Apprentissage machine: projet applications en recherche clinique

Thomas Ollard (doctorant), Joe de Keizer (doctorant), Rémi Lenain (doctorant),
Julie Drux (master 2) et Yohann Foucher

Séminaire du Laboratoire de Mathématique et Applications (UMR 7348)
Jeudi 17 octobre 2024

Plan

The problematic of the method choice for survival prediction

- The prediction of the probability that a subject experienced an event is often of interest.
- Several regressions can be used for right-censored data. :
  - Most of the studies use proportional hazard (PH)-based assumption.
  - Other models such as accelerated failure time (AFT) approaches are not frequent.
- Machine learning is increasingly used and avoids such modeling assumptions :
  - Random survival forests.
  - Survival neural networks.
  - Support-vector machines.
  - Etc.

A super learner (SL) allows us to combine regressions and algorithms

- In 2011, Polley and van der Laan proposed a survival SL (right censoring). [1]
- Two R packages are available :
  - The first one was proposed by Golmakani et al. (2020). It allows us to obtain the linear predictor of a PH regression.
  - The second one was developed by Westling et al. (2021) with additional learners : several parametric PH models, a generalized additive Cox regression, and a random survival forest.
- **We aimed to extend these packages to additional learners and loss functions.**

1. Super Learning for Right-Censored Data. In MJ van der Laan, S Rose (eds.), Targeted Learning, Springer Series in Statistics.

The survival SL is estimated by minimizing the cross-validated loss function

- $S_m(.)$ is the survival function obtained by the m^{th} learner ($m = 1, ..., M$).
- w_m is the corresponding weight with respect to $\sum_1^M w_m = 1$ and $0 \leq w_m \leq 1$.
- The sample is randomly divided into V cross-validated sub-samples.
- For each of the folds, one can estimate the M learners from the training subjects and predict $\tilde{S}_m(.)$ of the leaving subjects.
- The weights $\hat{w}_m$ are then obtained by minimizing the loss function, i.e., distance between the observations and the predictions $\tilde{S}_{sl}(t \mid Z) = \sum_{m=1}^M w_m \tilde{S}_m(t \mid Z)$.
- The final survival SL is obtained by $\hat{S}_{sl}(t \mid Z) = \sum_{m=1}^M \hat{w}_m \hat{S}_m(t \mid Z)$, where $\hat{S}_m(t \mid Z)$ are estimated on the entire sample.

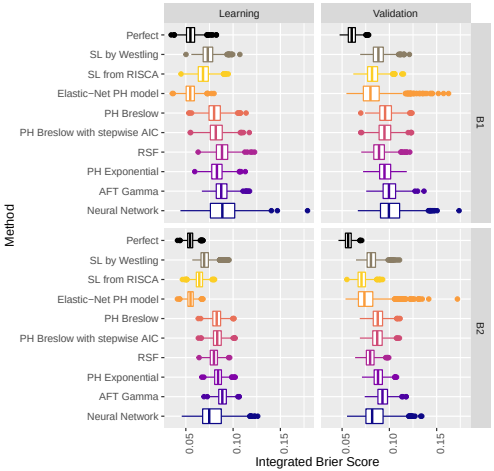
The implemented learners

- Parametric AFT models. (Weibull, Gamma and generalized Gamma distributions).
- Parametric PH models (Exponential or Gompertz distributions).
- Semiparametric PH models with a non-parametric baseline hazard function estimated by using the Breslow estimator (with an option for covariates selection by forward AIC-based selection).
- Penalized PH models (Lasso, Ridge, or Elastic-Net). The quantitative covariates are transformed with B-splines to relax the log-linear assumption.
- Random survival forests.
- Survival neural networks (NN). The linear predictor of the previous semi-parametric PH model is obtained by a single hidden layer network with non-linear activation functions.

The implemented loss functions

- The Brier Score (BS) for right-censored data and a prediction at time t.
- The negative binomial log-likelihood (BLL) for a prognostic at time t.
- The integrated BS and BLL up to the maximum follow-up time.
- The restricted integrated BS and BLL up to a time t.

The performances of the survival SL



Our recent implementation of neural network for survival data in R : the partial logistic regression approach (PLANN)

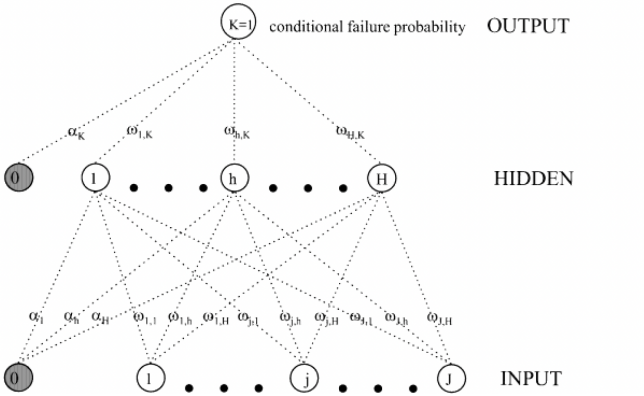


FIGURE – Biganzoli et al. Feed forward NN for t censored survival data. Stat Med. (1998)

The current developments of R packages

- The survival neural network (PLANN) :

`https://github.com/chupverse/survivalPLANN`
`https://cran.r-project.org/package=survivalPLANN`

- The survial super learner :

`https://github.com/chupverse/survivalSL`
`https://cran.r-project.org/package=survivalSL`

Plan

Context

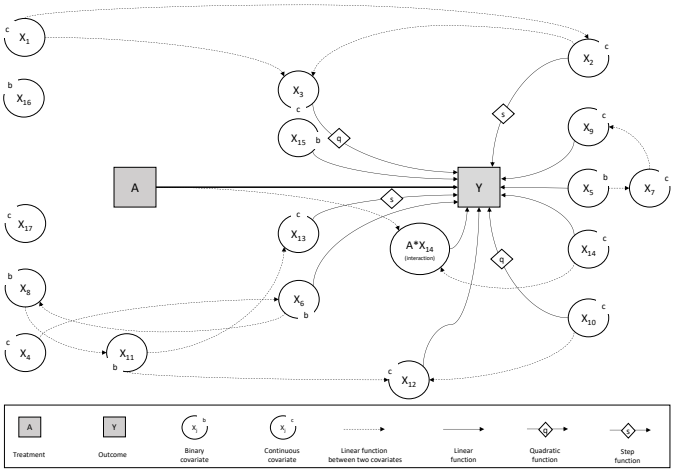
- In a randomized clinical trial with individual randomization (RCT), the random treatment allocation aims to balance the prognostic factors.
- But residual differences due to chance may persist, namely near-confounders.
- Three categories of methods for estimating marginal effects : propensity score, G-computation, and doubly robust estimators).
- Regarding our previous results in observational studies, [2] we hypothesized that G-computation associated with ML could be suitable for RCTs.

2. Le Borgne et al. G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes. Sci Rep (2021)

The G-computation

- Let (Y, A, X) the binary outcome , the 1 :1 randomized treatment, and the set of baseline prognostic factors.
- Fit the outcome model.
- Predict the potential outcomes as if the subjects were included in the two arms.
- Compute the difference of the mean of the potential outcomes.
- We considered several models and algorithms as outcome models : Lasso and Elasticnet logistic regression, neural networks, support vector machine, and super learner of the previous methods.

The simulations



The simulations

Sample size	mOR*	Method	Absolute mean bias				log mÔR				
			$\hat{\pi}_0$	$\hat{\pi}_1$	log mÔR	$\hat{\Delta}$	RMSE	VEB	Coverage	Error†	RSS
$n = 200$	1.9	Unadjusted	0.03%	-0.02%	0.0044	-0.05%	0.29	-0.43%	94.57%	42.58%	0.00%
		Elasticnet	-0.10%	0.03%	0.0108	0.13%	0.24	0.61%	94.54%	26.07%	31.07%
		Lasso	-0.11%	-0.07%	0.0071	0.04%	0.24	3.59%	95.24%	27.19%	27.53%
		Neural network	1.36%	-2.27%	-0.1439	-3.63%	0.27	-5.14%	93.07%	48.56%	-30.94%
		Support vector machine	0.80%	-1.78%	-0.1034	-2.59%	0.24	-9.94%	90.88%	31.22%	18.18%
		Super learner	1.71%	-0.58%	-0.0890	-2.29%	0.24	-7.09%	92.22%	29.16%	22.86%
	1.3	Unadjusted	0.00%	-0.06%	0.0000	-0.06%	0.29	-1.19%	91.96%	86.69%	0.00%
		Elasticnet	-0.11%	-0.08%	-0.0030	0.02%	0.24	-0.42%	95.04%	82.20%	32.90%
		Lasso	-0.13%	-0.12%	0.0025	0.01%	0.24	2.23%	94.65%	83.34%	30.28%
		Neural network	0.55%	-1.73%	-0.0899	-2.27%	0.23	3.16%	94.14%	92.03%	-79.15%
		Support vector machine	0.73%	-0.33%	-0.0415	-1.06%	0.21	-3.04%	94.80%	83.42%	25.90%
		Super learner	1.37%	0.20%	-0.0456	-1.17%	0.21	-3.01%	93.57%	83.94%	20.42%
	1.0	Unadjusted	-0.03%	0.02%	0.0020	0.05%	0.29	-1.38%	94.09%	5.93%	-
		Elasticnet	-0.13%	-0.08%	0.0020	0.05%	0.24	0.21%	94.39%	5.61%	-
		Lasso	-0.15%	-0.09%	0.0024	0.06%	0.24	2.72%	94.83%	5.17%	-
		Neural network	-0.05%	-1.25%	-0.0493	-1.20%	0.21	10.24%	95.76%	4.24%	-
		Support vector machine	0.58%	0.67%	0.0036	0.09%	0.20	-0.02%	94.20%	5.80%	-
		Super learner	1.10%	0.80%	-0.0131	-0.30%	0.20	-0.05%	94.01%	5.99%	-

Conclusions

- G-computation with Elasticnet and splines when appropriate represents a relevant approach for increasing the power of RCTs.
- This approach is particularly effective when the outcome model has significant predictive capacities.
- It remains robust even with very small sample sizes ($n = 60$).
- We should avoid complex outcome models when the sample size is small.

Plan

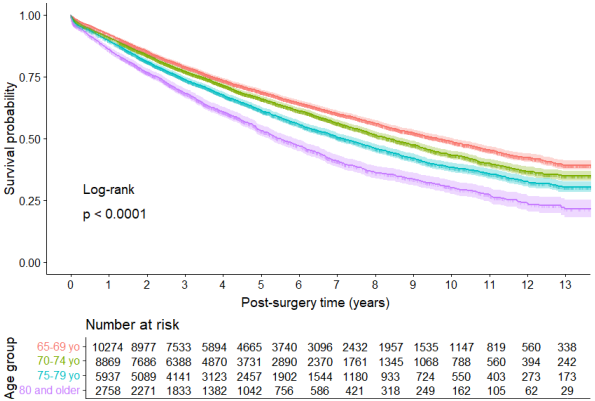
Context

- There has been an increase in the incidence of cancers but a decrease in associated mortality.
- Interest in predicting probability of cure and time to cure for the right to be forgotten and the patient information
- Two competing deaths : $Y = (E, P)$ due or not to the cancer.

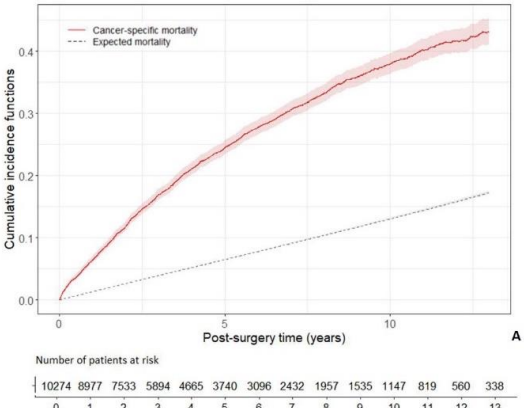
$$\lambda_O(t \mid X_P, X_E) = \tilde{\lambda}_P(t \mid X_P) + \lambda_E(t \mid X_E)$$

  - $\lambda_O(t \mid .)$ is the overall instantaneous hazard at time t.
  - $\tilde{\lambda}_P(t \mid X_P)$ is the population hazard and X_P the predictors : age, sex, and year of diagnosis.
  - $\lambda_E(t \mid X_E)$ the excess hazard and X_E the predictors : age, sex, and type of cancer.

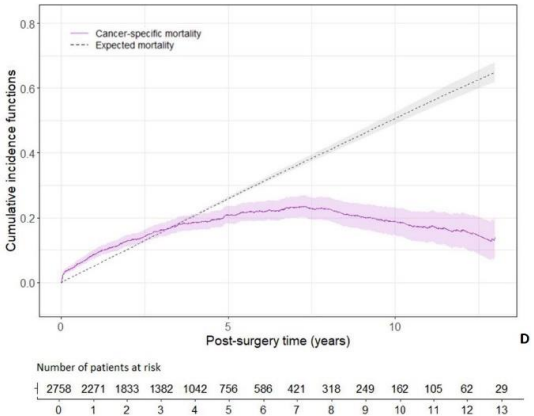
Current application : post-surgery lung cancer (French EPITHOR registry)



Cumulative incidence functions in patients in-between 60 and 69 years old



Cumulative incidence functions in patients older than 80 years



The current flexible strategy for time-to-cure estimation [3]

- Maximum likelihood estimation for each stratum s, c :

$$\lambda_E(t \mid sex = s, cancer = c, age = a) = \lambda_{b,s,c}(t) \exp\Big(\sum_i \beta_{s,c,i}(t) \mathbb{1}\{a \in A_i\}\Big)$$

  - $\lambda_{s,c}(t)$, strata-specific baseline hazard function by cubic splines (4 internal nodes).
  - $A_1 = [45; 55[$, $A_2 = [55; 65[$, $A_3 = [65; 75[$
  - $\beta_{s,c,i}(t)$, strata-specific function by cubic splines (2 internal nodes).

- Avoid assumptions of log-linearity and hazards proportionality.

3. Boussari et al. A new approach to estimate time-to-cure from cancer registries data. Cancer Epidemiology 2018.

Method limitations : small effective per stratum [4]

Cancer site	15-44 years		45-54 years		55-64 years		65-74 years	
	P (95% CI)	TTC (95% CI)	P (95% CI)	TTC (95% CI)	P (95% CI)	TTC (95% CI)	P (95% CI)	TTC (95% CI)
Breast	72 (71-74)	12.1 (11.5-12.7)	80 (80-81)	10.6 (10.0-11.1)	79 (78-79)	11.2 (10.6-11.8)	73 (71-74)	12.3 (11.7-12.9)
Central nervous system	38 (33-42)	10.4 (6.2-14.5)	19 (15-23)	10.8 (3.3-18.3)	NC	NC	NC	NC
Cervix uteri	77 (75-79)	6.5 (4.3-8.7)	63 (60-67)	8.5 (4.6-12.5)	49 (45-53)	11.0 (7.6-14.5)	46 (42-51)	10.7 (6.3-15.2)
Colon [a]	62 (58-66)	7.4 (3.8-11.1)	62 (59-64)	6.8 (4.5-9.2)	58 (56-60)	8.1 (6.1-10.1)	55 (53-57)	8.6 (6.9-10.2)
Colon and Rectum	62 (59-65)	7.6 (4.8-10.4)	59 (57-61)	8.1 (6.2-10.0)	58 (57-60)	8.5 (7.1-10.0)	53 (52-55)	9.2 (7.9-10.5)
Corpus uteri	81 (75-86)	6.8 (0.3-13.3)	77 (73-79)	8.4 (3.6-11.8)	75 (73-77)	8.8 (6.5-11.1)	63 (60-65)	10.9 (8.8-13.0)

FIGURE – Proportion of cured patients (P)and time-to-cure (TTC) in women.

4. Romain et al. Time-to-cure and cure proportion in solid cancers in France. A population based study. Cancer Epidemiology, 2019.

Our proposal : estimate $\lambda_{O,i}(.)$ by ML and deduce the other quantities

- Excess hazard function

$$\lambda_{E,i}(t \mid X_{Ei}) = \lambda_{O,i}(t \mid X_{Ei}, X_{Pi}) - \tilde{\lambda}_{P,i}(t \mid X_{Pi})$$

- Cumulative incidence functions ($k = E, P$) :

$$F_{k,i}(t \mid X_{Ei}, X_{Pi}) = Pr(T_i \leq t, Y_i = k) = \int_0^t S_{O,i}(u- \mid X_{Ei}, X_{Pi}) \tilde{\lambda}_{k,i}(u \mid X_{Pi}) du$$

- Probability of being cured, i.e. , dying from another cause :

$$P_i(X_{Ei}, X_{Pi}) = F_{P,i}(\infty \mid X_{Ei}, X_{Pi})$$

- Probability of being cured in patients alive at time t :

$$Pr(Y_i = P \mid T_i \geq t) = \frac{Pr(T_i \geq t \mid Y_i = P) Pr(Y_i = P)}{Pr(T_i \geq t)} = \frac{P_i(X_{Ei}, X_{Pi})}{S_{E,i}(t \mid X_{Ei})}$$

Application in patients with colon cancer in the Slovene registry (R data)

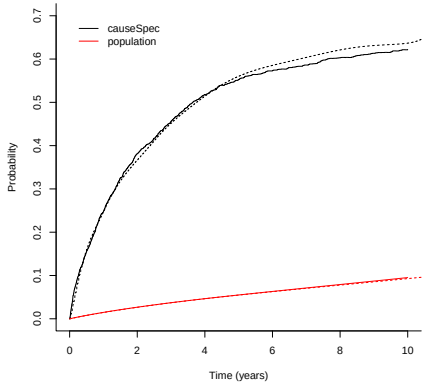


FIGURE – Marginal cumulative incidence functions $F_P(t)$ and $F_E(t)$. Solid line : the Pohar-Perm non-parametric estimation. Dotted line :the Survival Neural Network.

The current developments of R packages

- The survival neural network with relative survival :

`https://github.com/chupverse/survivalPLANN`
`https://cran.r-project.org/package=survivalPLANN`

- The flexible B-spline approach :

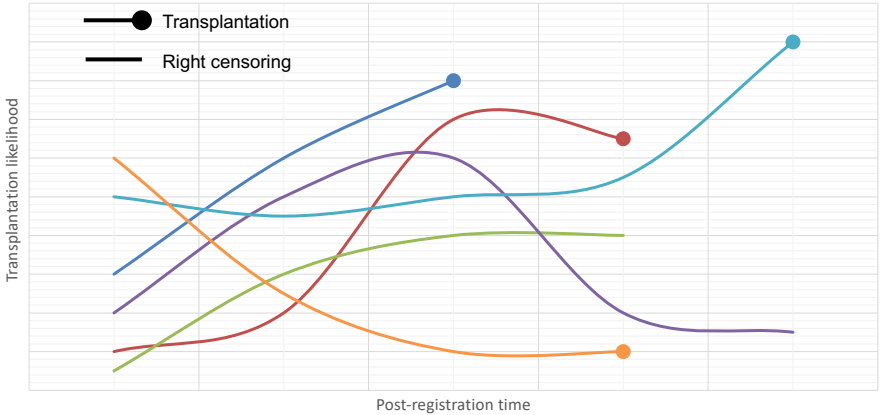
`https://github.com/chupverse/survivalNET`

Plan

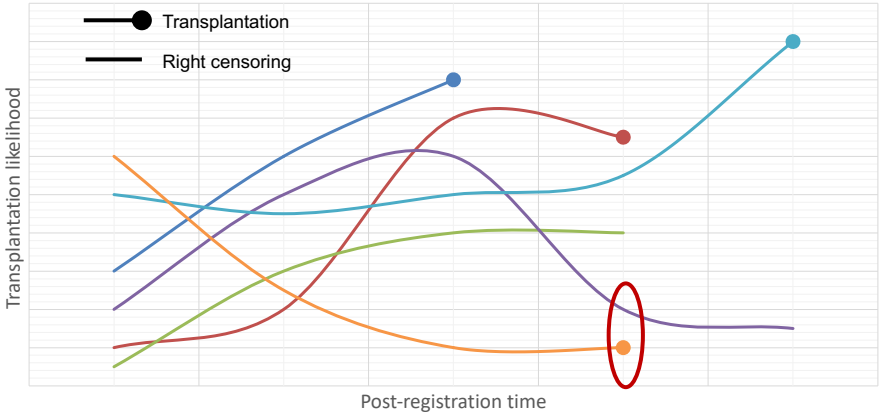
Context

- The estimation of the impact of a treatment/exposure versus no intervention from real life data.
- Example : the impact of transplantation compared to waiting in dialysis.
- Two main issues :
  - The unbalanced time-dependent confounders at the time of the decision.
  - The absence of event to determine the baseline in the non-intervention group.

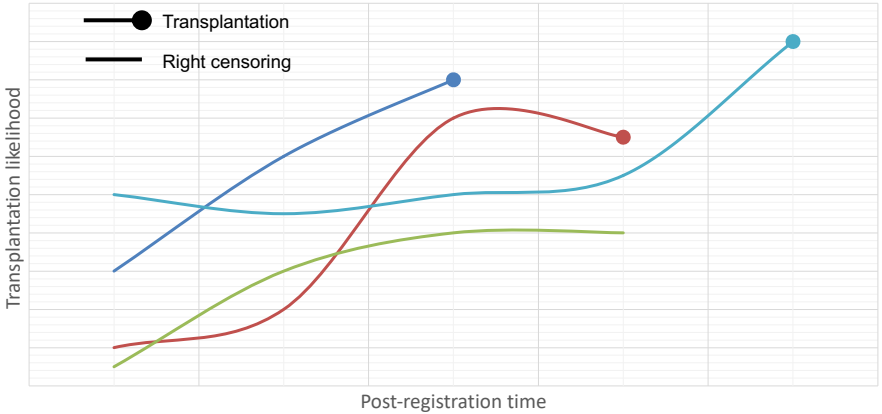
Matching on time-dependent propensity scores



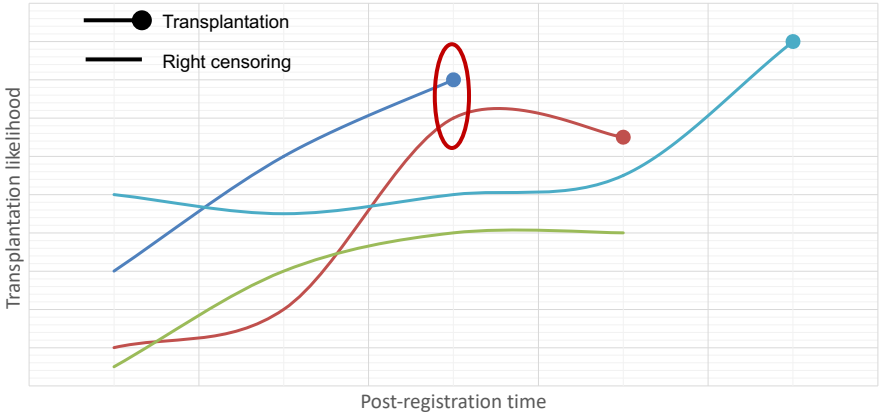
Matching on time-dependent propensity scores



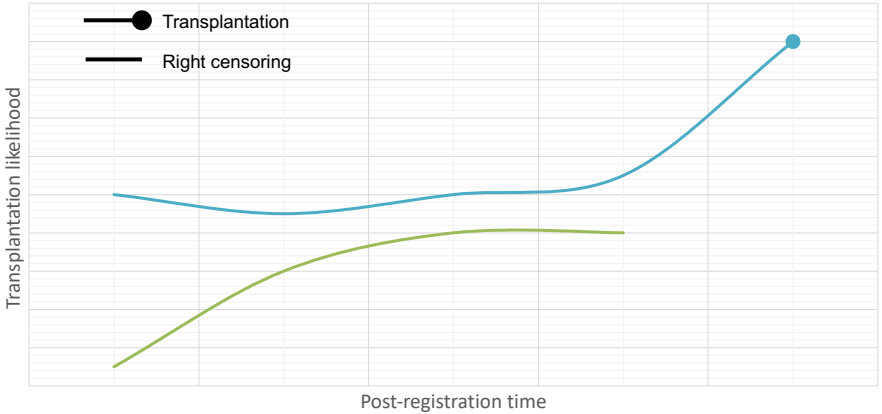
Matching on time-dependent propensity scores



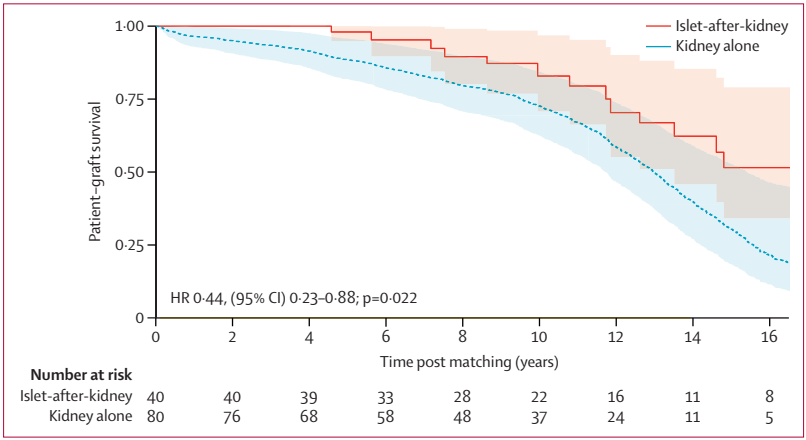
Matching on time-dependent propensity scores



Matching on time-dependent propensity scores



Islet-after-kidney transplantation versus kidney alone (type 1 diabetes) [5]



5. Maanaoui, et al. The Lancet Diabetes and Endocrinology, 2024.

The issue when modeling time-dependent propensity scores

- The time-dependent predictors : $X(t)$.
- The time-varying effects of the time-dependent predictors : $\beta(t)$.

$$\lambda(t \mid X(t)) = \lambda_0(t) \exp(\beta(t) X(t))$$

- The "manual" modeling of such an updated Cox model is complex.
- Its construction cannot be included in the estimation of confidence intervals by bootstrapping.

Our objective

- To describe the interest of survival SL for estimating the time-dependent propensity scores.

Acknowledgments

TEAM

Rémi Lenain
Nephrologist
Master Modélisation en Pharmacologie Clinique et Epidémiologie
PhD defence in december 2024

Joe de Keizer
Master Statistique et données du vivant
PhD student

Thomas Ollard
Master Statistique et données du vivant
PhD student

Julie Drux
Master Modélisation en Pharmacologie Clinique et Epidémiologie
Medical intern

Amina Belkebir-Mekki
Engineer from ENSAI
PhD beginning in november 2024

GRANTS

PEPR santé numérique
ANR-22-PESN-0003
SMATCH

Recherches sur le cancer en Sciences Humaines et Sociales, Épidémiologie, Santé publique Subventions doctorales
DOCSHS23-025

Année de recherche 2023-2024

ANR-10-COHO-002

AAP DATEA 2023