curves with multiple causes of death

Nora M. Villanueva*

Department of Statistics and O. R., SiDOR Group, University of Vigo, Spain

XI Xornadas de Usuarios de R 2024, Santiago de Compostela, Spain

October 24th 2024

*work jointly done with M. Sestelo, L. Meira-Machado and J. Roca-Pardiñas

Data from leukemia patients from the European Group for Blood and Marrow Transplantation (EBMT)*

- 8966 patients with some type leukemia
- Patients underwent bone marrow transplantation
- At risk of developing a variety of complications that compete with each other

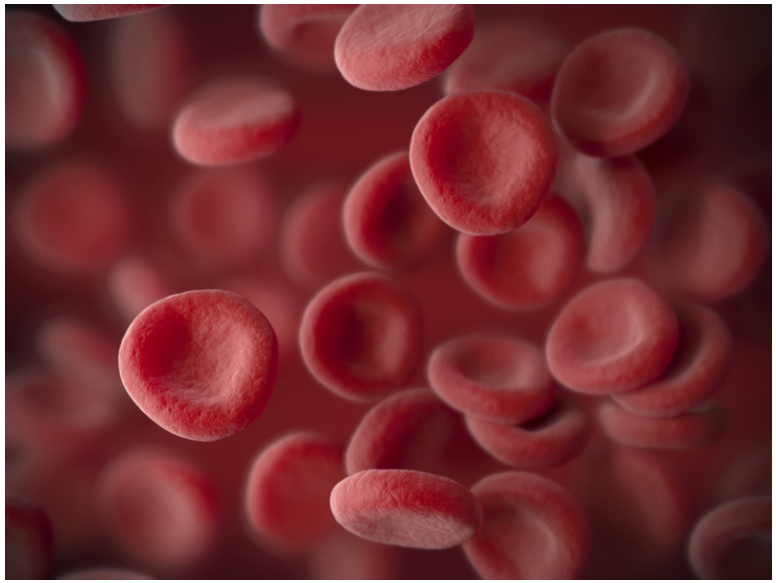
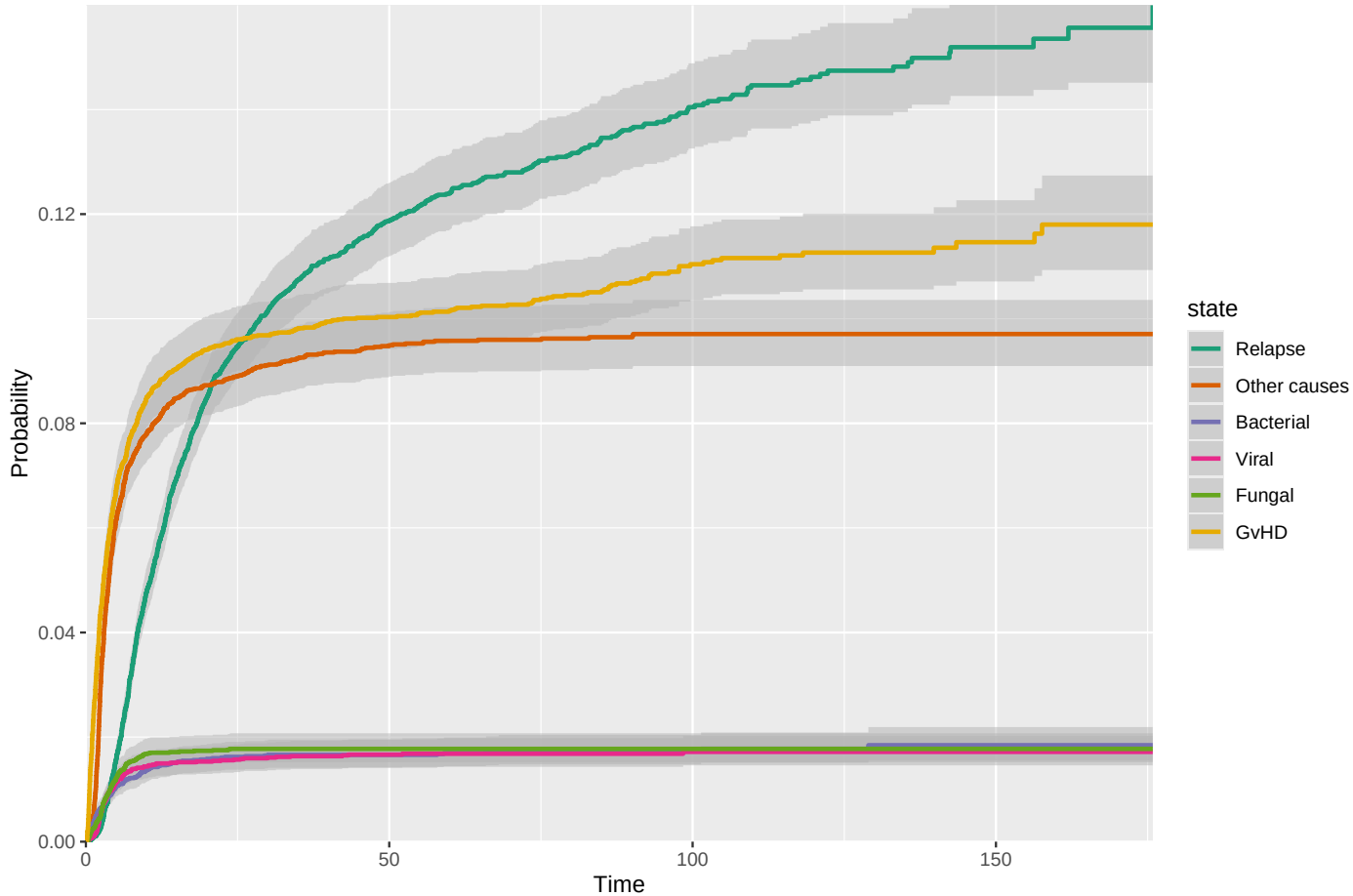


Table: Number of events for each cause of death.

Event	Relapse	GvHD	Bacterial	Viral	Fungal	Other	Censored
Number	1098	834	151	147	156	924	5656

- Dataset contain several variables:
  - Times (in days) from transplantation to death or last follow-up (`time`)
  - Status indicator (`status`): 0 = censored, 1 = relapse, etc.

* Fiocco et al. (2005), Wreede et al. (2011).



1. Are all these curves equal?

2. Can we identify groups in some way?

Methods to test for the equality of cumulative incidence curves

- Aly et al. (1994), Carriere and Kochar (2000), Kochar et al. (2002) and Sankaran et al. (2010), compare of the CIF among each other
- Gray (1988) compares the CIF for a particular type of failure among the different levels of a factor.

If the null hypothesis of equality of curves is rejected, at least one curve is different

- Can we perform groups? How many of them are there?
- There are no methodological papers proposing clusters of CIF for competing risk data
- We propose an approach that allows determining CIF groups with an automatic selection of their number

Some previous notation

- General random censorship model, in which n individuals mutually independent are observed.
- Let T_i $(i = 1, \ldots, n)$ be the lifetime corresponding to any J competing causes with $j = 1, ..., J$ and E_i the type of event with $E \in \{1, ..., J\}$. Here we consider the events to be deaths from different causes.
- Assuming that T_i is observed subject to a (univariate) random right-censoring variable C_i assumed to be independent of T_i
- Due to censoring we only observe $(\widetilde{T}_i, \Delta_i, \Delta_i E_i)$ where $\widetilde{T}_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$

The distribution of the lifetime, T, can be characterized by $S(t)$ given by

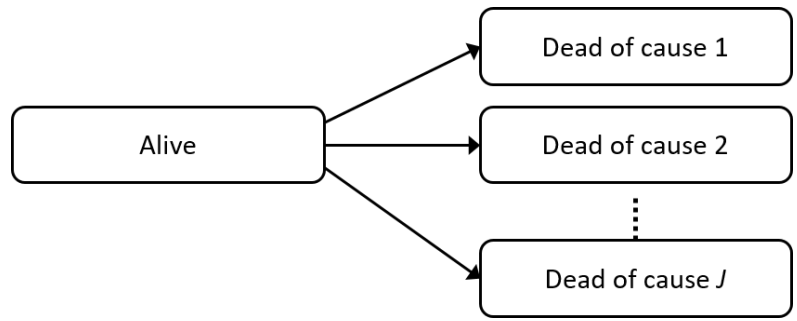
$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t).$$

This distribution may be derived from the following J cause-specific hazard functions

$$h_j(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, E = j \mid T > t)}{\Delta t}.$$

Certainly, one has

$$S(t) = \exp(-\int_0^t \sum_{j=1}^{J} h_j(t)).$$



It also holds that

$$S(t) = 1 - \sum_{j=1}^{J} F_j(t),$$

where $F_j(t) = P(T \leq t, E = j)$, is the CIF, which is the probability of dying from a particular cause, $E = j$, by time t while also being at risk of dying from other causes.

The cause-specific CIF can also be expressed as a function of the cause-specific hazards for all J causes as

$$F_j(t) = \int_0^t S(u) h_j(u) du.$$

Since the censoring time is assumed to be independent of the process, the survival function, $S(t) = P(T > t)$ may be consistently estimated by the Kaplan-Meier estimator (Kaplan and Meier, 1958)

$$\widehat{S}(t) = \prod_{\widetilde{T}_{(i)} \leq t} \left(1 - \frac{\Delta_i}{R(\widetilde{T}_{(i)})}\right),$$

where $\widetilde{T}_{(1)} \leq \cdots \leq \widetilde{T}_{(n)}$ denotes the ordered T-sample based on all n individuals, and $R(t) = \sum_{i=1}^{n} I(\widetilde{T}_i \geq t)$ indicates the number of individuals at risk just before time t.

An estimator of the cumulative incidence functions, CIF, is obtained directly from the previous equation by plug-in the Nelson-Aalen estimator and the product-limit estimator of survival (Geskus, 2011)

$$\widehat{F}_j(t) = \sum_{\widetilde{T}_{(i)} \leq t} \widehat{S}(\widetilde{T}_{(i)}^-) \frac{I(\Delta_{[i]} E_{[i]} = j)}{R(\widetilde{T}_{(i)})}.$$

If $H_0 : F_1(t) = F_2(t) = \ldots = F_J(t)$ for all $t > 0$ is rejected...

- We would like to asses if the levels $\{1, \ldots, J\}$ can be grouped in K groups $\{G_1, \ldots, G_K\}$ with $K < J$, so that
  - $F_i = F_j$ for all $i, j \in G_k$, for each $k = 1, \ldots, K$
  - $\{G_1, \ldots, G_K\}$ must be a partition of $\{1, \ldots, J\}$
  - $G_1 \cup \ldots \cup G_K = \{1, \ldots, J\}$ and $G_i \cap G_j = \emptyset$ for all $i \neq j \in \{1, \ldots, K\}$
- A procedure to test, for a given number K, the null hypothesis $H_0(K)$ is that at least exists a partition $\{G_1, \ldots, G_K\}$ so that all the conditions above are verified.

The testing procedure is based on the J-dimensional process

$$\widehat{\mathbf{U}}(t) = (\widehat{U}_1(t), \widehat{U}_2(t), \ldots, \widehat{U}_J(t))^t,$$

where, for $j = 1, \ldots, J$,

$$\widehat{U}_j(t) = \sum_{k=1}^{K} w(t)^{-1/5} [\widehat{F}_j(t) - \widehat{M}_k(t)] \; I_{\{j \in G_k\}}$$

and $\widehat{M}_k$ corresponds to the average of the CIF estimate $\widehat{F}_j$ for all $j \in G_k$, i. e.,

$$\widehat{M}_k(t) = \frac{1}{r_k} \sum_{j \in G_k} \widehat{F}_j(t), \qquad \text{where } r_k = \#G_k.$$

- Statistic tests

$$D_{CM} = \min_{G_1, \ldots, G_K} \sum_{j=1}^{J} \int_{\tau_{\widetilde{T}}} \hat{U}_j^2(t) dy,$$

$$D_{KS} = \min_{G_1, \ldots, G_K} \sum_{j=1}^{J} \int_{\tau_{\widetilde{T}}} |\hat{U}_j(t)| dy.$$

$*$ With $J = 30$ and $K = 5$, the total number of distinct assignments is $7.7\ 10^{18}$, following Jain and Dubes (1988), $R(J, K) = \frac{1}{K!} \sum_{i=1}^{K} (-1)^{K-i} \binom{K}{i} (i)^n$

The testing procedure is based on the J-dimensional process

$$\widehat{\mathbf{U}}(t) = (\widehat{U}_1(t), \widehat{U}_2(t), \ldots, \widehat{U}_J(t))^t,$$

where, for $j = 1, \ldots, J$,

$$\widehat{U}_j(t) = \sum_{k=1}^{K} w(t)^{-1/5} [\widehat{F}_j(t) - \widehat{M}_k(t)] \; I_{\{j \in G_k\}}$$

and $\widehat{M}_k$ corresponds to the average of the CIF estimate $\widehat{F}_j$ for all $j \in G_k$, i. e.,

$$\widehat{M}_k(t) = \frac{1}{r_k} \sum_{j \in G_k} \widehat{F}_j(t), \qquad \text{where } r_k = \#G_k.$$

- Statistic tests

$$D_{CM} = \min_{G_1, \ldots, G_K} \sum_{j=1}^{J} \int_{\tau_{\widetilde{T}}} \hat{U}_j^2(t) dy, \longrightarrow \text{ Kmeans}$$

$$D_{KS} = \min_{G_1, \ldots, G_K} \sum_{j=1}^{J} \int_{\tau_{\widetilde{T}}} |\hat{U}_j(t)| dy. \longrightarrow \text{ Kmedians}$$

$*$ With $J = 30$ and $K = 5$, the total number of distinct assignments is $7.7\ 10^{18}$, following Jain and Dubes (1988), $R(J, K) = \frac{1}{K!} \sum_{i=1}^{K} (-1)^{K-i} \binom{K}{i} (i)^n$

The testing procedure is based on the J-dimensional process

$$\widehat{\mathbf{U}}(t) = (\widehat{U}_1(t), \widehat{U}_2(t), \ldots, \widehat{U}_J(t))^t,$$

where, for $j = 1, \ldots, J$,

$$\widehat{U}_j(t) = \sum_{k=1}^{K} w(t)^{-1/5} [\widehat{F}_j(t) - \widehat{M}_k(t)] \; I_{\{j \in G_k\}}$$

and $\widehat{M}_k$ corresponds to the average of the CIF estimate $\widehat{F}_j$ for all $j \in G_k$, i. e.,

$$\widehat{M}_k(t) = \frac{1}{r_k} \sum_{j \in G_k} \widehat{F}_j(t), \qquad \text{where } r_k = \#G_k.$$

- Statistic tests

$$D_{CM} = \min_{G_1, \ldots, G_K} \sum_{j=1}^{J} \int_{\tau_{\widetilde{T}}} \hat{U}_j^2(t) dy, \longrightarrow \text{ Kmeans}$$

$$D_{KS} = \min_{G_1, \ldots, G_K} \sum_{j=1}^{J} \int_{\tau_{\widetilde{T}}} |\hat{U}_j(t)| dy. \longrightarrow \text{ Kmedians}$$

- Decision rule: we reject H_0 for large statistic values.
- Distribution of D? Bootstrap method (Efron, B., 1979, 1981)

The steps of the testing procedure, for a given K, are as follows

Step 1. Using the original sample, for $j = 1, \ldots, J$ and $i = 1, \ldots, n$, estimate the cumulative incidence functions F_j in a non parametric way and in a common grid, using each sample separately.

Then, using the proposed algorithms, obtain the "best" partition $\{G_1, \ldots, G_K\}$ and with it obtain the estimated curves $\widehat{M}_k$.

Step 2. Obtain the D value as explained before.

Step 3. Draw bootstrap samples using a pooled bootstrap procedure (i.e., bootstrap from the pooled combined partition sample given by the null hypothesis $H_0(K)$).

Step 4. Let D^{*b} be the test statistic obtained from the bootstrap samples $\{(\widetilde{T}_{ij}^{*b}, \Delta_{ij}^{*b}), i = 1, \ldots, n_j\}, j = 1, \ldots, J$

The decision rule consists of rejecting the null hypothesis if $D > D^{*(1-\alpha)}$, where $D^{*(1-\alpha)}$ is the empirical $(1-\alpha)$-percentile of values D^{*b} $(b = 1, \ldots, B)$ previously obtained.

Algorithm 1. K-cumulative incident curves algorithm

1. With $\{(\widetilde{T}_i, \Delta_i, \Delta_i E_i),\ i = 1, \ldots, n\}$, and using the Geskus estimator obtain $\hat{F}_j$.
2. Initialize with $K = 1$ and test $H_0(K)$:
   - 2.1. Obtain the "best" partition $\{G_1, \ldots, G_K\}$ by means of the K-means or K-medians algorithm.
   - 2.2. For $k = 1, \ldots, K$, estimate $\widehat{M}_k$ and retrieve the test statistic D.
   - 2.3. Generate B bootstrap samples and calculate D^{*b}, for $b = 1, \ldots, B$.
   - 2.4. **if** $D > D^{*(1-\alpha)}$ **then**

     reject $H_0(K)$

     $K = K + 1$

     go back to 2.1

     else

     accept $H_0(K)$

     end
3. The number K of groups of CIF is determined.

I. Testing one specific hypothesis $H_0(2)$

We have followed the algorithm from Beyersmann' Book (Beyersmann et al., 2012) to simulate the data, where the variable T was generated from the distribution

$$F(t) = 1 - S(t) = 1 - \exp\left(-\int_0^t h_1(u) + h_2(u) + \ldots + h_J(u)du\right).$$

Scenario with $J = 3$ cause-specific hazards

* $h_1(t) = 0.58/(t+4)$, $h_2(t) = 0.03 \times \log(t+1)$ and $h_3(t) = 0.03 \times \log(t+1+6a)$, with a being a constant.
* $C \sim U[0, c]$ with c being 40 and 20, leading to a proportion of censored data (when $a = 0$) of approximately 15% and 30%, respectively.

* Different values of a were considered, ranging from 0 to 0.4 Note that $a = 0$ corresponds to the null hypothesis $H_0(2)$ and when the value $a \neq 0$, the number of groups is three.

* 1000 trials at the significance levels of 0.05 and 0.10, and sample sizes of $n = 500, 1000$ and 1500.

* We apply the bootstrap method (500 bootstrap samples for type I errors and for the power under the alternative) to determine the critical values of the tests.

Table: Experiment I. Estimated type I errors of testing $H_0(2)$ based on the test statistics D_{CM} and D_{KS} when the distribution of the censoring time C is $U(0, 40) \sim$ 15% censoring or $U(0, 20) \sim$ 30% censoring.

			D_{CM}		D_{KS}	
C	n	α:	**0.05**	**0.10**	**0.05**	**0.10**
$U(0, 40)$	500		0.02	0.06	0.03	0.07
	1000		0.04	0.07	0.05	0.08
	1500		0.03	0.07	0.04	0.08
$U(0, 20)$	500		0.02	0.06	0.03	0.06
	1000		0.04	0.08	0.04	0.09
	1500		0.03	0.06	0.04	0.07

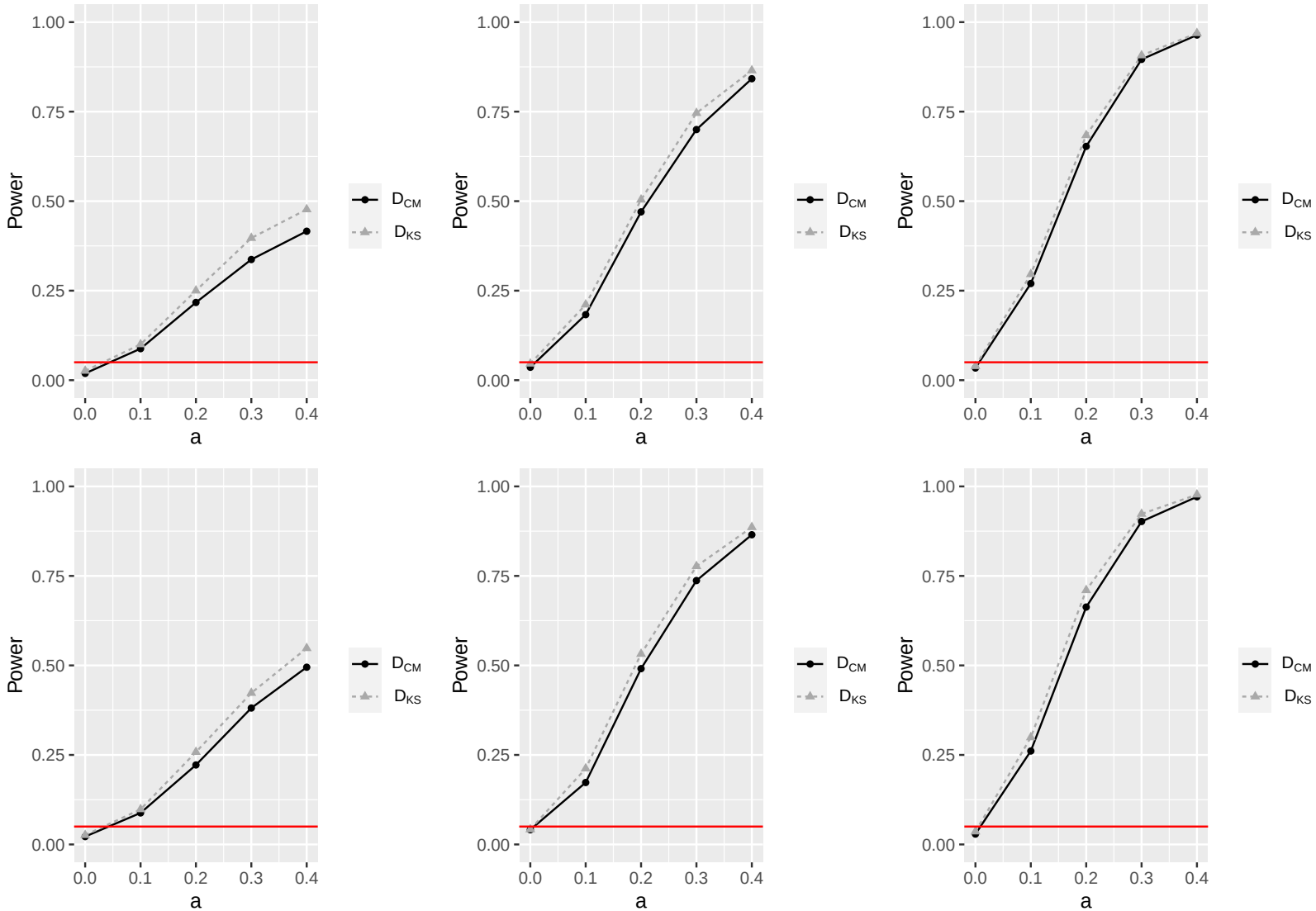


Figure: Rejection probabilities of the two tests for nominal level 5% (red line).

II. Assessing the K-cumulative incident curves algorithm

* Same scenario than the previous one, but taking into account $a = 0$.

* There are $J = 3$ cause-specific hazards but two of them are equal.

* The censoring variable C and the remainder parameters were generated and kept as previously.

* Results of this simulation refers to the number of times (of 1000 repetitions) that Algorithm 1 selects the number of groups using a nominal level of 5%.

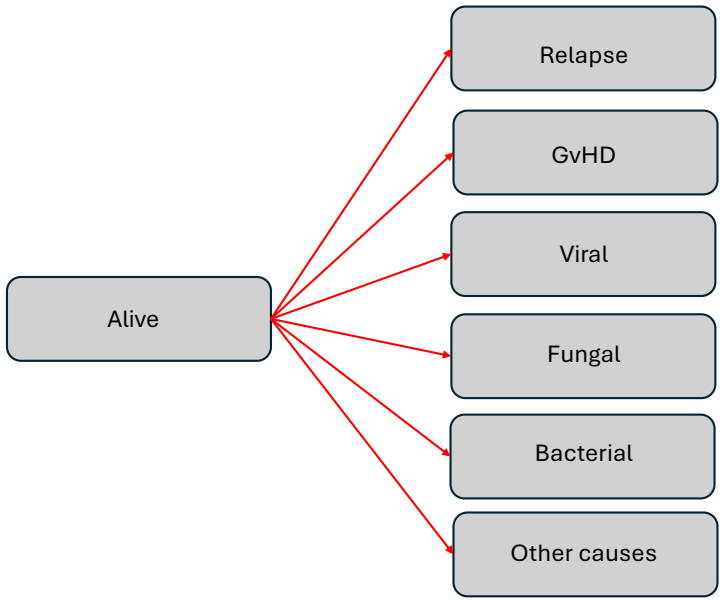
Table: Experiment II. Number of times (of 1000 trials) that Algorithm 1 selects the number of groups using a nominal level of 5%.

		Number of groups					
		D_{CM}			D_{KS}		
C	n	**1**	**2**	**3**	**1**	**2**	**3**
$U(0,40)$	500	0	**985**	15	0	**976**	24
	1000	0	**963**	37	0	**956**	44
	1500	0	**968**	32	0	**962**	38
$U(0,20)$	500	0	**975**	25	0	**974**	26
	1000	0	**965**	35	0	**955**	45
	1500	0	**973**	27	0	**967**	33

$*$ Note that, in order to perform correctly, the algorithm must reject $H_0(1)$ and then, accept $H_0(2)$.

Data from leukemia patients from the European Group for Blood and Marrow Transplantation (EBMT)*

- 8966 patients with some type leukemia
- Patients underwent bone marrow transplantation
- At risk of developing a variety of complications that compete with each other



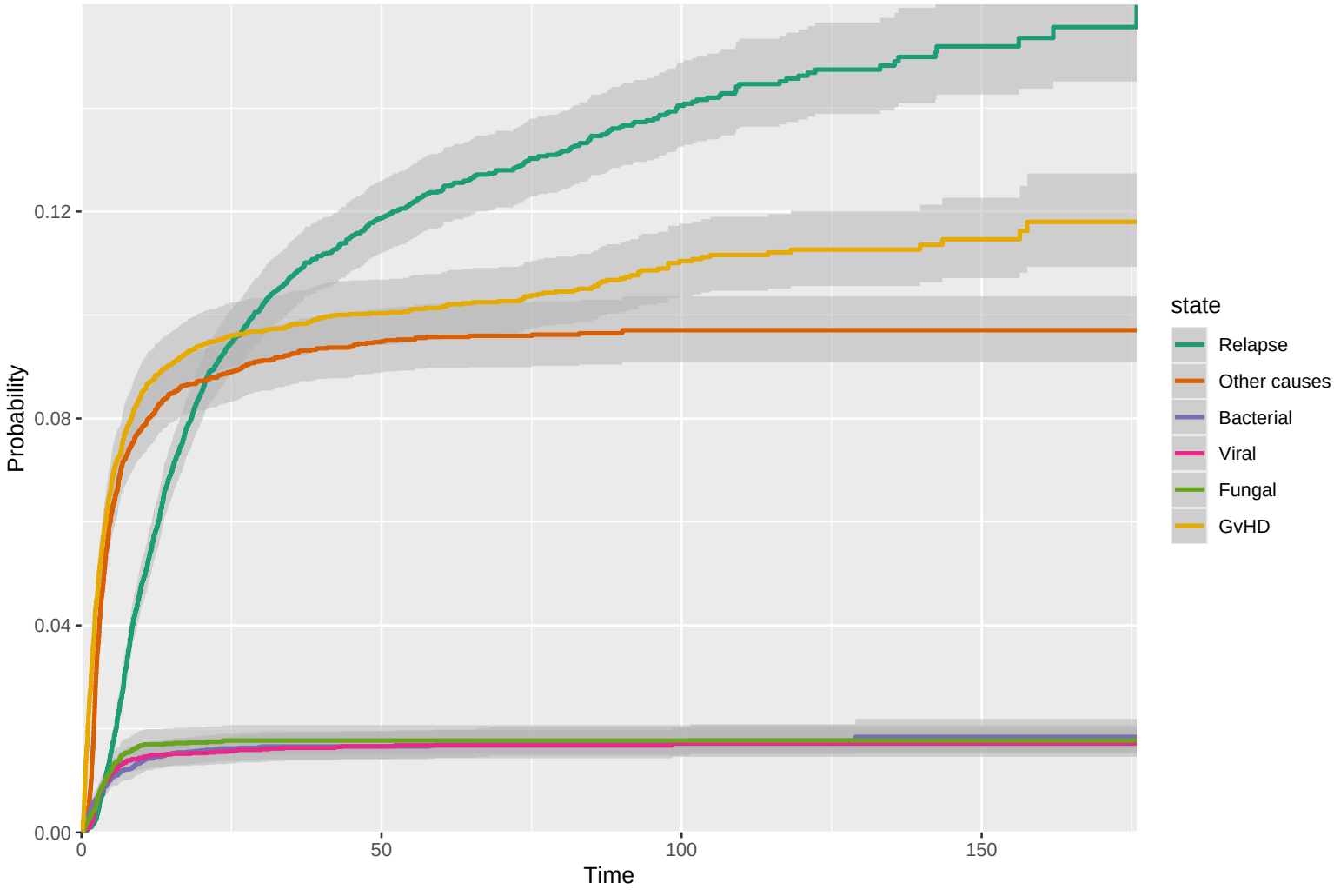
* Fiocco et al. (2005), Wreede et al. (2011).

- Dataset contain several variables:
  - Times (in days) from transplantation to death or last follow-up (`time`)
  - Status indicator (`status`): 0 = censored, 1 = relapse, etc.

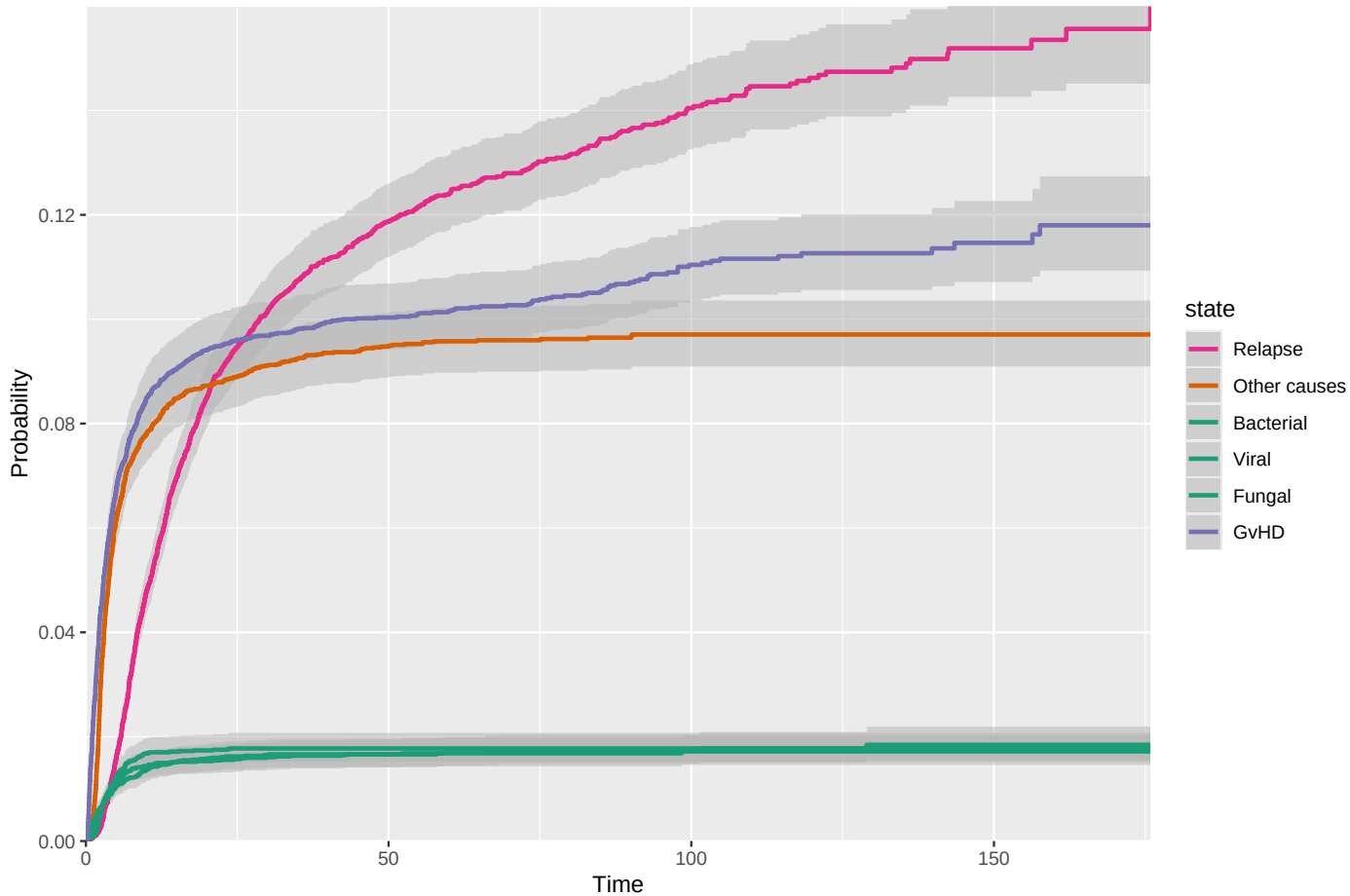
- clustcurv package is a shortcut for "clustering curves" that allows users determining groups of multiple curves with an automatic selection of their number
- The package works for survival, regression and cumulative incidence functions
- In view of the high computational cost entailed in these methods, parallelization techniques are included to become feasible and efficient onto real situations
- The package can be downloaded from github https://github.com/noramvillanueva/clustcurv/tree/CIF
- Starting with the analysis

```
R> devtools::install_github("noramvillanueva/clustcurv", ref = "CIF")
R> library(clustcurv)
R> library(mstate)
R> data(ebmt2)
R> table(ebmt2)
  Relapse      GvHD Bacterial     Viral    Fungal     Other  Alive
     1098       834       151       147       156       924      5656
```

- Two main types of functionalities:
  - to determine groups of curves, given a number K, with `kregcurves()`, `ksurvcurves()` or `kcifcurves()` functions
  - to determine groups of curves with the automatic selection of their number with `regclustcurves()`, `survclustcurves()` or `cifclustcurves()` functions
- Numerical and graphical summaries can be obtained by using the generic functions `print()`, `summary()` and `autoplot()`


Probability
0.00
0.04
0.08
0.12
0
50
100
150
Time
state
Relapse
Other causes
Bacterial
Viral
Fungal
GvHD

```
R> out2 <- cifclustcurves (time = ebmt2$time, status = ebmt2$status,
kbin = 50, nboot = 200, algorithm = "kmeans",
cluster = TRUE, seed = 300716)
R> autoplot(out2 , groups_by_colour = TRUE)
```



∗ $H_0(4)$: p-value $= 0.13$

- A new procedure is proposed that let us, not only testing the equality of cumulative incident curves but also grouping them if they are not equal.

- Simulation studies show that our test controls type I error rate quite well under all situations considered, and also the power performance for the alternative is good.

* Software in the form of an R package has been developed and is freely available from GitHub (soon on CRAN).

- The contributions of this talk are based on:

Sestelo, M., Meira-Machado, L., Villanueva, N. M., and Roca-Pardiñas, J. (2024). A method for determining groups in cumulative incidence curves in competing risk data. *Biometrical Journal*, **66**, 2300084.

Villanueva, N. M., Sestelo, M., and Meira-Machado and Roca-Pardiñas, J. (2021). clustcurv: An R package for Determining Groups in Multiple Curves. *The R Journal*, 12(1):164–183.

Villanueva, N. M., Sestelo, M., and Meira-Machado, L. (2019). A method for determining groups in multiple survival curves. *Statistics in Medicine*, **38**, 866–877.

- Beyersmann, J., Allignol, A., and Schumacher, M. (2012). Competing Risks and Multistate Models in R.

- Efron B.(1981) Censored Data and the Bootstrap. *Journal of the American Statistical Association*, **76 (374)**:312–319.

- Geskus, R. B. (2011). Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring. *Biometrics*, **67**, 39–49.

- Kalbfleisch, J. D. and Prentice, R. L. (1980). The statistical analysis of time failure data. John Wiley and Sons New York.

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

- Macqueen J. B. (1967). Some methods of classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

Comparison of cumulative incidence curves with multiple causes of death

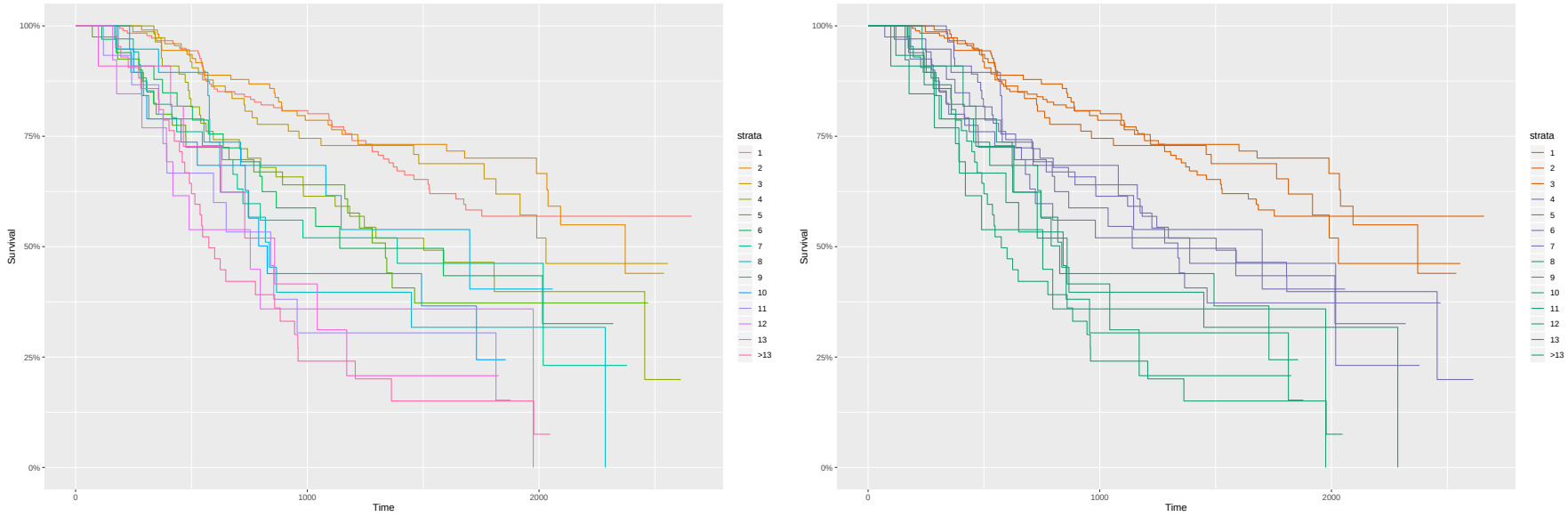
Nora M. Villanueva*

Department of Statistics and O. R., SiDOR Group, University of Vigo, Spain

XI Xornadas de Usuarios de R 2024, Santiago de Compostela, Spain

October 24th 2024

*work jointly done with M. Sestelo, L. Meira-Machado and J. Roca-Pardiñas



1. Are all these curves equal?

2. Can we identify groups in some way?

Villanueva, N. M., Sestelo, M., and Meira-Machado, L. (2019). A Method for Determining Groups in Multiple Survival Curves. *Statistics in Medicine*, 38:366–377.

Villanueva, N. M., Sestelo, M., Ordóñez, C., and Roca-Pardiñas, J. (2020). An automatic procedure to determine groups of nonparametric regression curves. *arXiv: 2012.15278*.

Sestelo, M., Meira-Machado, L., Villanueva, N. M., and Roca-Pardiñas, J. (2024). A method for determining groups in cumulative incidence curves in competing risk data. *Biometrical Journal*, 66, 2300084.

Since the censoring time is assumed to be independent of the process, the survival function, $S(t) = P(T > t)$ may be consistently estimated by the Kaplan-Meier estimator (Kaplan and Meier, 1958)

$$\widehat{S}(t) = \prod_{\widetilde{T}_{(i)} \leq t} \left(1 - \frac{\Delta_i}{R(\widetilde{T}_{(i)})}\right),$$

where $\widetilde{T}_{(1)} \leq \cdots \leq \widetilde{T}_{(n)}$ denotes the ordered T-sample based on all n individuals, and $R(t) = \sum_{i=1}^{n} I(\widetilde{T}_i \geq t)$ indicates the number of individuals at risk just before time t.

The Kaplan-Meier product-limit estimator can also be expressed using Kaplan-Meier weights (Meira-Machado and Sestelo, 2019)

$$\widehat{S}(t) = 1 - \sum_{i=1}^{n} w_i I(\widetilde{T}_{(i)} \leq t),$$

where

$$w_i = \frac{\Delta_{[i]}}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{[j]}}{n-j+1}\right]$$

If $H_0 : F_1(t) = F_2(t) = \ldots = F_J(t)$ for all $t > 0$ is rejected...

- We would like to asses if the levels $\{1, \ldots, J\}$ can be grouped in K groups $\{G_1, \ldots, G_K\}$ with $K < J$, so that
  - $S_i = S_j$ for all $i, j \in G_k$, for each $k = 1, \ldots, K$
  - $\{G_1, \ldots, G_K\}$ must be a partition of $\{1, \ldots, J\}$
  - $G_1 \cup \ldots \cup G_K = \{1, \ldots, J\}$ and $G_i \cap G_j = \emptyset$ for all $i \neq j \in \{1, \ldots, K\}$

- A procedure to test, for a given number K, the null hypothesis $H_0(K)$ is that at least exists a partition $\mathbf{G}_0 = \{G_1, \ldots, G_P\}$ with $P \leq K$ so that all the conditions above are verified.

- The alternative hypothesis $H_1(K)$ is that for any partition $\mathbf{G}_1 = \{G_1, \ldots, G_L\}$ with $L > K$, not exists another partition $\mathbf{G}_0$ verifying $\#\mathbf{G}_0 < \#\mathbf{G}_1$ where

$$\#\{G_1, \ldots, G_{\mathcal{K}}\} = 1 + \sum_{k_2=2}^{\mathcal{K}} \left(\prod_{k_1 < k_2} I\{G_{k_1} \neq G_{k_2}\} \right)$$

  and, for definition, $G_{k_1} \neq G_{k_2}$ is verified if $S_i \neq S_j$ for all $(i, j) \in G_{k_1} \times G_{k_2}$.

The steps of the testing procedure, for a given K, are as follows

Step 1. Using the original sample, for $j = 1, \ldots, J$ and $i = 1, \ldots, n$, estimate the cumulative incidence functions F_j in a non parametric way and in a common grid, using each sample separately.

Then, using the proposed algorithms, obtain the "best" partition $\{G_1, \ldots, G_K\}$ and with it obtain the estimated curves $\widehat{M}_k$.

Step 2. Obtain the D value as explained before.

Step 3. Draw bootstrap samples using a pooled bootstrap procedure (i.e., bootstrap from the pooled combined partition sample given by the null hypothesis $H_0(K)$). To this end, proceed with the following steps:

3.1 Compute the sample size for each possible cause of death j as $n_j = \sum_{i=1}^{n} I(\Delta_i E_i = j)$ and the sample size for each pooled combined partition sample from G_k as $m_k = \sum_{j=1}^{J} n_j I_{\{j \in G_k\}}$.

3.2 For $b = 1, \ldots, B$, draw $(\widetilde{T}_1^{*b}, \Delta_1^{*b}, \Delta_1^{*b} E_1^{*b}), (\widetilde{T}_2^{*b}, \Delta_2^{*b}, \Delta_2^{*b} E_2^{*b}), \ldots, (\widetilde{T}_n^{*b}, \Delta_n^{*b}, \Delta_n^{*b} E_n^{*b})$ by independent sampling m_k times with replacement from $\hat{R}_k$, the empirical distribution function putting mass $m_k^{-1} w_j$ at each point $(\widetilde{T}_i, \Delta_i, \Delta_i E_i = j)$ with $j \in G_k$.

Drawing from $\hat{R}_k$ involves sampling from the observed data points corresponding to the cause of death partition G_k. Specifically, each draw is performed by randomly selecting a data point $(\widetilde{T}_i, \Delta_i, \Delta_i E_i = j)$, where j belongs to the cause of death partition G_k. The number of times each cause of death j contributes to the sampling is proportional to its weight w_j in the pooled combined partition sample.

To provide further clarification, it is important to note that, in accordance with $H_0(K)$, each new bootstrap partition of size m_k must be balanced concerning the sample size (n_j) of each cause of death included within G_k. This balance is essential because it ensures that the cumulative incidence of causes of death within each group remains similar. Therefore, w_j serves as a weight for each competing cause of failure j which rescales the initial mass m_k^{-1} using a ratio of the form $w_j = p_0/p_j$ with $p_0 = 1/r_k$ and $p_j = n_j/m_k$ thus achieving a balanced sample within each cluster.

Step 4. Let D^{*b} be the test statistic obtained from the bootstrap samples $\{(\widetilde{T}_{ij}^{*b}, \Delta_{ij}^{*b}), i = 1, \ldots, n_j\}, j = 1, \ldots, J$

The decision rule consists of rejecting the null hypothesis if $D > D^{*(1-\alpha)}$, where $D^{*(1-\alpha)}$ is the empirical $(1-\alpha)$-percentile of values D^{*b} $(b = 1, \ldots, B)$ previously obtained.