Supplementary Material for: Non-Parametric Individual Treatment Effect Estimation for Survival Data with Random Forests

Sami Tabib, Denis Larocque[*]
Department of Decision Sciences, HEC Montréal, Canada

Contents

1 Estimating the mean survival time by integrating the Kaplan-Meier estimate 2

2 Split control and allowable splits 3
2.1 Parameters used in the simulation study 3

3 Description of the Individual DGPs Used in the Simulation Study 4
3.1 DGP 1 6
3.2 DGP 2 6
3.3 DGP 3 7
3.4 DGP 4 7
3.5 DGP 5 8

4 Global Simulation Results 8
4.1 MSE Results 9
4.2 C-Index Results 10

5 Detailed Simulation Results 11
5.1 MSE Results 12
5.2 C-Index Results 17

6 Estimated treatment effects for AFT models fitted with cumulative deciles samples for all methods for the data examples 22

7 Kaplan-Meier curves for the top 2 ITE deciles for all methods for the data examples 24
7.1 Colon data 25
7.2 German breast cancer data 27

[*] Corresponding author. Department of Decision Sciences, HEC Montréal, 3000 chemin de la Côte–Sainte–Catherine, Montréal (Québec), Canada, H3T 2A7. E-mail:denis.larocque@hec.ca

1 Estimating the mean survival time by integrating the Kaplan-Meier estimate

In the calculation of the splitting criterion (Section 2.2 of the article), we need to estimate the mean survival time by integrating the Kaplan-Meier estimate. Let $\hat{S}_{KM}(\cdot)$ be a generic Kaplan-Meier (KM) estimate and let t_{max} be the largest value where it is defined. If $\hat{S}_{KM}(t_{max}) = 0$, then computing the mean survival time by integrating the KM is straightforward. However, it may happen that the KM is undefined past a certain value, that is $\hat{S}_{KM}(t_{max}) > 0$. One easy way to compute the mean survival time would be to set the KM at 0 after t_{max}, but we use a more sophisticated method in this paper. Basically, we add a tail from the exponential distribution at the end of the KM, as illustrated in Figure 1.

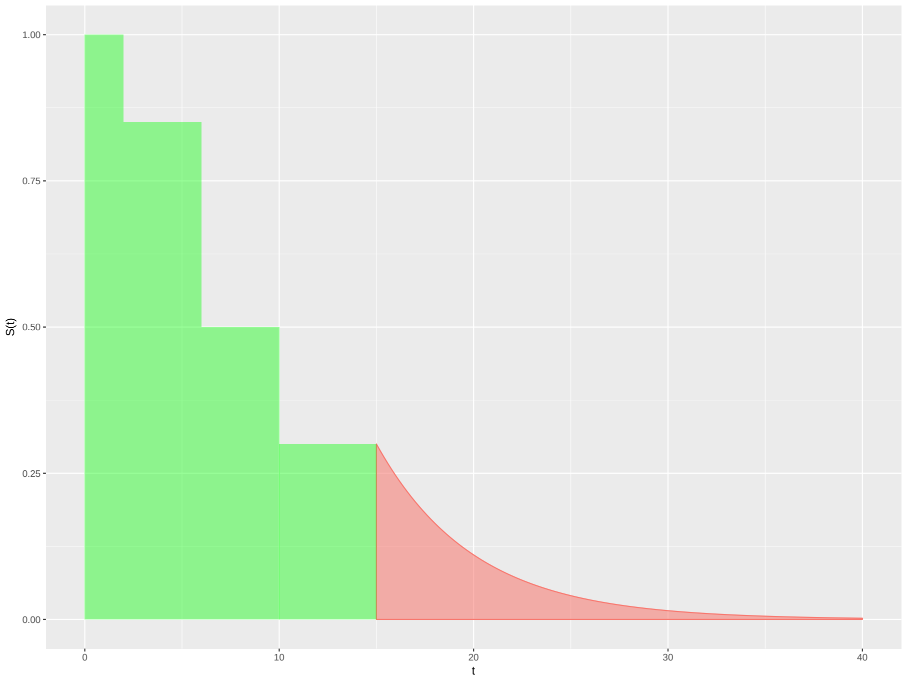

Figure 1: Kaplan-Meier extension when it is undefined past a certain value.

In details, let $S_\lambda(t) = \exp(-\lambda t)$ be the survival function from the exponential distribution with parameter λ. When $\hat{S}_{KM}(t_{max}) > 0$, let $\hat{\lambda}$ be the value such that $\hat{S}_{KM}(t_{max}) = S_{\hat{\lambda}}(t_{max})$. The KM is extended past t_{max} by $S_{\hat{\lambda}}(t)$, that is we define $\hat{S}_{KM}(t) = S_{\hat{\lambda}}(t)$ for all $t > t_{max}$. We can then compute the mean survival time by integrating the extended KM. This amounts to integrating the original KM up to t_{max}, and add $\int_{t_{max}}^{\infty} S_{\hat{\lambda}}(t)dt = \exp(\hat{\lambda} t_{max})/\hat{\lambda}$.

2 Split control and allowable splits

As explained in Section 2.2 of the article, the best split is the one maximizing $\Delta\tau$ among all allowable splits. Here we describe what we mean by allowable splits, and what are the possible criteria to stop splitting nodes. Due to the nature of ITE modeling, we need to ensure having enough observations of both the treatment and control groups in each children node to be able to compute the Kaplan-Meier estimates. For that reason, we add several parameters to our algorithm.

Pre-splitting conditions. They are used to decide if we attempt to split a node or not. All conditions must be met to attempt splitting:

- $N_P \geq$ minP. The parent node should have a minimum number of observations.
- $N_P^T \geq$ minPT. The parent node should have a minimum number of treatment observations.
- $N_P^C \geq$ minPC. The parent node should have a minimum number of control observations.
- Node depth $\geq$ maxdepth. We control the tree depth with this parameter.

Post-Splitting conditions. They are used to decide if a split is allowable or not. All conditions must be met to consider the candidate split allowable:

- $\min(N_L, N_R) \geq$ minLR. Each children node should have a minimum number of observations.
- $\min(N_L^T, N_R^T) \geq$ minLRT. Each children node should have a minimum number of treatment observations.
- $\min(N_L^C, N_R^C) \geq$ minLRC. Each children node should have a minimum number of control observations.
- Tpl $\leq \frac{N_L^T}{N_L} \leq$ Tpu, and Tpl $\leq \frac{N_R^T}{N_R} \leq$ Tpu. The proportion of treatment observations in each children node must be within a given range.
- Each child node must have one uncensored observation in each of the treatment and the control groups. That is, we require having at least one event for each of the Kaplan-Meier estimate that we need to compute.

2.1 Parameters used in the simulation study

Here are the tree growth control parameters for the proposed method that are used for all simulations:

- Minimum observations in parent node to try the split is set to minP=100 when $n_{train} = 1000$ and minP=50 when $n_{train} = 500$.
- Minimum treatment observations in parent node to try the split is set to minPT=20.
- Minimum control observations in parent node to try the split is set to minPC=20.
- Minimum observations in child nodes is set to minLR=50 when $n_{train} = 1000$ and minLR=25 when $n_{train} = 500$.
- Minimum treatment observations in a child node is set to minLRT=10.
- Minimum control observations in a child node is set to minLRC=10.
- Maximum tree depth is set to maxdepth=10.

For simulations with the same proportion of treatment and control (50% each), we use the following parameters:

- Treatment proportion should be $\geq$ Tpl = 30% of observations in the child node.
- Treatment proportion should be $\leq$ Tpu= 70% of observations in the child node.

For simulations with 25% of treatment and 75% of control, we use the following parameters:

- Treatment proportion should be $\geq$ Tpl = 10% of observations in the child node.
- Treatment proportion should be $\leq$ Tpu = 40% of observations in the child node.

3 Description of the Individual DGPs Used in the Simulation Study

Here is a detailed description of the five DGPs used in the simulation study. To get an insight about these DGPs, Figures 2 and 3 present histograms of the survival time for the control and treatment groups and of the ITE function for typical samples.

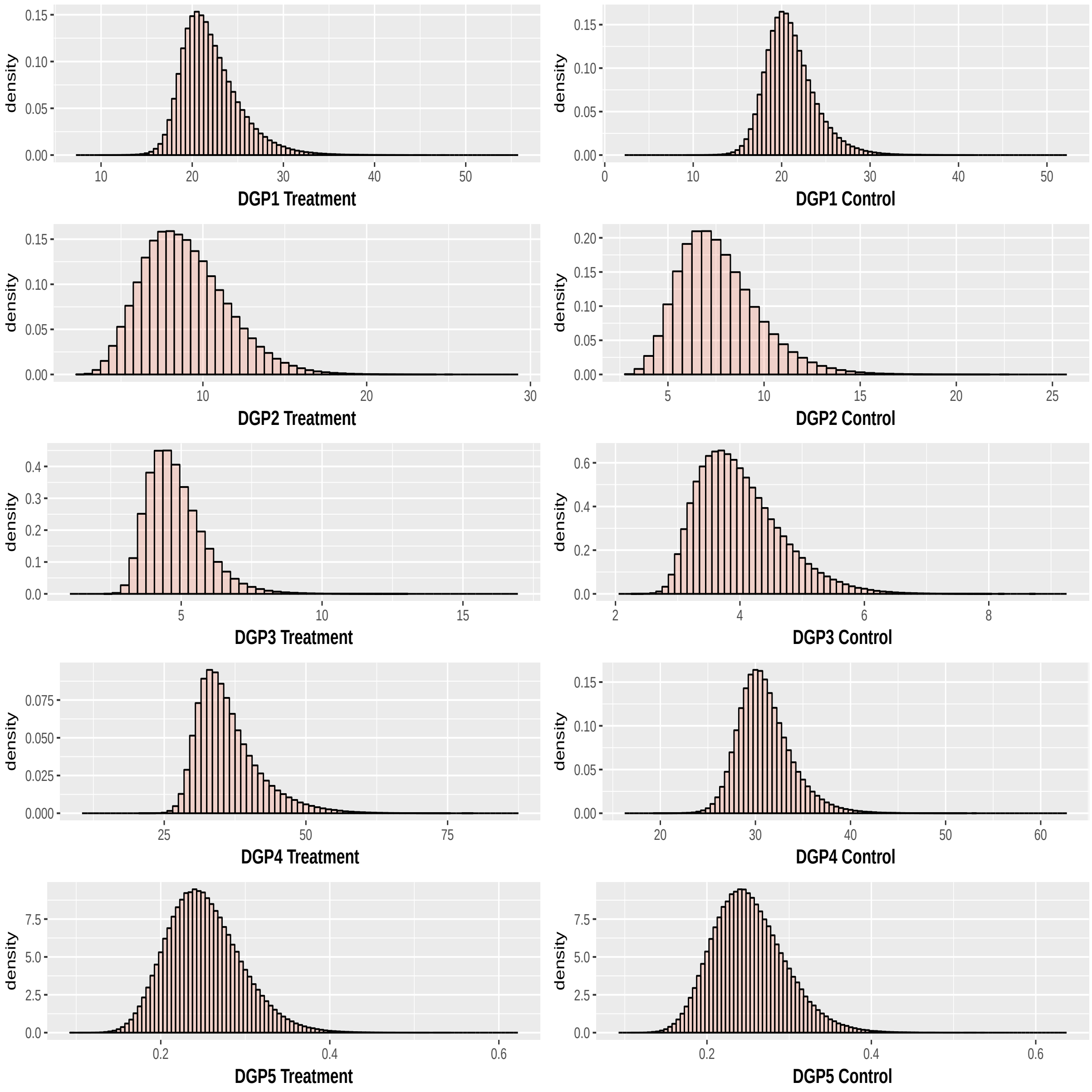

Figure 2: Typical distributions of the observed time for the DGPs

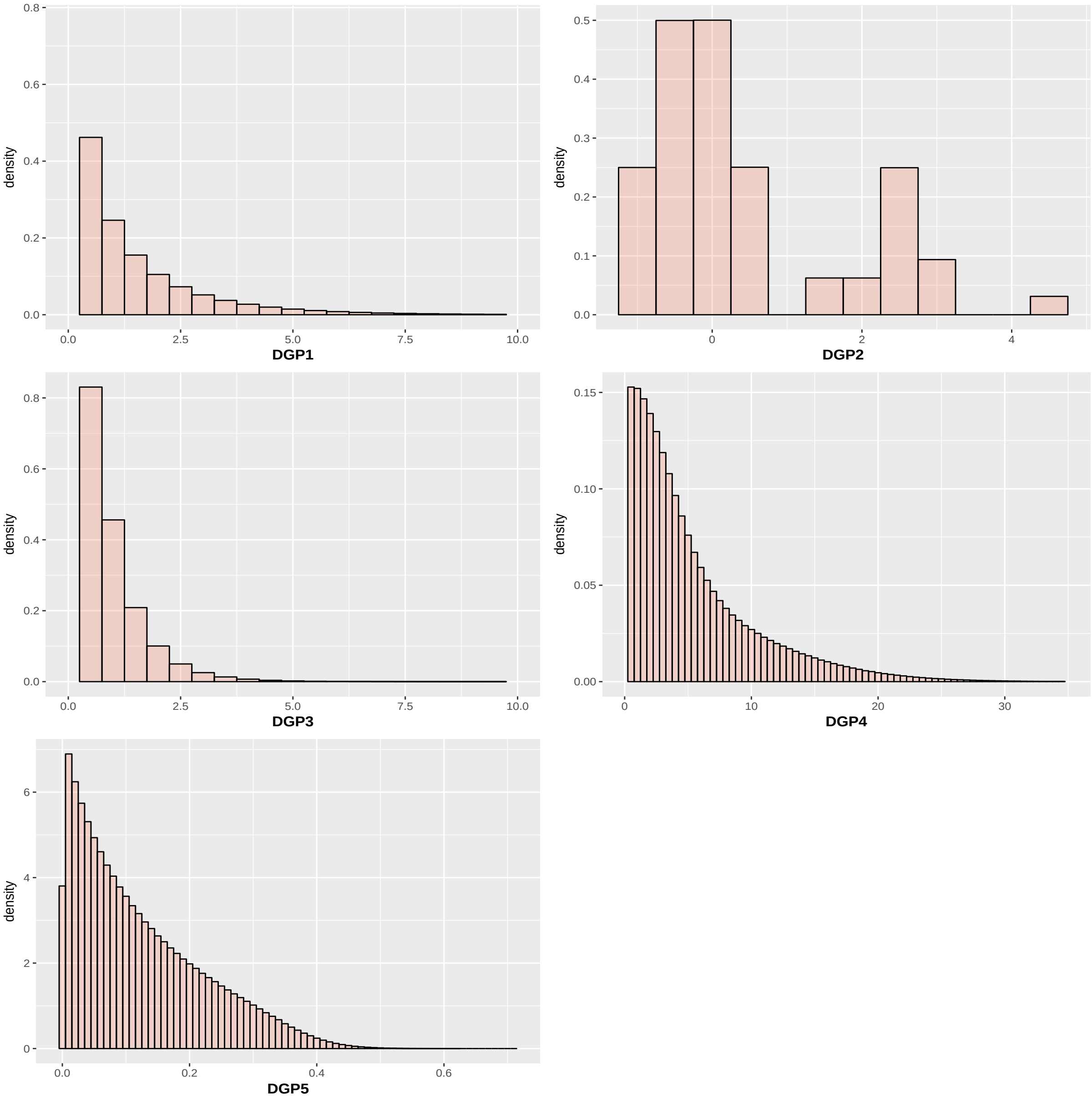

Figure 3: Typical distributions of the ITE function for the DGPs

For the first four DGPs, ten covariates $\boldsymbol{X} = (X_1, X_2, \ldots, X_{10})$ are available. They are independent and normally distributed with mean 0 and standard deviation 1. The true survival time of a control group observation, for which $W = 0$, is generated according to

$$Y = c(\boldsymbol{X}) + \epsilon,$$

where $c(\boldsymbol{X})$ is a function of the covariates and ϵ is an independent error term. The true survival time of a treatment group observation, for which $W = 1$, is generated according to

$$Y = c(\boldsymbol{X}) + \tau(\boldsymbol{X}) + \epsilon,$$

where $c(\cdot)$ and ϵ are as above, and where $\tau(\boldsymbol{X})$ is the ITE function that we want to estimate. The specific choices of $c(\cdot)$, $\tau(\cdot)$, and the error distribution are specified below.

3.1 DGP 1

This DGP represents a simple ITE given by the squared value of X_1. The control model is given by

$$c(\boldsymbol{X}) = 20 + X_1 + X_2 + X_3 + X_4 + X_2^2 + X_3 \times X_4.$$

The ITE function is

$$\tau(\boldsymbol{X}) = X_1^2.$$

The error term ϵ is normally distributed with mean 0 and standard deviation 1. For this DGP, the covariates $X_5, X_6, X_7, X_8, X_9, X_{10}$ are not used, but they are still provided as potential covariates. Thus, they are noise covariates.

3.2 DGP 2

This DGP is a function of two tree shaped functions $Tree_1$ and $Tree_2$. Their structures are represented in Figures 4 and 5. The control model is given by

$$c(\boldsymbol{X}) = 24/10 + Tree_1(\boldsymbol{X})/5.$$

The ITE function is

$$\tau(\boldsymbol{X}) = -7/4 + Tree_1(\boldsymbol{X})/20 + Tree_2(\boldsymbol{X})/4.$$

The error term ϵ has the gamma distribution with a shape of 4 and a scale of 1 (`rgamma`(1,4,1) in R). For this DGP, the covariates X_8, X_9, X_{10} are not used and are the noise covariates.

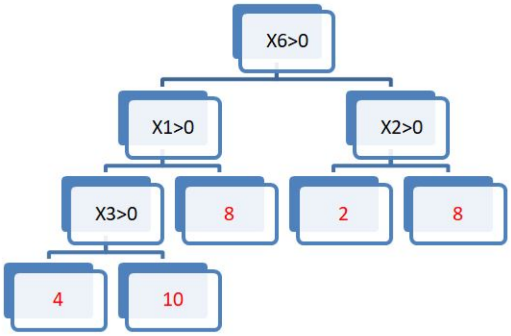

Figure 4: Tree 1 structure for DGP 2

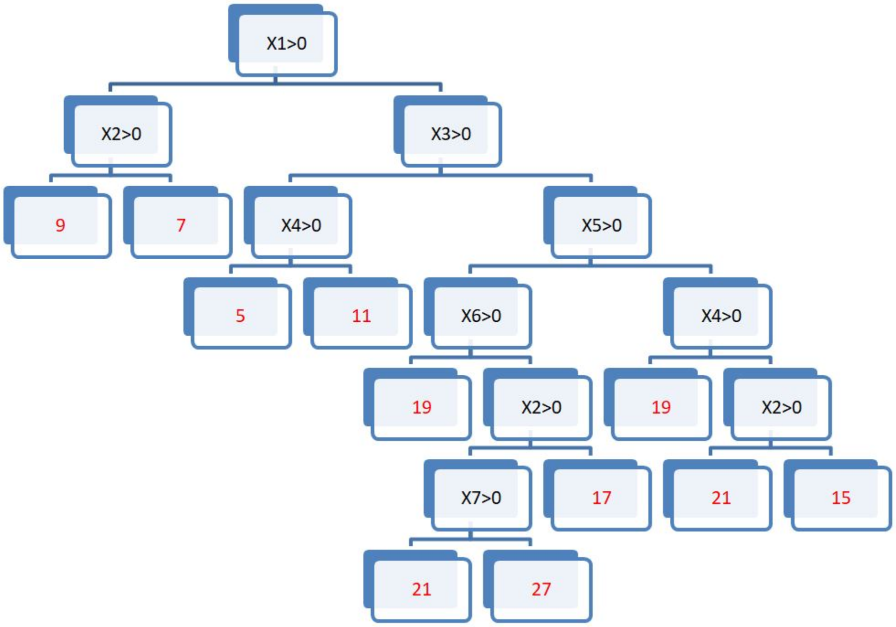

Figure 5: Tree 2 structure for DGP 2

3.3 DGP 3

This DGP is also characterized by a complex ITE function. The control model is given by

$$c(\boldsymbol{X}) = 3 + (X_1 + X_2 + X_3 + X_4 + X_2^2 + X_3 \times X_4)/10.$$

The ITE function is

$$\tau(\boldsymbol{X}) = (X_7 + 2 \times X_5 + 4 \times X_1^2 + |X_2| + X_1 \times X_2 + |X_3^3| + e^{X_8})/10.$$

The error term ϵ has the Weibull distribution with a shape of 1.5 and a scale of 1 (`rweibull`(1,1.5,1) in R). For this DGP, the covariates X_6, X_9, X_{10} are not used and are the noise covariates.

3.4 DGP 4

This DGP has a complex ITE function made by a tree with non-constant terminal node means. The control model is the following:

$$c(\boldsymbol{X}) = 30 + X_1 + X_2 + X_3 + X_4 + X_2^2 + X_3 \times X_4.$$

The ITE function is

$$\tau(\boldsymbol{X}) = \begin{cases} |X_1^2 + 3 \times X_8| & \text{if } X_1 > 0 \\ |X_4^2 + 10 \times |X_5|| & \text{if } X_1 \leq 0 \text{ and } X_4 > 0 \\ |2 \times |X_4| \times X_1 + 8 \times X_{10}| & \text{if } X_1 \leq 0 \text{ and } X_4 \leq 0 \end{cases}$$

The error term ϵ is normally distributed with mean 0 and standard deviation 1. For this DGP, the covariates X_6, X_7, X_9 are not used and are the noise covariates.

3.5 DGP 5

This is an AFT DGP. Ten independent covariates $\boldsymbol{X} = (X_1, X_2, \ldots, X_{10})$ from a uniform distribution on (0,1) are available. The true log survival time of a control group observation, for which $W = 0$, is generated according to

$$\log(Y) = c_0(\boldsymbol{X}) + \epsilon,$$

and the true log survival time of a treatment group observation, for which $W = 1$, is generated according to

$$\log(Y) = c_0(\boldsymbol{X}) + \tau_0(\boldsymbol{X}) + \epsilon,$$

where ϵ is the error term form the logistic distribution with mean 0 and scale $\sigma = \sqrt{3}/(10\pi)$. The functions c_0 and τ_0 are

$$c_0(\boldsymbol{X}) = -2 + .2 * (X_1 + X_2 + X_3 + X_4 + X_5 + X_6),$$

and

$$\tau_0(\boldsymbol{X}) = |X_1^2 - X_8|.$$

The ITE function is thus

$$\tau(\boldsymbol{X}) = (\exp(\tau_0(\boldsymbol{X})) - 1)\exp(c_0(\boldsymbol{X}))E[\exp(\epsilon)],$$

and $E[\exp(\epsilon)] = 1.005$. For this DGP, the covariates X_7, X_9, X_{10} are not used and are the noise covariates.

4 Global Simulation Results

Here are the global simulations results. See Section 3.3 of the article for an explanation of the performance criteria. Figures 6 and 7 present the results for the percent increase in MSE for training sample sizes of 1000 (Figure 6) and 500 (Figure 7). In fact, Figure 6 is the same as Figure 1 in the article and is reproduced here to ease the comparison with the $n_{train} = 500$ case. As mentioned in the article, the proposed and BAFT methods outperform the others, globally. The relative ordering are the same for both training sample sizes, but all methods are slightly closer in the $n_{train} = 500$ case. Figures 8 and 9 present the results for the percent decrease in C-index for training sample sizes of 1000 (Figure 8) and 500 (Figure 9). In fact, Figure 8 is the same as Figure 2 in the article. According to this criterion, the proposed and BAFT methods outperform the others again. Again also, the differences among all methods are slightly smaller in the $n_{train} = 500$ case. The 2AFT and Interaction methods have basically the same performance. Hence, using a single AFT model with interactions between the treatment indicator and the covariates perform the same as using two separate AFT models, one for the treatment observations and one for the control observations. The next section explores the results in details by looking at each DGP separately.

4.1 MSE Results

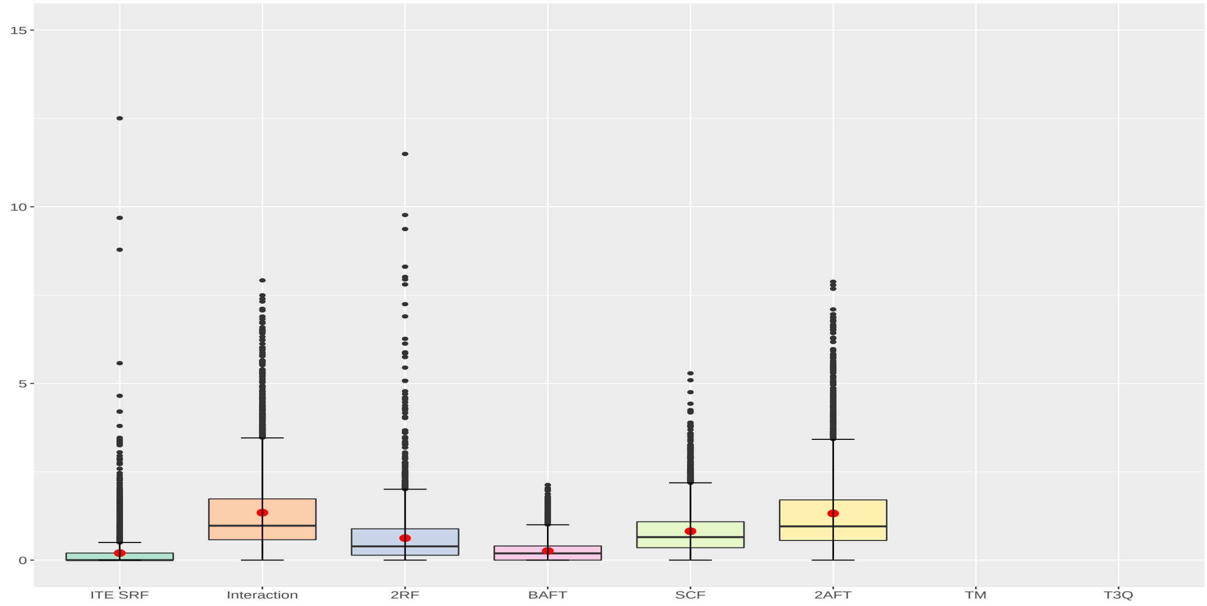

Figure 6: Global MSE simulation results with a training sample size of 1000. The box-plots represent the distribution of the % increase in MSE with respect to the best performer of the run for the 4000 runs.

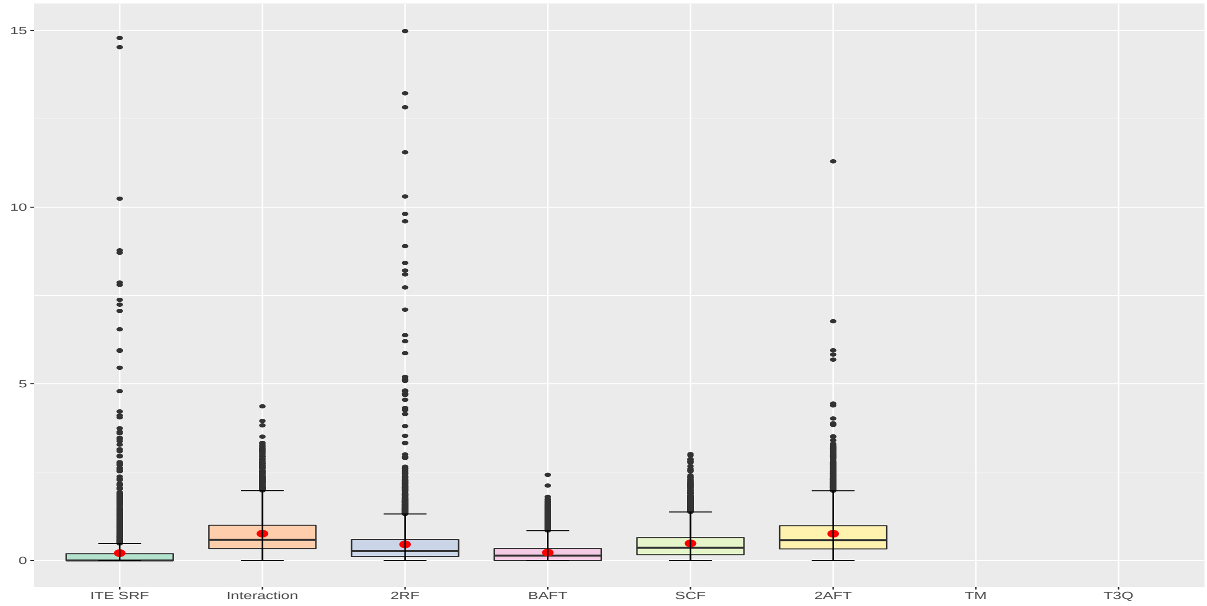

Figure 7: Global MSE simulation results with a training sample size of 500. The box-plots represent the distribution of the % increase in MSE with respect to the best performer of the run for the 4000 runs.

4.2 C-Index Results

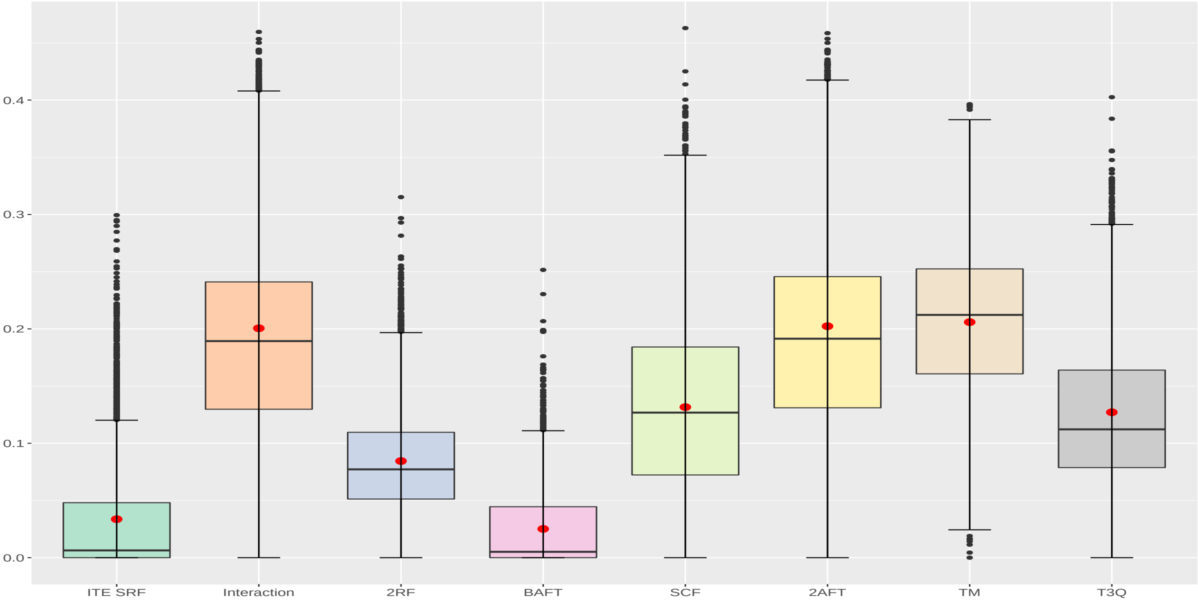

Figure 8: Global C-index simulation results with a training sample size of 1000. The box-plots represent the distribution of the % decrease in C-Index with respect to the best performer of the run for the 4000 runs.

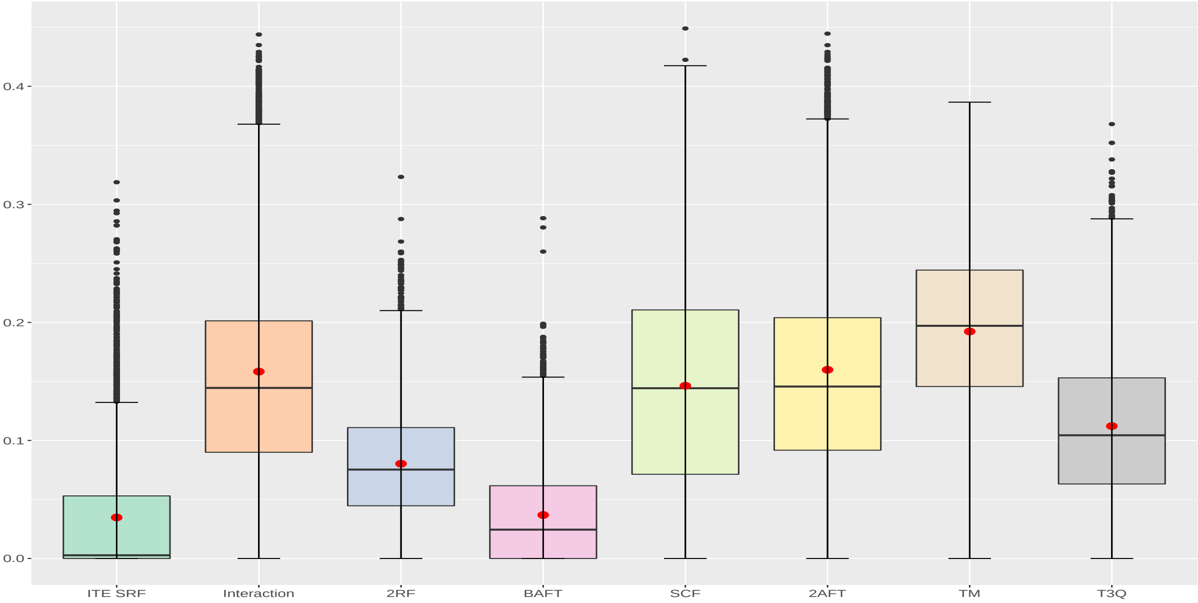

Figure 9: Global C-index simulation results with a training sample size of 500. The box-plots represent the distribution of the % decrease in C-Index with respect to the best performer of the run for the 4000 runs.

5 Detailed Simulation Results

To gain more insights, we examine in details all scenarios. The results are presented in Figures 10 to 19 for the MSE and 20 to 29 for the C-index. These plots present the distribution of the raw MSE and C-Index over the 100 runs for each scenario and each method. Each figure corresponds to a DGP, the left column is for the case of 50% treatment and 50% control, and the right one is for the case of 25% treatment and 75% control. In addition, each row corresponds to a censoring rate, ranging from 0% to 60%.

The results for $n_{train} = 500$ and $n_{train} = 1000$ are very similar. For a given configuration, the MSE for the $n_{train} = 1000$ are slightly lower than the ones for the $n_{train} = 500$. This is expected since more data are available. Likewise, the C-index for the $n_{train} = 1000$ are slightly higher than the ones for the $n_{train} = 500$, for a given configuration. In the following discussion, we focus on the $n_{train} = 1000$ cases unless a notable difference occurs between the two sample sizes.

For DGP1 in Figure 10, the BAFT method has the smallest MSE for all censoring rate and treatment control combinations, and the proposed method comes in second place. Likewise, the BAFT method has the highest C-index, followed by the proposed method (see Figure 20). This DGP has the simplest ITE function which is related to a single covariate. The AFT interaction and 2AFT models have the poorest results across all DGP1 simulations. With this DGP, the TM and T3Q methods have a similar performance for the C-index. For each method, we can see that the MSE is generally increasing and the C-index is generally decreasing when the censoring level increases. This is expected since less information is available as the censoring level increases. Moreover, we also notice that the results are better when we have an equal proportion of treatment and control observations. The ordering of the methods does not change when comparing the 50/50 treatment/control proportions to the 25/75 proportions, but the MSE are slightly higher and the C-index slightly lower in the 25/75 case. Similarly, the ordering of the methods does not change much when the censoring rate varies.

The results for DGP2, which has a more complex ITE function, are presented in Figures 12 and 22. This time the proposed method is generally the best one alone, or very close to the best one, depending on the scenario. Contrarily to DGP1, the ordering of the methods depends on the scenario. For example, the SCF is the best one along with the proposed method for the 50/50 treatment/control proportions with no censoring (upper left plot), but its performance deteriorates more rapidly when the censoring rate increases. On the contrary, the BAFT method becomes relatively more competitive as the censoring rate increases.

The results for DGP3, which also involves a complex ITE function, are presented in Figures 14 and 24. The proposed method is generally the best one alone, or tied with another one depending on the scenario, for the MSE criterion. For the C-index, except for low censoring and the 50/50 treatment/control proportions cases, the proposed method is the best one. This time, the sample size makes a small difference in the rankings since the proposed method is always the best one in the $n_{train} = 500$ case (see Figure 25 for example). This time, the SCF is not performing as well as for DGP2. A larger difference is seen between the TM and T3Q methods, the latter one performing better.

The results for DGP4 are presented in Figures 16 and 26. The proposed method has the best performance overall. The 2RF and BAFT methods come in second place depending on the scenario. The SCF method seems more affected by the censoring as we can notice a large decrease in its C-index when moving from 0% of censoring to 20%. A large difference is again noted between the TM and T3Q methods, the latter one still performing better.

The results for DGP5 are presented in Figures 18 and 28. The proposed method has the best performance in all scenarios except in the 0% censoring and 50/50 treatment/control proportions case, where BAFT is better. As for the last 2 DGPs, TM performs clearly better than T3Q.

These results show that the proposed method is a strong competitor for estimating the ITE function with censored data in a variety of situations. The proposed methods and BAFT are the two best ones with very stable performances across all scenarios. However, except for the interaction, 2AFT and TM models, each method comes in first or second place for at least one scenario. Hence, it is not possible to declare that a single method is the best one in all cases. Moreover, the fact that the interaction and 2AFT models do not perform well was expected. The DGPs have complex links between the covariates and the response so a simple main effects plus treatment interactions AFT model is clearly at a disadvantage. One way to improve this model would be to add other interactions in the AFT model, but choosing which one is not straightforward. One of the strengths of forest based methods is that they are able to automatically model complex DGPs.

5.1 MSE Results

Here are the detailed simulation results presented in Section 3.3 of the article. The next 10 figures show the distribution of the MSE of each method in each scenario.

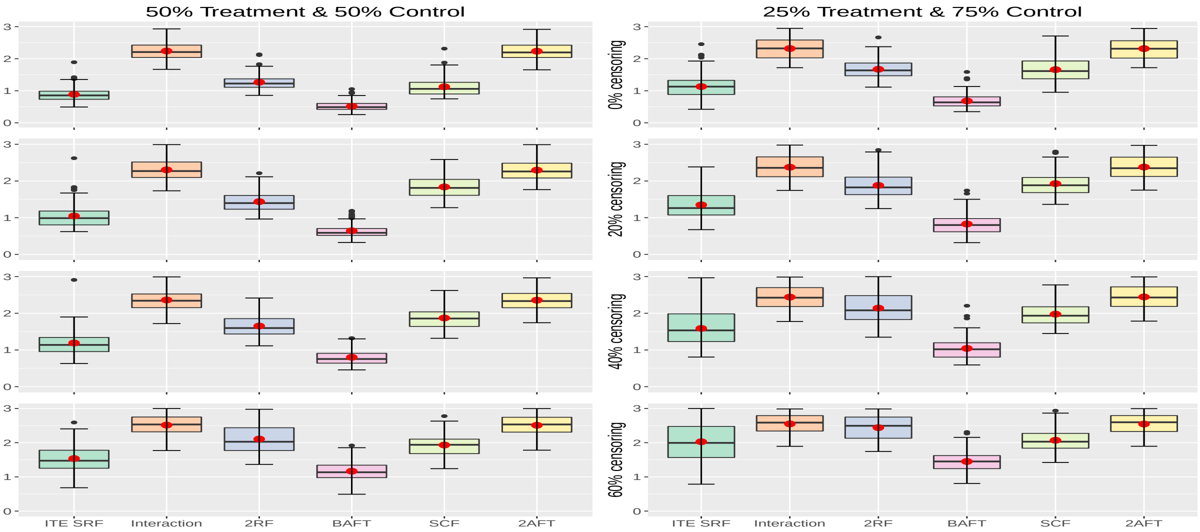

Figure 10: DGP1 MSE results with $n_{train} = 1000$. The box-plots represent the distribution of the MSE for the 100 runs.

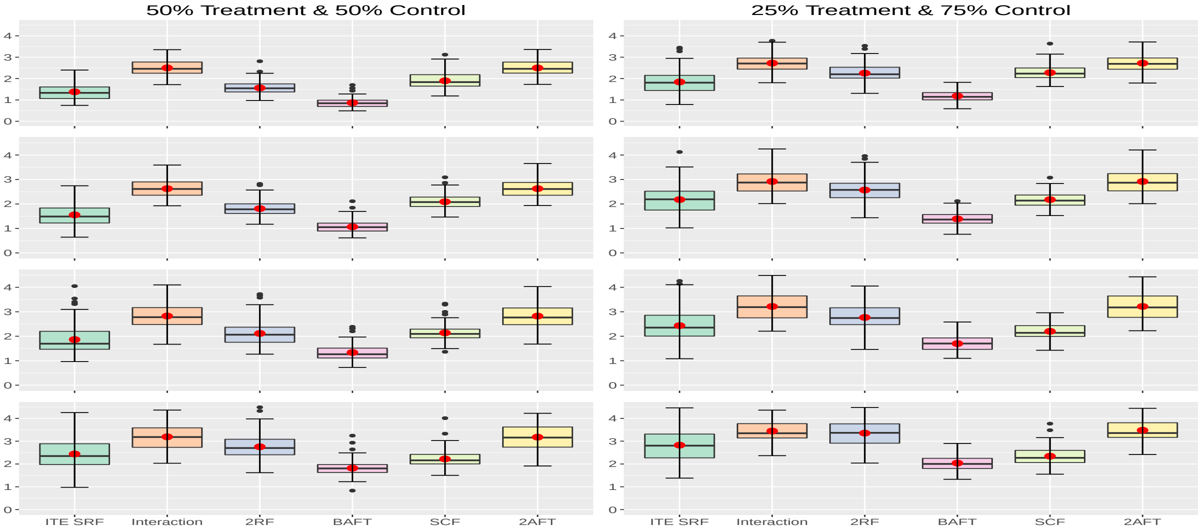

Figure 11: DGP1 MSE results with $n_{train} = 500$. The box-plots represent the distribution of the MSE for the 100 runs.

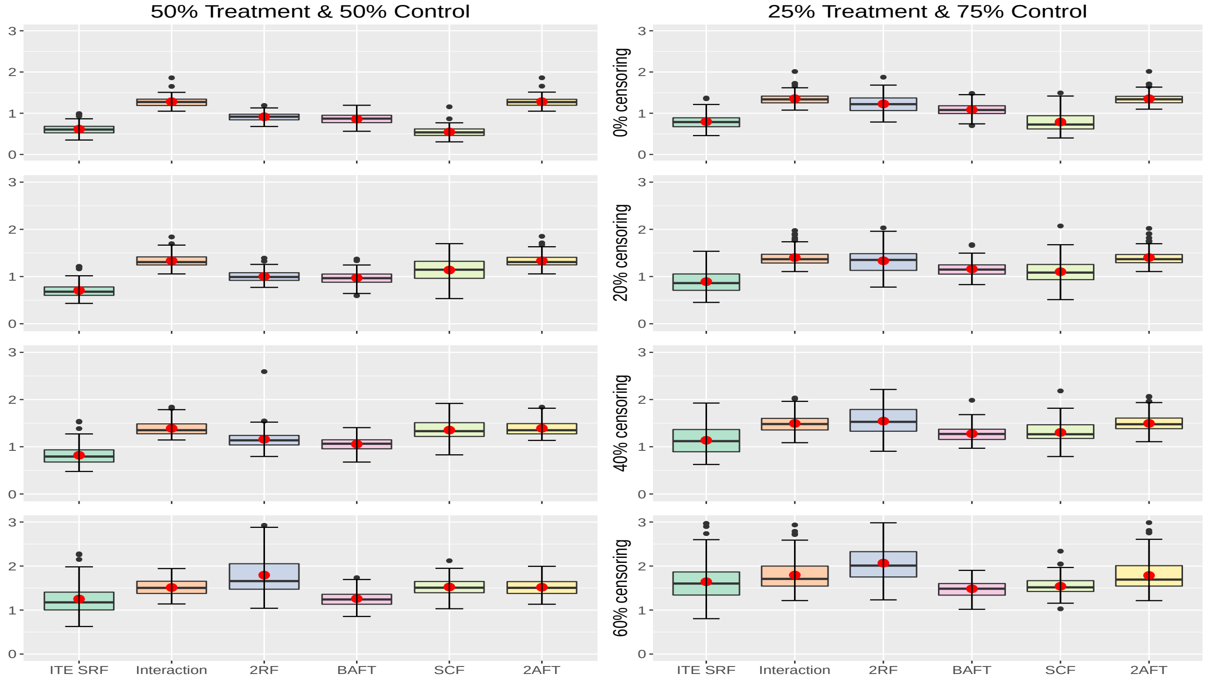

Figure 12: DGP2 MSE results with $n_{train} = 1000$. The box-plots represent the distribution of the MSE for the 100 runs.

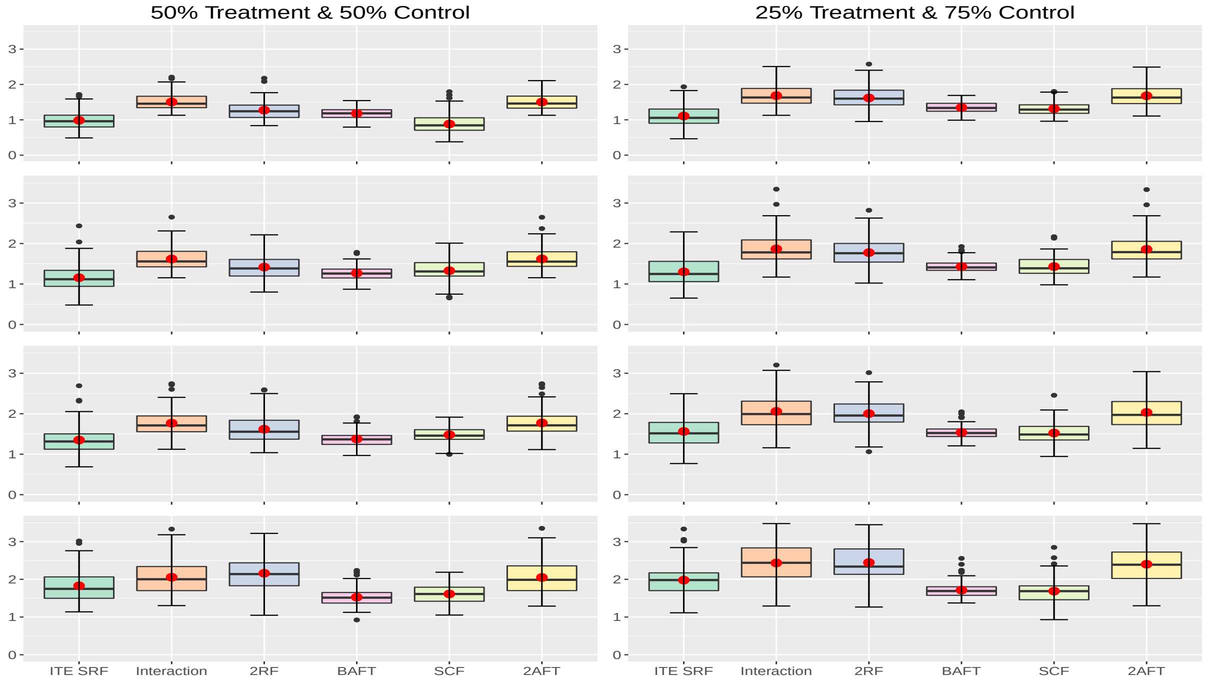

Figure 13: DGP2 MSE results with $n_{train} = 500$. The box-plots represent the distribution of the MSE for the 100 runs.

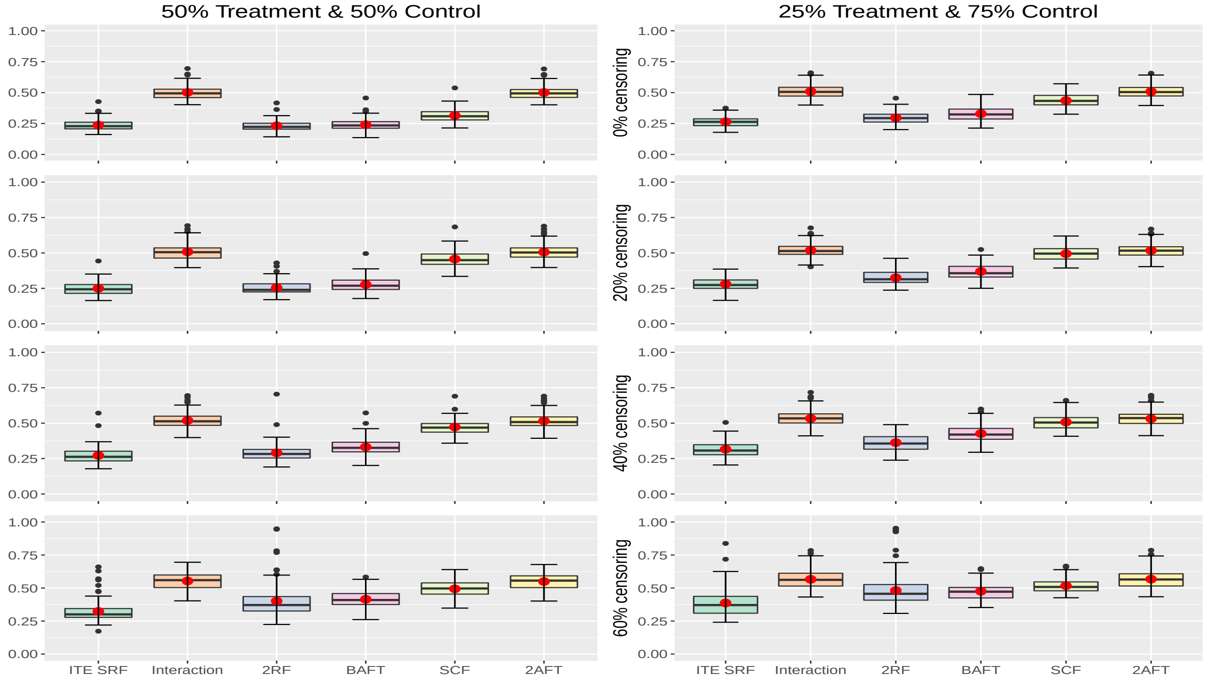

Figure 14: DGP3 MSE results with $n_{train} = 1000$. The box-plots represent the distribution of the MSE for the 100 runs.

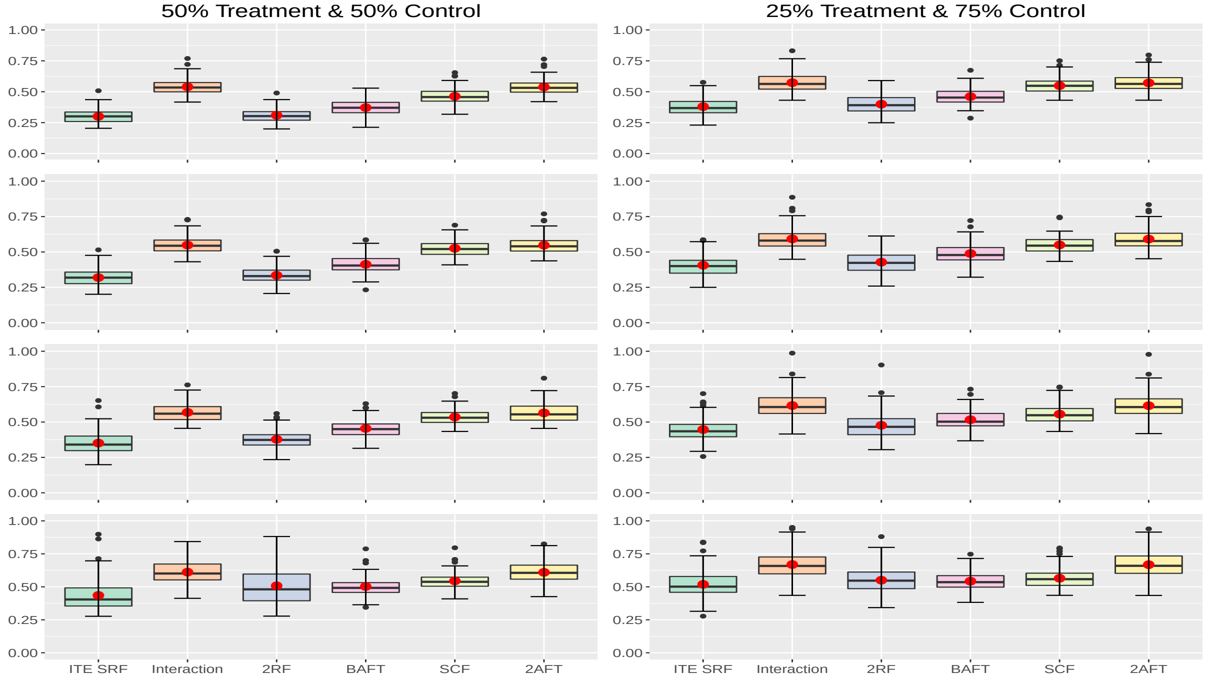

Figure 15: DGP3 MSE results with $n_{train} = 500$. The box-plots represent the distribution of the MSE for the 100 runs.

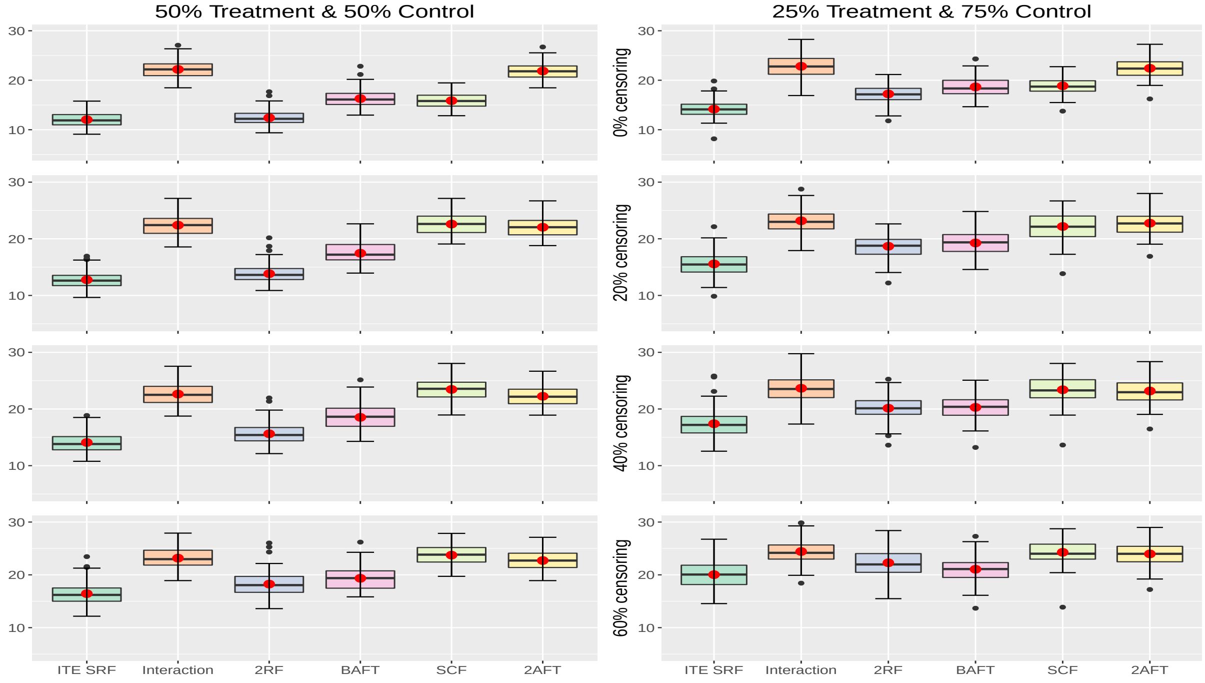

Figure 16: DGP4 MSE results with $n_{train} = 1000$. The box-plots represent the distribution of the MSE for the 100 runs.

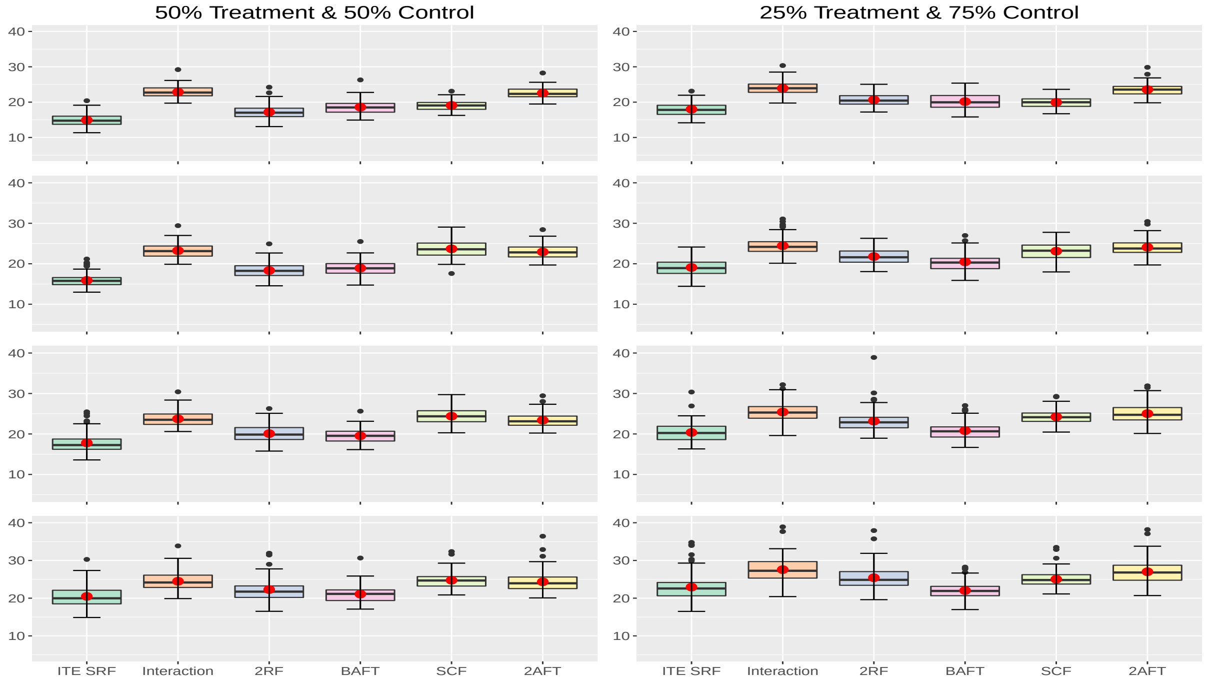

Figure 17: DGP4 MSE results with $n_{train} = 500$. The box-plots represent the distribution of the MSE for the 100 runs.

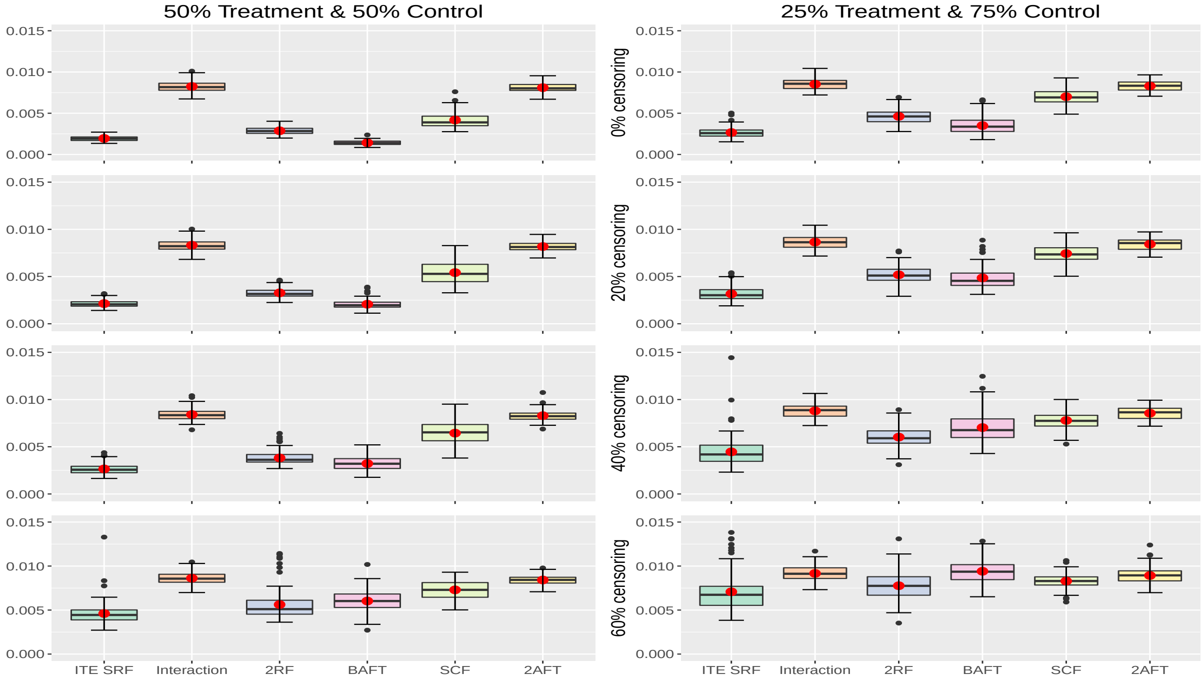

Figure 18: DGP5 MSE results with $n_{train} = 1000$. The box-plots represent the distribution of the MSE for the 100 runs.

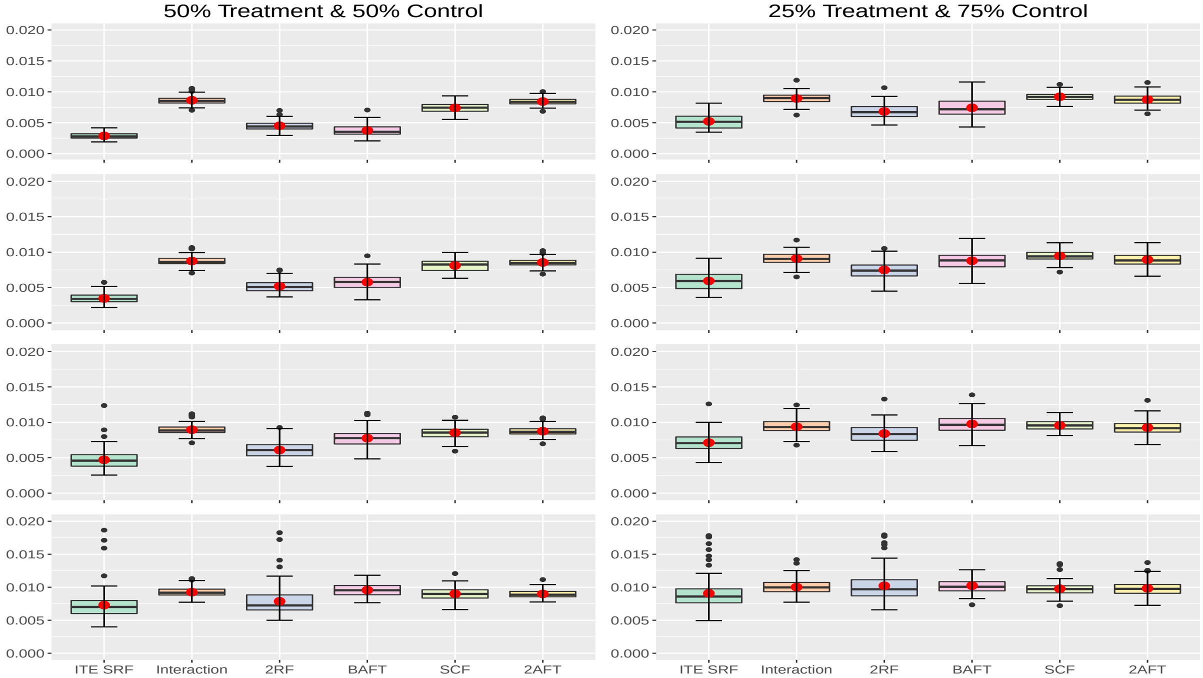

Figure 19: DGP5 MSE results with $n_{train} = 500$. The box-plots represent the distribution of the MSE for the 100 runs.

5.2 C-Index Results

Here are the detailed simulation results presented in Section 3.3 of the article. The next 10 figures show the distribution of the C-index of each method in each scenario.

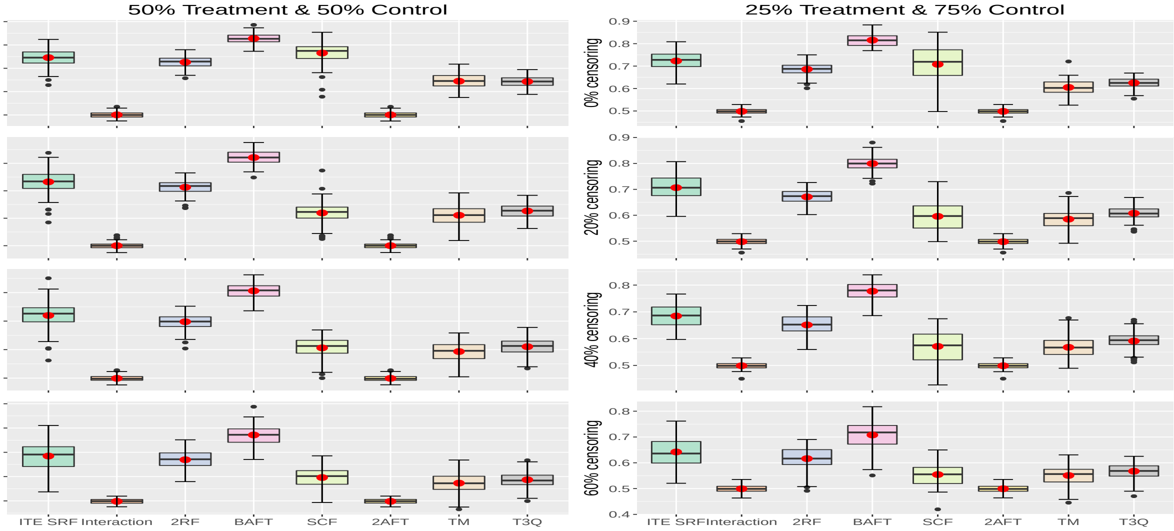

Figure 20: DGP1 C-index results results with $n_{train} = 1000$. The box-plots represent the distribution of the C-Index for 100 runs.

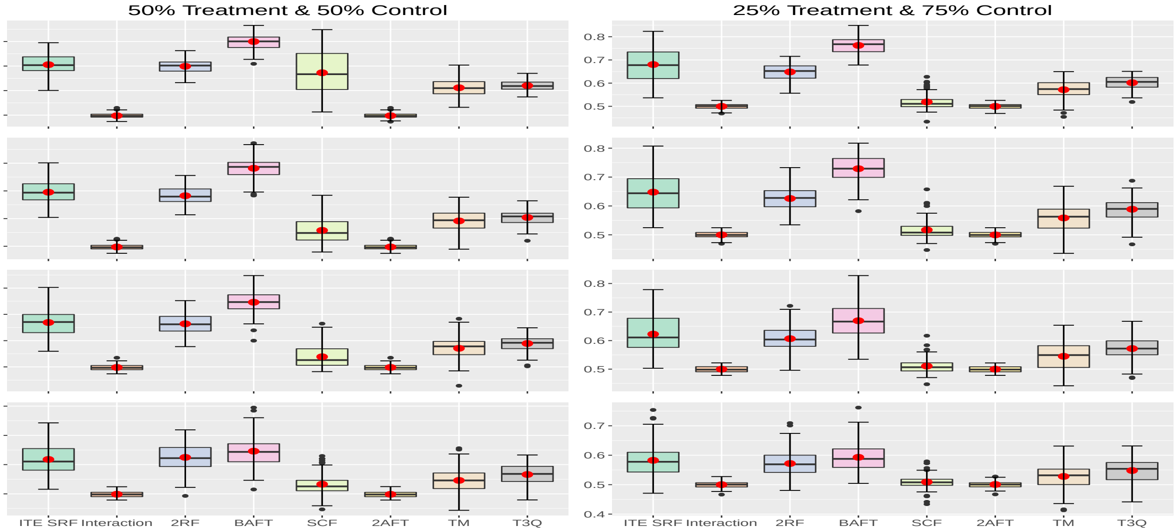

Figure 21: DGP1 C-index results results with $n_{train} = 500$. The box-plots represent the distribution of the C-Index for 100 runs.

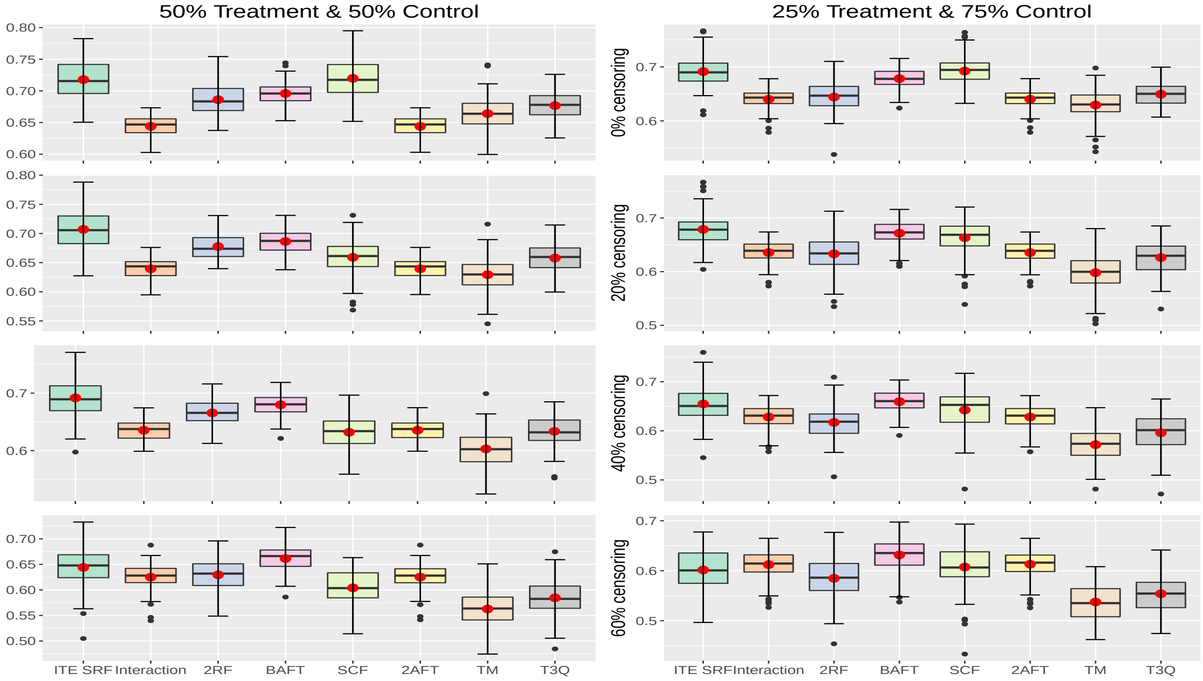

Figure 22: DGP2 C-index results results with $n_{train} = 1000$. The box-plots represent the distribution of the C-Index for the 100 runs.

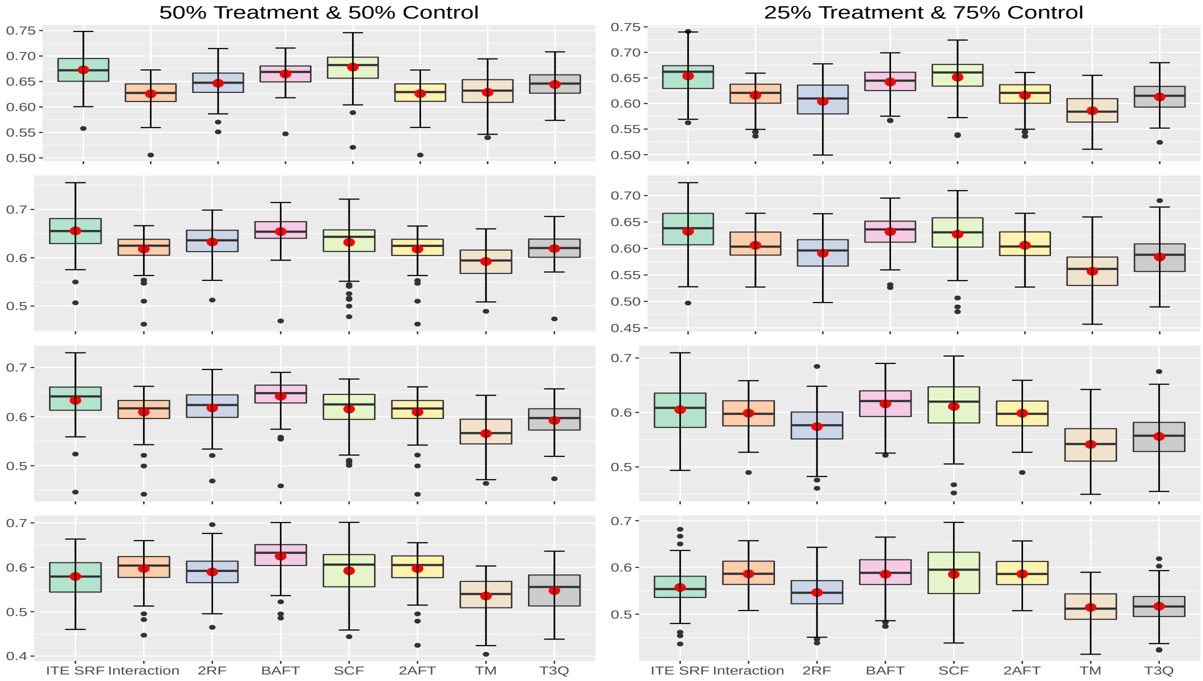

Figure 23: DGP2 C-index results results with $n_{train} = 500$. The box-plots represent the distribution of the C-Index for the 100 runs.

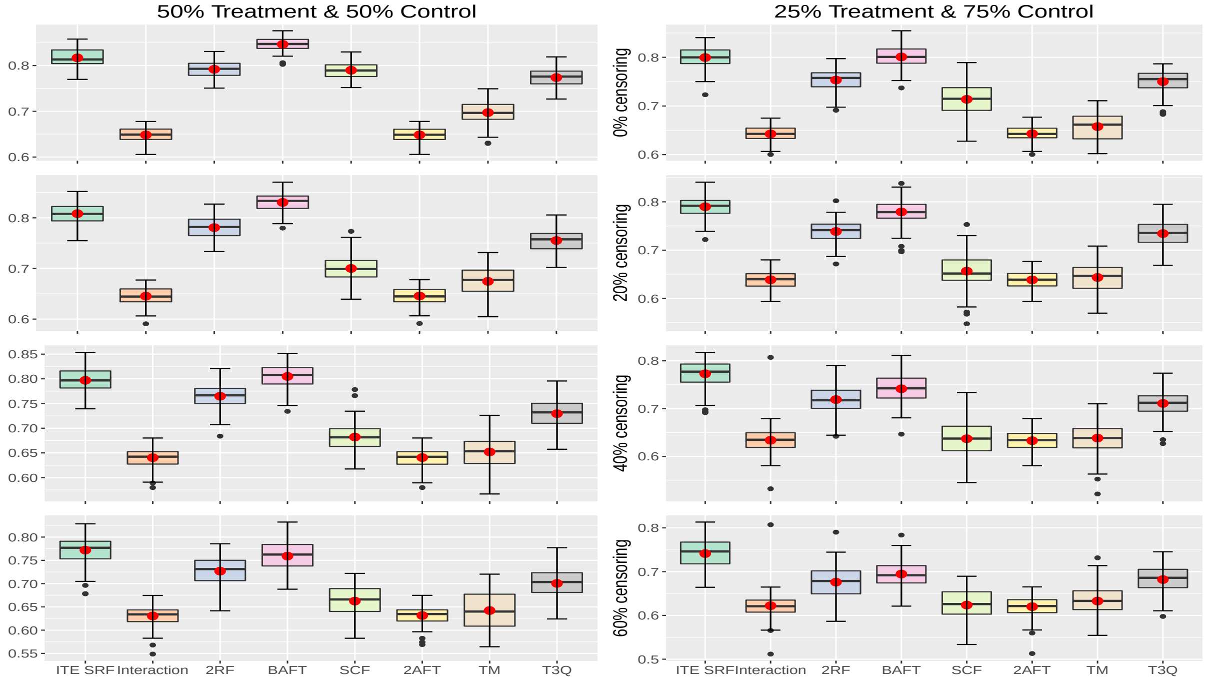

Figure 24: DGP3 C-index results results with $n_{train} = 1000$. The box-plots represent the distribution of the C-Index for the 100 runs.

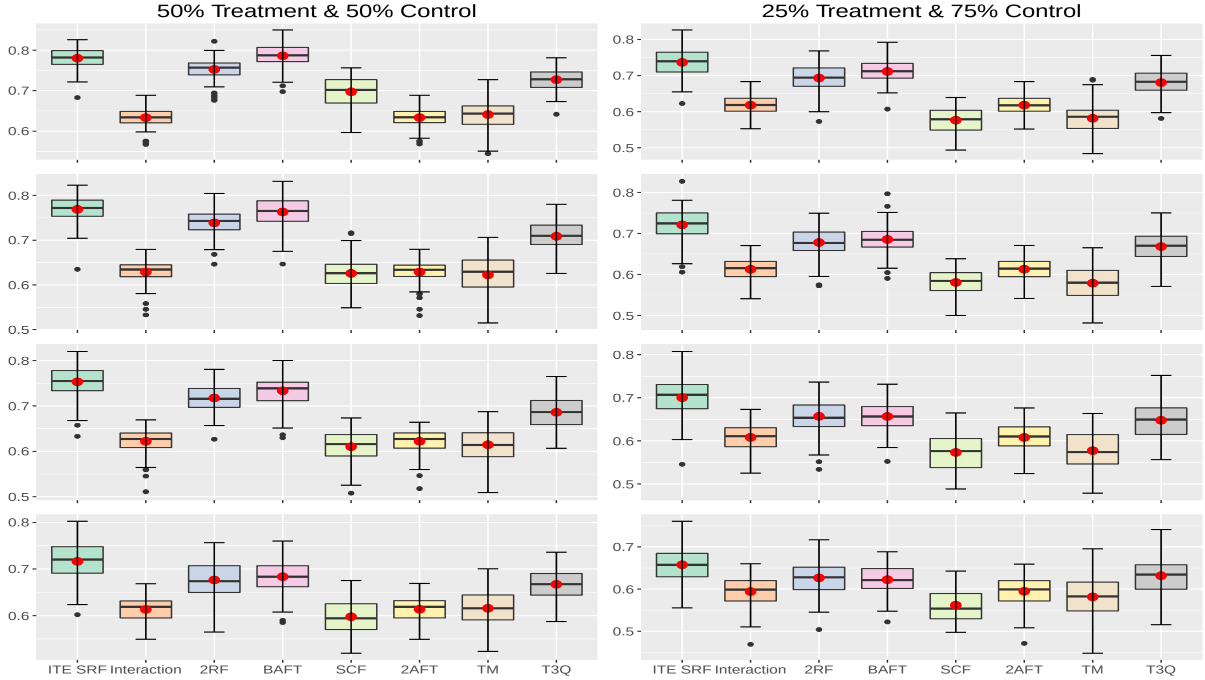

Figure 25: DGP3 C-index results results with $n_{train} = 500$. The box-plots represent the distribution of the C-Index for the 100 runs.

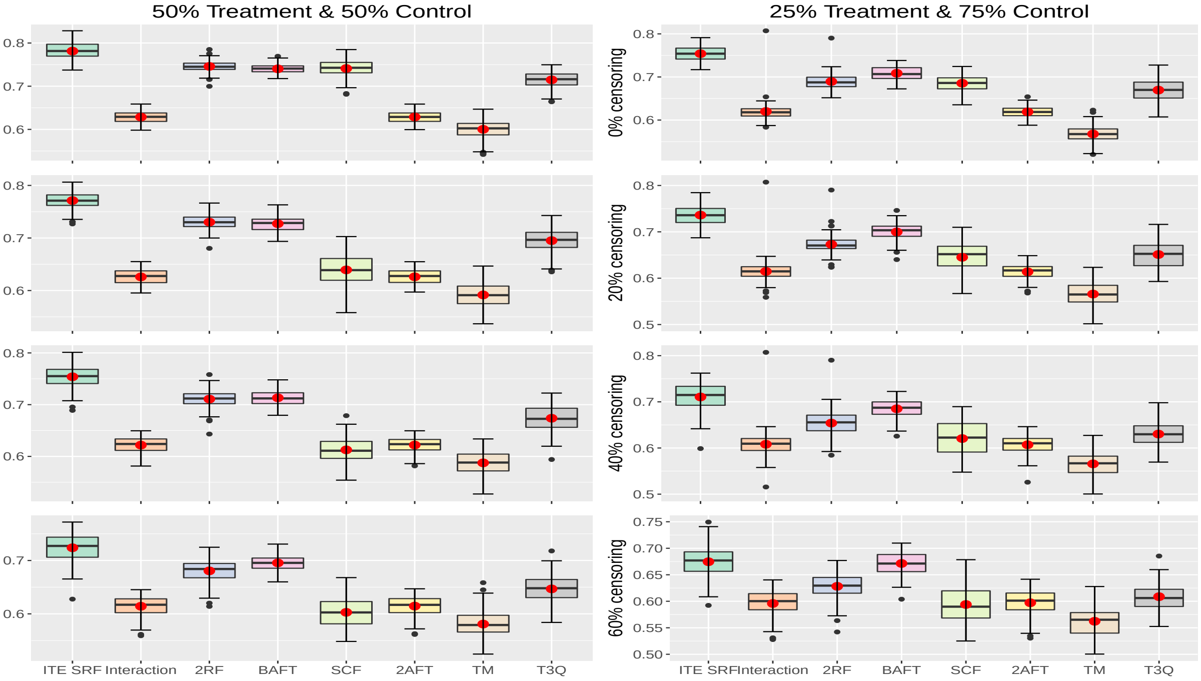

Figure 26: DGP4 C-index results results with $n_{train} = 1000$. The box-plots represent the distribution of the C-Index for the 100 runs.

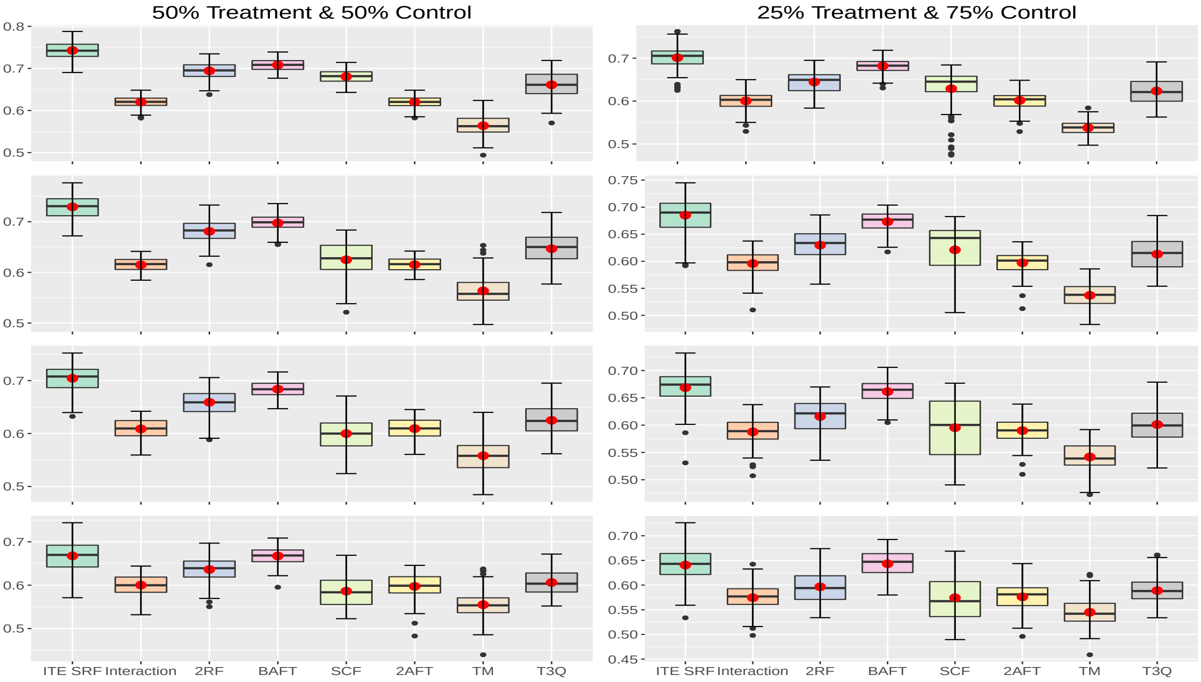

Figure 27: DGP4 C-index results results with $n_{train} = 500$. The box-plots represent the distribution of the C-Index for the 100 runs.

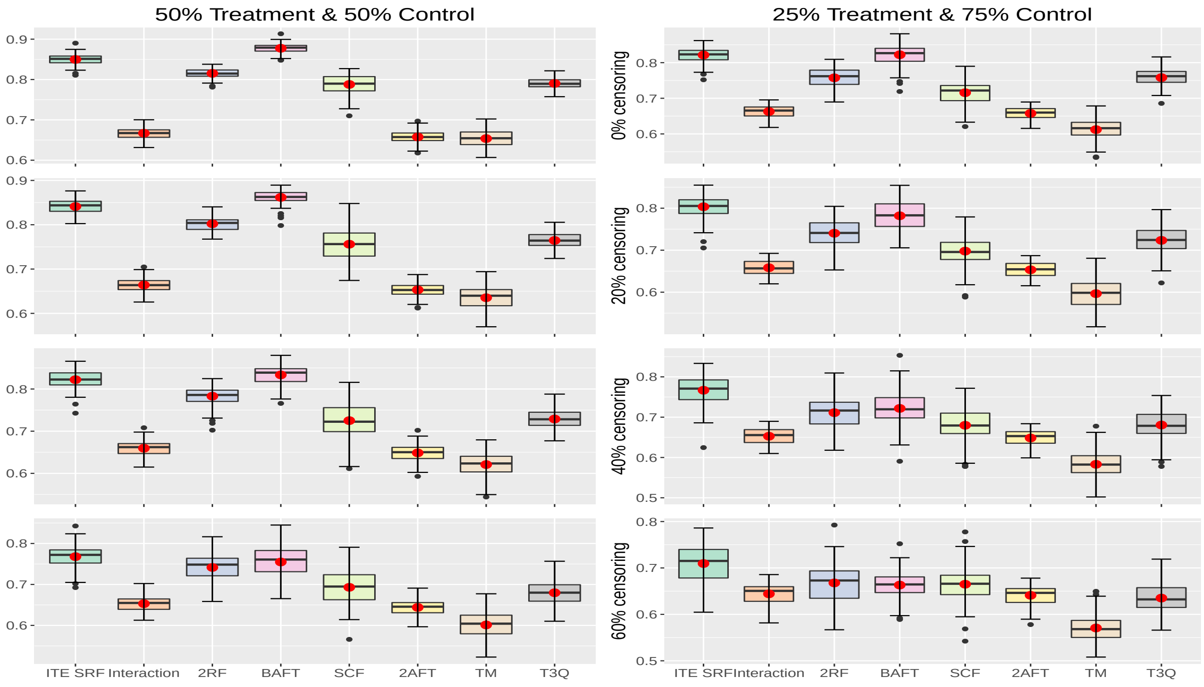

Figure 28: DGP5 C-index results results with $n_{train} = 1000$. The box-plots represent the distribution of the C-Index for the 100 runs.

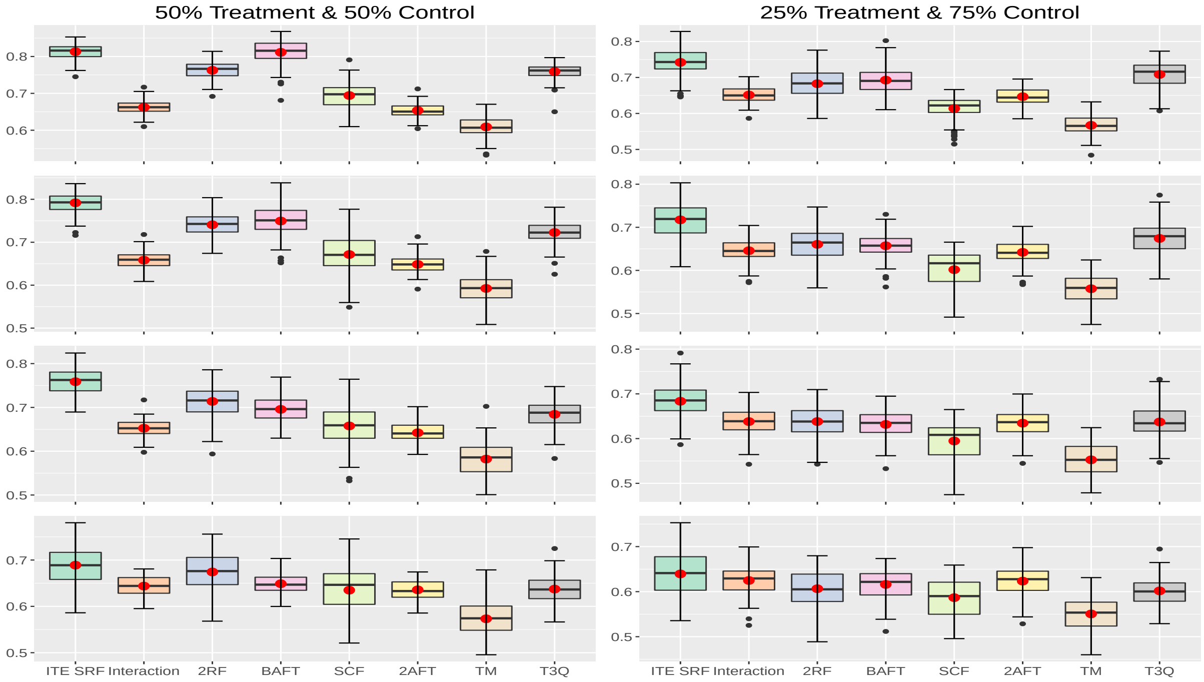

Figure 29: DGP5 C-index results results with $n_{train} = 500$. The box-plots represent the distribution of the C-Index for the 100 runs.

6 Estimated treatment effects for AFT models fitted with cumulative deciles samples for all methods for the data examples

The two following tables present the information used to prepare Figures 4 and 6 in the article. As a reminder, for each of the eight methods, the estimated ITE are ranked from the largest to the smallest values and grouped in deciles. For a given method, the first decile contains the top 10% of observations with the largest estimated ITE. The second decile contains the next 10% of observations and so on. The tenth decile contains the 10% of observations with the smallest ITE. For each method, we fit AFT models to samples formed by consecutive cumulated deciles. The first model is fit using the first decile only (10% of the data). The second model is fit using the first two deciles (20% of the data), and so on. The last model is fit to the whole sample (100% of the data). The tables contain the estimated treatment effects for these models along with the corresponding one-sided p-values to test if the parameter is larger than 0.

		ITE SRF		Interaction		2RF		BAFT		SCF		2AFT		TM		T3Q	
Deciles	Sample	β	pval	β	pval	β	pval	β	pval	β	pval	β	pval	β	pval	β	pval
1	60	1.021	0.009	-0.531	0.966	0.376	0.110	-0.364	0.925	-0.469	0.913	-0.722	0.994	0.324	0.180	-0.336	0.768
1 to 2	120	0.500	0.033	-0.306	0.928	-0.018	0.525	-0.193	0.836	-0.199	0.792	-0.254	0.873	0.430	0.035	-0.281	0.858
1 to 3	180	0.208	0.161	-0.050	0.615	0.175	0.223	-0.100	0.713	-0.265	0.904	-0.060	0.624	0.188	0.174	0.078	0.343
1 to 4	240	0.163	0.183	-0.127	0.805	0.145	0.213	-0.159	0.840	-0.160	0.832	0.073	0.324	0.031	0.431	0.206	0.117
1 to 5	300	0.007	0.483	-0.164	0.870	0.108	0.232	-0.140	0.826	-0.080	0.702	0.034	0.410	-0.015	0.535	0.183	0.133
1 to 6	360	0.101	0.241	-0.092	0.746	0.116	0.197	0.074	0.294	-0.104	0.772	-0.071	0.696	0.106	0.236	0.188	0.109
1 to 7	420	0.024	0.429	-0.120	0.831	0.061	0.318	-0.021	0.565	-0.021	0.566	-0.036	0.607	0.125	0.177	0.122	0.182
1 to 8	480	0.015	0.451	-0.099	0.793	0.066	0.292	-0.012	0.539	0.019	0.437	-0.071	0.717	0.040	0.373	0.043	0.360
1 to 9	540	0.017	0.440	-0.104	0.815	0.001	0.498	-0.020	0.568	0.027	0.406	-0.040	0.635	0.016	0.446	-0.013	0.545
1 to 10	599	-0.005	0.518	-0.005	0.518	-0.005	0.518	-0.005	0.518	-0.005	0.518	-0.005	0.518	-0.005	0.518	-0.005	0.518

Table 1: Estimated treatment effects and one-sided p-values for cumulative deciles for the colon data

		ITE SRF		Interaction		2RF		BAFT		SCF		2AFT		TM		T3Q	
Deciles	Sample	β	pval	β	pval	β	pval	β	pval	β	pval	β	pval	β	pval	β	pval
1	69	0.264	0.181	0.917	0.023	0.279	0.181	-0.150	0.696	0.479	0.042	1.023	0.013	0.492	0.061	-0.073	0.574
1 to 2	137	0.659	0.005	0.610	0.011	0.527	0.015	0.081	0.339	0.321	0.044	0.521	0.021	0.523	0.022	0.472	0.047
1 to 3	206	0.444	0.020	0.402	0.032	0.522	0.002	0.383	0.016	0.095	0.286	0.546	0.009	0.310	0.057	0.227	0.139
1 to 4	274	0.396	0.017	0.505	0.004	0.531	0.001	0.288	0.031	0.197	0.108	0.394	0.015	0.386	0.010	0.248	0.078
1 to 5	343	0.387	0.006	0.453	0.001	0.458	0.000	0.335	0.009	0.299	0.016	0.410	0.002	0.247	0.046	0.259	0.054
1 to 6	411	0.381	0.004	0.353	0.004	0.396	0.001	0.363	0.002	0.299	0.007	0.358	0.004	0.216	0.050	0.281	0.021
1 to 7	480	0.382	0.001	0.364	0.002	0.296	0.005	0.327	0.003	0.300	0.006	0.342	0.003	0.164	0.087	0.289	0.009
1 to 8	548	0.385	0.000	0.279	0.008	0.254	0.010	0.311	0.002	0.353	0.001	0.258	0.012	0.283	0.005	0.285	0.006
1 to 9	617	0.301	0.002	0.298	0.002	0.298	0.002	0.367	0.000	0.359	0.000	0.281	0.004	0.296	0.002	0.286	0.003
1 to 10	686	0.309	0.001	0.309	0.001	0.309	0.001	0.309	0.001	0.309	0.001	0.309	0.001	0.309	0.001	0.309	0.001

Table 2: Estimated treatment effects and one-sided p-values for cumulative deciles for the GBSG2 data

7 Kaplan-Meier curves for the top 2 ITE deciles for all methods for the data examples

The following figures present the Kaplan-Meier curves of the control and treatment groups for the whole sample and for the top 2 deciles for all methods. The ones for the proposed methods are already in the article (Figures 3 and 5) and are repeated here for completeness.

We must interpret these plots with caution because they show raw estimated survival curves without controlling for the covariates effects. But still, the plots are coherent with the estimated treatment effects from the AFT models (that control for the other covariates) reported in Figures 4 and 6 of the article and Tables 1 and 2 of this document.

For the colon data, the proposed and TM methods are the only two with a treatment effect significantly greater than 0 when considering the top 2 ITE deciles. The survival curves for the top 2 ITE deciles in Figures 30 (the same as Figure 3 in the article) and 36 show that they are able to separate successfully the treatment and control groups. But it is interesting to see that they did not find the same subgroups. In fact, only about a third (39 out of the 120) observations in their first 2 deciles are common. This is shown by the fact that the top 2 deciles survival curves of the proposed method are higher than the ones of the TM method. This does not mean that one solution is better than the other. Both are separating the control and treatment groups very well. This shows one feature of methods that try to maximize the ITE. The identified subjects are not guaranteed to have a very high survival rate if treated. It only means that their survival rate should be much higher if we treat them compared to not treating them. Hence, the choice of the subgroups could be based on additional considerations. Some methods, namely BAFT, SCF, 2AFT, Interaction and T3Q, even have a negative estimated treatment effect for their top 2 ITE deciles (see Figure 4 in the article). The survival curves for the treatment group of these methods are also under the ones of the control group, showing that they were not successful (at least when considering their top 2 ITE deciles) at findings subjects that could potentially benefit the most from the treatment.

For the German breast cancer data, all methods have a positive estimated treatment effect at the top 2 ITE deciles (see Figure 6 in the article). But the BAFT method have an estimated treatment effect that is lower than the overall (whole sample) treatment effect. The top 2 ITE deciles survival curves indicate that the identified subgroups are not necessarily the same from one method to another.

7.1 Colon data

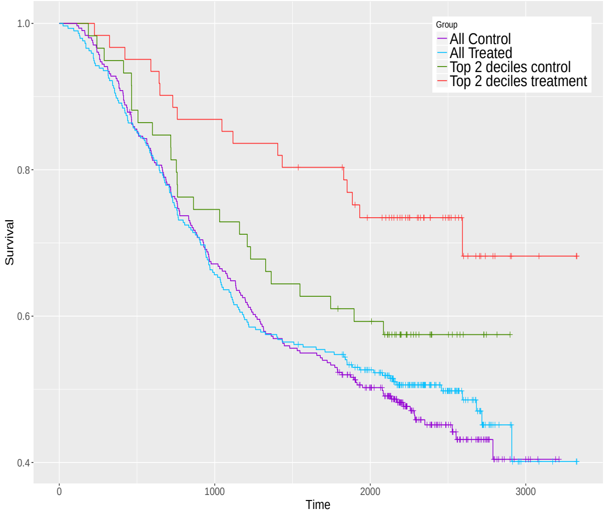

Figure 30: Colon Death - ITE SRF

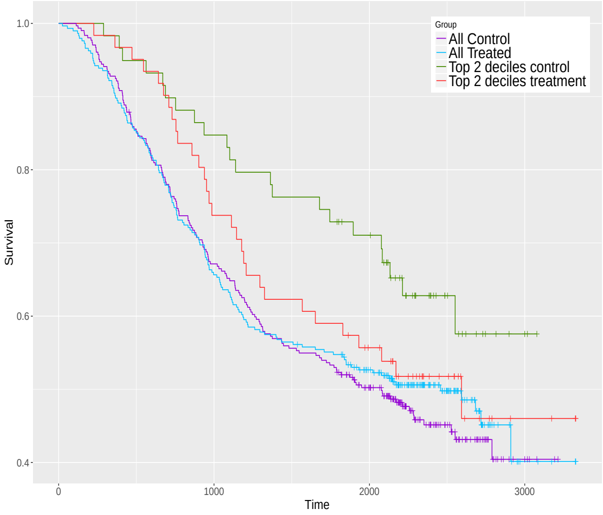

Figure 31: Colon Death - Interaction

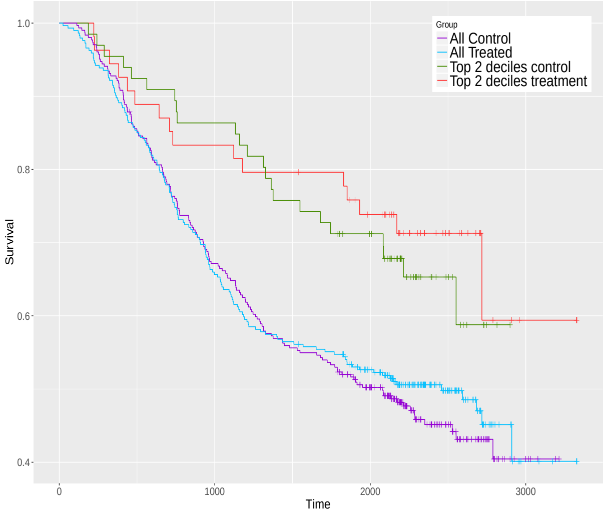

Figure 32: Colon Death - 2RF

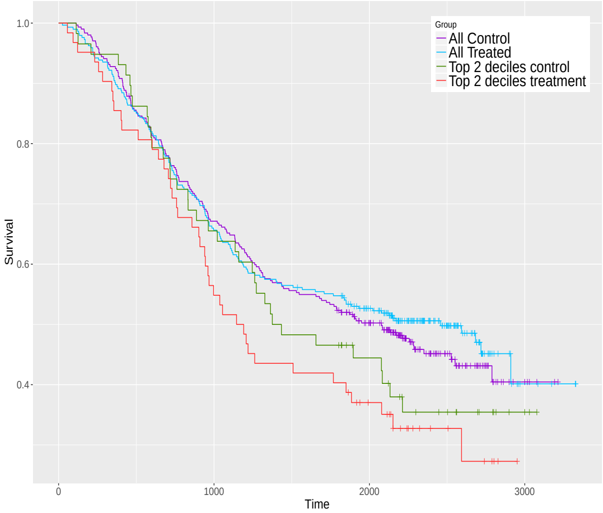

Figure 33: Colon Death - BAFT

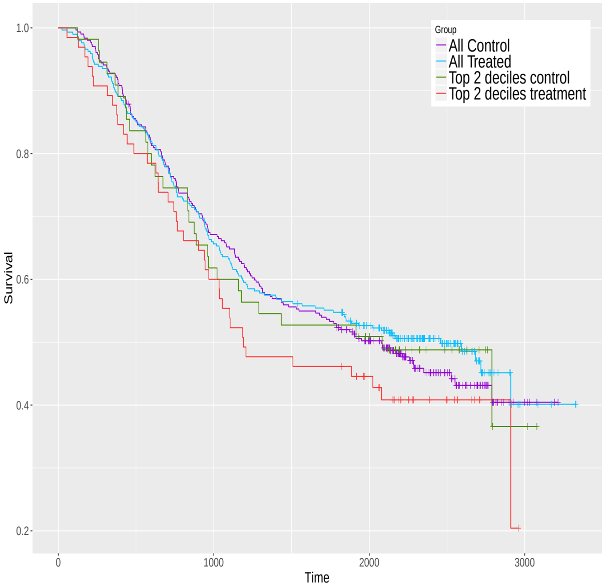

Figure 34: Colon Death - SCF

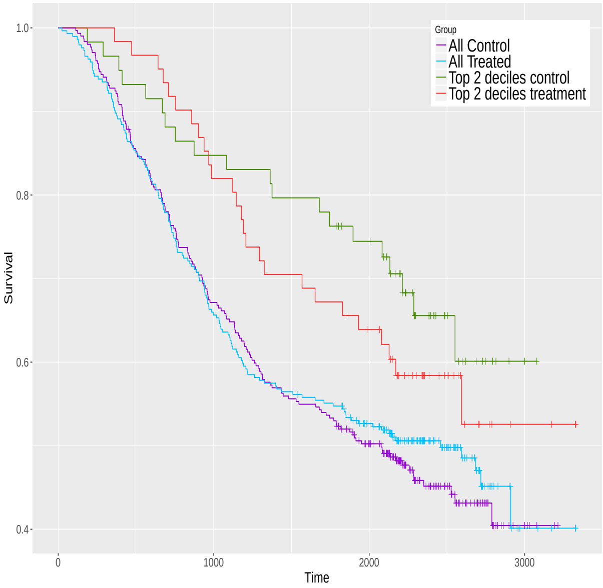

Figure 35: Colon Death - 2AFT

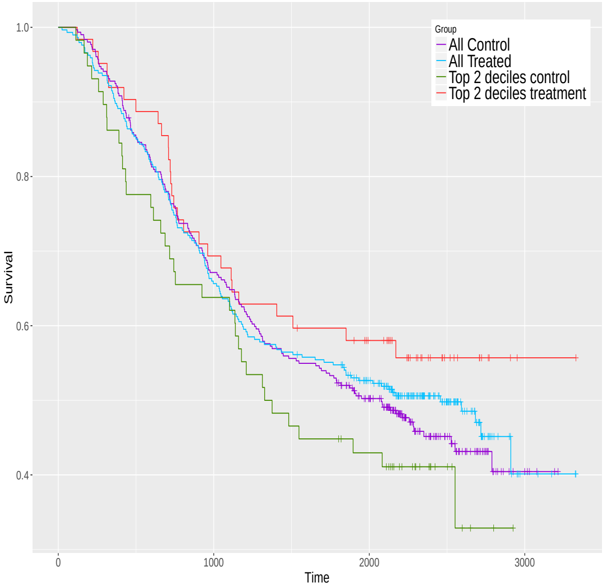

Figure 36: Colon Death - TM

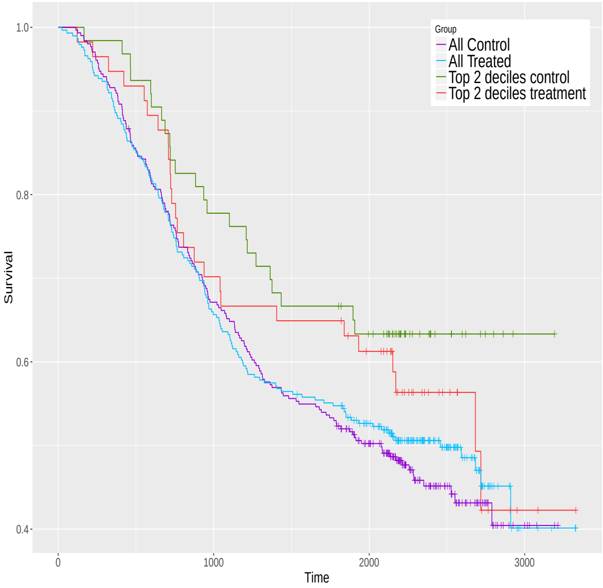

Figure 37: Colon Death - T3Q

7.2 German breast cancer data

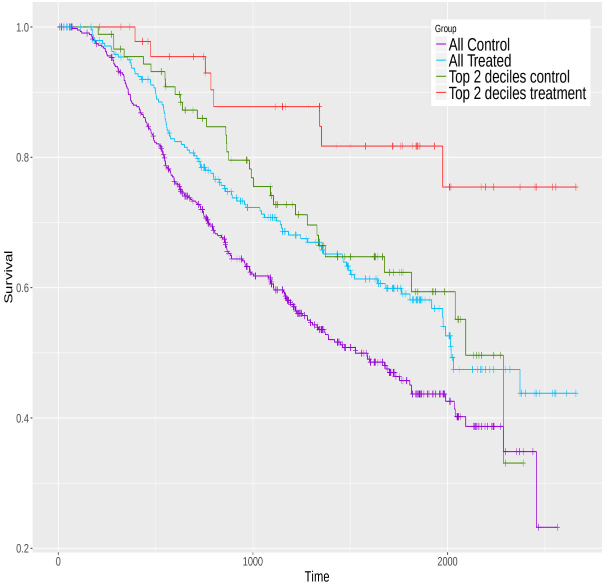

Figure 38: GBSG2 - ITE SRF

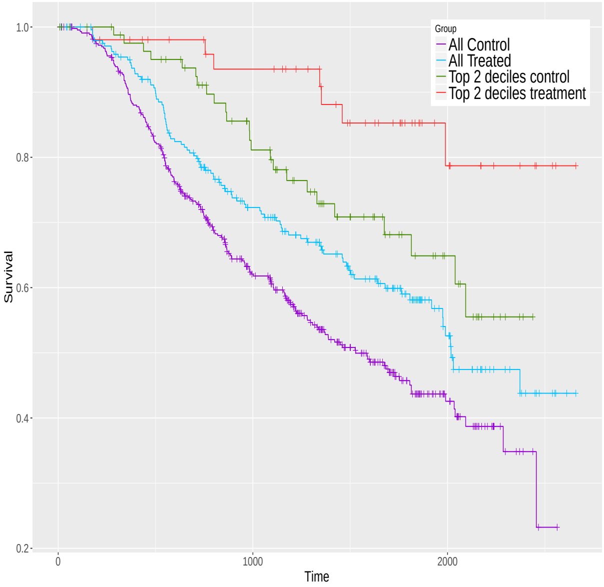

Figure 39: GBSG2 - Interaction

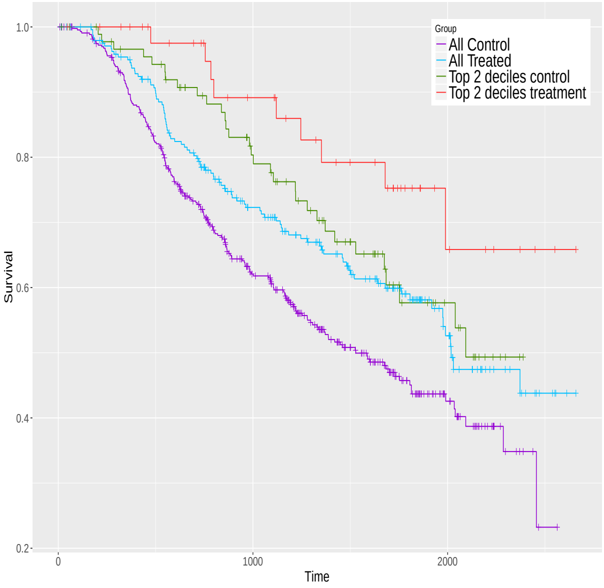

Figure 40: GBSG2 - 2RF

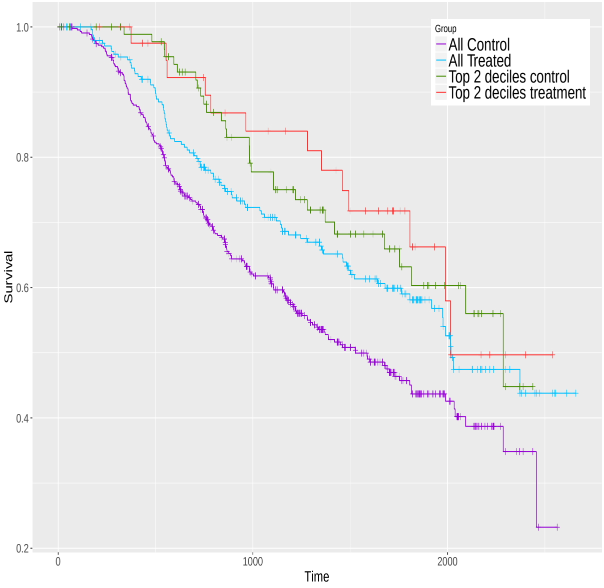

Figure 41: GBSG2 - BAFT

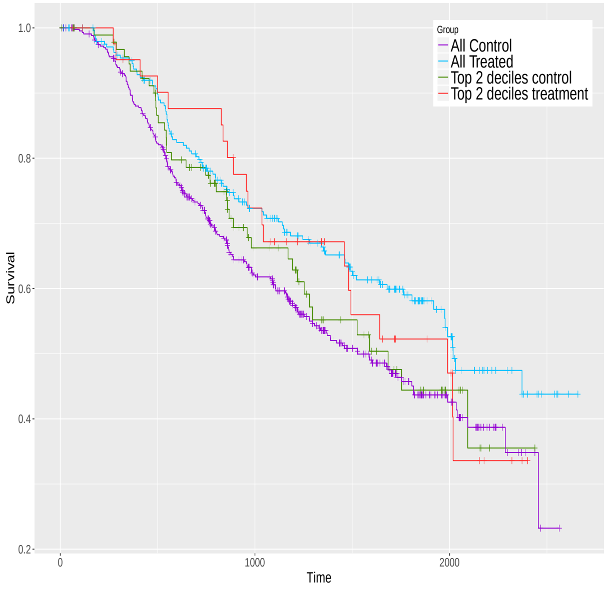

Figure 42: GBSG2 - SCF

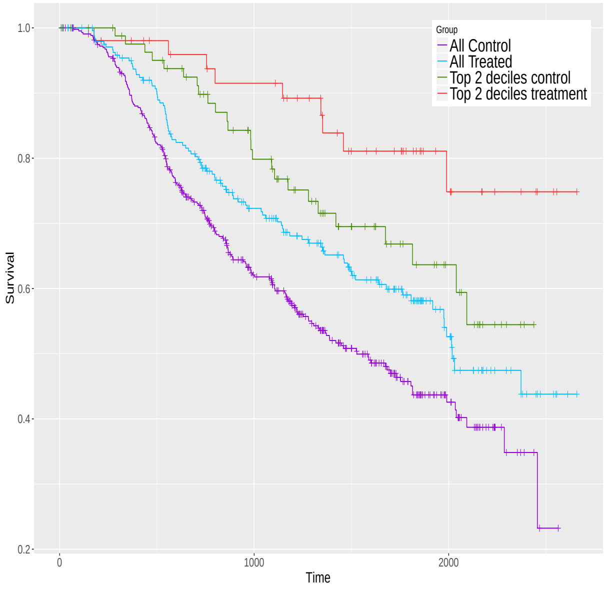

Figure 43: GBSG2 - 2AFT

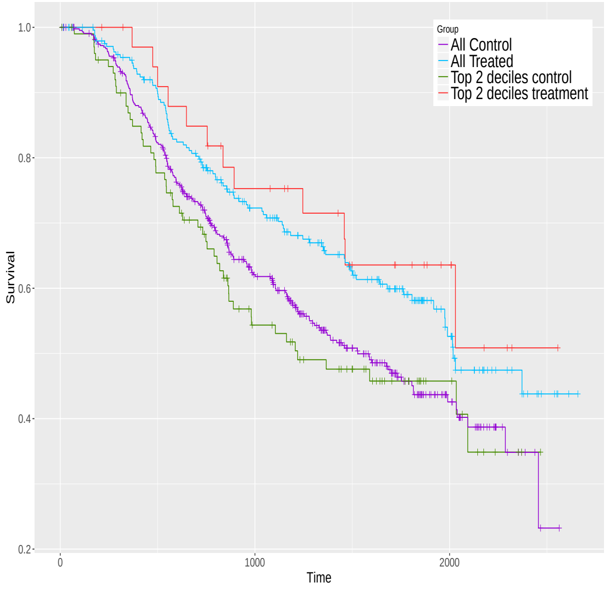

Figure 44: GBSG2 - TM

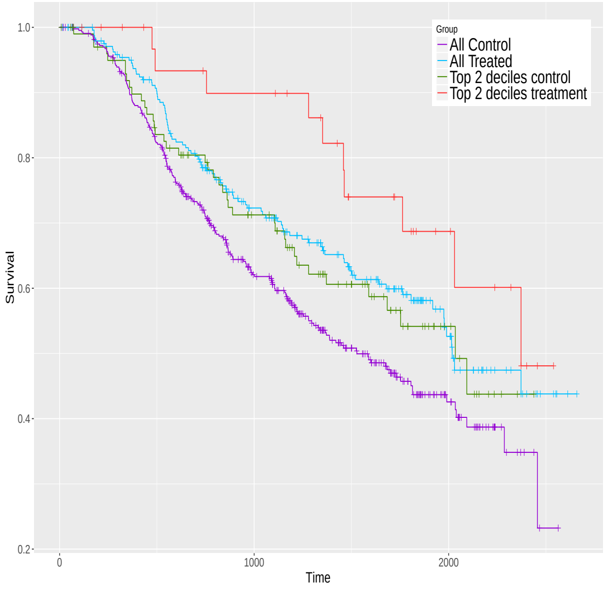

Figure 45: GBSG2 - T3Q