Pseudo-Observations and Super Learner for the Estimation of the Restricted Mean Survival Time

Ariane Cwiling [1] & Vittorio Perduca [1] & Olivier Bouaziz [1]

[1] *Université Paris Cité, MAP5, UMR 8145, France*

ariane.cwiling@u-paris.fr *vittorio.perduca@parisdescartes.fr*
olivier.bouaziz@parisdescartes.fr

Résumé. Il peut être utile pour les cliniciens de prédire le délai avant l'apparition d'un événement tel qu'une rechute, l'apparition d'un cancer ou le décès d'un patient. En présence de données censurées à droite, il est naturel de considérer plutôt un délai restreint en raison des problèmes d'estimation de la queue de distribution. Le problème de prédiction est alors équivalent à l'estimation du temps de survie moyen restreint. À cette fin, nous proposons un nouveau modèle de régression flexible et facile à utiliser, basé sur les pseudo-observations et le super learner. Pour prouver la validité théorique de cette méthode, nous présentons une nouvelle définition des pseudo-observations.

Mots-clés. Analyse de survie, RMST, pseudo-observations, super learner, prédiction.

Abstract. It can be relevant for clinicians to have access to a prediction of the time to an event such as a relapse, a cancer occurrence, or the death of a patient. When predicting the time to event based on right-censored data, it is natural to rather consider a restricted time because of tail estimation issues. The prediction task is then equivalent to the estimation of the restricted mean survival time (RMST). To that aim, we propose a new flexible and easy to use regression model based on pseudo-observations and super learning. To prove the theoretical validity of this method, we present a new definition of the pseudo-observations.

Keywords. Survival analysis, RMST, pseudo-observations, super learner, prediction.

1 Introduction

Survival analysis on right-censored data typically involves estimating hazard and survival functions. However, there is a growing interest in directly predicting the time to event, particularly in medical contexts such as predicting relapse, cancer occurrence, or patient death. Due to tail estimation challenges, predicting up to a relevant fixed time horizon, or equivalently estimating the restricted mean survival time (RMST), is preferred. The RMST is a clinically meaningful quantity that has gained attention for its simplicity and interpretability. In particular, pseudo-observations, introduced by Andersen et al. (2004), have enabled the application of a large range of prediction models by transforming incomplete observed times into data that can be handled as uncensored. This approach facilitates the use of various prediction models adapted to uncensored data, from generalized linear models (Andersen et al., 2004) to neural networks (Zhao, 2021). The best prediction model in a defined library can be selected with cross-validation (Van Der Laan and Dudoit, 2003). A simple improvement to this method is not to select the best model but the best combination of them, called super learning (Van Der Laan et al., 2007). Previous research has explored super learning for analyzing censored data, such as the survival super learner proposed by Golmakani and Polley (2020), albeit with limitations regarding the proportional hazard assumption. Another example is the super learner with Inverse Probability Censoring Weight (IPCW) loss (Keles et al., 2004 ; Devaux et al., 2022), which requires consistent censoring estimation. In this study, we investigate the predictive ability of super learning applied to pseudo-observations for right-censored data, which remains unexplored. To adapt the convergence result of the super learner from Van Der Laan et al. (2007), we introduce a new type of pseudo-observations, called *split* pseudo-observations, primarily for theoretical purposes, as they exhibit similar practical performance to classic pseudo-observations.

2 Pseudo-observations

In the context of right-censored data, we denote by T^* the variable of interest, C the censoring time, $T = T^* \wedge C$ the observed variable and $\delta = \mathbb{1}\{T^* \leq C\}$ the censoring indicator. An observation is represented by the vector $O = (T, \delta, Z)$ where $Z \in \mathbb{R}^d$ is a covariate vector. We note $S(t \mid Z) = \mathbb{P}(T^* > t \mid Z)$ the survival function of T^* conditionally on the covariates Z. Let $\tau_H = \inf\{t > 0 : \mathbb{P}(T > t \mid Z) = 0 \text{ a.s.}\}$. The RMST is defined for a fixed time horizon $\tau \leq \tau_H$, conditionally on the covariates, as

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

Given this definition, the RMST can be estimated for instance by integrating an estimator of the survival function between 0 and τ, or by regressing the restricted event times on covariates. In the second case, censoring must be taken into account since the times T^* are not observed for all individuals. This can be achieved by using pseudo-observations. Consider censored observations $D_n = \{O_i = (T_i, \delta_i, Z_i)\}$, $i = 1, \dots, n$. Classical pseudo-observations are computed in the following way, for a given $\tau < \infty$,

$$\Gamma_i := n \int_0^\tau \hat{S}(t)dt - (n-1) \int_0^\tau \hat{S}^{-i}(t)dt, \ i = 1, \dots, n, \tag{1}$$

where $\hat{S}$ is the Kaplan-Meier estimator of the survival function computed on all data and $\hat{S}^{-i}$ is the same estimator computed on all data but the i-th. The interest of pseudo-observations for regression purposes lies in the following result by Jacobsen and Martinussen (2016),

$$\mathbb{E}[\Gamma_i \mid Z_i] = \mathbb{E}[T^* \wedge \tau \mid Z_i] + \mathbb{E}[\xi_n \mid Z_i], \tag{2}$$

where $\xi_n = o_{\mathbb{P}}(1)$. This result is valid under the following independent censoring assumption.

Assumption 1 (Independent censoring)**.** *The censoring variable C and the pair of variables (T^*, Z) are independent.*

Pseudo-observations are, by construction, correlated with each other, which makes it difficult to study their theoretical properties. To deal with this issue, we propose a new type of pseudo-observations, called *split pseudo-observations*. The idea is to split the data in two subsets D_{n_1} and D_{n_2} of size n_1 and $n_2 = n - n_1$, respectively. The former is used to compute the Kaplan-Meier estimator and the latter for the pseudo-observations. We then define a new type of pseudo-observations as follows :

$$\Gamma_i(D_{n_1}) = \Gamma_{O_i}(D_{n_1}) := (n_1 + 1) \int_0^\tau \hat{S}_{D_{n_1},O_i}(t)dt - n_1 \int_0^\tau \hat{S}_{D_{n_1}}(t)dt, \ i = 1, \ldots, n_2, \tag{3}$$

where O_i represents an observation from D_{n_2}, $\hat{S}_{D_{n_1}}$ is the Kaplan-Meier estimator of the survival function computed on the n_1 data points in D_{n_1} and $\hat{S}_{D_{n_1},O_i}$ is the same estimator computed on the $n_1 + 1$ data points obtained by adding O_i to the sample D_{n_1}. The main advantage of this construction is that the new pseudo-observations constructed for all observations in D_{n_2} are independent conditionally on D_{n_1}. A result similar to Equation (2) can then be easily derived for those split pseudo-observations. Under Assumption 1 :

$$\mathbb{E}[\Gamma_i(D_{n_1}) \mid Z_i, D_{n_1}] = \mathbb{E}[T^* \wedge \tau \mid Z_i] + \mathbb{E}[\xi_{n_1} \mid Z_i, D_{n_1}], \tag{4}$$

where $\xi_{n_1} = o_{\mathbb{P}}(1)$. In Section 3 we establish a convergence result for the super learner coupled with split pseudo-observations. Simulation results are presented in Section 4. In practice, we observe that split and traditional pseudo-observations are very similar, and the choice between them has minimal impact on the prediction quality. Therefore, split pseudo-observations are mostly used for theoretical purposes while traditional pseudo-observations are mostly used in applications.

3 Super Learning

The super learner algorithm is based on cross-validation. During this process, a new dataset is constructed where each observation is paired with a set of predictions from all candidate learners. The most effective algorithm, determined by comparing predictions to observations using a specific loss function, is designated as the discrete super learner. On the other hand, the continuous super learner (simply termed as the "super learner" in what follows) derives the optimal combination of models using a user-chosen algorithm to assign weights to all candidate learners. Theorem 1 in Dudoit and Van Der Laan (2005) shows that the discrete super learner performs asymptotically as well as, or better, than any of the candidate learners, when using a quadratic loss - the mean squared error (MSE). The extension of this result to the continuous super learner is immediate, for it only involves considering the minimum cross-validated risk predictor as based on a parametric regression, as outlined by Van Der Laan et al. (2007). Our aim is to adapt those results to our method, using the MSE as the loss function. This loss allows to compare a wide range of models without making any specific model assumption. Besides, it is well known that, under this loss, the best prediction model is the conditional expectation of the variable of interest, which, in our case, results in estimating the RMST. In order to take into account right-censored data, a first approach is to use an IPCW loss (Keles et al., 2004 ; Devaux et al.,

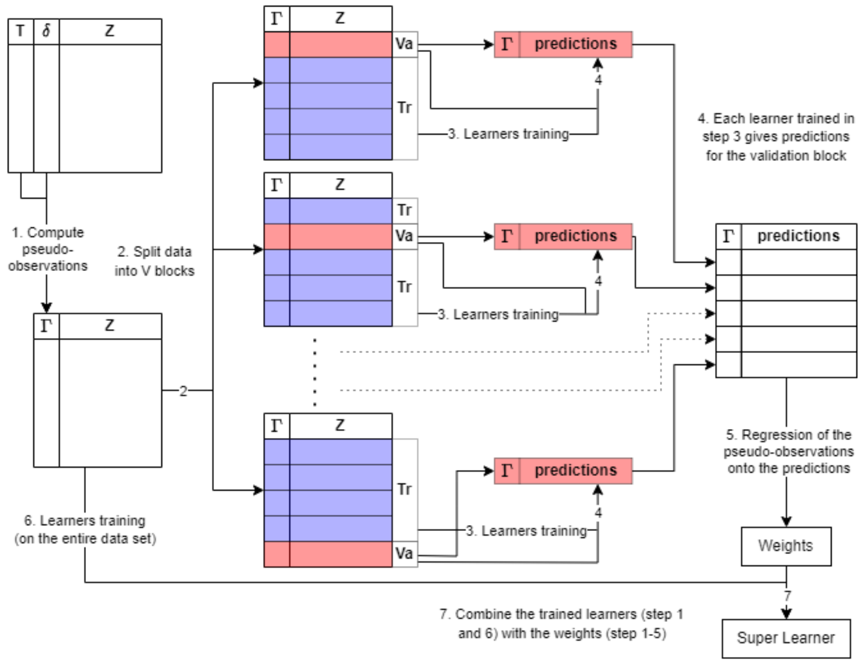


FIGURE 1 – Diagram of the super learner based on standard pseudo-observations for right-censored data, see Equation (1). Pseudo-observations are computed once and for all at the beginning.

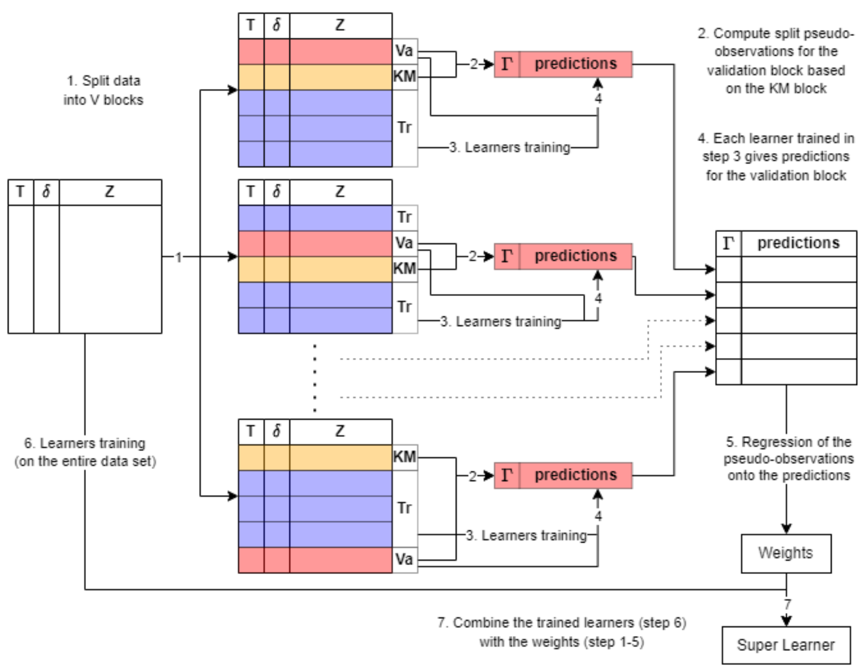


FIGURE 2 – Diagram of the super learner based on split pseudo-observations for right-censored data, see Equation (3). Pseudo-observations are computed during the cross-validation step, for each validation block, based on an additional subset of the data (KM).

2022). Another idea, proposed by Golmakani and Polley (2020), is to minimize the negative log partial likelihood. Our approach consists in feeding pseudo-observations directly into the super learner, using a standard quadratic loss applied on those pseudo-observations. This idea is motivated by the asymptotic result in Equation (2). Our algorithm is therefore identical to the super learner described in Van Der Laan et al. (2007), with an additional first step for computing pseudo-observations. A diagram illustrating the method is provided in Figure 1. However, the dependence structure of the pseudo-observations makes it difficult to provide theoretical results with this method. This is why we also introduce a different algorithm, described in Figure 2, for which we can prove the classical optimality result for the super learner by taking advantage of the conditional independence structure of the split pseudo-observations. The algorithm is slightly different in that case, as it requires using a subset of the data to compute pseudo-observations in the validation set. In practice, we will show in Section 4 that applying the super learner to standard or split pseudo-observations does not make a significant difference in terms of predictive performance. Thus, the theoretical results below are provided with regard to the algorithm based on split pseudo-observations (Figure 2), yet we recommend the use of standard pseudo-observations in practice (Figure 1) for ease of use, lower computational cost and allocation of more data to learners training.

Formally, consider triplets $\{(T_i^*, C_i, Z_i)\}$, $i = 1, \ldots, n$, $T_i^* \in \mathbb{R}^+$ the time to event, $C_i^* \in \mathbb{R}^+$ the censoring time, $Z_i \in \mathbb{R}^d$ the covariates, of joint law $P(t) = \mathbb{P}(T^* \leq t, C \leq c, Z \leq z)$. We only observe a transformation of these triplets defined as the data set $D_n = \{O_i = (T_i, \delta_i, Z_i) = (T_i^* \wedge C_i, \mathbb{1}\{T_i^* \leq C_i\}, Z_i)\}$, $i = 1, \ldots, n$, $T_i \in \mathbb{R}^+$, $\delta_i \in \{0, 1\}$, $Z_i \in \mathbb{R}^d$. We consider K_n candidate estimators for the RMST $\widehat{\psi}_k$, $k = 1, \ldots, K_n$, among which we wish to select the best in terms of risk, using cross-validation. Combining split pseudo-observations with cross-validation imposes to divide the data in three subsets instead of two. Observations are divided according to an independent random vector $B_n = (B_n(i) : i = 1, \ldots, n) \in \{0, 1, 2\}^n$ into a first training set $\{O_i : i, B_n(i) = 0\}$ of size $n_0 = n - \lfloor np_{1,n} \rfloor - \lfloor np_{2,n} \rfloor$ for $p_{1,n}, p_{2,n}, p_{1,n} + p_{2,n} \in (0, 1)$, a second training set $\{O_i : i, B_n(i) = 1\}$ of size $n_1 = \lfloor np_{1,n} \rfloor$ and a validation set $\{O_i : i, B_n(i) = 2\}$ of size $n_2 = \lfloor np_{2,n} \rfloor$. The first training set is used to construct the candidate estimators of the RMST. The second training set is used to compute the Kaplan-Meier estimator which in turn is used for the computation of the pseudo-observations. We refer to this set as the Kaplan-Meier (KM) set. Pseudo-observations are computed for the data in the validation set. We denote $P^0_{B_n}$, $P^1_{B_n}$ and $P^2_{B_n}$ the empirical distributions of the three subsets. Several cross-validation schemes, i.e. distributions for B_n, exist. We focus on V-fold cross-validation, where data are divided into V subsets, or folds, of approximately same size. One by one, each fold serves as a validation set while the remaining folds constitute the training sets. The associated distribution of B_n assigns a mass of $1/V$ to each of the V binary vectors. In this context, Equation (4) can be rewritten as

$$\mathbb{E}[\Gamma_O(P^1_{B_n}) \mid Z, P^1_{B_n}, B_n] = \mathbb{E}[T^* \wedge \tau \mid Z] + \mathbb{E}[\xi_{n_1} \mid Z, P^1_{B_n}, B_n], \tag{5}$$

where $\Gamma_O(P^1_{B_n})$ is the split pseudo-observation constructed for the observation O based on the distribution $P^1_{B_n}$ of the KM set. Consider the quadratic loss function for pseudo-observations

$$L^{\text{po}} : (\psi, P^1_{B_n}, O) \mapsto \big(\Gamma_O(P^1_{B_n}) - \psi(Z)\big)^2, \text{ for a parameter } \psi, P^1_{B_n} \sim P^{\otimes n_1}, O \sim P.$$

The quantity of interest is the *risk* of the parameter ψ for the distributions P and $P^1_{B_n}$,

$$\Theta^{\mathrm{po}}(\psi, P^1_{B_n}, P) = \int L^{\mathrm{po}}(\psi, P^1_{B_n}, o)dP(o).$$

Given this definition, the (unknown) risk minimizer is defined as

$$\psi_1^* = \psi_1^*(P^1_{B_n}, P) = \arg\min_{\psi\in\Psi} \Theta^{\mathrm{po}}(\psi, P^1_{B_n}, P) = \arg\min_{\psi\in\Psi} \int L^{\mathrm{po}}(\psi, P^1_{B_n}, o)dP(o),$$

and characterizes the optimal risk

$$\theta_1^* = \Theta^{\mathrm{po}}(\psi_1^*, P^1_{B_n}, P) = \min_{\psi\in\Psi} \Theta^{\mathrm{po}}(\psi, P^1_{B_n}, P) = \min_{\psi\in\Psi} \int L^{\mathrm{po}}(\psi, P^1_{B_n}, o)dP(o).$$

The cross-validated risk estimator for the k-th candidate learner is defined as

$$\begin{aligned}\widehat{\theta}^{\mathrm{po}}_n(k) &= \mathbb{E}_{B_n}\Theta^{\mathrm{po}}\big(\widehat{\psi}_k(P^0_{B_n}), P^1_{B_n}, P^2_{B_n}\big)\\ &= \mathbb{E}_{B_n}\int L^{\mathrm{po}}\big(\widehat{\psi}_k(P^0_{B_n}), P^1_{B_n}, o\big)dP^2_{B_n}(o)\\ &= \mathbb{E}_{B_n}\frac{1}{n_2}\sum_{i:B_n(i)=2} L^{\mathrm{po}}\big(\widehat{\psi}_k(P^0_{B_n}), P^1_{B_n}, O_i\big).\end{aligned}$$

We wish to select the learner that minimizes this risk. This cross-validated selector is denoted

$$\widehat{k}^{\mathrm{po}} = \arg\min_{k\in\{1,\ldots,K_n\}} \widehat{\theta}^{\mathrm{po}}_n(k).$$

Optimality results are based on the comparison between the cross-validation selector and the selector which for each given data set makes the best choice, knowing the true data distribution (Van Der Laan and Dudoit, 2003). This cross-validated oracle selector minimizes the cross-validated conditional risk

$$\widetilde{\theta}^{\mathrm{po}}_n(k) = \mathbb{E}_{B_n}\Theta^{\mathrm{po}}\big(\widehat{\psi}_k(P^0_{B_n}), P^1_{B_n}, P\big) = \mathbb{E}_{B_n}\int L^{\mathrm{po}}\big(\widehat{\psi}_k(P^0_{B_n}), P^1_{B_n}, o\big)dP(o),$$

and is denoted

$$\widetilde{k}^{\mathrm{po}} = \arg\min_{k\in\{1,\ldots,K_n\}} \widetilde{\theta}^{\mathrm{po}}_n(k).$$

We compared the risk differences $\widetilde{\theta}^{\mathrm{po}}_n(\widehat{k}^{\mathrm{po}}) - \theta_1^*$ and $\widetilde{\theta}^{\mathrm{po}}_n(\widetilde{k}^{\mathrm{po}}) - \theta_1^*$ to demonstrate optimality in a manner similar to Theorem 1 in Dudoit and Van Der Laan (2005).

Theorem 1. *Let $O_1, \ldots, O_n$ be a random sample from a data generating distribution P. Each $O_i = (T_i, \delta_i, Z_i)$ consists of a univariate outcome $T_i \in \mathbb{R}^+$, a binary censoring indicator $\delta_i \in \{0,1\}$ and a covariate vector $Z_i \in \mathbb{R}^d$. Let $\{\widehat{\psi}_k : k = 1, \ldots, K_n\}$ denote a sequence of K_n candidate estimators for the RMST, $\mathbb{E}[T^* \wedge \tau \mid Z]$. If we consider the quadratic loss function $L^{po}(\psi, P^1_{B_n}, O) = \big(\Gamma_O(P^1_{B_n}) - \psi(Z)\big)^2$, then the risk minimizer $\psi_1^*(Z) = \mathbb{E}\big[\Gamma_O(P^1_{B_n}) \mid Z, P^1_{B_n}, B_n\big]$ is asymptotically equivalent to the RMST (see Equation* (5)*). Suppose that*

$$|\Gamma_O(P^1_{B_n})| \leq M < \infty \text{ and } \sup_{Z,\psi\in\Psi} |\psi(Z)| \leq M < \infty \text{ almost surely,}$$

where the supremum is taken over a support of the distribution of Z. Suppose that Assumption 1 holds.

Finite sample result. *Let $M_1 = 8M^2$, $M_2 = 16M^2$ and $c(M,\gamma) = 2(1+\gamma)^2(M_1/3 + M_2/\gamma)$. For all $\gamma > 0$,*

$$0 \leq \mathbb{E}[\widetilde{\theta}_n^{po}(\widehat{k}^{po}) - \theta_1^*] \leq (1+2\gamma)\mathbb{E}[\widetilde{\theta}_n^{po}(\widetilde{k}^{po}) - \theta_1^*] + 2c(M,\gamma)\frac{1+\log(K_n)}{n_2}.$$

Asymptotic result.

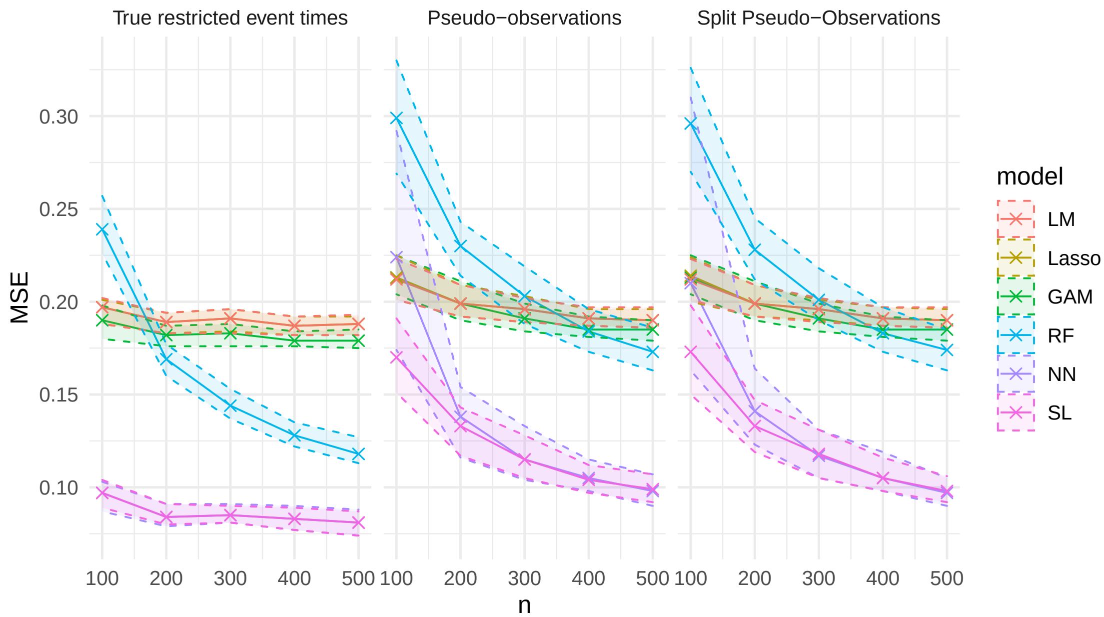
$$\text{If} \quad \frac{\log(K_n)}{n_2(\widetilde{\theta}_n^{po}(\widetilde{k}^{po}) - \theta_1^*)} \xrightarrow[n\to\infty]{\mathbb{P}} 0, \quad \text{then} \quad \frac{\widetilde{\theta}_n^{po}(\widehat{k}^{po}) - \theta_1^*}{\widetilde{\theta}_n^{po}(\widetilde{k}^{po}) - \theta_1^*} \xrightarrow[n\to\infty]{\mathbb{P}} 1.$$

4 Simulations

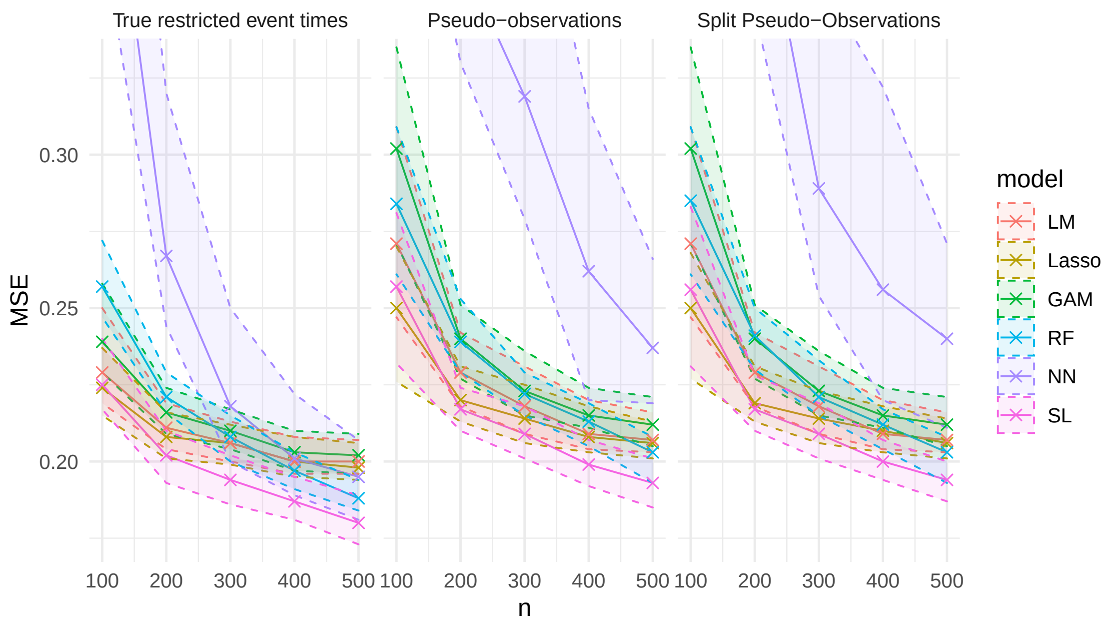
In this section, we present some simulation results that validate our approach. Data are simulated according to two different settings, 1 and 2, corresponding respectively to schemes B and C in Cwiling (2023). Both rely on the Cox model, with different levels of complexity.

Combining standard pseudo-observations with super learning simply implies computing pseudo-observations on the whole data set and feeding them into the super learner as outcomes (see Figure 1). Standard prediction models can be directly applied to the pseudo-observations, thus we chose as candidate learners the linear model (LM), the Lasso, the Generalized Additive Model (GAM), the Random Forest (RF) and the Neural Network (NN). On the other hand, split pseudo-observations require a different methodology as illustrated in Figure 2. In particular, the learners applied to the training set need to handle right-censored observations. In order to be able to compare this methodology with the previous one, we therefore chose to consider candidate learners based on pseudo-observations, combined with LM, Lasso, GAM, RF, and NN.

We simulated data of increasing size $n \in \{100, 200, 300, 400, 500\}$ on which we applied both types of pseudo-observations based super learners. For comparison, we added the classic super learner trained on the true restricted event times. All super learners were trained using 6-folds cross-validation. The MSEs of the super learner and of every candidate learners were computed on an independent test set of size 1000. We repeated the process 80 times for each data size n, simulating new data sets each time. The performances of the candidate learners were evaluated from their last training on the whole data set. The results displayed in Figure 3 show similar performances for both types of pseudo-observations, regardless of the simulation scheme, indicating no significant difference in practice. Figure 3 also illustrates the asymptotic result from Theorem 1, stating that the super learner applied to split pseudo-observations performs asymptotically as well as or better than any of the candidate learners. This also seems to be the case with standard pseudo-observations, even though we did not provide any theoretical result in that case. Thus, in practice, we recommend to compute standard pseudo-observations once and for all and feed them into the super learner. This reduces the complexity of the method and allows to allocate more data to the training of candidate learners.

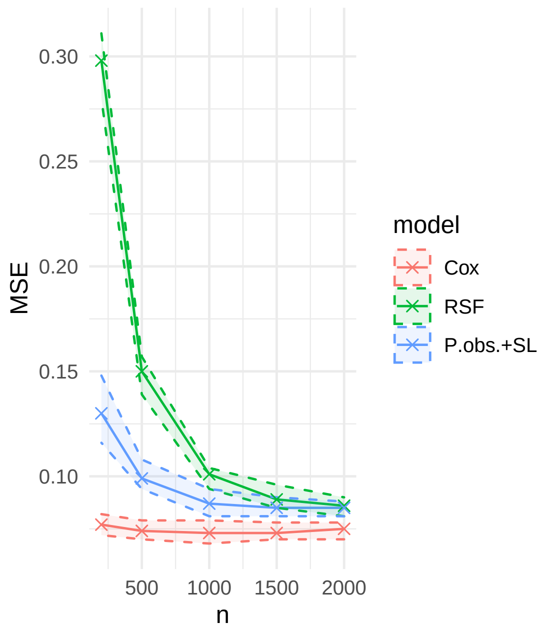


(a) Simulation scheme 1

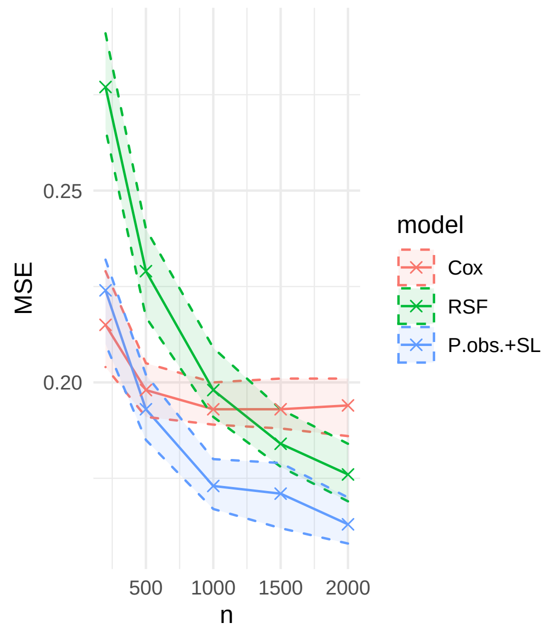


(b) Simulation scheme 2

FIGURE 3 – Application of the super learner on standard and split pseudo-observations and on true restricted event times, for data sets of size n. The MSEs of the super learner and of every candidate learners were computed on a test set of size 1000. The process was repeated 80 times for each data size n, with new data sets each time. Values of the median (cross), first and third quantiles (dashed lines) are reported.



(a) Simulation scheme 1



(b) Simulation scheme 2

FIGURE 4 – Prediction of the restricted time to event with the Cox model (no interaction between covariates is included), the RSF and pseudo-observations combined with super learning. The algorithms were trained on a data set of size n. The MSEs were computed on a test set of size 1000. The process was repeated 80 times for each data size n, with new data sets each time. Values of the median (cross), first and third quantiles (dashed lines) are reported.

As a consequence, in the next simulations, we only selected the method with traditional pseudo-observations and compared its predictive performance with other survival methods. Our method was implemented with the same candidate learners and same number of folds as previously, and compared to two widely used survival methods, namely the Cox model (including no interaction between covariates) and the Random Survival Forests (RSF). These methods are used to estimate the survival function, from which the estimation of the RMST is derived by integrating the survival curve between 0 and τ. We simulated data of increasing size $n \in \{200, 500, 1000, 1500, 2000\}$ on which we applied the three methods. The MSEs were computed on an independent test set of size 1000. We repeated the process 80 times for each data size n, simulating new data sets each time. The results are displayed in Figure 4. The first simulation scheme consists in simulating time to events according to a Cox model without interactions between covariates. Hence, as expected, Figure 4a shows a better performance for the Cox model compared to the other methods. However, the RSF and our method show similar performances for large sample sizes, with our approach providing a better predictive performance than the RSF on the overall. In the second simulation scheme, interactions between covariates are included in the Cox model used to simulate the data. The results are presented in Figure 4b for the Cox model (without interactions), the RSF and our method. It is seen that our method outperforms the Cox model for $n \geq 500$ and the RSF for all sample sizes.

5 Conclusion

This study introduces a novel approach for predicting the restricted time to event from right-censored data by combining pseudo-observations with the super learner. Numerical experiments demonstrated that our method performs at least as well as the best candidate learner and competes favorably with classical survival methods like the Cox model or RSF. The independent censoring assumption and its computation cost are the main limitations of the method. On the other hand, the combination of pseudo-observations with the super learner is a straightforward method and its good results on simulated and real data motivates its use in practice. To take into account the dependence structure of the pseudo-observations, we proposed the conditionally independent split pseudo-observations which allowed us to extend the convergence result of the super learner to right-censored data. The potential behind the split pseudo-observations goes beyond its application to the super learner : it can be used in various fields to address theoretical issues related to their dependence structure.

Bibliographie

Andersen, P. K., Hansen, M. G., and Klein, J. (2004), Regression Analysis of Restricted Mean Survival Time Based on Pseudo-Observations, *Lifetime Data Analysis*, 10(4), pp. 335-350.

Cwiling, A., Perduca, V. and Bouaziz, O. (2023), A Comprehensive Framework for Evaluating Time to Event Predictions using the Restricted Mean Survival Time.

Devaux, A., Genuer, R., Peres, K. and Proust-Lima, C. (2022), Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history : a landmark approach, *BMC Medical Research Methodology*, 22(1), pp. 188.

Dudoit, S. and Van Der Laan, M. J. (2005), Asymptotics of cross-validated risk estimation in estimator selection and performance assessment, *Statistical Methodology*, 2(2), pp. 131-154.

Golmakani, M. K. and Polley, E. C. (2020), Super Learner for Survival Data Prediction, *The International Journal of Biostatistics*, 16(2), pp. 20190065.

Jacobsen, M. and Martinussen, T. (2016), A Note on the Large Sample Properties of Estimators Based on Generalized Linear Models for Correlated Pseudo-observations, *Scandinavian Journal of Statistics*, 43(3), pp. 845-862.

Keles, S., Van Der Laan, M. J. and Dudoit, S. (2004), Asymptotically optimal model selection method with right censored outcomes, *Bernoulli*, 10(6).

Van Der Laan, M. J. and Dudoit, S. (2003), Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated adaptive Epsilon-Net Estimator : Finite Sample Oracle Inequalities and Examples, *U.C. Berkeley Division of Biostatistics Working Paper Series.*

Van Der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007), Super Learner, *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Zhao, L. (2021), Deep neural networks for predicting restricted mean survival times, *Bioinformatics*, 36(24), pp. 5672-5677.