Machine learning for survival data prediction: Application of the super learner on pseudo-observations

Ariane Cwiling, Vittorio Perduca, Olivier Bouaziz

Laboratoire MAP5, Université Paris Cité

November 17th, 2022

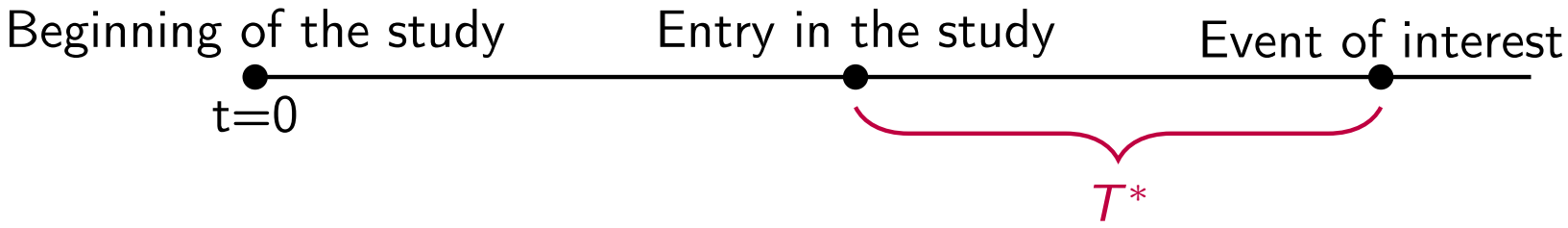
Université Paris Cité

data intelligence institute of Paris

Survival analysis

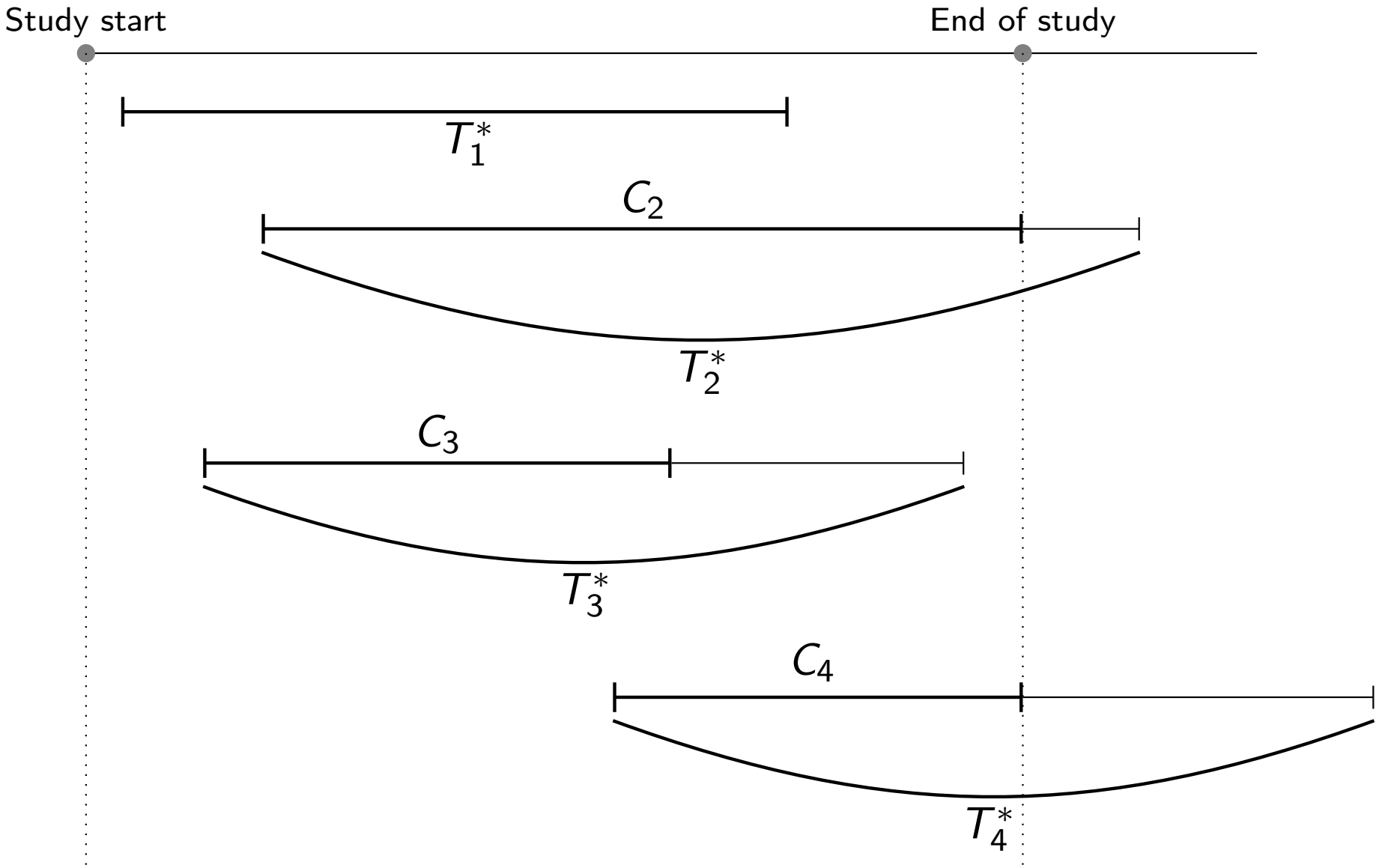
▷ In healthcare, time periods are often of interest: time to relapse of Leukemia patients, time to onset of cancer, time to death...

▷ Formally, we seek to study a positive continuous time to event variable T^*, the time difference between study entry and event of interest.



▷ Data is collected through long-term studies ⇒ Problem: We do not always witness the event of interest because of censoring (end of follow-up, dropout or delayed entry...).

Background in time to event analysis: right censoring



Notations

▷ For an individual $i \in \{1, \ldots, n\}$, we denote

- T_i^* the time to event of interest
- C_i the censoring time
- $T_i = \min(C_i, T_i^*)$ the observed time
- $\Delta_i = \mathbb{1}(T_i^* \leq C_i)$ the censoring status
- Z_i the covariates (in $\mathbb{R}^d$).

▷ Assumption: $C_i \perp\!\!\!\perp (T_i^*, Z_i)$

▷ The outcome of interest (T_i, Δ_i) is in $\mathbb{R}^+ \times \{0, 1\}$ and is incomplete because of censoring.

⚠ Censored data $\neq$ Missing data

Literature review: Survival models

▷ Survival function : $S(t) = \mathbb{P}(T^* > t)$

- Kaplan-Meier estimator (1958)

↪ Conditionally on covariates : $S(t \mid Z) = \mathbb{P}(T^* > t \mid Z)$

- Cox model (1975)
- Random survival forests (Ishwaran, 2008)

▷ RMST (Restricted Mean Survival Time) : $\mu_\tau^* = \mathbb{E}[T^* \wedge \tau] = \int_0^\tau S(t)dt$

↪ Conditionally on covariates :

$$\mu_\tau(Z) = \mathbb{E}[T^* \wedge \tau \mid Z] = \int_0^\tau S(t \mid Z)dt$$

- Proportional and semi-parametric proportional hazards model (Karrison, 1987, Zucker, 1998)
- Pseudo-observations and GLM (Andersen et al., 2004) or neural networks (Zhao, 2021)

Objective

Aim

▷ Predict the RMST (Restricted Mean Survival Time) conditionally on the covariates:

$$\mu_\tau(Z) = \mathbb{E}[T^* \wedge \tau \mid Z] = \int_0^\tau S(t \mid Z)dt$$

Question

How performing is the method combining pseudo-observations with the super learner ?

▷ Pseudo-observations : Andersen, 2004

▷ Super learner : Van der Laan et al., 2007

Pseudo-observations and conditional RMST

▷ Pseudo-observations are a transformation of censored data into data that can be handled as uncensored. They are defined by

$$\hat{\mu}_{\tau,i} = n \int_0^\tau \hat{S}(t)dt - (n-1) \int_0^\tau \hat{S}^{-i}(t)dt$$

where $\hat{S}$ is the Kaplan-Meier estimator using all n data and $\hat{S}^{-i}$ is the same estimator without the i-th subject.

▷ Jacobsen and Martinussen (2016) showed that

$$\mathbb{E}\left[\hat{\mu}_{\tau,i} \mid Z_i\right] = \mathbb{E}[T_i^* \wedge \tau \mid Z_i] + o_{\mathbb{P}}(1).$$

The idea is then to replace the incompletely observed $T_i^* \wedge \tau$ by $\hat{\mu}_{\tau,i}$ and regress them against the covariates.

Motivation behind the super learner

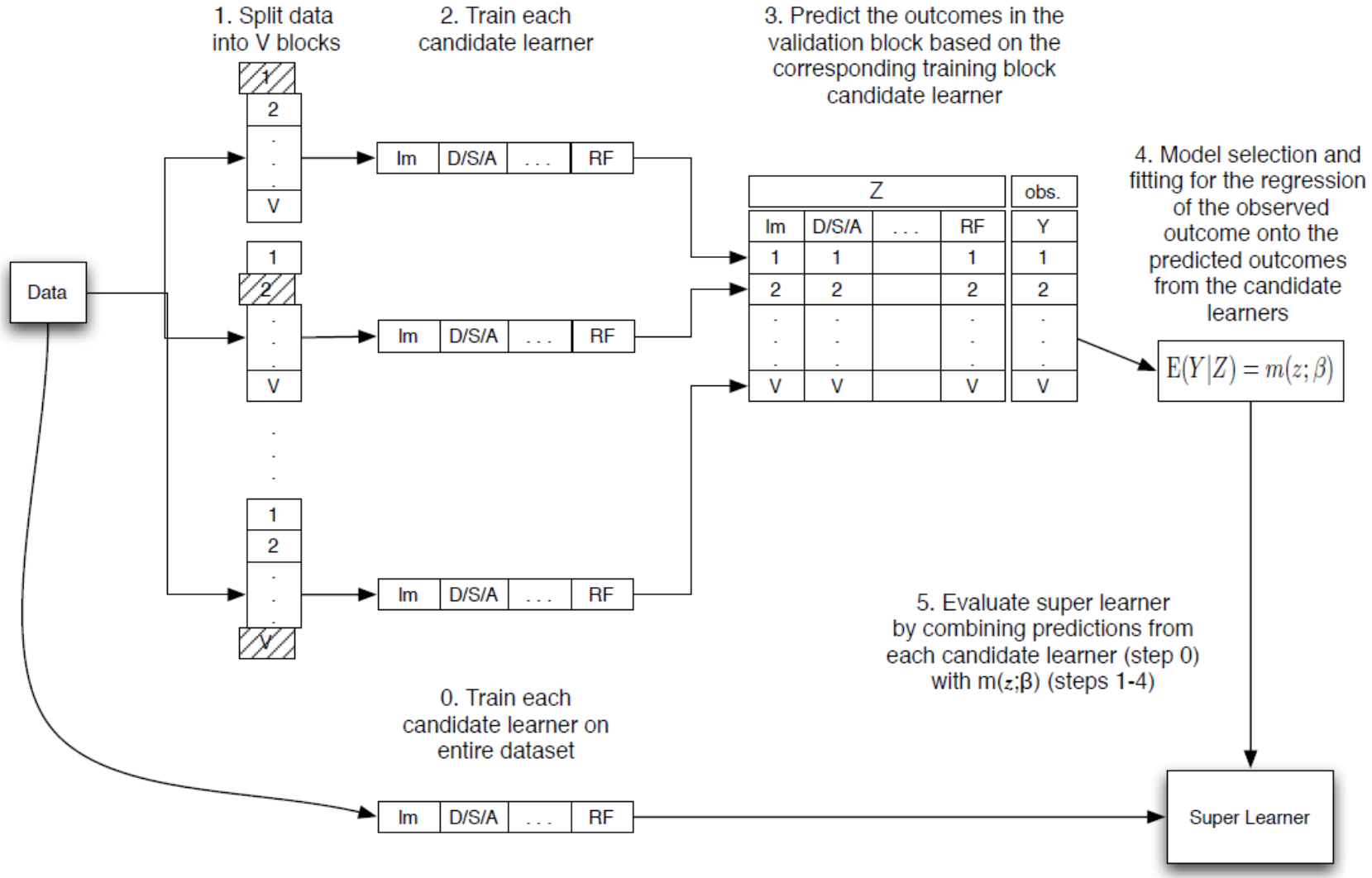
▷ Which regression algorithm to use ?

↪ Cross-validation can select an optimal regression method in a list of candidate learners: Discrete Super Learner

▷ Can we draw information from a whole library of learners ?

↪ Fit a weighted combination of many learners: Continuous Super Learner

Super Learner (Van der Laan et al., 2007)



Interest of the Super Learner

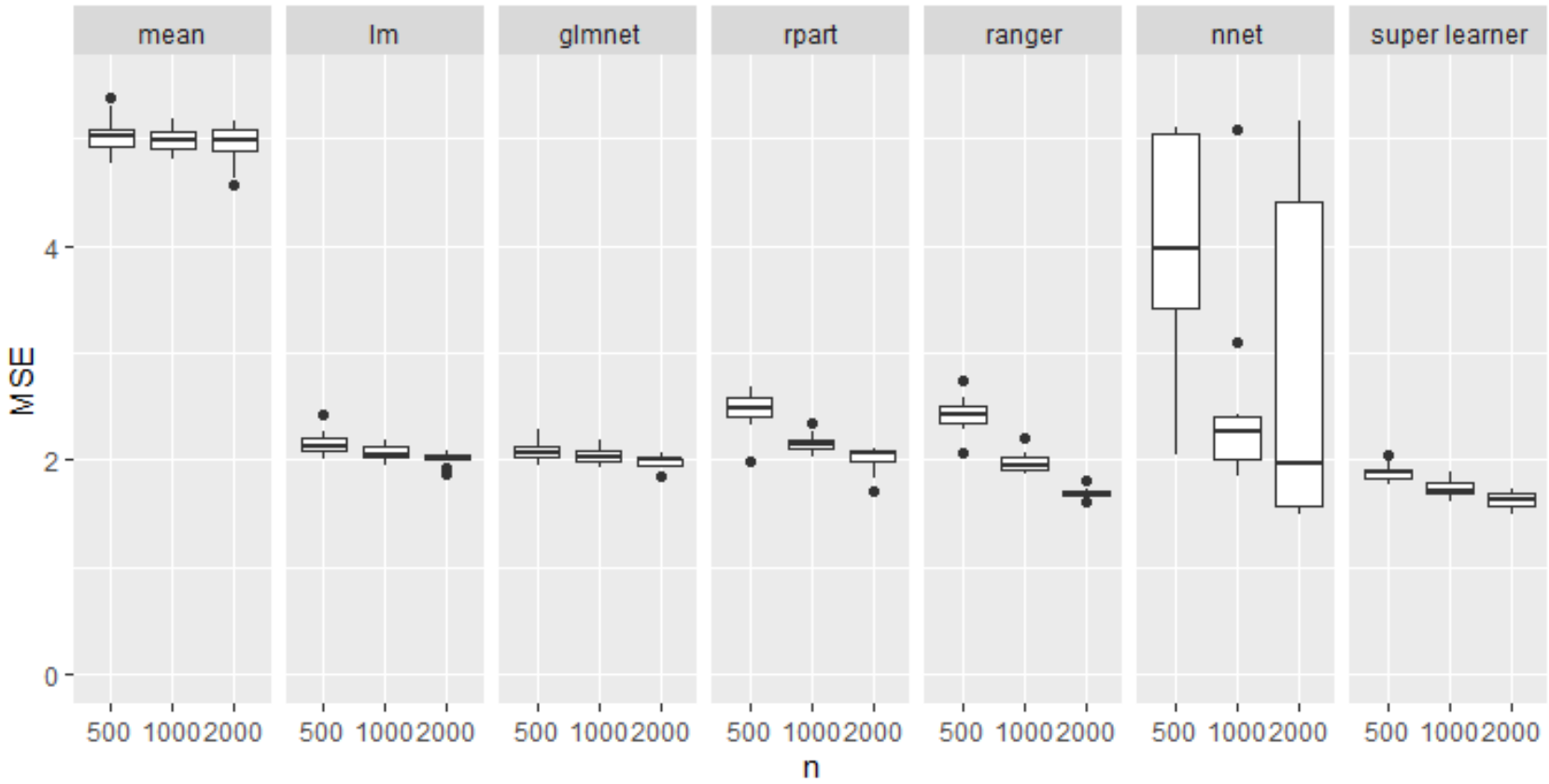
▷ Van der Laan et al. (2007) showed that the super learner performs asymptotically as well as or better than any of the candidate learners.

▷ Golmakani and Polley (2020) adapted the super learner to handle censored data and apply survival estimation methods.

▷ By combining pseudo-observations with the super learner, we can take advantage of a large variety of learners usually used in the context of uncensored data.

Performance on simulations: $\mathbb{E}\left[\left(T^* \wedge \tau - \hat{\mu}_\tau(Z)\right)^2\right]$

▷ Linear model with interactions for event times, and uniform censoring ($\approx 33\%$ censoring).



Evaluation methods for real datasets

▷ The classic MSE

$$\mathbb{E}\left[\left(T^* \wedge \tau - \hat{\mu}_\tau(Z)\right)^2 \;\middle|\; D_{\text{train}}\right]$$

can not be computed on real datasets as we do not know event times for censored data.

▷ IPCW (Inverse Probability Censoring Weights) approach: We assign censoring weights to the observations.

↪ These weights are built on an estimator $\hat{G}$ of the censoring survival function

$$G(t \mid Z) = \mathbb{P}(C > t \mid Z).$$

IPCW estimation of the MSE

▷ We adapted an IPCW estimator from Gerds and Schumacher (2006) called WRSS (Weighted Residual Sums of Squares) to estimate the MSE

$$MSE(\tau, \hat{\mu}_\tau, S) = \mathbb{E}\left[\left(T^* \wedge \tau - \hat{\mu}_\tau(Z)\right)^2 \,\middle|\, D_{\text{train}}\right],$$

resulting in the following estimator:

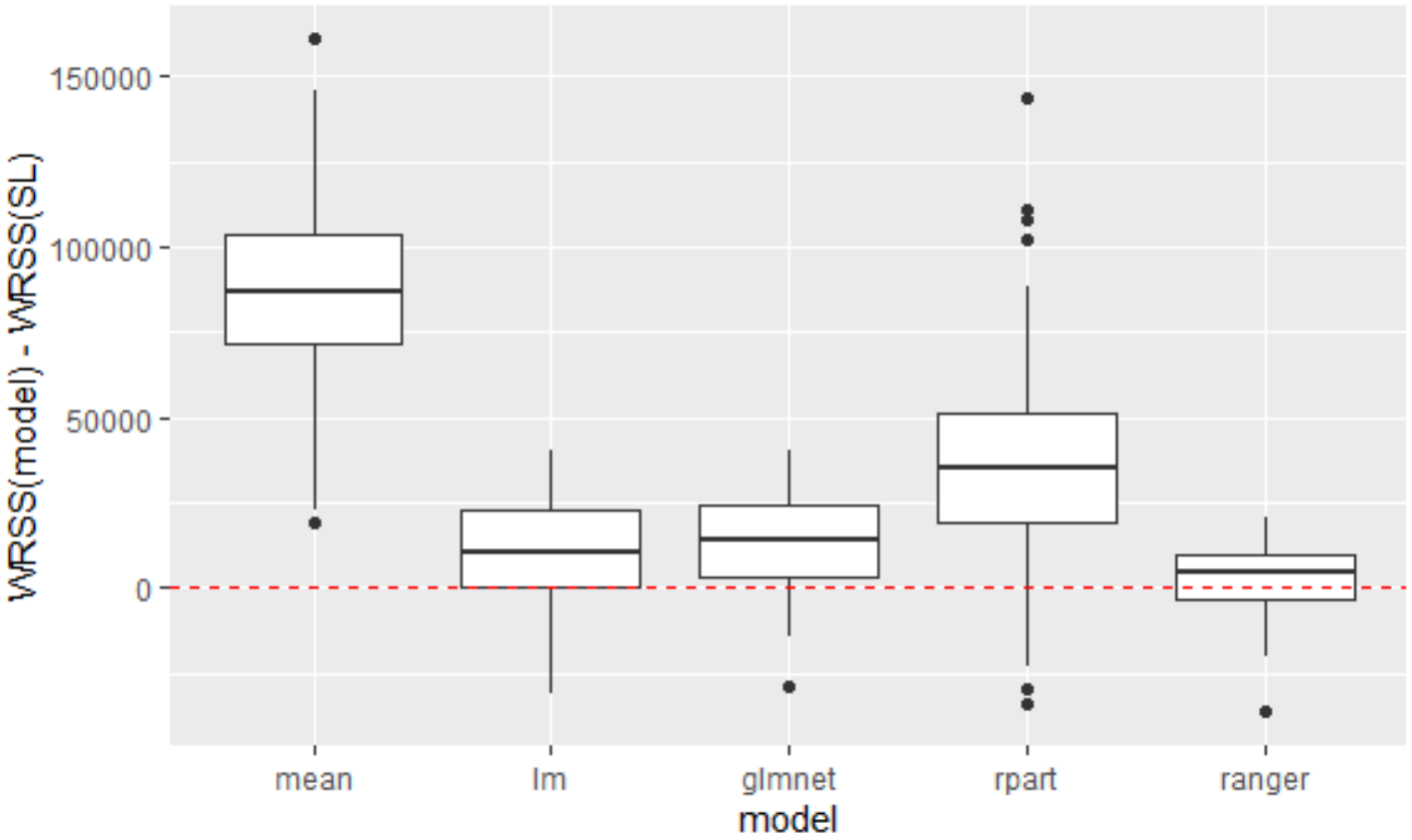
$$WRSS(\tau, \hat{\mu}_\tau, \hat{G}) = \frac{1}{n}\sum_{i=1}^{n}\left(T_i \wedge \tau - \hat{\mu}_\tau(Z_i)\right)^2 \hat{\omega}_i$$

$$\hat{\omega}_i = \frac{\mathbb{1}(T_i \leq \tau)\Delta_i}{\hat{G}(T_i- \mid Z_i)} + \frac{\mathbb{1}(T_i > \tau)}{\hat{G}(\tau \mid Z_i)}$$

▷ We proved that this estimator is consistent under certain conditions.

Breast cancer dataset

▷ German Breast Cancer Study: data from patients with primary node positive breast cancer (1984-1989). Available on R.
↪ 686 data, 8 covariates, 56% censoring



Conformal inference for uncensored data

▷ When the outcome is not censored (consider T^* known), a method to build prediction intervals is to use conformal inference. The goal is to create a prediction band $C \subseteq \mathbb{R}^d \times \mathbb{R}$ based on the observations such that

$$\mathbb{P}\big(T^*_{n+1} \wedge \tau \in C(Z_{n+1})\big) \geq 1 - \alpha.$$

▷ Several algorithms exist, we focused on the split conformal algorithm:

Algorithm 1 Split conformal prediction

1: **Input:** Data (T_i^*, Z_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$, split coefficient $\rho \in (0, 1)$
2: **Output:** Prediction band, over $z \in \mathbb{R}^d$
3: Randomly split $\{1, \ldots, n\}$ into subsets $\mathcal{I}_1, \mathcal{I}_2$ of sizes $n_1 = \rho n$, $n_2 = (1-\rho)n$
4: $\hat{\mu}_\tau = \mathcal{A}(\{(T_i^*, Z_i) : i \in \mathcal{I}_1\})$
5: $R_i^* = |T_i^* \wedge \tau - \hat{\mu}_\tau(Z_i)|, i \in \mathcal{I}_2$
6: $d =$ the kth smallest value in $\{R_i^* : i \in \mathcal{I}_2\}$, where $k = \lceil n_2(1-\alpha) \rceil$, i.e. the $(1-\alpha)$-quantile of the empirical c.d.f. of the residuals defined for all $t \in \mathbb{R}$ by $\hat{\mathcal{R}}_n(t) = 1/n_2 \sum_{i \in \mathcal{I}_2} \mathbb{1}(R_i^* \leq t)$
7: Return $C_{\text{split}}(z) = [\hat{\mu}(z) - d, \hat{\mu}(z) + d]$ for all $z \in \mathbb{R}^d$

Conformal IPCW

▷ We propose a new algorithm combining split conformal and IPCW:

Algorithm 2 Split conformal IPCW prediction

1: **Input:** Data (T_i, Δ_i, Z_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$ for the RMST, regression algorithm $\mathcal{B}$ for the censoring function G, split coefficient $\rho \in (0, 1)$
2: **Output:** Prediction band, over $z \in \mathbb{R}^d$
3: $\hat{G} = \mathcal{B}(\{(T_i, \Delta_i, Z_i) : i \in \{1, \ldots, n\}\})$
4: Randomly split $\{1, \ldots, n\}$ into subsets $\mathcal{I}_1, \mathcal{I}_2$ of sizes $n_1 = \rho n$, $n_2 = (1-\rho)n$
5: $\hat{\mu}_\tau = \mathcal{A}(\{(T_i, \Delta_i, Z_i) : i \in \mathcal{I}_1\})$
6: $R_i = |T_i \wedge \tau - \hat{\mu}_\tau(Z_i)|$ and $\hat{\omega}_i = \frac{\mathbb{1}(T_i \leq \tau)\Delta_i}{\hat{G}(T_i - | Z_i)} + \frac{\mathbb{1}(T_i > \tau)}{\hat{G}(\tau | Z_i)}$, $i \in \mathcal{I}_2$
7: $d =$ the $(1-\alpha)$-quantile of the empirical c.d.f. of the residuals defined for all $t \in \mathbb{R}$ by $\hat{\mathcal{R}}_n^{\hat{G}}(t) = 1/(\sum_{i \in \mathcal{I}_2} \hat{\omega}_i) \sum_{i \in \mathcal{I}_2} \mathbb{1}(R_i \leq t)\hat{\omega}_i$
8: Return $C_{\text{split}}^{\hat{G}}(z) = [\hat{\mu}_\tau(z) - d, \hat{\mu}_\tau(z) + d]$ for all $z \in \mathbb{R}^d$

▷ We proved that for a new i.i.d. pair (T_{n+1}^*, Z_{n+1})

$$\mathbb{P}\big(T_{n+1}^* \wedge \tau \in C_{\text{split}}^{\hat{G}}(Z_{n+1})\big) \xrightarrow[n \to \infty]{} 1 - \alpha.$$

Conformal IPCW

▷ Simulations: $n = 3000$ data allocated at 60% to $\mathcal{I}_1$ and at 40% to $\mathcal{I}_2$.

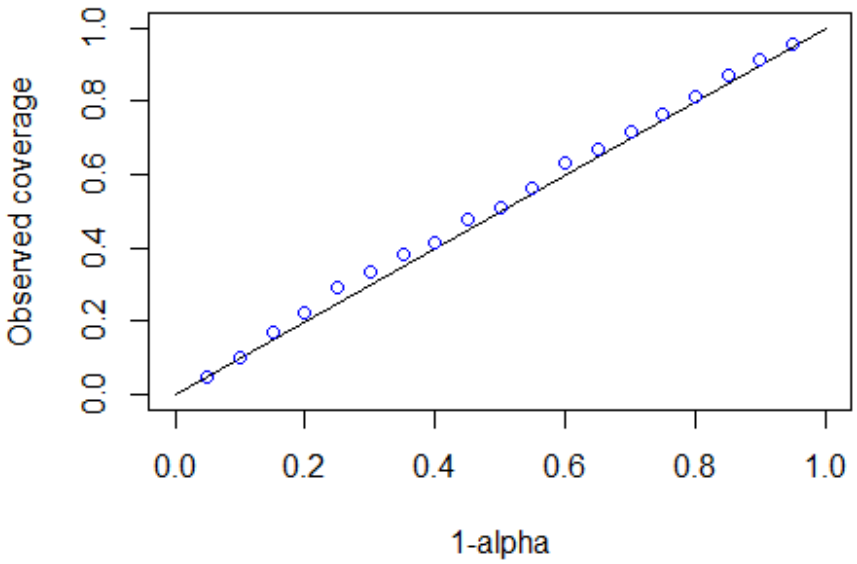


Figure 1: Coverage level $1 - \alpha$ given to the algorithm against observed coverage on an independent set.

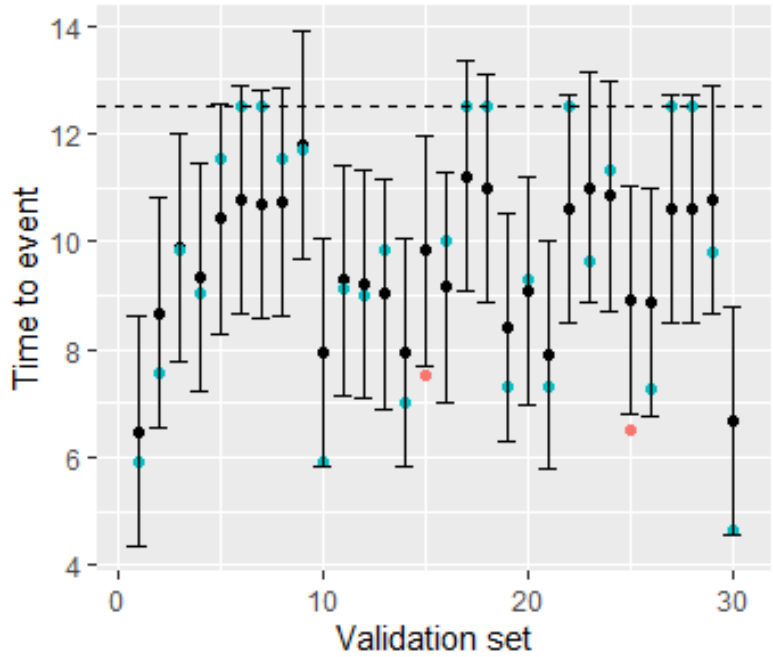


Figure 2: Prediction intervals at coverage level 90% compared to true event times restricted to τ.

▷ We established that fitting a GLM on pseudo-observations or on true event times is asymptotically equivalent in terms of MSE.

↪ Simulation results give us good confidence in the extension of this result to the super learner.

▷ We adapted the estimator from Gerds and Schumacher (2006) to approximate the MSE for real data sets, and proved that their convergence theorem holds.

↪ Though results on real data appear mixed, it may be due to the small size of the dataset.

▷ We combined the split conformal algorithm with the IPCW approach to compute prediction intervals for censored data. We proved that this procedure is asymptotically valid.

↪ We are currently working on extending this algorithm to the estimation of the importance of variables in a prediction model.

▷ We established that fitting a GLM on pseudo-observations or on true event times is asymptotically equivalent in terms of MSE.

↪ Simulation results give us good confidence in the extension of this result to the super learner.

▷ We adapted the estimator from Gerds and Schumacher (2006) to approximate the MSE for real data sets, and proved that their convergence theorem holds.

↪ Though results on real data appear mixed, it may be due to the small size of the dataset.

▷ We combined the split conformal algorithm with the IPCW approach to compute prediction intervals for censored data. We proved that this procedure is asymptotically valid.

↪ We are currently working on extending this algorithm to the estimation of the importance of variables in a prediction model.

Thank you for your attention!

Appendix

Additional slides

Pseudo-observations and GLM for conditional RMST

▷ Let g be an invertible function and suppose for all $k = 1, \ldots, n$,

$$T_k^* \wedge \tau = g(Z_k^T \beta_0) + \epsilon_k, \quad \mathbb{E}[\epsilon_k \mid Z_k] = 0, \quad \mathbb{E}[\epsilon_k^2 \mid Z_k] = \sigma^2$$

▷ Given an estimator $\hat{\beta}_n$ of β_0 and a test individual $\{T_t^* \wedge \tau, Z_t\}$ independent from the training set, the mean squared prediction error can be decomposed into

$$\mathbb{E}[(T_t^* \wedge \tau - g(Z_t^T \hat{\beta}_n))^2 \mid Z_t]$$
$$= Bias(g(Z_t^T \hat{\beta}_n) \mid Z_t)^2 + Var(g(Z_t^T \hat{\beta}_n) \mid Z_t) + \sigma^2$$

▷ Classic setting without censoring: using a GLM estimator, we fit the model on the true event times $T_k^* \wedge \tau$.

- $Bias(g(Z_t^T \hat{\beta}_n) \mid Z_t) = 0$
- $Var(g(Z_t^T \hat{\beta}_n) \mid Z_t) \xrightarrow[n \to \infty]{} 0$

▷ Finally : $\mathbb{E}[(T_t^* \wedge \tau - g(Z_t^T \hat{\beta}_n))^2 \mid Z_t] \xrightarrow[n \to \infty]{} \sigma^2$.

Pseudo-observations and GLM for conditional RMST: new results

▷ Let g be an invertible function and suppose for all $k = 1, \ldots, n$,

$$T_k^* \wedge \tau = g(Z_k^T \beta_0) + \epsilon_k, \quad \mathbb{E}[\epsilon_k \mid Z_k] = 0, \quad \mathbb{E}[\epsilon_k^2 \mid Z_k] = \sigma^2$$

▷ Given an estimator $\hat{\beta}_n$ of β_0 and a test individual $\{T_t^* \wedge \tau, Z_t\}$ independent from the training set, the mean squared prediction error can be decomposed into

$$\mathbb{E}[(T_t^* \wedge \tau - g(Z_t^T \hat{\beta}_n))^2 \mid Z_t]$$
$$= Bias(g(Z_t^T \hat{\beta}_n) \mid Z_t)^2 + Var(g(Z_t^T \hat{\beta}_n) \mid Z_t) + \sigma^2$$

▷ Setting with censoring: using a GLM estimator, we fit the model on the pseudo-observations $\hat{\mu}_{\tau,k}$.

- $Bias(g(Z_t^T \hat{\beta}_n) \mid Z_t) \xrightarrow[n \to \infty]{} 0$
- $Var(g(Z_t^T \hat{\beta}_n) \mid Z_t) \xrightarrow[n \to \infty]{} 0$

▷ Finally : $\mathbb{E}[(T_t^* \wedge \tau - g(Z_t^T \hat{\beta}_n))^2 \mid Z_t] \xrightarrow[n \to \infty]{} \sigma^2$.

IPCW estimation of the MSE

▷ Corollary of Gerds' and Schumacher's theorem: Let $\mathcal{T}$ be a point in time where $G(\mathcal{T} \mid Z) \geq \epsilon > 0$ almost surely. Let $\mathcal{G}$ be a model for the censoring distribution and $\hat{G}_n$ be uniformly consistent i.e.

$$\sup_{G \in \mathcal{G}} \left\{ \int_{\mathbb{R}} \int_0^{\mathcal{T}} \{\hat{G}_n(s \mid z) - G(s \mid z)\} P^X(ds, \cdot, dz) \right\} \xrightarrow[n \to \infty]{} 0.$$

Then if the survival model is consistent and correctly specified, if the censoring model is correctly specified, if the estimator is bounded and under conditional independence,

$$\sup_{\tau \leq \mathcal{T}} |WRSS(\tau, \hat{\mu}_\tau, \hat{G}_n) - MSE(\tau, \mu_\tau, S)| \xrightarrow[n \to \infty]{} 0.$$

Conformal IPCW

▷ Let $q_{1-\alpha}$ be the $(1-\alpha)$-quantile of the true c.d.f of the residuals defined for all $t \in \mathbb{R}$ by

$$\mathcal{R}(t) = \mathbb{P}(R^* \leq t \mid D(\mathcal{I}_1)),$$

where $R^* = |T^* - \hat{\mu}_\tau(Z)|$ and $D(\mathcal{I}_1) = \{(T_i, \Delta_i, Z_i), i \in \mathcal{I}_1\}$.

▷ Let $\hat{q}_{1-\alpha}^{\hat{G}}$ be the $(1-\alpha)$-quantile of the empirical c.d.f. defined for all $t \in \mathbb{R}$ by

$$\hat{\mathcal{R}}_n^{\hat{G}}(t) = \frac{1}{\sum_{i \in \mathcal{I}_2} \hat{w}_i} \sum_{i \in \mathcal{I}_2} \mathbb{1}(R_i \leq t) \hat{w}_i.$$

▷ We established that if $\hat{G}$ is a consistent estimator, then for all $\alpha \in (0,1)$ and conditionally on $D(\mathcal{I}_1)$

$$\hat{q}_{1-\alpha}^{\hat{G}} \xrightarrow[n_2 \to \infty]{a.s.} q_{1-\alpha}$$

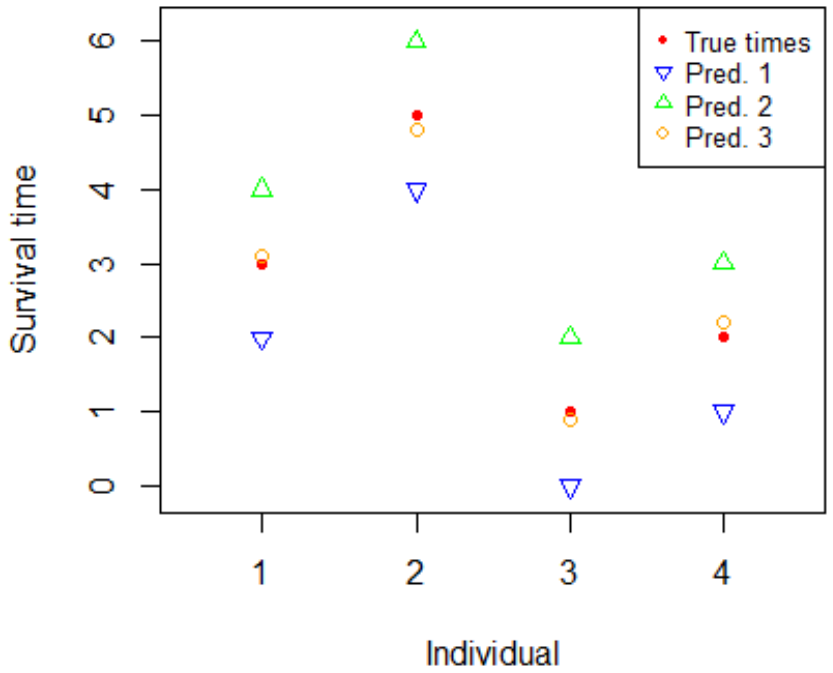
hence for a new i.i.d. pair (T_{n+1}^*, Z_{n+1})

$$\mathbb{P}\big(T_{n+1}^* \wedge \tau \in C(Z_{n+1})\big) \xrightarrow[n \to \infty]{} 1 - \alpha.$$

Weights in the Super Learner

▷ Weights are computed using non-negative least squares based based on the Lawson-Hanson algorithm and normalized so weights sum to one.

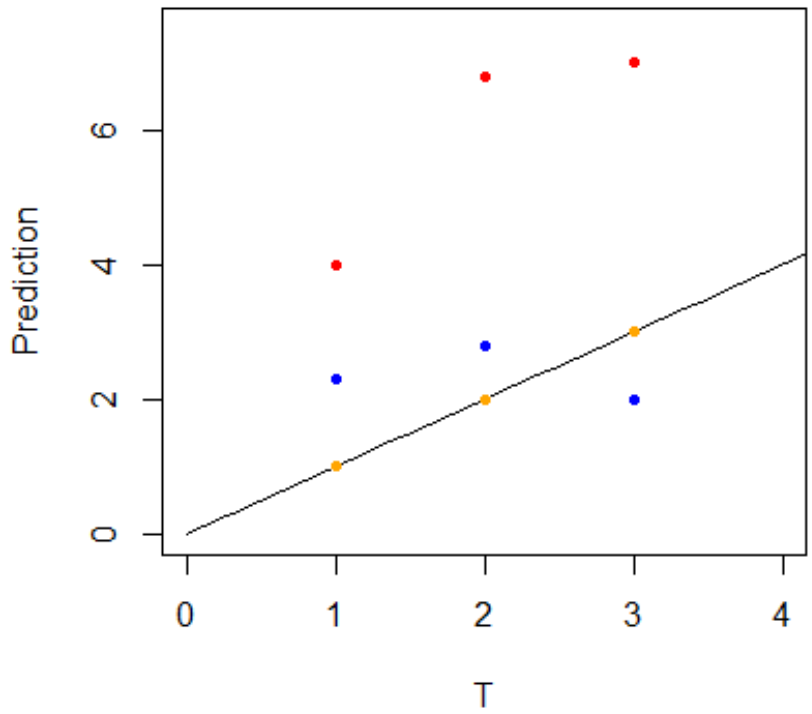
▷ Here the weights are 0.5 for prediction 1, 0.5 for prediction 2 and 0 for prediction 3. Indeed, even if prediction 3 is the closest to reality, the optimal combination is $0.5 * \text{prediction } 1 + 0.5 * \text{prediction } 2$.



Limits of the C-index

▷ The output of the C-index can be misleading: it does not capture the quantitative difference between times and predictions. It can be very encouraging when it should not and vice versa.

▷ Recent papers like the article from Blanche et al. (2018) question the validity of the C-index.



Simulation model