Relative survival analysis and dependency assumptions: recent contributions

IME 2025 @ Tartu

Oskar Laverny[1]

July 4, 2025

Based on 2023+ joint works with: R. Alhajal, R. Giorgi, N. Grafféo & F.J Rubio

Partly funded by Cancéropôle PACA

[1] Aix Marseille University, INSERM, IRD, SESSTIM, ISSPAM, Marseille, France.

Table of content

Introduction to relative survival

Introduction to relative survival

Reminder on survival analysis

Survival analysis: notations, hypotheses, goal

A **standard survival analysis**[1] problem is described by the following variables:

Random Variable	Name	Observed ?
O	"Overall" lifetime	No
C	"Censoring" time	No
$T = O \wedge C$	Event time	Yes
$\Delta = \mathbb{1}\{T \leq C\}$	Event status	Yes

Dependency: The standard assumption is $O \perp\!\!\!\perp C$.

Sample: We assume $(O_i, C_i, T_i, \Delta_i)_{i \in 1,\ldots,n}$ to be a n-sample of (O, C, T, Δ).

Filtration: $\mathcal{F}_t = \sigma\left\{(T_i, \Delta_i) : T_i \leq t, \; \forall i \in 1, .., n\right\}.$

Goal: Estimation the distribution of O, say by it's hazard $\partial \Lambda_O(t) = -\partial \ln S_O(t)$.

[1] P. K. Andersen. *Counting Process Models for Life History Data: A Review*. Oslo: Universitetet i Oslo. Matematisk Institutt, 1984. ISBN: 978-82-553-0561-3.

In standard survival analysis, we define the following stochastic processes:

$$
\begin{aligned}
N(t) &= \mathbb{1}\{O \leq t, O \leq C\} && \textit{(Uncensored deaths process)} \\
Y(t) &= \mathbb{1}\{O \geq t, C \geq t\} && \textit{(At-risk process)} \\
M(t) &= N(t) - \int_0^t Y(s)\partial\Lambda_O(s) && \textit{(Martingale)}
\end{aligned}
$$

We similarly defined individual versions N_i, Y_i, M_i. From them, we can derive the **Nelson-Aalen estimator**:

$$\partial\widehat{\Lambda}_O(t) = \frac{\sum_i \partial N_i(t)}{\sum_i Y_i(t)}.$$

Facts: $\widehat{\Lambda}_O$ is unbiaised, convergent, assymptotically gaussian, has explicit DM decomposition...

Outputs: Survival curves, confidence intervals, log-rank tests, etc...

Classical extension: competitive risks

Introduction to relative survival

Relative survival analysis

The relative survival context

In population-based studies and/or **cancer registries**, the specific cause of death is often unidentified, unreliable or even unavailable.

Random Variable	Name	Observed ?
E	"Excess" lifetime	No
P	"Population" lifetime	No, but $\mathcal{L}(P_i)$ are known.
$O = E \wedge P$	"Overall" lifetime	No
C	"Censoring" time	No
$T = E \wedge P \wedge C = O \wedge C$	Event time	Yes
$\Delta = \mathbb{1}\{T \leq C\}$	Event status	Yes
$\Gamma = \mathbb{1}\{E \leq P\}$	Cause of death	No

Dependency: Assume C, E and P to be mut. $\perp\!\!\!\perp$; while $\mathcal{L}(P_i)$ are known from life tables.

Goal: Estimate the distribution of E, say by it's hazard $\partial\Lambda_E(t) = -\partial \ln S_E(t)$.

Remark: With the missing cause of death indicatrix, we cannot use directly competing risks analysis..

Same framework: We keep the same $(T_i, \Delta_i)_{i=1,\dots,n}$ sample and same filtration, plus the $\mathcal{L}(P_1), \dots, \mathcal{L}(P_n)$ information.

Previous stochastic processes:

$$\begin{aligned} N(t) &= \mathbb{1}\{O \le t, O \le C\} && \textit{(Uncensored deaths process)} \\ Y(t) &= \mathbb{1}\{O \ge t, C \ge t\} && \textit{(At-risk process)} \\ M(t) &= N(t) - \int_0^t Y(s)\partial\Lambda_O(s) && \textit{(Martingale)} \end{aligned}$$

New stochastic processes:

$$\begin{aligned} N_E(t) &= \mathbb{1}\{E \le t, E \le C\} && \textit{(Excess uncensored deaths process, new)} \\ Y_E(t) &= \mathbb{1}\{E \ge t, C \ge t\} && \textit{(Excess at-risk process, new)} \end{aligned}$$

We similarly defined individual versions N_i, Y_i, M_i, N_{E_i} and Y_{E_i}.

Issue: N_{E_i} and Y_{E_i} are not observable!

1. Integrate out the variable P, using the independence assumption:

$$\mathbb{E}\left(\partial N(t) \mid E, C\right) = \partial N_E(t) S_P(t) + Y_E(t) \partial S_P(t)$$
$$\mathbb{E}\left(Y(t) \mid E, C\right) = S_P(t) Y_E(t)$$

2. Invert the system, denoting $w(t) = S_P(t)^{-1}$, to get:

$$\partial N_E(t) = \mathbb{E}\left(w(t)\partial N(t) - w(t) Y(t) \partial \Lambda_P(t) \mid E, C\right)$$
$$Y_E(t) = \mathbb{E}\left(w(t) Y(t) \mid E, C\right)$$

3. Drop conditional expectations to get observables:

$$\partial \widetilde{N}_E(t) = w(t)\partial N(t) - w(t) Y(t) \partial \Lambda_P(t)$$
$$\widetilde{Y}_E(t) = w(t) Y(t)$$

Result: The Pohar Perme estimator $\partial \widetilde{\Lambda}_E(t) = \frac{\sum_i \partial \widetilde{N}_{E,i}(t)}{\sum_i \widetilde{Y}_{E,i}(t)}$ is unbiaised, convergent, assymp. Gaussian...

[2] Maja Pohar Perme, Janez Stare, and Jacques Estève. "On Estimation in Relative Survival". In: *Biometrics* 68.1 (Mar. 2012), pp. 113–120. ISSN: 0006-341X, 1541-0420. DOI: 10.1111/j.1541-0420.2011.01640.x. (Visited on 11/05/2023).

Relaxing the independence assumption

(In)Dependence assumptions

Assumptions (Standard relative survival assumptions[3])

(i) $C_i \perp\!\!\!\perp (E_i, P_i)\ \forall i$

(ii) $\mathcal{L}(P_i)$ are known from life tables (but diff from each other)

Assumptions (Dependence structure of (E, P))

The $\mathcal{H}_\mathcal{C}$ hypothesis states that all couples (E_i, P_i) have the same survival copula $\mathcal{C}$:

$$\mathcal{H}_\mathcal{C} : \forall i \in 1, ..., n,\ S_{O_i}(t) = \mathcal{C}\left(S_E(t), S_{P_i}(t)\right) \tag{1}$$

Example: Denoting Π the independence copula, $\mathcal{H}_\Pi \iff \forall i\ E_i \perp\!\!\!\perp P_i$ was assumed in previous literature.

Issue: It would be reasonable to assume that $\mathcal{C} \neq \Pi$..

Remark: $\mathcal{C}$ is not identifiable !!

[3] Maja Pohar Perme, Janez Stare, and Jacques Estève. "On Estimation in Relative Survival". In: *Biometrics* 68.1 (Mar. 2012), pp. 113–120. ISSN: 0006-341X, 1541-0420. DOI: 10.1111/j.1541-0420.2011.01640.x. (Visited on 11/05/2023).

Relaxing the independence assumption

Estimation of the excess hazard

Define the following constants:

$$a_i(t) = \mathbb{P}\left(P_i \geq t \mid E_i = t\right) = \mathcal{C}_1\left(S_E(t), S_{P_i}(t)\right)$$
$$b_i(t) = \mathbb{P}\left(P_i = t \mid E_i \geq t\right) = \mathcal{C}_2\left(S_E(t), S_{P_i}(t)\right)\frac{-\partial S_{P_i}(t)}{S_E(t)}$$
$$c_i(t) = \mathbb{P}\left(P_i \geq t \mid E_i \geq t\right) = \mathcal{C}(S_E(t), S_{P_i}(t))\frac{1}{S_E(t)},$$

Then we can integrate out P, solve the system and drop conditional expectations as previously to obtain:

$$\partial\widetilde{\Lambda}_E(t) = \frac{\sum_{i=1}^n \frac{\partial N_i(t)}{a_i(t)} - \frac{b_i(t)Y_i(t)}{a_i(t)c_i(t)}}{\sum_{i=1}^n \frac{Y_i(t)}{c_i(t)}}.$$

Problem: $\partial\widetilde{\Lambda}_E(t)$ is unbiaised, convergent, assymp. Gaussian... but still not observable since constants depend on unknow S_E!

Exception: Under $\mathcal{H}_\Pi$, $\widetilde{\Lambda}_E(t)$ is observable as we already saw.

A differential equation to be solved

Definition (Generalized PPE)

We call *generalized Pohar Perme estimator* the solution $\widehat{\Lambda}_E$ of the differential equation

$$\partial\widehat{\Lambda}_E(t) = \frac{\sum_{i=1}^n \partial\widehat{N}_{E,i}(t)}{\sum_{i=1}^n \widehat{Y}_{E,i}(t)}, \text{ where:} \tag{2}$$

$$\widehat{N}_{E,i}(t) = \frac{\partial N_i(t)}{\widehat{a}_i(t)} - \frac{\widehat{b}_i(t)Y_i(t)}{\widehat{a}_i(t)\widehat{c}_i(t)}, \qquad \widehat{a}_i(t) = \mathcal{C}_1\left(\widehat{S}_E(t), S_{P_i}(t)\right),$$

$$\widehat{Y}_{E,i}(t) = \frac{Y_i(t)}{\widehat{c}_i(t)}, \qquad \widehat{b}_i(t) = \mathcal{C}_2\left(\widehat{S}_E(t), S_{P_i}(t)\right)\frac{-\partial S_{P_i}(t)}{\widehat{S}_E(t)},$$

$$\widehat{S}_E(t) = \exp\left\{-\widehat{\Lambda}_E(t)\right\}, \qquad \widehat{c}_i(t) = \frac{\mathcal{C}\left(\widehat{S}_E(t), S_{P_i}(t)\right)}{\widehat{S}_E(t)}.$$

Remark: Under $\mathcal{H}_\Pi$, $\mathcal{C}(u,v) = uv, \mathcal{C}_1(u,v) = v$ and $\mathcal{C}_2(u,v) = u$, and the differential equation is separable, no need to solve at each time step in the original Pohar Perme estimator.

Relaxing the independence assumption

Second order, asymptotics, tests

Variance estimation

DM decomposition: $\widetilde{\Lambda}_E(t) = \Lambda_E(t) + \Xi(t)$, where $\partial\Xi(t) = \frac{\sum_{i=1}^n \frac{1}{a_i(t)} \partial M_i(t)}{\sum_{i=1}^n \frac{Y_i(t)}{c_i(t)}}$.

Property (Variance of $\widetilde{\Lambda}_E(t)$)

$$\mathrm{Var}\left(\widetilde{\Lambda}_E(t)\right) = \mathbb{E}\left(\left[\Xi\right](t)\right) = \mathbb{E}\left(\int_0^t \frac{\sum_{i=1}^n \frac{1}{a_i(t)^2} \partial N_i(t)}{\left(\sum_{i=1}^n \frac{Y_i(t)}{c_i(t)}\right)^2}\right)$$

Thus, a good estimator for the variance of $\widetilde{\Lambda}_E(t)$ is simply $\left[\Xi\right](t)$.

Definition (Estimator of $\widetilde{\Lambda}_E(t)$'s variance)

$$\widetilde{\sigma}_E^2(t) = \left[\Xi\right](t) = \int_0^t \frac{\sum_{i=1}^n \frac{1}{a_i(t)^2} \partial N_i(t)}{\left(\sum_{i=1}^n \frac{Y_i(t)}{c_i(t)}\right)^2} \quad \text{and} \quad \widehat{\sigma}_E^2(t) = \int_0^t \frac{\sum_{i=1}^n \frac{1}{\widehat{a}_i(t)^2} \partial N_i(t)}{\left(\sum_{i=1}^n \frac{1}{\widehat{c}_i(t)} Y_i(t)\right)^2}$$

Under $\mathcal{H}_\Pi$, $\widetilde{\sigma}_E^2(t)$ is feasible, already obtained in previous litterature. However, under $\mathcal{H}_C$, $\widetilde{\sigma_E^2}(t)$ is not feasible, and thus we propose to use the straightforward plug-in estimator $\widehat{\sigma}_E^2(t)$.

Let $G = \{g_1, .., g_r\}$ be a partition of $1, ..., n$. We want to check the hypothesis:

$$(H_0) : \forall g \in G, \forall i \in g,\ \Lambda_{E_i} = \Lambda_E.$$

Let us denote $\widetilde{Y}_{E,g} = \sum_{i \in g} \widetilde{Y}_{E,i}$ for any group $g \in G$, and $\widetilde{Y}_{E,\bullet} = \sum_{g \in G} \widetilde{Y}_{E,g}$. Similarly, denote $\widetilde{N}_{E,g} = \sum_{i \in g} \widetilde{N}_{E,i}$ and $\widetilde{N}_{E,\bullet} = \sum_{g \in G} \widetilde{N}_{E,g}$.

Define finally the vectors $\boldsymbol{R}(t), \boldsymbol{Z}(t)$, the matrix $\boldsymbol{\Gamma}(t)$ and the test statistic $\widetilde{\chi}(T)$ by:

$$R_g(t) = \frac{\widetilde{Y}_{E,g}(t)}{\widetilde{Y}_{E,\bullet}(t)}$$

$$Z_g(t) = \widetilde{N}_{E,g}(t) - \int_0^t R_g(s) \partial \widetilde{N}_{E,\bullet}(s)$$

$$\Gamma_{g,h}(t) = \sum_{\ell \in G} \int_0^t \left(\delta_{\ell,g} - R_g(s)\right) \left(\delta_{\ell,h} - R_h(s)\right) \sum_{i \in \ell} \frac{\partial N_i(s)}{a_i(s)^2}.$$

$$\widetilde{\chi}(T) = \boldsymbol{Z}(T)' \boldsymbol{\Gamma}(T)^{-1} \boldsymbol{Z}(T)$$

Log-rank test (2/2)

Property

Under (H_0), assuming the existence of an $\epsilon > 0$ such that $a_i(t) > \epsilon$ and $c_i(t) > \epsilon$ over $t \in [0, T]$, we have

$$\widetilde{\chi}(T) \xrightarrow[n\to\infty]{\mathcal{D}} \texttt{Chi2}\left(|G| - 1\right).$$

Lemma (Elements of proofs, using Robolledo's Martingale CLT)

Let $T < \infty$. Under (H_0), assuming that there exists an $\epsilon > 0 :\ a_i(t) > \epsilon, c_i(t) > \epsilon$ over $t \in [0, T]$, the following points hold over $t \in [0, T]$,

(i) $\boldsymbol{Z}$ *is a centered local square integrable martingale,*

(ii) $\mathrm{Cov}(\boldsymbol{Z}(t)) = \mathbb{E}\left(\boldsymbol{\Gamma}(t)\right)$,

(iii) $n^{-1}\boldsymbol{\Gamma}(t) \xrightarrow[n\to\infty]{\mathbb{P}} \boldsymbol{V}(t)$, $\boldsymbol{V}$ *is deterministic, and both* $\boldsymbol{\Gamma}(t)$ *and* $\boldsymbol{V}(t)$ *are semi-definite positives,*

(iv) $n^{-\frac{1}{2}}\boldsymbol{Z}(t) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \boldsymbol{V}(t)\right)$,

(v) $\mathrm{Ker}(\boldsymbol{V}(t)) = \mathrm{Vect}(1)$.

Relaxing the independence assumption

Short example

Dataset: Colorectal cancer

The dataset we have consists of french patients with colorectal cancer, well described in Wolski & Al[4]. See also this page of `NetSurvival.jl`'s documentation.

Characteristics of the dataset:

- 10 years of follow-up before administrative censoring
- Demographic covariates to fetch P_i's distribution: age, sex, date of birth.
- Extra covariates: the primary tumor location, left or right.

Main question on this data: Does the tumor location affect significantly the net survival ?

State of the art: Previous literature, restricted to $\mathcal{H}_\Pi$, conclude that it does not. But $\mathcal{H}_\Pi$ is known to be false..

[4] Anna Wolski, Nathalie Graffféo, Roch Giorgi, and the CENSUR working survival group. "A Permutation Test Based on the Restricted Mean Survival Time for Comparison of Net Survival Distributions in Non-Proportional Excess Hazard Settings". In: *Statistical Methods in Medical Research* 29.6 (June 2020), pp. 1612–1623. ISSN: 0962-2802, 1477-0334. DOI: 10.1177/0962280219870217. (Visited on 12/13/2023).

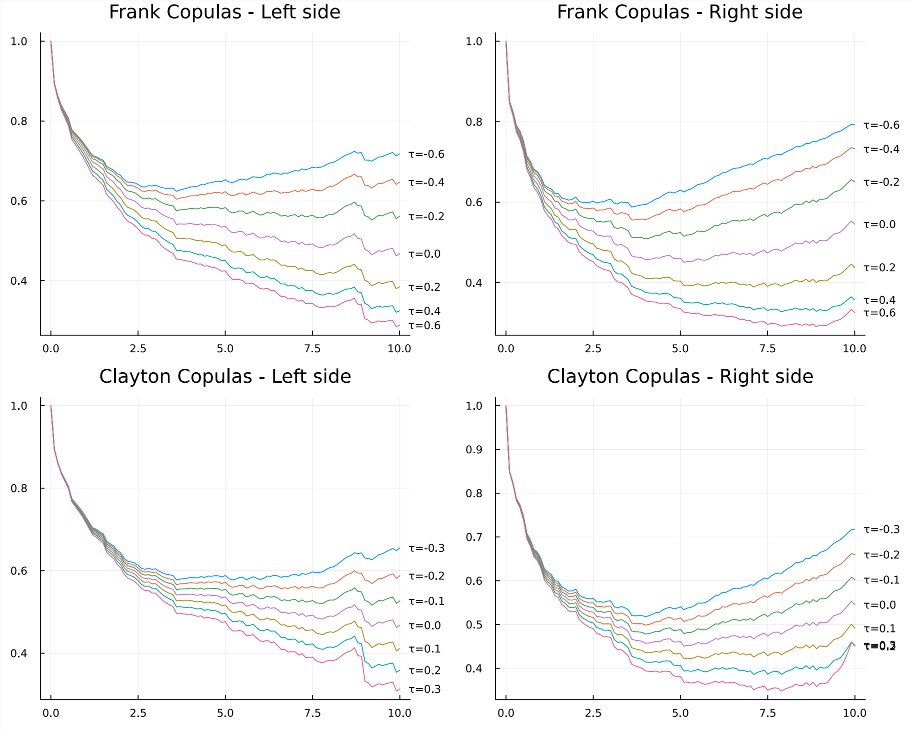

Figure 1: $\widehat{S}_E$ for several $\mathcal{H}_\mathcal{C}$. Data was split w.r.t. tumor location (left or right), and several copulas $\mathcal{C}$ are proposed: Frank copulas (top), Clayton copulas (bottom), with varying Kendall τ. In each graph, $\tau = 0 \iff \mathcal{C} = \Pi$

Tests results for several Frank copulas.

Table 1: Obtained p-value for the generalized log-rank-type test for $\mathcal{C} = \texttt{Frank}(\tau)$, at various horizons T (in years).

τ	$T = 3$		$T = 5$		$T = 8$		$T = 10$	
-0.6	0.05266	+	0.20128		0.90222		0.66067	
-0.5	0.03689	*	0.13102		0.77497		0.75530	
-0.4	0.02417	*	0.07991	+	0.64116		0.85883	
-0.3	0.01476	*	0.04493	*	0.49968		0.98195	
-0.2	0.00845	**	0.02329	*	0.35883		0.86804	
-0.1	0.00461	**	0.01127	*	0.23305		0.69194	
0.0	0.00244	**	0.00522	**	0.13575		0.50419	
0.1	0.00129	**	0.00240	**	0.07163	+	0.33148	
0.2	0.00070	***	0.00114	**	0.03537	*	0.19859	
0.3	0.00040	***	0.00058	***	0.01724	*	0.11324	
0.4	0.00025	***	0.00034	***	0.00889	**	0.06671	+
0.5	0.00018	***	0.00023	***	0.00533	**	0.04642	*
0.6	0.00015	***	0.00021	***	0.00435	**	0.04985	*

Remarks on these results

(i) We enforced the same copula on both left and right side...

(ii) Experts think that the true dependence structures should be concordant ($\tau > 0$) in this dataset.

(iii) Same kind of results with Claytons and Gumbels.

Recall: Non-identifibility of $\mathcal{C}$ because of the missing indicatrix...

Testing for misclassified deaths in cancer registries

Relative survival context: cancer registries

Data Source: Population-based studies and/or cancer registries.

Random Variable	Name	Observed ?
E	"Excess" lifetime	No
P	"Population" lifetime	No, but known distribution.
$O = E \wedge P$	"Overall" lifetime	No
C	"Censoring" time	No
$T = O \wedge C$	Event time	Yes
$\Delta = \mathbb{1}\{T \leq C\}$	Event status	Yes
$\Gamma \overset{?}{=} \mathbb{1}\{E \leq P\}$	Cause of death	Yes, but potentially corrupted.

Dependency: Assume C, E and P to be mut. $\perp\!\!\!\perp$; while $\mathcal{L}(P_i)$ are known from life tables.

Problem: The reported Γ's might be wrong.

Goal: Test the null hypothesis $\mathcal{H}_0 : \forall i \;\; \Gamma_i = \mathbb{1}\{E_i \leq P_i\}$.

Dataset, filtration, stochastic processes...

Observations: Let $(T_i, \Delta_i, \Gamma_i)_{i=1,\dots,n}$ be an observed, i.i.d., n-sample.

Filtered probability space: $(\Omega, \mathcal{A}, \{\mathcal{F}_t, t \in \mathbb{R}_+\}, \mathbb{P})$ with $\mathcal{F}_t = \sigma\{(T_i, \Delta_i, \Gamma_i) : T_i \leq t, \forall i \in 1, .., n\}$.

Previous stochastic processes:

$$
\begin{aligned}
N(t) &= \mathbb{1}\{O \leq t, O \leq C\} && \textit{(Uncensored deaths process)} \\
Y(t) &= \mathbb{1}\{O \geq t, C \geq t\} && \textit{(At-risk process)} \\
N_E(t) &= \mathbb{1}\{E \leq t, E \leq C\} && \textit{(Excess uncensored deaths process)} \\
Y_E(t) &= \mathbb{1}\{E \geq t, C \geq t\} && \textit{(Excess at-risk process)}
\end{aligned}
$$

New stochastic processes:

$$
\begin{aligned}
N^e(t) &= \Gamma N(t) && \textit{(Uncensored deaths process -- excess part)} \\
N^p(t) &= (1 - \Gamma) N(t) && \textit{(Uncensored deaths process -- pop part)},
\end{aligned}
$$

We similarly defined individual versions $N_i, Y_i, N_{E_i}, Y_{E_i}$ and N_i^e, N_i^p

Testing for misclassified deaths in cancer registries

Estimators of the excess hazard

Without using Γ, we have the Pohar Perme estimator:

Definition (Pohar Perme estimator)

Without using the cause of death, we can estimate the excess hazard by:

$$\partial\widehat{\Lambda_E}(t) = \frac{\sum_{i=1}^{n} w_i(t)\partial N_i(t) - w_i(t)Y_i(t)\partial\Lambda_{P_i}(t)}{\sum_{i=1}^{n} w_i(t)Y_i(t)}, \text{ where: } w_i(t) = S_{P_i}(t)^{-1}.$$

Property (Facts on $\partial\widehat{\Lambda_E}$)

This estimator is unbiased and convergent. Its Doob-Meyer decomposition writes:

$$\widehat{\Lambda_E}(t) = \Lambda_E(t) + \widehat{\Xi}(t), \text{ where } \widehat{\Xi}(t) = \int_0^t \frac{\sum_{i=1}^{n} w_i(s)\partial M_i(s)}{\sum_{i=1}^{n} w_i(s)Y_i(s)} \text{ is a martingale.}$$

Including Γ: competitive risks estimator

Assuming the reliability of $\Gamma_1, ...\Gamma_n$, due to the independence, we can once again integrate P's out to get:

$$\mathbb{E}\left(\partial N_i^e(t) \mid E_i, C_i\right) = S_{P_i}(t)\partial N_{E,i}(t).$$

Definition (Weighted Kaplan-Meier)

Using the cause of death, we can estimate the excess hazard by:

$$\partial\widetilde{\Lambda_E}(t) = \frac{\sum_{i=1}^n w_i(t)\partial N_i^e(t)}{\sum_{i=1}^n w_i(t)Y_i(t)}. \tag{3}$$

Property (Facts on $\partial\widetilde{\Lambda_E}$)

This estimator is unbiased and convergent. Its Doob-Meyer decomposition writes:

$$\widetilde{\Lambda_E}(t) = \Lambda_E(t) + \widetilde{\Xi}(t), \text{ where } \widetilde{\Xi}(t) = \int_0^t \frac{\sum_{i=1}^n w_i(s)\partial M_i^e(s)}{\sum_{i=1}^n w_i(s)Y_i(s)},$$

Testing for misclassified deaths in cancer registries

Testing strategy

Proposed Test statistic

We have two estimators $\widehat{\Lambda_E}$ and $\widetilde{\Lambda_E}$ of the same hazard function Λ_E. Consider:

$$Z = \widehat{\Lambda_E} - \widetilde{\Lambda_E}.$$

Theorem (Assymptotical test)

The stochastic process Z is centered, asymptotically Gaussian, and, denoting $[Z]$ its quadratic variation process,

$$\widehat{\chi^2}(t) = \frac{Z(t)^2}{[Z](t)} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{Chi2}(1).$$

Note that we have explicit expression for Z and $[Z]$:

$$Z(t) = \int_0^t \frac{\sum_{i=1}^n w_i(s)\left\{\partial N_i^p(s) - Y_i(s)\partial\Lambda_{P_i}(s)\right\}}{\sum_{i=1}^n w_i(s)Y_i(s)} \text{ and } [Z](t) = \int_0^t \frac{\sum_{i=1}^n w_i(s)^2\partial N_i^p(s)}{\left(\sum_{i=1}^n w_i(s)Y_i(s)\right)^2}.$$

A few remarks...

$$Z(t) = \int_0^t \frac{\sum_{i=1}^n w_i(s)\left\{\partial N_i^p(s) - Y_i(s)\partial\Lambda_{P_i}(s)\right\}}{\sum_{i=1}^n w_i(s)Y_i(s)} \text{ and } [Z](t) = \int_0^t \frac{\sum_{i=1}^n w_i(s)^2 \partial N_i^p(s)}{\left(\sum_{i=1}^n w_i(s)Y_i(s)\right)^2}.$$

Remark 1: We are in fact testing the martingality of N_i^p's around Λ_{P_i}'s, which is logic.

Remark 2: If the integrals against ∂N_i^p's are discrete, their compensators are continuous, alike in the Pohar Perme estimator[5]

Remark 3: The test is log-rank inspired, but since Z is Gaussian, other tools could be developed from it.

---

[5] **perme2012estimation**.

Testing for misclassified deaths in cancer registries

Simulations

Simulation framework

We sample $M = 1000$ datasets of $N = 2000$ patients:

- P is extracted from the Slovenian national life table with the folowing demographics:
  - The sex is $\texttt{Uniform\{M,F\}}$.
  - Age at diagnosis is $\texttt{Uniform[45,75)}$.
  - Date of diagnosis is $\texttt{Uniform[1990,2010)}$.
- $E \sim \texttt{Exponential}(10)$ is the excess lifetime
- $C \sim 15 \wedge \texttt{Exponential}(20)$ is the censoring time.

With several modalities on the cause of death reporting:

- a is the rate of truly dead by cancer are wrongly reported dead by other causes.
- b is the rate of truly dead by other causes are wrongly reported dead by cancer.
- Γ is correctly reported when $a = b = 0$, and 100% wrong when $a = b = 1$.

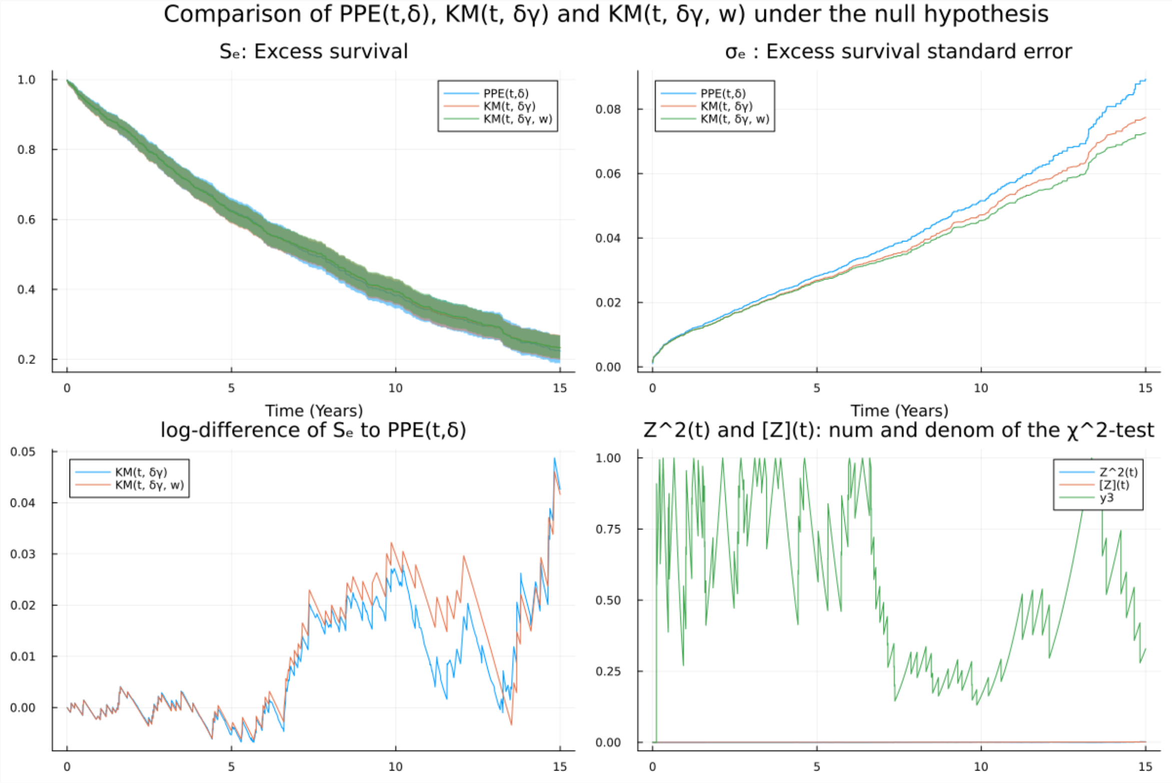

Figure 2: Under the null hypothesis: derivation process of the test statistic

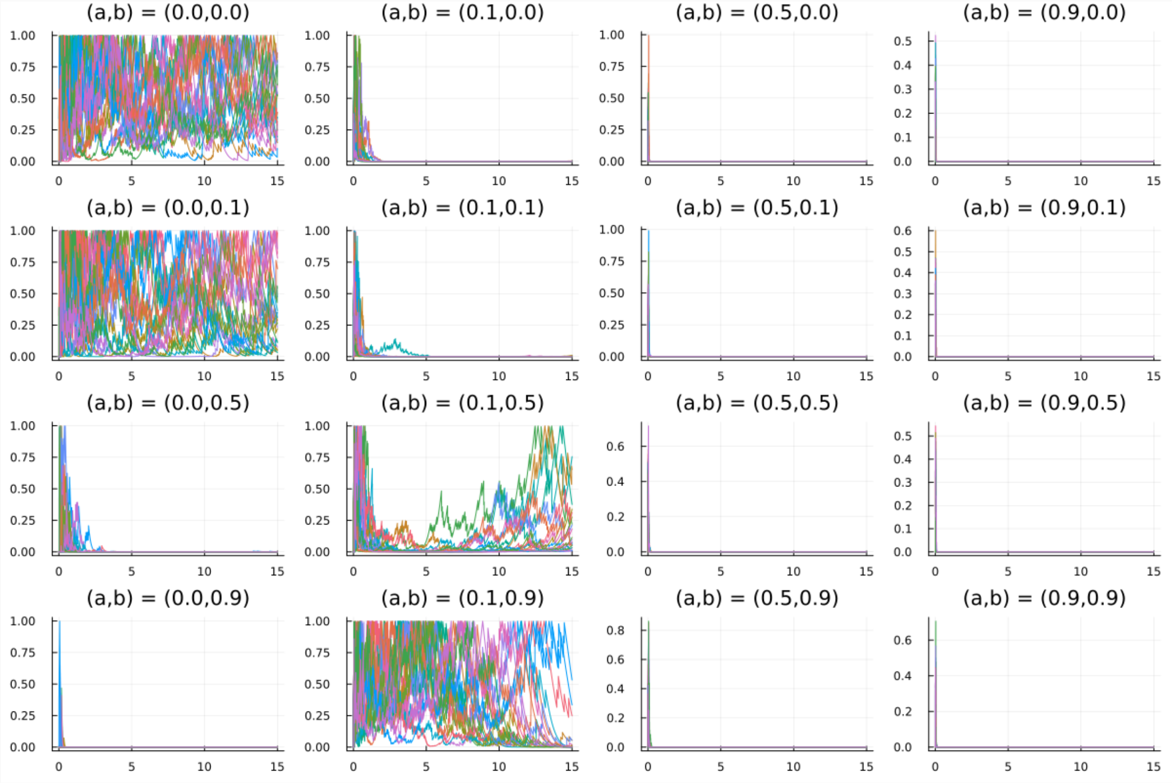

Figure 3: Process $p(t) = S_{\texttt{Chi2(1)}}(\widehat{\chi^2}(t))$ giving the p-value of the test along the time frame.

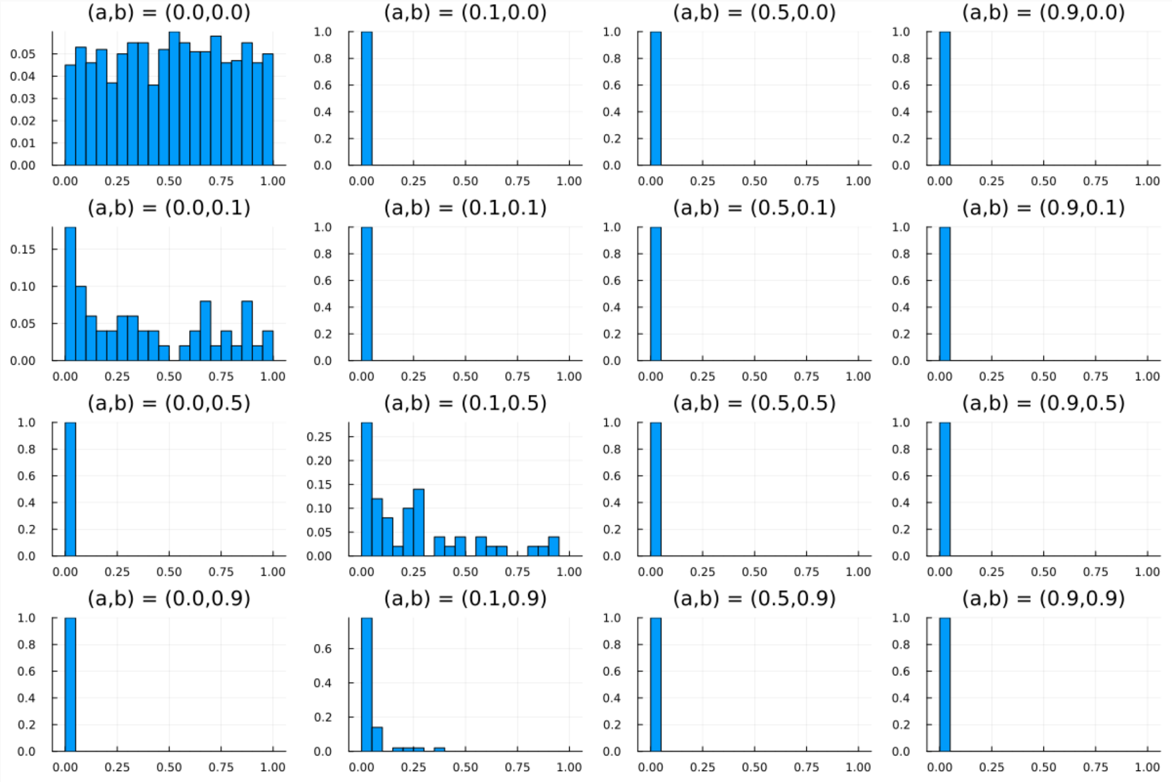

Figure 4: Histogram of the p-value at the end time point $T = 15$.

Conclusion and perspectives

Conclusion

So far:

(i) The relative survival field assumes untrustable Γ's, and thus relies on $\mathcal{H}_{\Pi} : E \perp\!\!\!\perp P$.

(ii) The dependence structure is unidentifiable without trusting Γ, and Γ is untestable without assuming a dependence structure.

(iii) However, even small dependencies ($\tau = 0.2$ or 0.3) can have large impact on results of estimators and tests, and thus on public health decisions.

(iv) Our new test verify consistency between observed Γ and a potential $\mathcal{H}_{\mathcal{C}}$, but not much more.

Shameless propaganda:

(i) Several papers available online, full code at `JuliaSurv/NetSurvival.jl`.

(ii) The `JuliaSurv` community.

(iii) NetPlus & LostLife projects on $L_1, ..., L_n$ i.i.d such that $O_i = P_i - L_i$.

Thanks !