Causal Survival Analysis: Practical Recommendations

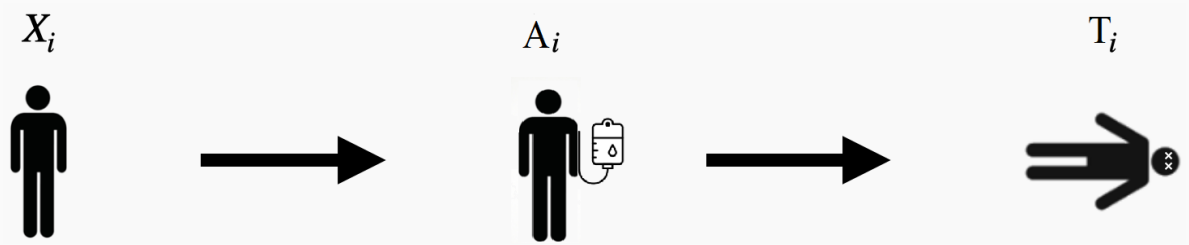
Supervisors: Julie Josse (Inria) & Bernard Sebastien (Sanofi R&D)

Joint work with: Clément Berenfeld & Imke Mayer

Charlotte VOINOT - PhD with Sanofi R&D and Inria (National Institute for Research in Digital Science and Technology)

May 23, 2025

Introduction to Causal Inference

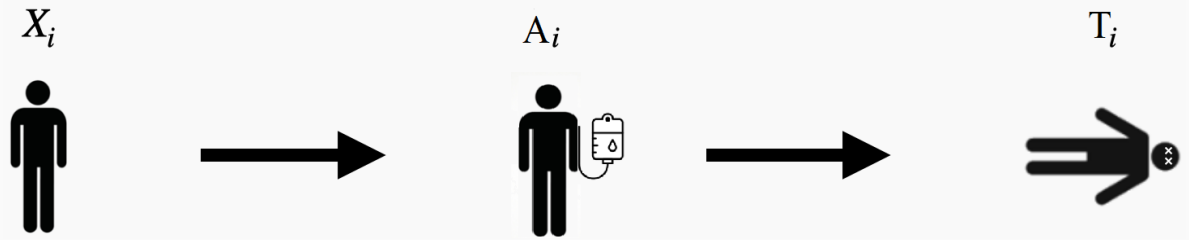


"People who are treated tend to have a smaller life expectancy"

$$\Rightarrow N \text{ i.i.d. } (\underbrace{X_i}_{\text{covariates}} , \overbrace{A_i}^{\text{treatment}} , \underbrace{Y_i}_{\text{outcome}}) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R} \times \mathbb{R}$$

Covariates		Treatment	Potential outcomes		Outcomes
X_1	X_2	A	$Y(0)$	$Y(1)$	Y
1	24	1	?	200	200
2	52	0	100	?	100
1	33	1	?	200	200

Introduction to Causal Inference



"People who are treated tend to have a smaller life expectancy"

$$\Rightarrow N \textbf{ i.i.d. } (\underbrace{X_i}_{\textbf{covariates}} , \overbrace{A_i}^{\textbf{treatment}} , \underbrace{Y_i}_{\textit{outcome}}) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R} \times \mathbb{R}$$

Covariates		**Treatment**	**Potential outcomes**		**Outcomes**
X_1	X_2	A	$Y(0)$	$Y(1)$	Y
1	24	1	?	200	200
2	52	0	100	?	100
1	33	1	?	200	200

Potential outcomes

$$\Rightarrow N \textbf{ i.i.d. } (\underbrace{X_i}_{\textbf{covariates}} , \overbrace{T_i}^{\textbf{treatment}} , \underbrace{Y_i}_{\textit{outcome}}) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R} \times \mathbb{R}$$

Covariates		**Treatment**	**Potential outcomes**		**Outcomes**
X_1	X_2	A	$Y(0)$	$Y(1)$	Y
1	24	1	?	200	200
2	52	0	100	?	100
1	33	1	?	200	200

Our goal is to compute the individual causal effect of the treatment:

$$\Delta_i = Y_i(1) - Y_i(0)$$

However we can never observed Δ_i (only one observed outcome/indiv)

Average Treatment Effect

Our goal is to compute the individual causal effect of the treatment:

$$\Delta_i = Y_i(1) - Y_i(0)$$

In order to fixe the fundamental problem of causal inference define the Average Treatment Effect.

Average Treatment Effect (ATE)

$$\tau = \mathbb{E}[\Delta] = \mathbb{E}[Y(1) - Y(0)]$$

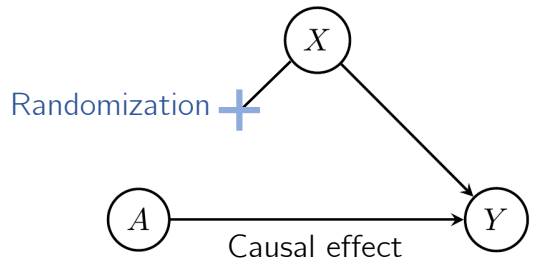
τ is also referred as the risk difference.
$\Rightarrow$ depends on the population

The ATE is the difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten the treatment.

Introduction to Causal Inference

Causal inference aims to estimate the effect of a treatment A (comparing all patients receiving A versus all patients receiving the control) on a fully observed outcome Y

RCT
We intervene (by randomization)

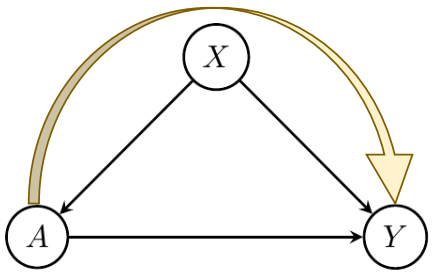


- ATE (Average Treatment Effect) can be derived by a simple difference in mean.

$$\tau_{ATE} = \mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$$

Observational study
We observe (collection)

Open confounding path



ex: Severe patients are more likely to have the treatment but they are also more likely to die.
→ Naively, we will observe that the treatment kill people.

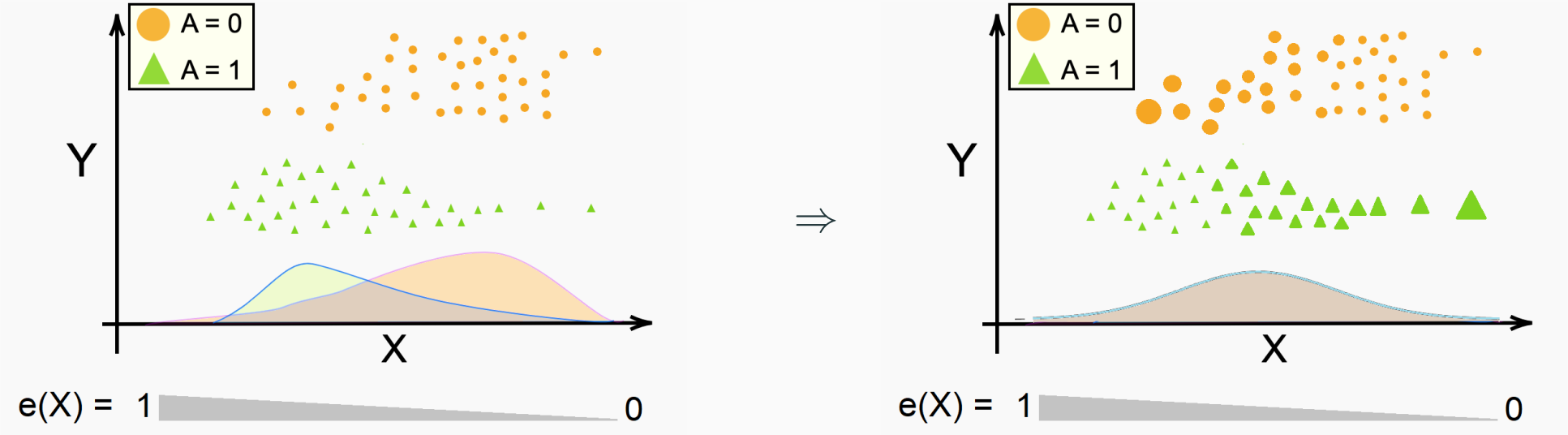
- ATE has to be derived using more sophisticated estimator (IPW, G-formula, AIPW).

- A is the treatment: $A \in \mathrm{A} = \{0,1\}$.
- X is a p-dimensional vector of baseline covariates: $X \subset \mathbb{R}^p$.
- Y is the outcome: $Y \in \mathbb{R}$.

The Inverse Probability Weighting (IPW) Estimator

IPW estimator is one of the most widely used estimators in classical causal inference.

Key Idea: We assign higher weights to underrepresented groups and lower weights to overrepresented ones (**mimic a RCT**)



The Inverse Probability Weighting (IPW) Estimator

IPW estimator is one of the most widely used estimators in classical causal inference.

Key Idea: We assign higher weights to underrepresented groups and lower weights to overrepresented ones (**mimic a RCT**)

Formula:

$$\hat{\tau}_{\mathrm{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{e(X_i)} - \frac{(1-A_i)Y_i}{1-e(X_i)}$$

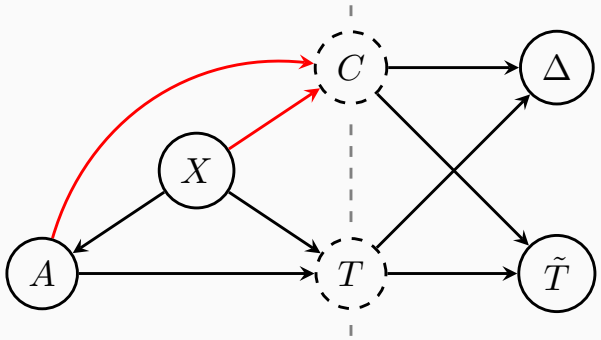
where $e(X) = P(A = 1|X)$ is the **propensity score** adjusted on X.

Properties: Consistent under classical causal assumptions:

- **Stable Unit Treatment Value Assumption:** $Y = A.Y(1) + (1-A).Y(0)$.
- **Overlap:** $0 < e(X) < 1$ almost surely.
- **Unconfoundedness:** $A \perp\!\!\!\perp (Y(0), Y(1)) \mid X$.

Causal Survival analysis: Causal inference and Survival analysis

Causal inference  Survival analysis



$\Rightarrow$ n **i.i.d.** $(\underbrace{X_i}_{\textbf{covariates}} , \underbrace{A_i}_{\textbf{treatment}} , \underbrace{T_i}_{\textit{outcome}}) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R}^+$

Covariates		**Treatment**	**Censoring**	**Status**	**Outcomes**		
X_1	X_2	**A**	**C**	Δ	$T(0)$	$T(1)$	$\tilde{T}$
1	24	1	?	1	?	200	200
2	52	0	?	1	100	?	100
1	33	1	200	0	?	?	200

Potential outcome framework of Rubin 1974

In grey, the observed data: $(X_i, A_i, \Delta_i, \widetilde{T}_i)$ with $\widetilde{T}_i = min(T_i, C_i)$

Causal effect in survival analysis

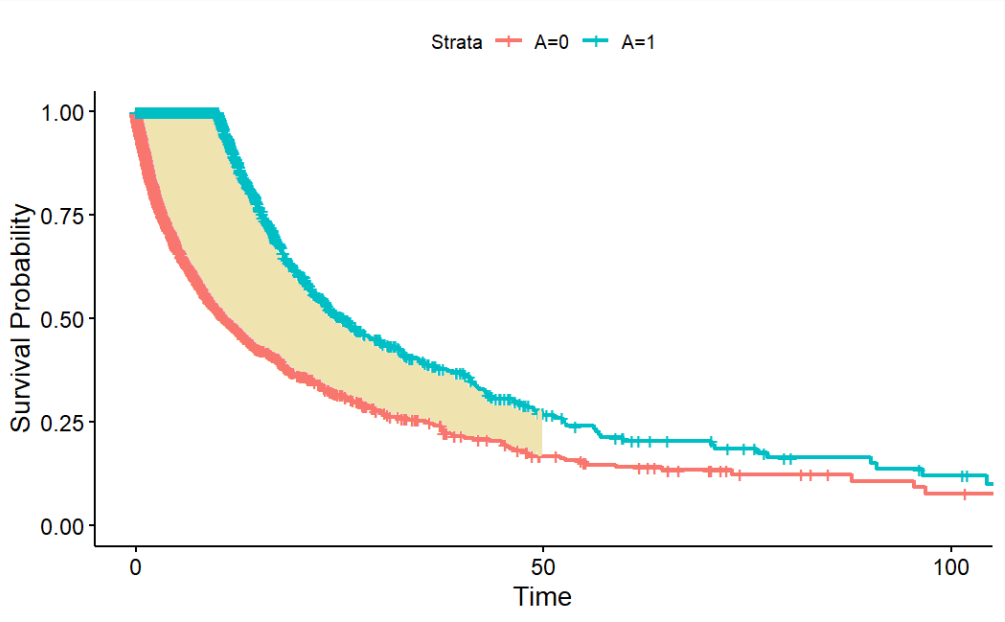
Restricted Mean Survival Time (RMST)

$$\mathrm{RMST}(\tau) = \mathbb{E}[min(T, \tau)] = \int_0^{\tau} \hat{S}(t)dt$$

RMST can be defined as a measure of average survival from time 0 to time τ a **fixed time horizon**

Average treatment effect in survival analysis

$$\hat{\theta}_{RMST}(\tau) = \mathrm{RMST}_1(\tau) - \mathrm{RMST}_0(\tau)$$



$\hat{\theta}_{RMST}(\tau = 50) = 10$ means that on average the treatment increases the survival time by 10 days at 50 days.

Figure 1: Plot of stratified kaplan meier survival function and the representation of $\theta_{RMST}(\tau = 50)$ (in yellow)

Why using RMST instead of Hazard Ratio ?

Limitations of the Hazard Ratio (HR):

- Assumes **proportional hazards**.
- Hazard ratios **are not causally interpretable** (even in RCT unless there is no treatment effect) due to a built-in selection bias. [Martinussen and Vansteelandt 2013; Martinussen, Vansteelandt, and Andersen 2020]
- **Not collapsible**.[Huitfeldt, M. Stensrud, and Suzuki 2019; Greenland, Robbins, and Pearl 1999]
- Difficult to interpret clinically. [Hernán 2010]

Advantages of Restricted Mean Survival Time (RMST):

- **Causal measure** as it is the extension of Average Treatment Effect. [Royston and Parmar 2013]
- Does not necessarily require the assumption of proportional risks.
- Provides an **absolute measure**.
- **Clinically interpretable**: "How much longer does a patient survive on average within a given time frame?"

Built-in selection bias HR

Here is the definition of Hazard ratio:

$$\exp(\beta) = \frac{\log P(T^1 > t)}{\log P(T^0 > t)}$$

Under proportional hazard assumption, it goes:

$$\begin{aligned}\exp(\beta) &= \frac{\lim\limits_{h\to 0} P(t \leqslant T < t+h | T \geqslant t, A = 1)}{\lim\limits_{h\to 0} P(t \leqslant T < t+h | T \geqslant t, A = 0)} \\ &= \frac{\lim\limits_{h\to 0} P(t \leqslant T^1 < t+h | T^1 \geqslant t)}{\lim\limits_{h\to 0} P(t \leqslant T^0 < t+h | T^0 \geqslant t)}\end{aligned}$$

First, the right-hand expression shows that $\exp(\beta)$ contrasts the hazard functions **with** and **without** intervention for two separate groups of individuals who survive time $t > 0$.
If it exists a treatment effect then, population **with** and **without** intervention cannot be compared anymore.

Built-in selection bias HR

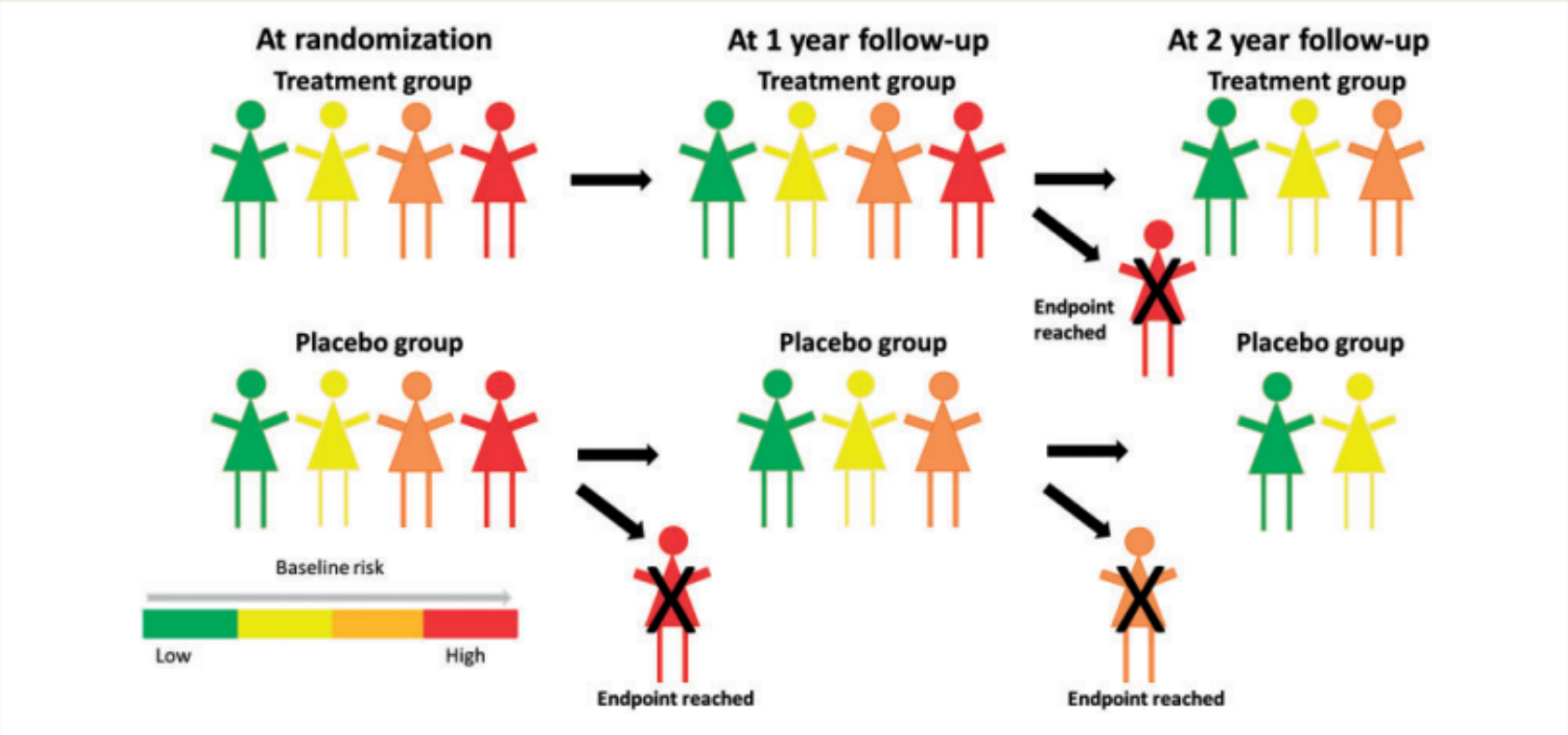


Figure 2: This schematic drawing illustrates the built-in selection bias in population-level hazard ratios: by definition, the population-level hazard ratio at a given time point is based on individuals who survived up to that time point, thereby it is a comparison between the unbalanced groups (from M. J. Stensrud et al. 2019)

Several ways of estimating $\hat{\theta}_{RMST}(\tau)$ in Observational studies

Estimators based on Censoring unbiased transformations

Estimators based on survival functions

- Kaplan meier like estimates
- G-formula

Doubly robust estimator

Estimators based on Censoring unbiased transformation

The basic requirement is to create an fully observable variable T^* such that $E(T^*|X,A) = E(T \wedge \tau|X,A)$. [Fan and Gijbels 1994]

IPC transformation [Koul, Susarla, and Ryzin 1981]:

$$T^*_{\mathrm{IPCW}} = \frac{\Delta^\tau}{G(\widetilde{T} \wedge \tau|X,A)} \widetilde{T} \wedge \tau$$

with $G(\widetilde{T} \wedge \tau|X,A) := \mathbb{P}(C \geqslant t|X,A)$ being the left limit of the **conditional survival function of the censoring**. $\Delta^\tau := I\{T \wedge \tau \leqslant C\}$ is the censoring indicator of the restricted time.

Buckley-James transformation [Buckley and James 1979]:

$$T^*_{\mathrm{BJ}} = \Delta^\tau(\widetilde{T} \wedge \tau) + (1 - \Delta^\tau) Q_S(\widetilde{T} \wedge \tau|X,A)$$

where, for $t \leqslant \tau$, $Q_S(t|X,A) := \mathbb{E}[T \wedge \tau|X,A,T \wedge \tau > t]$ estimated with the **conditional survival function**.

Estimators based on Censoring unbiased transformation

The basic requirement is to create an fully observable variable T^* such that $E(T^*|X, A) = E(T \wedge \tau|X, A)$. [Fan and Gijbels 1994]

IPC transformation [Koul, Susarla, and Ryzin 1981]:

$$T^*_{\text{IPCW}} = \frac{\Delta^\tau}{G(\widetilde{T} \wedge \tau|X, A)} \widetilde{T} \wedge \tau$$

with $G(\widetilde{T} \wedge \tau|X, A) := \mathbb{P}(C \geqslant t|X, A)$ being the left limit of the **conditional survival function of the censoring**. $\Delta^\tau := I\{T \wedge \tau \leqslant C\}$ is the censoring indicator of the restricted time.

Buckley-James transformation [Buckley and James 1979]:

$$T^*_{\text{BJ}} = \Delta^\tau(\widetilde{T} \wedge \tau) + (1 - \Delta^\tau) Q_S(\widetilde{T} \wedge \tau|X, A)$$

where, for $t \leqslant \tau$, $Q_S(t|X, A) := \mathbb{E}[T \wedge \tau|X, A, T \wedge \tau > t]$ estimated with the **conditional survival function**.

Estimators based on Censoring unbiased transformation

The basic requirement is that $E(T^*|X, A) = E(T \wedge \tau|X, A)$.

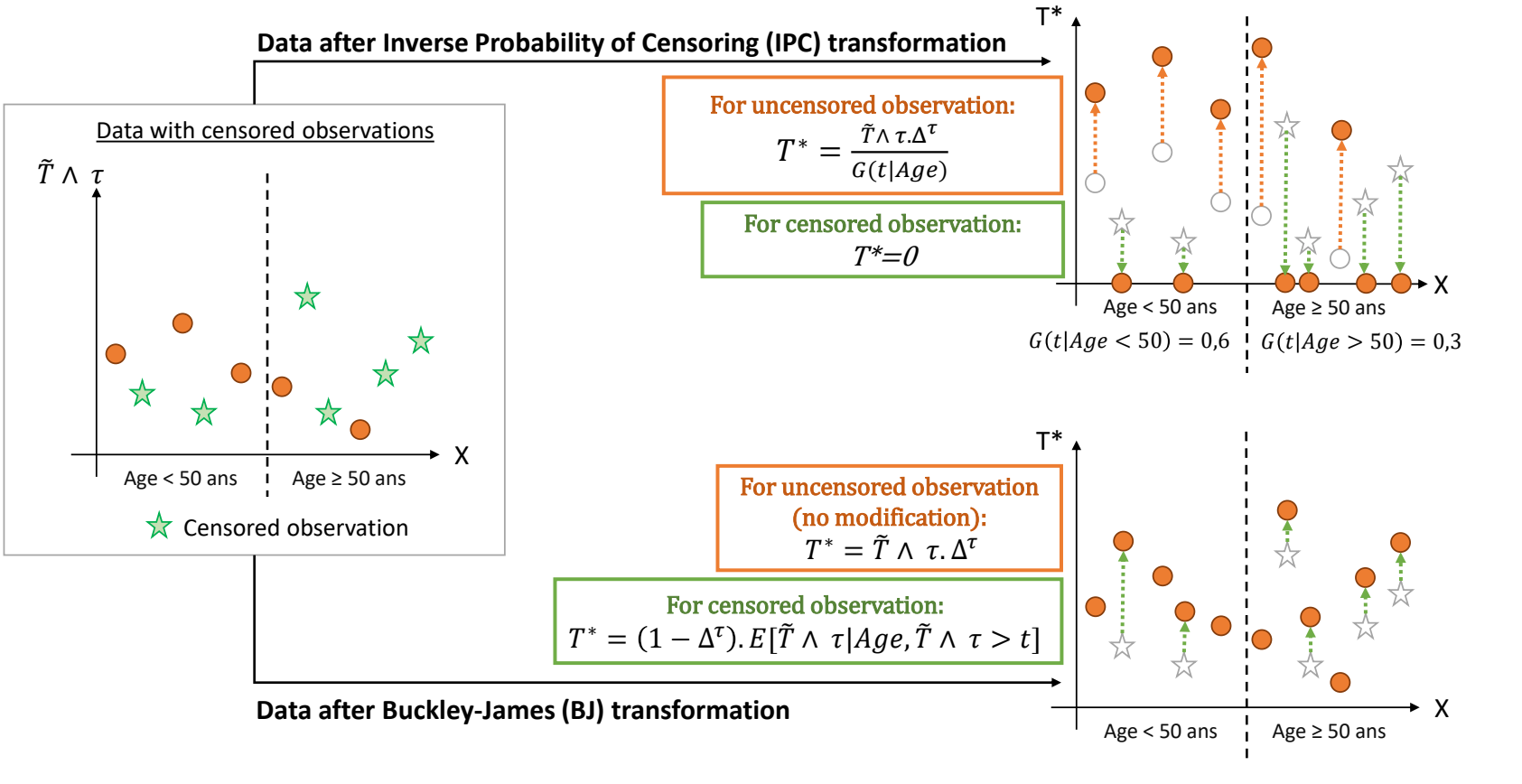


Figure 3: Illustration of Inverse-Probability-of-Censoring and Buckley-James transformations

Estimators based on Censoring unbiased transformation

IPTW-IPCW and IPTW-BJ estimator of ATE

It combines censoring unbiased transformation and IPW estimator:

$$\widehat{\theta}_{\mathrm{RMST}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i}{\hat{e}(X_i)} - \frac{1-A_i}{1-\hat{e}(X_i)}\right).T_i^*$$

with n corresponds to the number of observations, T^* corresponds to the censoring transformation (IPCW or BJ) and $e(X_i)$ the propensity score (probability of being treated).

Estimates are consistent under standard causal inference, survival and estimation assumptions :

- **Stable Unit Treatment Value Assumption:** $T = A.T(1) + (1-A).T(0)$.
- **Overlap:** $0 < e(X) < 1$ almost surely.
- **Unconfoundedness:** $A \perp\!\!\!\perp (T(0), T(1)) \mid X$.
- **Conditionally independent censoring:** $C \perp\!\!\!\perp (T_0, T_1) \mid X, A$
- The quantities used to define the estimators **are consistently estimated**.

Several ways of estimating $\hat{\theta}_{RMST}(\tau)$ in Observational studies

Estimators based on Censoring unbiased transformations

Estimators based on survival functions

- Kaplan meier like estimates
- G-formula

Doubly robust estimator

Estimator based on survival functions

IPW estimator can be extended to causal survival analysis

IPTW adjusted Kaplan-Meier estimator

$$\widehat{S}_{\mathrm{IPTW}}(t|A=a) = \prod_{t_k \leqslant t} \left(1 - \frac{D_k^{\mathrm{IPTW}}(a)}{N_k^{\mathrm{IPTW}}(a)}\right).$$

with $D_k^{\mathrm{IPTW}}(a) := \sum_{i=1}^{n} \left(\frac{a}{\widehat{e}(X_i)} + \frac{1-a}{1-\widehat{e}(X_i)}\right) .I(\widetilde{T}_i = t_k, A_i = a)$ and
$N_k^{\mathrm{IPTW}}(a) := \sum_{i=1}^{n} \left(\frac{a}{\widehat{e}(X_i)} + \frac{1-a}{1-\widehat{e}(X_i)}\right) .I(\widetilde{T}_i \geqslant t_k, A_i = a)$ and
$\hat{e}(X_i) = P(A_i = 1|X_i)$, the propensity score.
Then, the RMST estimator can be derived as:

$$\hat{\theta}_{RMST} = \int_0^{\tau} \hat{S}_{IPTW}(t, A=1) - \hat{S}_{IPTW}(t, A=0)dt$$

- Estimates are consistent under the some standard causal inference, survival and estimation assumptions (Unconfoundedness, Consistency, Overlap, Independent Censoring and that the quantities are consistently estimated).

Estimator based on survival functions

Even if censoring transformation refers to a fully observed data, weighted survival function using transformation can be found in articles.

IPTW-IPCW adjusted Kaplan-Meier estimator

$$\widehat{S}_{\text{IPTW}-\text{IPCW}}(t|A=a) = \prod_{t_k \leqslant t} \left(1 - \frac{D_k(a)}{N_k(a)}\right).$$

with $D_k(a) = \sum_{i=1}^{n} \left(\frac{a}{\widehat{e}(X_i)} + \frac{1-a}{1-\widehat{e}(X_i)}\right) \cdot \frac{\Delta_i^{\tau}}{\hat{G}(\tilde{T}\wedge\tau|X_i=x,A_i=a)} . I(\widetilde{T}_i = t_k, A_i = a)$ and $N_k(a) := \sum_{i=1}^{n} \left(\frac{a}{\widehat{e}(X_i)} + \frac{1-a}{1-\widehat{e}(X_i)}\right) \cdot \frac{\Delta_i^{\tau}}{\hat{G}(\tilde{T}\wedge\tau|X_i=x,A_i=a)} . I(\widetilde{T}_i \geqslant t_k, A_i = a)$ and $\hat{e}(X_i) = P(A_i = 1|X_i)$, the estimated propensity score.

Then, the RMST estimator can be derived as:

$$\hat{\theta}_{RMST} = \int_0^{\tau} \hat{S}_{IPTW-IPCW}(t, A=1) - \hat{S}_{IPTW-IPCW}(t, A=0)dt$$

- Estimates are consistent under the some standard causal inference, survival and estimation assumptions (Unconfoundedness, Consistency, Overlap, Conditionally Independent Censoring and that the quantities are consistently estimated).

Estimator based on survival functions

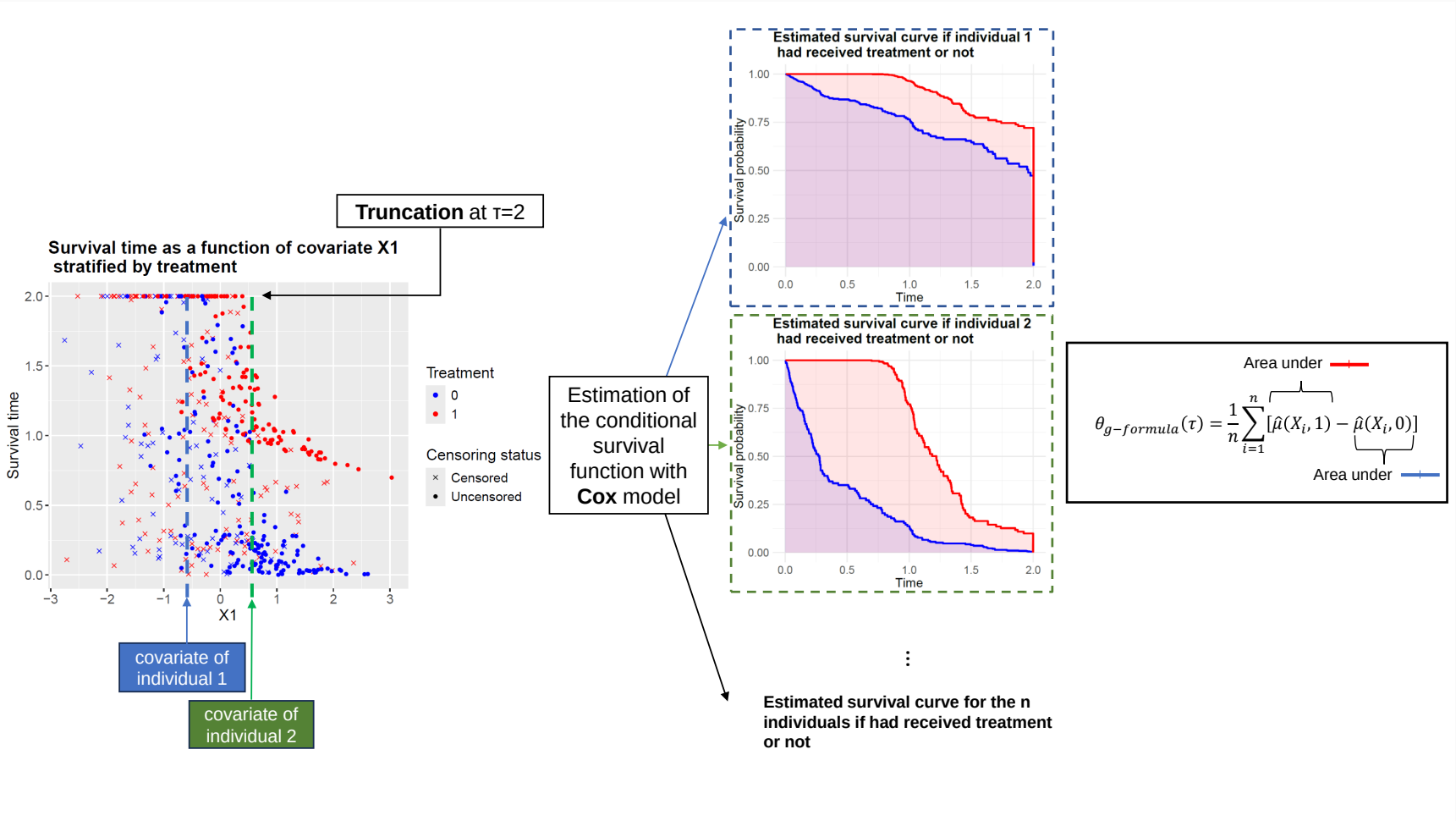
G-formula estimator

$$\widehat{\theta}_{\mathrm{G-formula}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(X_i, 1) - \widehat{\mu}(X_i, 0).$$

with $\widehat{\mu}_a \triangleq \mathbb{E}[T \wedge \tau \mid X = x, A = a] = \int_0^\tau S(t|X = x, A = a)dt$ the integral of the **conditional survival function** truncated at τ.

- Various potential methods to estimate the **conditional survival function**:
  - using a parametric model (Weibull, Exponential)
  - using a semi-parametric model (Cox model)
  - using a "model free" (Survival Forest)
- Estimates are consistent under the some standard causal inference, survival and estimation assumptions (Unconfoundedness, Consistency, Overlap, Conditionally Independent Censoring and that the quantities are consistently estimated).

Estimator based on survival functions



Two possibilities for estimating $\hat{S}(t|x,a)$:

- S-learner: **Fit one model** on all data with adjustment on X and A.
- T-learner: **Fit two models** (stratified analysis) on data with A=1 and with A=0 and adjustment on X.

Several ways of estimating $\hat{\theta}_{RMST}(\tau)$ in Observational studies

Estimators based on Censoring unbiased transformations

Estimators based on survival functions

- Kaplan meier like estimates
- G-formula

Doubly robust estimator

Doubly robust estimator

Augmented estimator: AIPTW-AIPCW [Ozenne et al. 2020]

$$\Delta^*_{\mathrm{QR}}(G, S, \mu, e) := \left(\frac{A}{e(X)} - \frac{1-A}{1-e(X)} \right) \left(T^*_{\mathrm{DR}}(G, S) - \mu(X, A) \right) + \mu(X, 1) - \mu(X, 0)$$

$$\widehat{\theta}_{\mathrm{AIPTW-AIPCW}} := \frac{1}{n} \sum_{i=1}^{n} \Delta^*_{\mathrm{QR}}(\widehat{G}, \widehat{S}, \widehat{\mu}, \widehat{e}).$$

$\Rightarrow$ 3 nuisance parameters to compute :

- Censoring model : $C \sim A + X$
- Propensity score model : $A \sim X$
- Conditional survival : $T \sim A + X$

Called doubly robust because the estimator is consistent if among the three quantities, conditional survival **or** propensity and censoring model are consistently estimated. Also, this estimator is consistent in the context of Unconfoundedness, Consistency, Overlap and Conditionally Independent Censoring.

Nuisance parameters for each estimator

Estimator	Context of application	Outcome model	Censoring model	Treatment model	Robustness
IPTW-KM	Obs. & Indep. cens.			(e)	No
IPCW-IPTW-KM	Obs & Conditionally Indep. cens.		(G)	(e)	No
G-formula		(μ)			No
IPTW-BJ		(Q_S)		(e)	No
AIPTW-AIPCW		(Q_S, μ)	(G)	(e)	Yes

All the models required for computing θ_{RMST} are referred to as **nuisance parameters**, as they are used to estimate quantities that may affect the convergence of the final estimator.

$\Rightarrow$ Consistency of the nuisance parameter estimators implies consistency of the RMST estimator except AIPTW-AIPCW estimator which requires either censoring and treatment or survival well specified

Three types of simulation of observational study

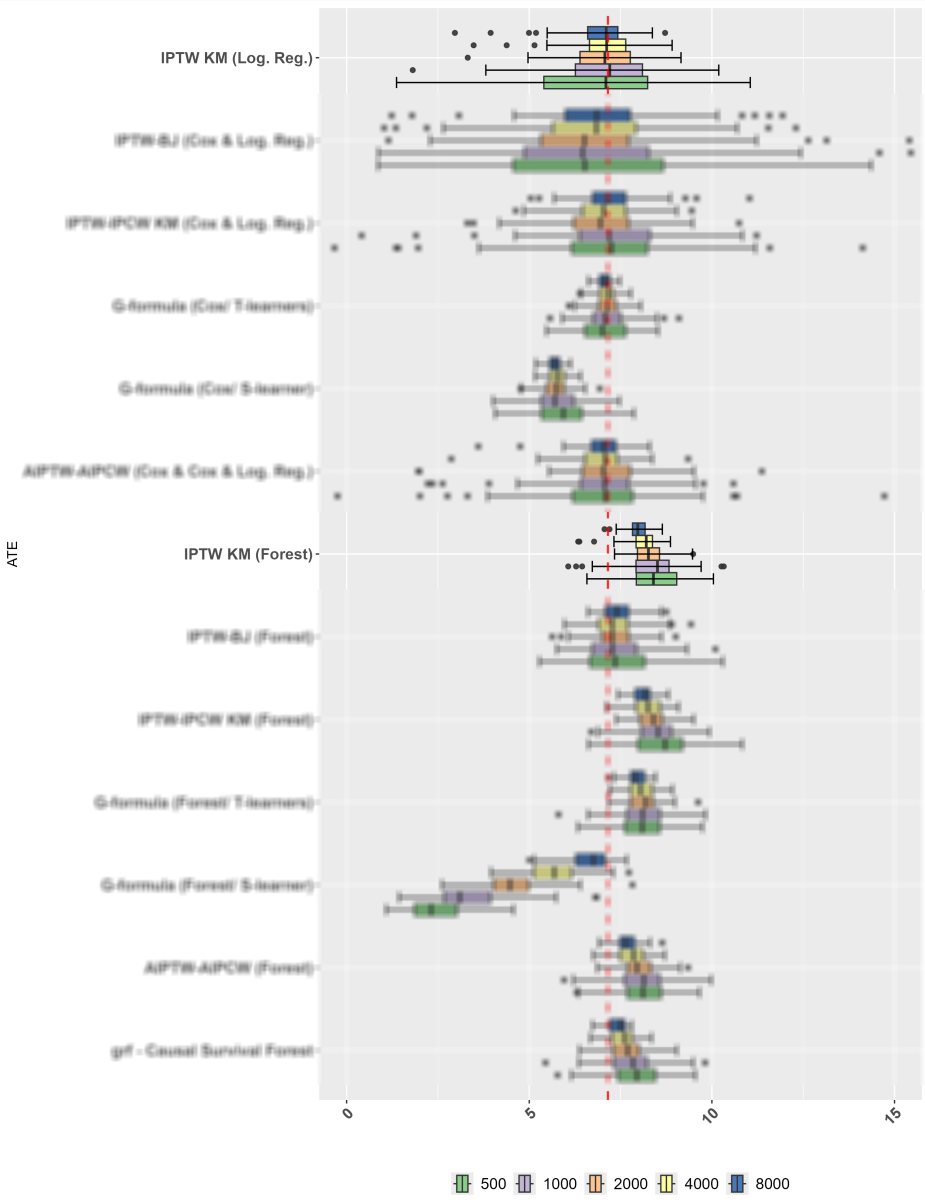
1 & 2 - **Observational study with independent censoring (1) or dependent censoring (2)** to assess the bias and variability of θ_{RMST} estimators under various context and sample size of observational studies:

- Assuming the working models are well specified : 4 predictors normally distributed
- T(0) $\sim$ exponential distribution with parameter=linear combination of predictors (without interaction) and T(1)= T(0) +10.
- Propensity score defined as linear combination of predictors on logistic scale
- **Independent censoring:** Censoring $\sim$ exponential distribution with parameter = constant.
  Dependent censoring: Censoring $\sim$ exponential distribution with parameter = linear combination of predictors (without interaction)

3 - **Observational study with dependent censoring (Assuming Model misspecifications):** also exponential distribution but the parameter of T(0) depends on interactions of the predictors. Those interactions are ignored in the working models to assess the performance of θ_{RMST} estimators in this context.

Obs & Independent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



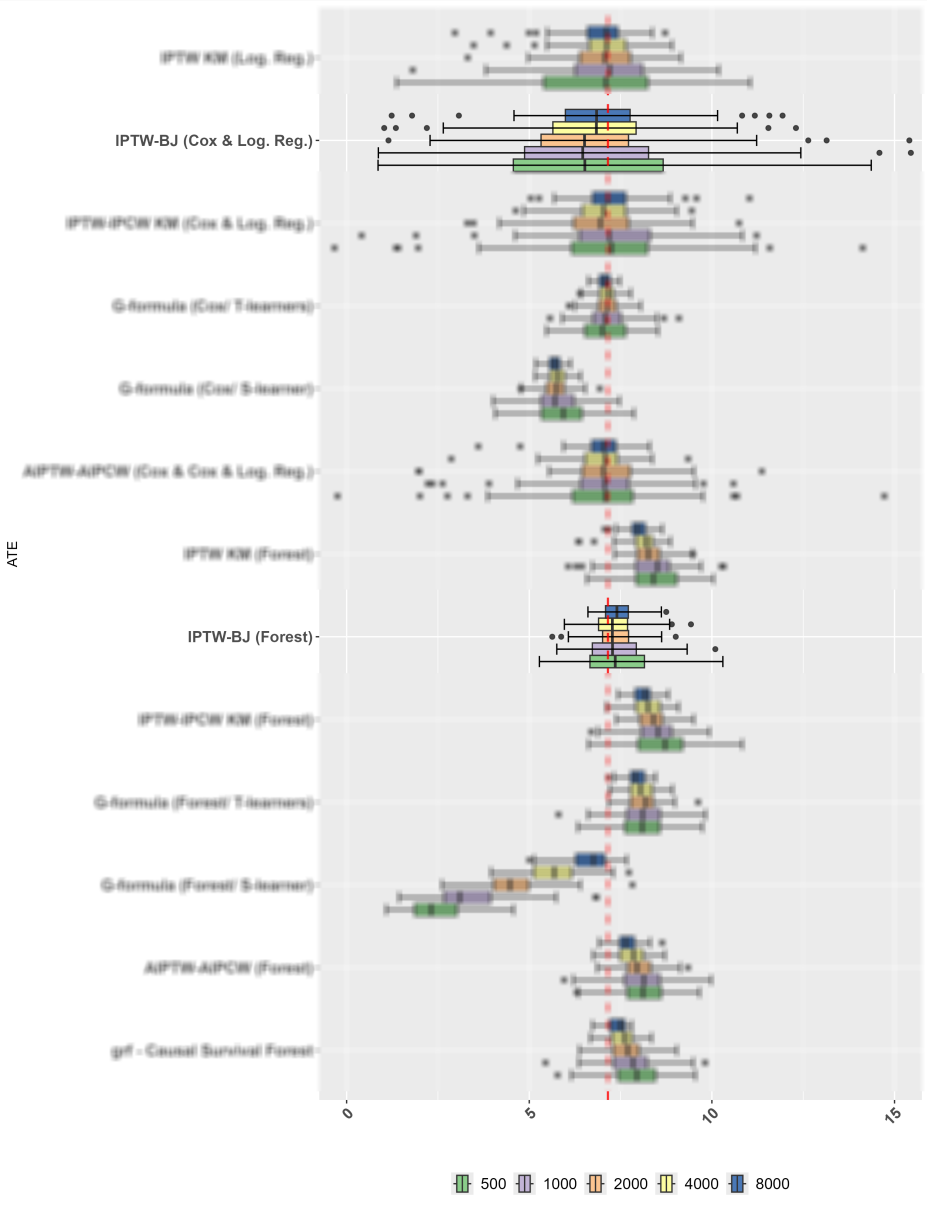
In this simulation, all models are well specified and we present only estimators we expect to converge

1. IPTW KM with propensity score estimated using logistic regression needs less observations to converge. When we use probability forest for the propensity score, IPTW KM needs much more data to converge.
2. In contrary, IPTW-BJ estimated with Cox model and logistic regression needs much more data than IPTW-BJ estimated with probability and random forest.
3. Same as 1 for IPTW-IPCW (Forest converges at high sample size).
4. A simple difference in adding the treatment (S-learner vs T-learners) in the adjustment set in Cox model leads to a non-proportional hazard (warning on the use of S-learner). In forest version, S-learner seems to converge at smaller sample size (no stratification of analysis, then learn on all data).
5. Same as 1 for AIPTW-AIPCW (Forest converges slower).
6. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It converges with few sample size as all the other forest version (needs to be prioritize).

- To conclude, in average forest versions need more data to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox.

Obs & Independent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



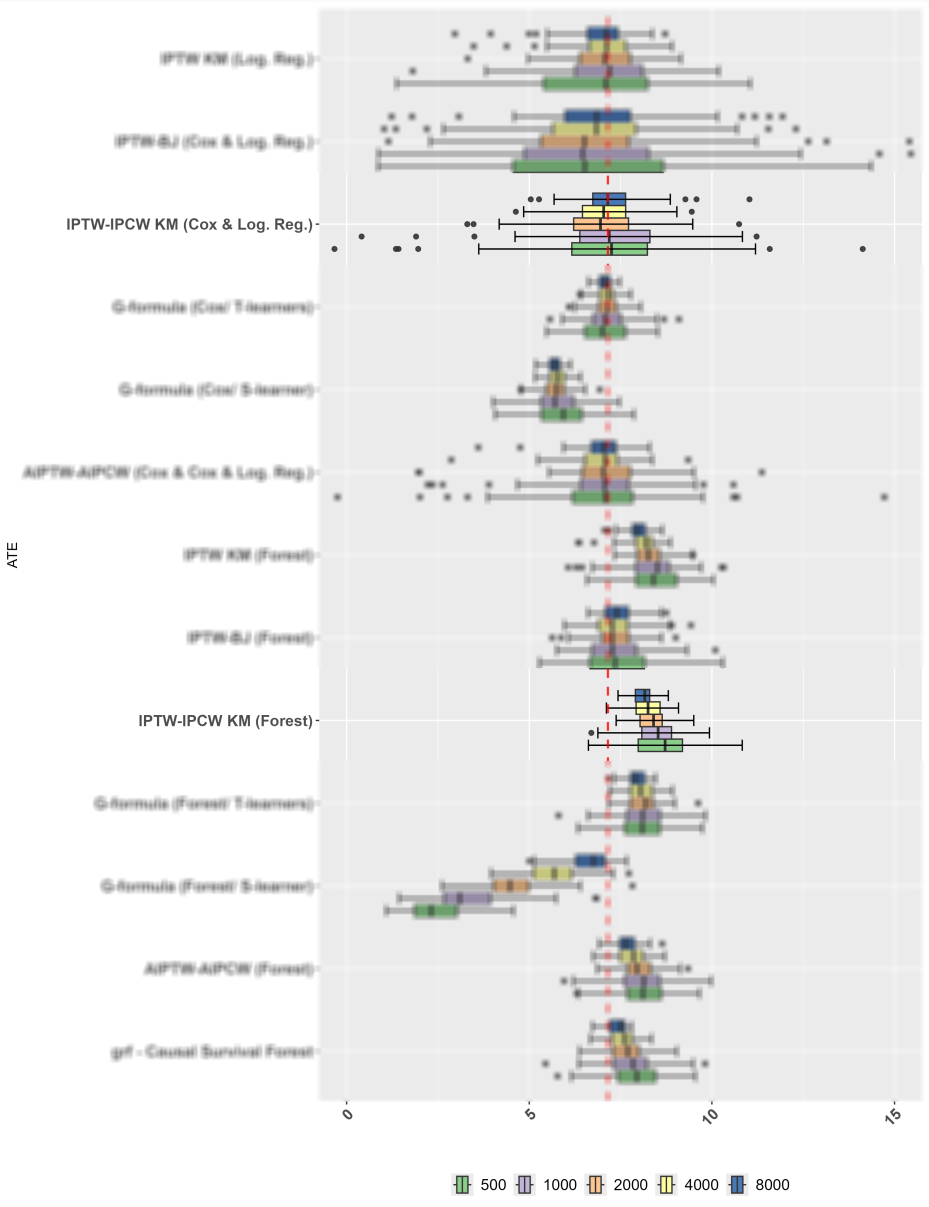
In this simulation, all models are well specified and we present only estimators we expect to converge

1. IPTW KM with propensity score estimated using logistic regression needs less observations to converge. When we use probability forest for the propensity score, IPTW KM needs much more data to converge.
2. In contrary, IPTW-BJ estimated with Cox model and logistic regression needs much more data than IPTW-BJ estimated with probability and random forest.
3. Same as 1 for IPTW-IPCW (Forest converges at high sample size).
4. A simple difference in adding the treatment (S-learner vs T-learners) in the adjustment set in Cox model leads to a non-proportional hazard (warning on the use of S-learner). In forest version, S-learner seems to converge at smaller sample size (no stratification of analysis, then learn on all data).
5. Same as 1 for AIPTW-AIPCW (Forest converges slower).
6. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It converges with few sample size as all the other forest version (needs to be prioritize).

- To conclude, in average forest versions need more data to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox.

Obs & Independent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



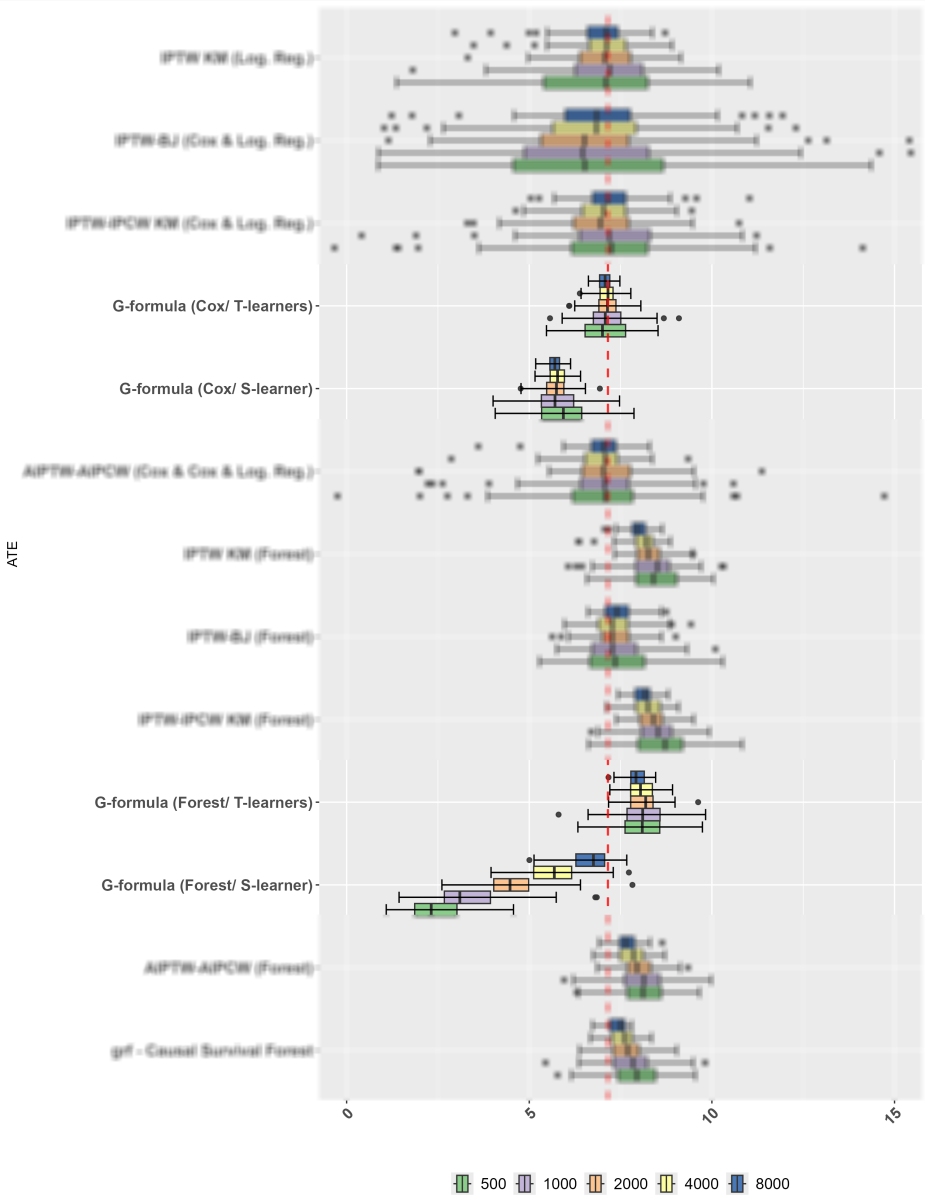
In this simulation, all models are well specified and we present only estimators we expect to converge

1. IPTW KM with propensity score estimated using logistic regression needs less observations to converge. When we use probability forest for the propensity score, IPTW KM needs much more data to converge.
2. In contrary, IPTW-BJ estimated with Cox model and logistic regression needs much more data than IPTW-BJ estimated with probability and random forest.
3. Same as 1 for IPTW-IPCW (Forest converges at high sample size).
4. A simple difference in adding the treatment (S-learner vs T-learners) in the adjustment set in Cox model leads to a non-proportional hazard (warning on the use of S-learner). In forest version, S-learner seems to converge at smaller sample size (no stratification of analysis, then learn on all data).
5. Same as 1 for AIPTW-AIPCW (Forest converges slower).
6. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It converges with few sample size as all the other forest version (needs to be prioritize).

- To conclude, in average forest versions need more data to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox.

Obs & Independent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



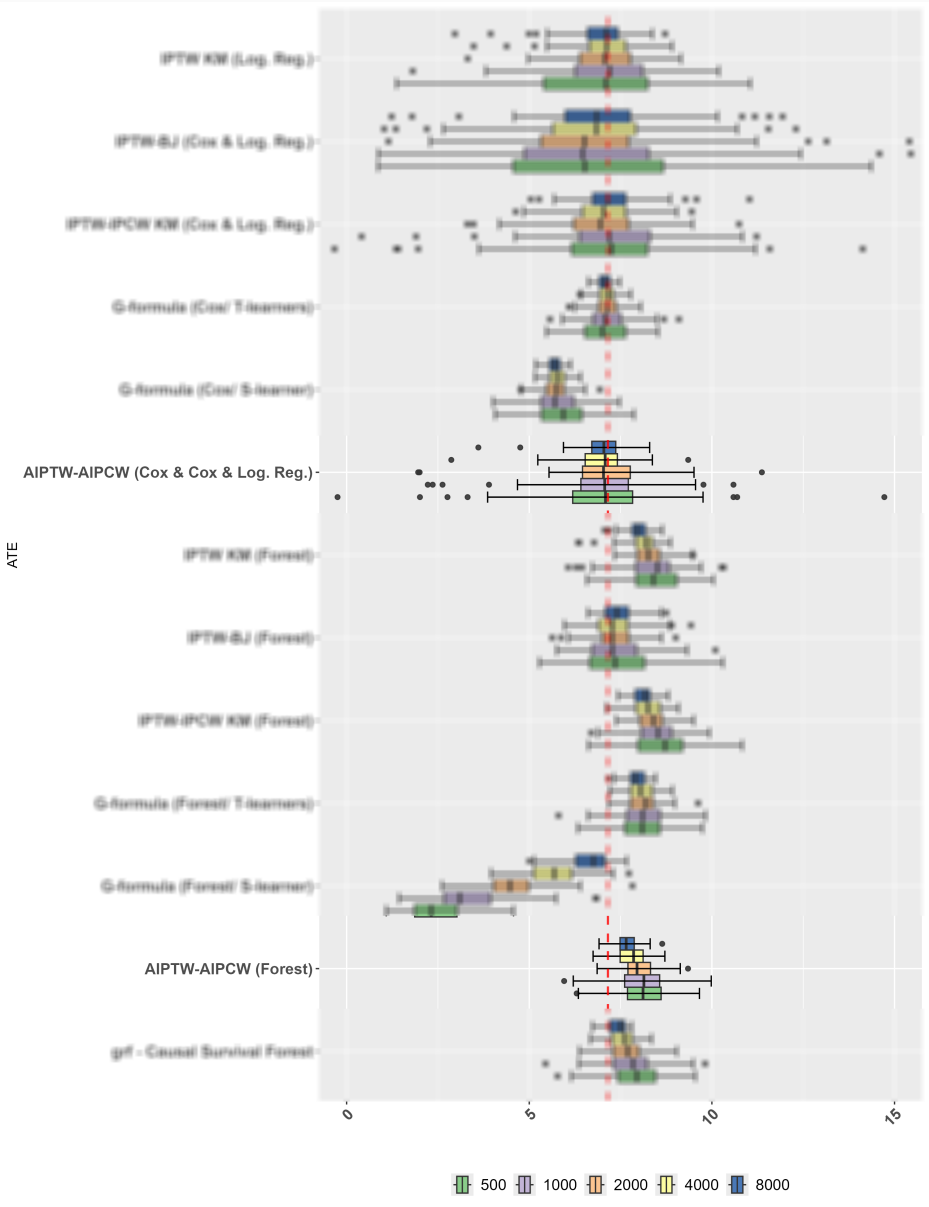
In this simulation, all models are well specified and we present only estimators we expect to converge

1. IPTW KM with propensity score estimated using logistic regression needs less observations to converge. When we use probability forest for the propensity score, IPTW KM needs much more data to converge.
2. In contrary, IPTW-BJ estimated with Cox model and logistic regression needs much more data than IPTW-BJ estimated with probability and random forest.
3. Same as 1 for IPTW-IPCW (Forest converges at high sample size).
4. A simple difference in adding the treatment (S-learner vs T-learners) in the adjustment set in Cox model leads to a non-proportional hazard (warning on the use of S-learner). In forest version, S-learner seems to converge at smaller sample size (no stratification of analysis, then learn on all data).
5. Same as 1 for AIPTW-AIPCW (Forest converges slower).
6. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It converges with few sample size as all the other forest version (needs to be prioritize).

- To conclude, in average forest versions need more data to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox.

Obs & Independent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



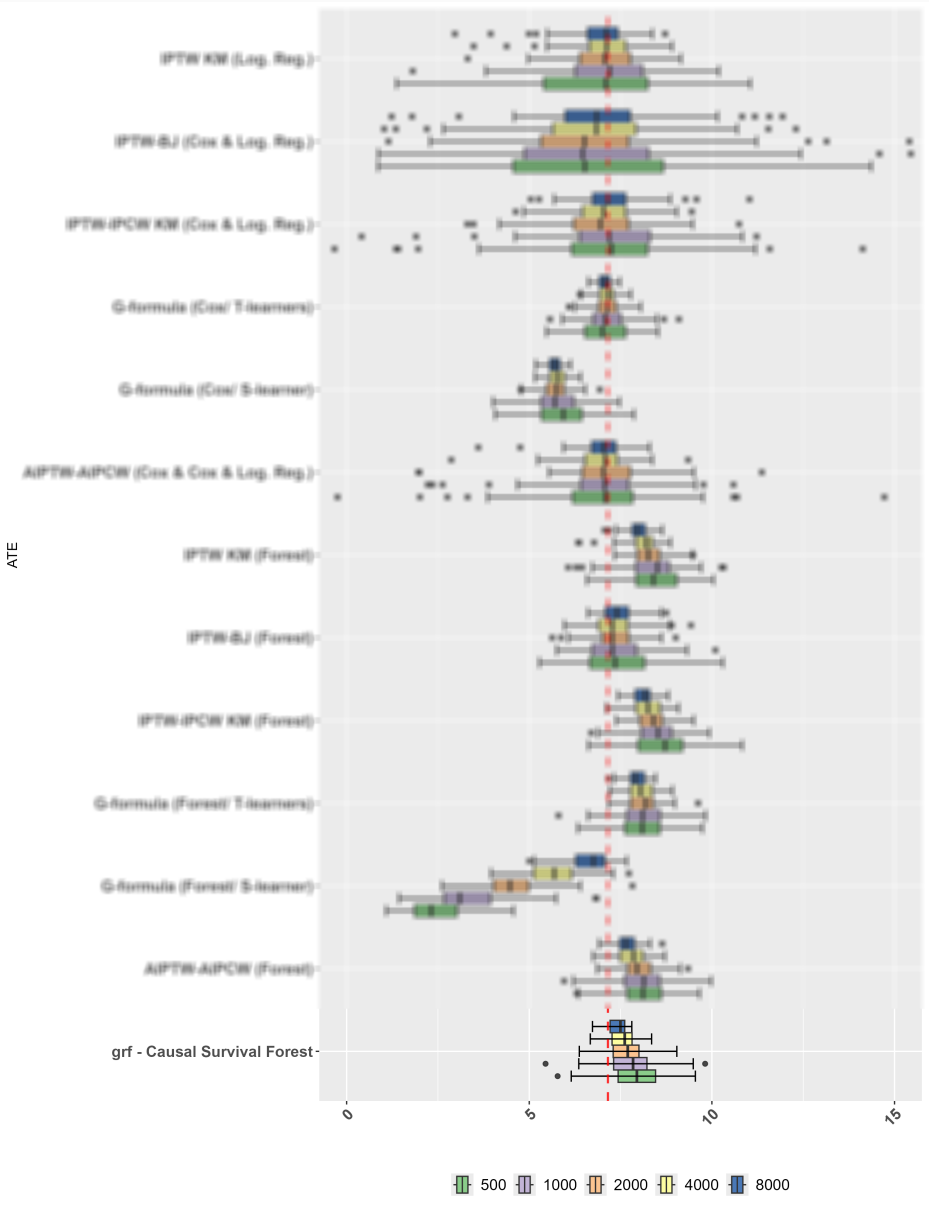
In this simulation, all models are well specified and we present only estimators we expect to converge

1. IPTW KM with propensity score estimated using logistic regression needs less observations to converge. When we use probability forest for the propensity score, IPTW KM needs much more data to converge.
2. In contrary, IPTW-BJ estimated with Cox model and logistic regression needs much more data than IPTW-BJ estimated with probability and random forest.
3. Same as 1 for IPTW-IPCW (Forest converges at high sample size).
4. A simple difference in adding the treatment (S-learner vs T-learners) in the adjustment set in Cox model leads to a non-proportional hazard (warning on the use of S-learner). In forest version, S-learner seems to converge at smaller sample size (no stratification of analysis, then learn on all data).
5. Same as 1 for AIPTW-AIPCW (Forest converges slower).
6. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It converges with few sample size as all the other forest version (needs to be prioritize).

- To conclude, in average forest versions need more data to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox.

Obs & Independent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



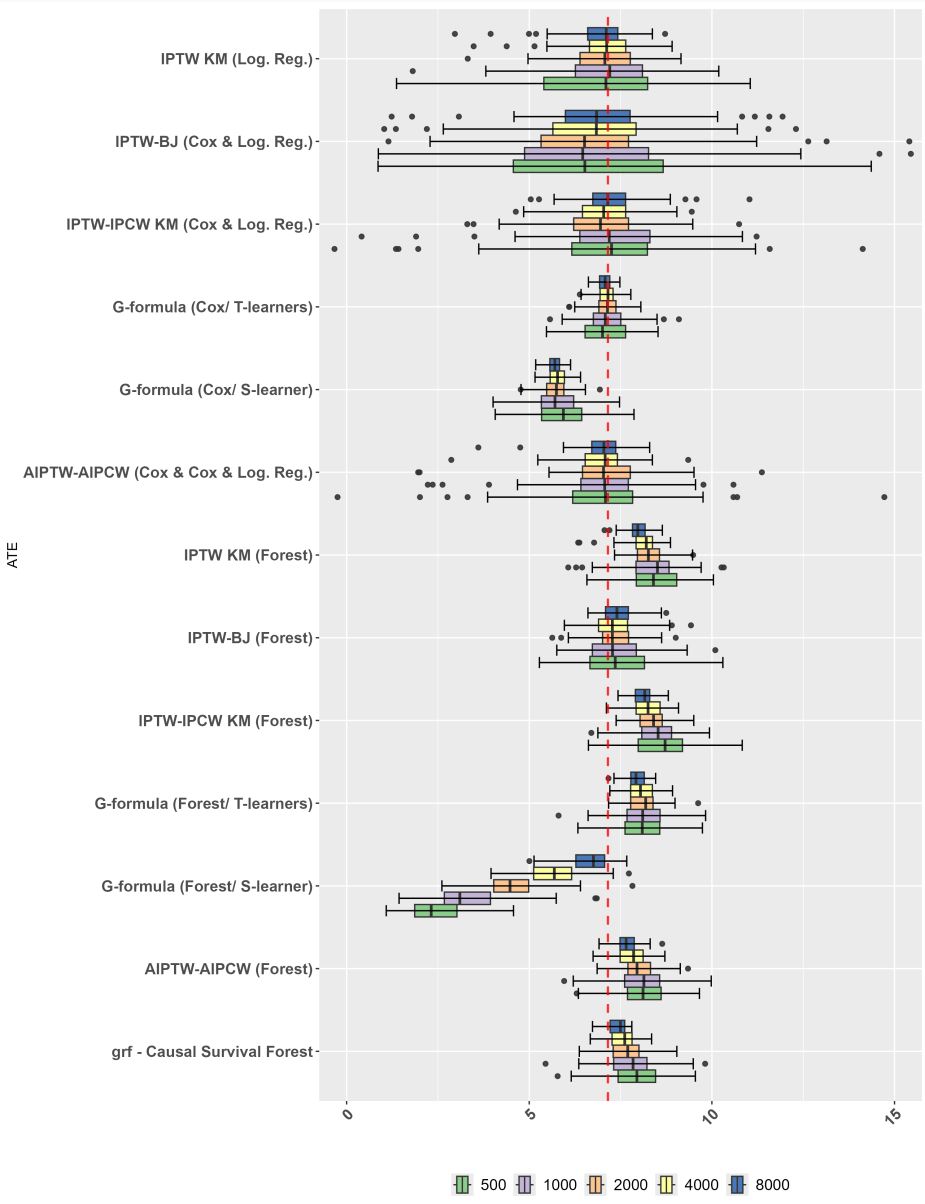
In this simulation, all models are well specified and we present only estimators we expect to converge

1. IPTW KM with propensity score estimated using logistic regression needs less observations to converge. When we use probability forest for the propensity score, IPTW KM needs much more data to converge.
2. In contrary, IPTW-BJ estimated with Cox model and logistic regression needs much more data than IPTW-BJ estimated with probability and random forest.
3. Same as 1 for IPTW-IPCW (Forest converges at high sample size).
4. A simple difference in adding the treatment (S-learner vs T-learners) in the adjustment set in Cox model leads to a non-proportional hazard (warning on the use of S-learner). In forest version, S-learner seems to converge at smaller sample size (no stratification of analysis, then learn on all data).
5. Same as 1 for AIPTW-AIPCW (Forest converges slower).
6. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It converges with few sample size as all the other forest version (needs to be prioritize).

- To conclude, in average forest versions need more data to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox.

Obs & Independent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



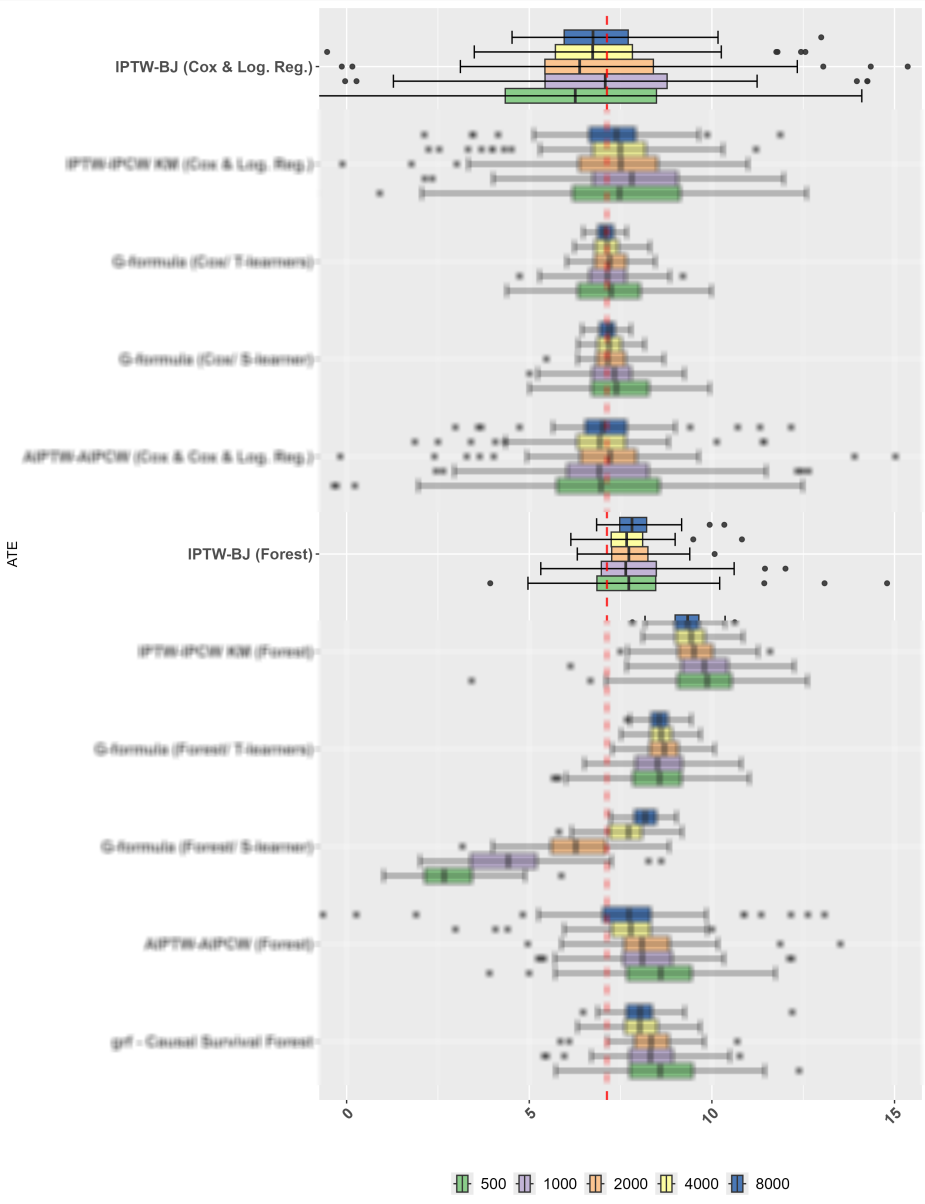
In this simulation, all models are well specified and we present only estimators we expect to converge

1. IPTW KM with propensity score estimated using logistic regression needs less observations to converge. When we use probability forest for the propensity score, IPTW KM needs much more data to converge.
2. In contrary, IPTW-BJ estimated with Cox model and logistic regression needs much more data than IPTW-BJ estimated with probability and random forest.
3. Same as 1 for IPTW-IPCW (Forest converges at high sample size).
4. A simple difference in adding the treatment (S-learner vs T-learners) in the adjustment set in Cox model leads to a non-proportional hazard (warning on the use of S-learner). In forest version, S-learner seems to converge at smaller sample size (no stratification of analysis, then learn on all data).
5. Same as 1 for AIPTW-AIPCW (Forest converges slower).
6. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It converges with few sample size as all the other forest version (needs to be prioritize).

- To conclude, in average forest versions need more data to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox.

Obs & Dependent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



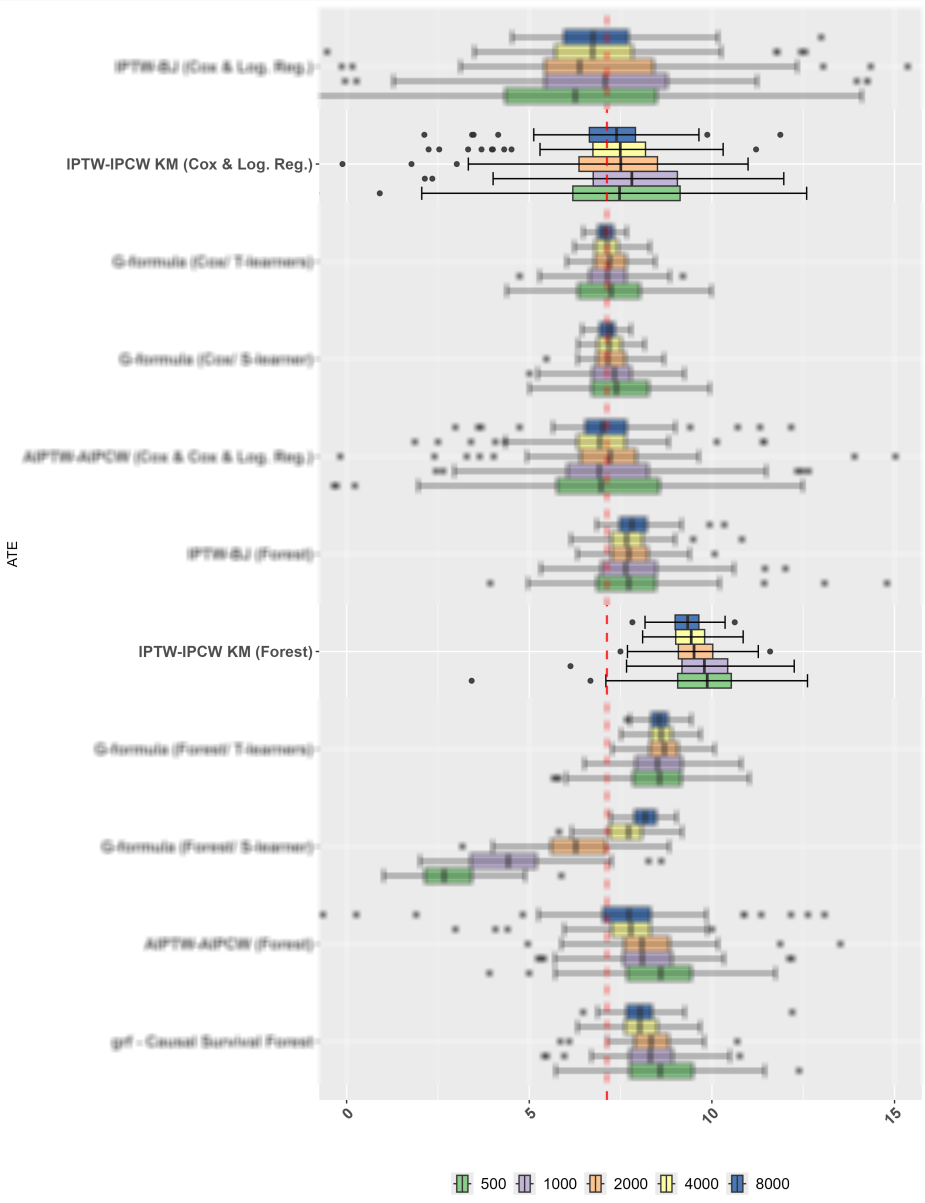
In this simulation, all models are well specified and we present only estimators we expect to converge

1. When introducing more complexity, IPTW-BJ estimated with Cox model and logistic regression has high variability and tends to converge at high sample size. When we use forest for the nuisance models, IPTW-BJ has less variability but always struggle to converge.
2. In contrary, IPTW-IPCW estimated with Cox model and logistic regression converge with few sample size than IPTW-IPCW estimated with forest (completely biased).
3. G-formula with Cox model converge at really small sample size. Forest versions tends to converge with much more sample size.
4. Same as 2 and 3 for AIPTW-AIPCW (Forest converges slower).
5. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It has a lower variability but still biased at 8,000 sample size.

- To conclude, in average forest versions need much more observations to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox (Same as before).

Obs & Dependent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



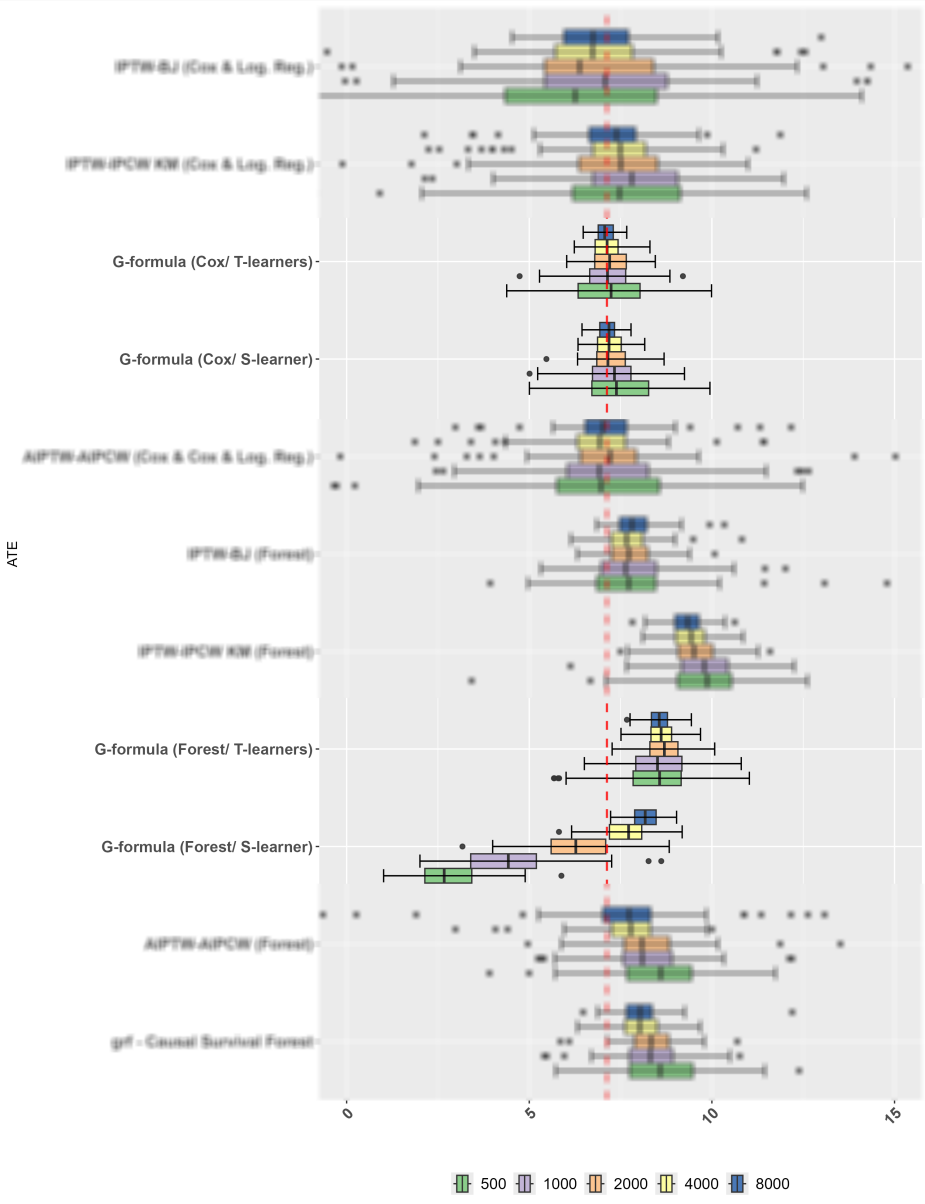
In this simulation, all models are well specified and we present only estimators we expect to converge

1. When introducing more complexity, IPTW-BJ estimated with Cox model and logistic regression has high variability and tends to converge at high sample size. When we use forest for the nuisance models, IPTW-BJ has less variability but always struggle to converge.
2. In contrary, IPTW-IPCW estimated with Cox model and logistic regression converge with few sample size than IPTW-IPCW estimated with forest (completely biased).
3. G-formula with Cox model converge at really small sample size. Forest versions tends to converge with much more sample size.
4. Same as 2 and 3 for AIPTW-AIPCW (Forest converges slower).
5. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It has a lower variability but still biased at 8,000 sample size.

- To conclude, in average forest versions need much more observations to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox (Same as before).

Obs & Dependent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



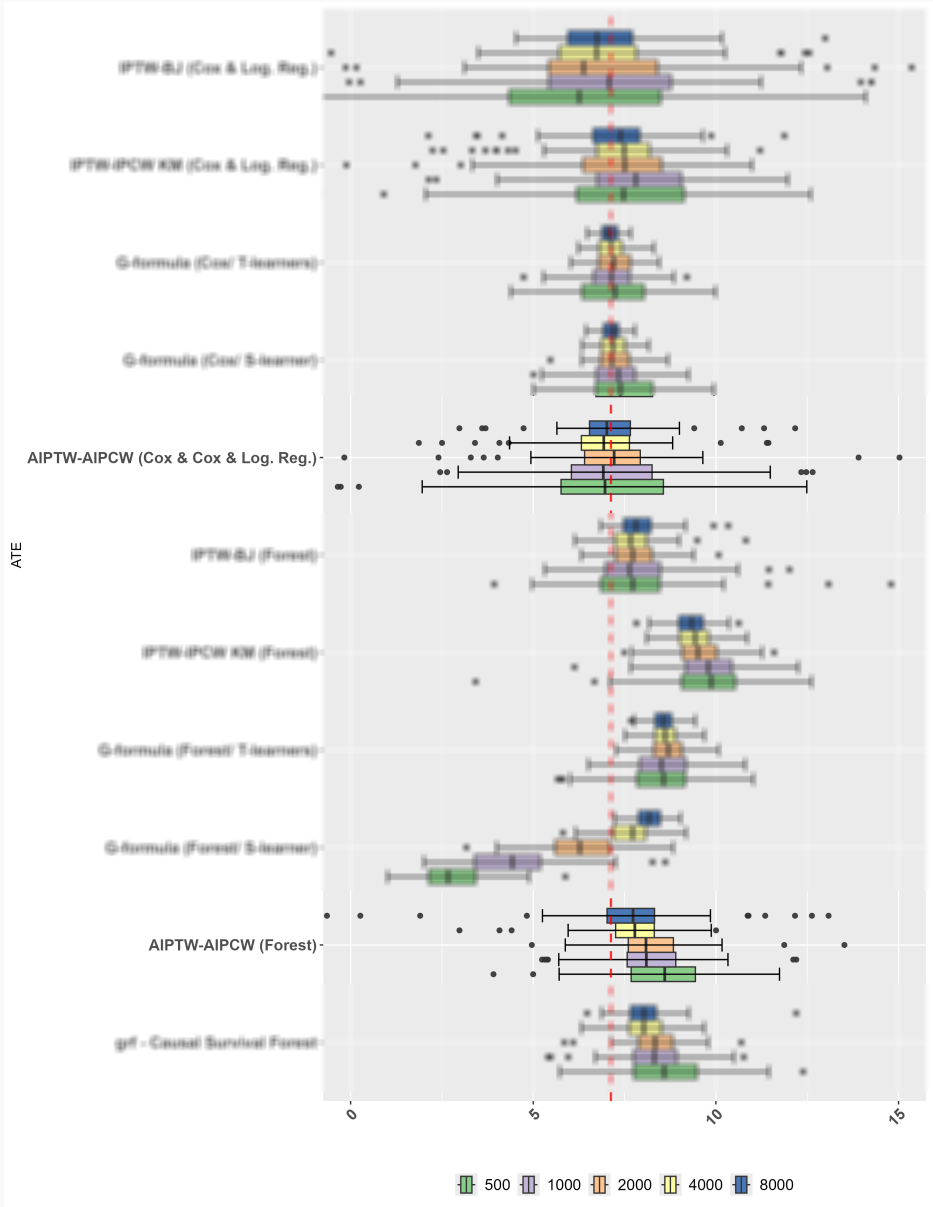
In this simulation, all models are well specified and we present only estimators we expect to converge

1. When introducing more complexity, IPTW-BJ estimated with Cox model and logistic regression has high variability and tends to converge at high sample size. When we use forest for the nuisance models, IPTW-BJ has less variability but always struggle to converge.
2. In contrary, IPTW-IPCW estimated with Cox model and logistic regression converge with few sample size than IPTW-IPCW estimated with forest (completely biased).
3. G-formula with Cox model converge at really small sample size. Forest versions tends to converge with much more sample size.
4. Same as 2 and 3 for AIPTW-AIPCW (Forest converges slower).
5. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It has a lower variability but still biased at 8,000 sample size.

- To conclude, in average forest versions need much more observations to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox (Same as before).

Obs & Dependent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



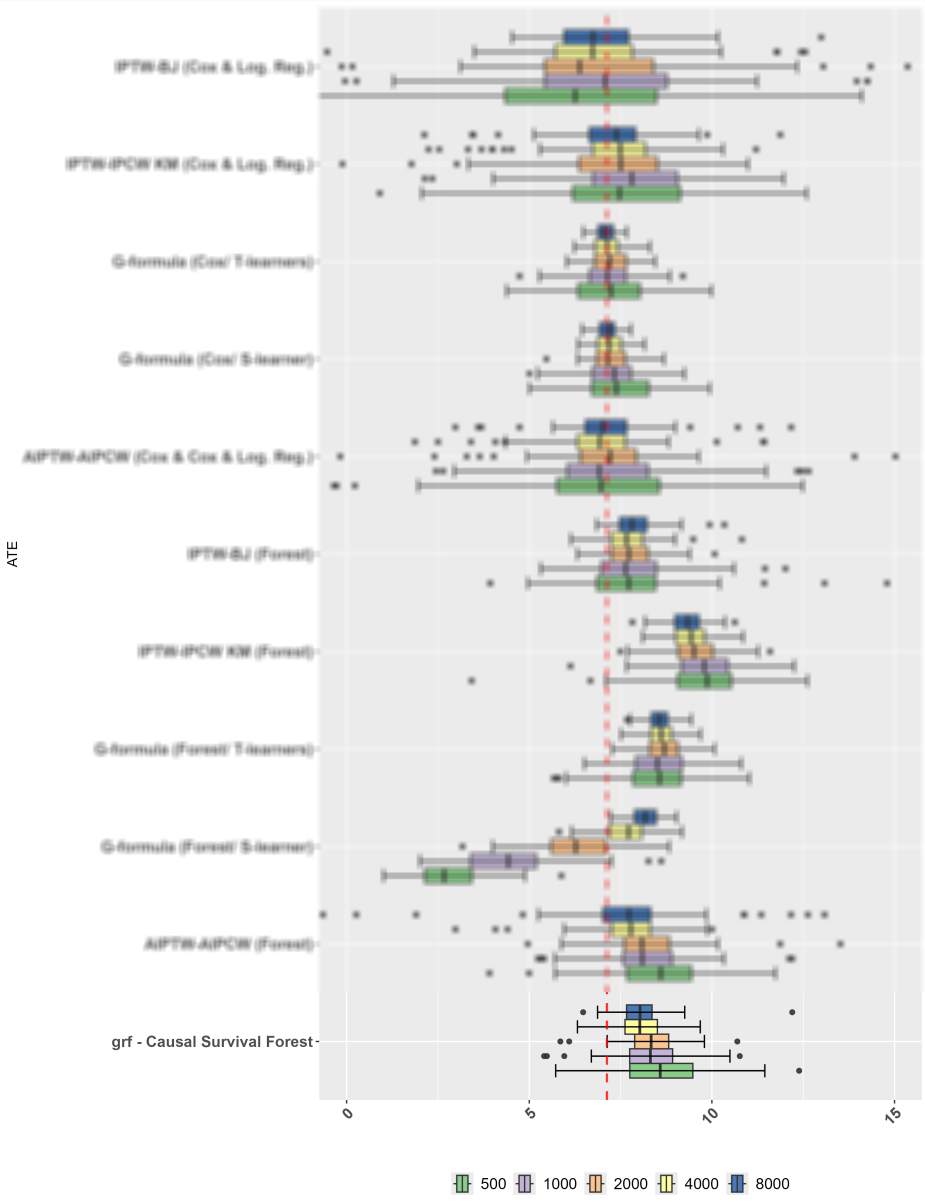
In this simulation, all models are well specified and we present only estimators we expect to converge

1. When introducing more complexity, IPTW-BJ estimated with Cox model and logistic regression has high variability and tends to converge at high sample size. When we use forest for the nuisance models, IPTW-BJ has less variability but always struggle to converge.
2. In contrary, IPTW-IPCW estimated with Cox model and logistic regression converge with few sample size than IPTW-IPCW estimated with forest (completely biased).
3. G-formula with Cox model converge at really small sample size. Forest versions tends to converge with much more sample size.
4. Same as 2 and 3 for AIPTW-AIPCW (Forest converges slower).
5. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It has a lower variability but still biased at 8,000 sample size.

- To conclude, in average forest versions need much more observations to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox (Same as before).

Obs & Dependent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



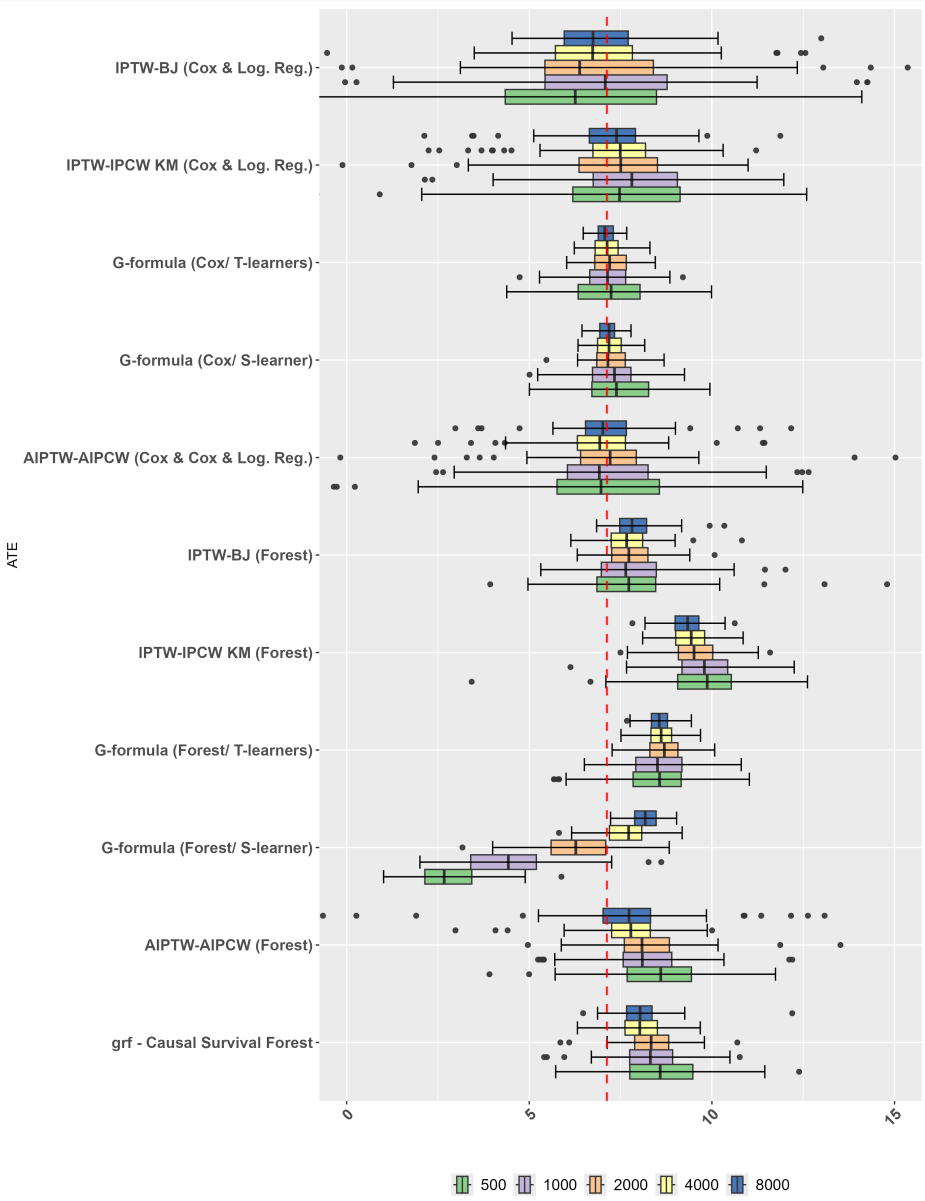
In this simulation, all models are well specified and we present only estimators we expect to converge

1. When introducing more complexity, IPTW-BJ estimated with Cox model and logistic regression has high variability and tends to converge at high sample size. When we use forest for the nuisance models, IPTW-BJ has less variability but always struggle to converge.
2. In contrary, IPTW-IPCW estimated with Cox model and logistic regression converge with few sample size than IPTW-IPCW estimated with forest (completely biased).
3. G-formula with Cox model converge at really small sample size. Forest versions tends to converge with much more sample size.
4. Same as 2 and 3 for AIPTW-AIPCW (Forest converges slower).
5. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It has a lower variability but still biased at 8,000 sample size.

- To conclude, in average forest versions need much more observations to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox (Same as before).

Obs & Dependent censoring ((semi)-parametric DGP)

100 studies of various sample sizes are simulated.



In this simulation, all models are well specified and we present only estimators we expect to converge

1. When introducing more complexity, IPTW-BJ estimated with Cox model and logistic regression has high variability and tends to converge at high sample size. When we use forest for the nuisance models, IPTW-BJ has less variability but always struggle to converge.
2. In contrary, IPTW-IPCW estimated with Cox model and logistic regression converge with few sample size than IPTW-IPCW estimated with forest (completely biased).
3. G-formula with Cox model converge at really small sample size. Forest versions tends to converge with much more sample size.
4. Same as 2 and 3 for AIPTW-AIPCW (Forest converges slower).
5. Causal Survival Forest (available in grf package in R) is a optimized version of AIPTW-AIPCW (Forest). It has a lower variability but still biased at 8,000 sample size.

- To conclude, in average forest versions need much more observations to converge. The better estimator in this simulation is G-formula (T-learners) estimated with Cox (Same as before).

Conclusions

RMST difference is the causal quantity for treatment effect assessment in observational studies (can also be useful in RCT as alternative to HR)

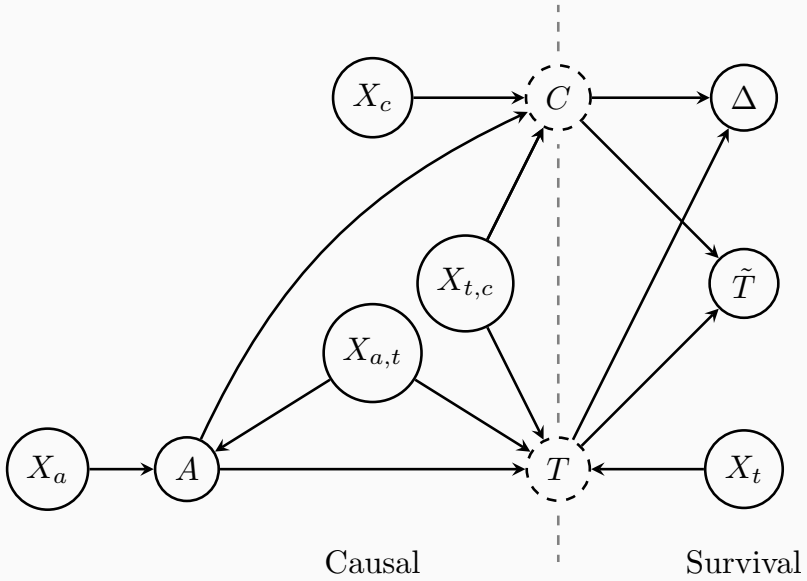
RECOMMENDATIONS

- **G-formula**: Accurate but sensitive to model misspecification.
  – T-learners preferred with Cox (avoid treatment form assumptions).
  – Forest versions underperform vs. AIPTW-AIPCW (Forest) or causal survival forests.
  – Simple to implement (1 model), but no full R package.
- **AIPTW-AIPCW**: Robust and stable, but complex (3 models needed).
- **Forest-based methods**: Need large samples; useful under model uncertainty (esp. for AIPTW-AIPCW).
- **Causal survival forest**: Most reliable flexible method.
  – Requires large sample sizes, easy in R (`grf`).

LIMITATIONS

- Simulations use large samples and simple data-generating mechanisms

What's next ?



$\Rightarrow$ Focus on **variable selection**:

$X_{a,t}$ (confounders) and $X_{t,c}$ (dependent censoring) are the minimum sufficient set but what happen if we include variable related only to the outcome (X_t) or treatment (X_a) or censoring (X_c) ?

Thank you for your attention !
Scan it if you are interested to go through the details

Arxiv article (submitted)

Github repository

Feel free to give me some feedback or contact me for any question:
charlotte.voinot@sanofi.com

References

Buckley and James (Dec. 1979). "Linear regression with censored data". In: *Biometrika* 66.3, pp. 429–436.

Fan, Jianqing and Irene Gijbels (1994). *Local polynomial modelling and its applications*.

Greenland, S., J. M. Robbins, and J. Pearl (Jan. 1999). "Confounding and collapsibility in causal inference". In: *Statistical Science* 14, pp. 29–46.

Hernán, Miguel A (Jan. 2010). "The hazards of hazard ratios". en. In: *Epidemiology* 21.1, pp. 13–15.

Huitfeldt, A., M. Stensrud, and E. Suzuki (Jan. 2019). "On the collapsibility of measures of effect in the counterfactual causal framework". In: *Emerging Themes in Epidemiology* 16, pp. 1–12.

References ii

Koul, H., V. Susarla, and J. Van Ryzin (1981). "Regression Analysis with Randomly Right-Censored Data". In: *The Annals of Statistics* 9.6, pp. 1276–1288.

Martinussen, Torben and Stijn Vansteelandt (July 2013). "On collapsibility and confounding bias in Cox and Aalen regression models". en. In: *Lifetime Data Anal.* 19.3, pp. 279–296.

Martinussen, Torben, Stijn Vansteelandt, and Per Andersen (Oct. 2020). "Subtleties in the interpretation of hazard contrasts". In: *Lifetime Data Analysis* 26.

Ozenne, Brice Maxime Hugues et al. (2020). "On the estimation of average treatment effects with right-censored time to event outcome and competing risks". In: *Biometrical Journal* 62.3, pp. 751–763.

References iii

Royston, Patrick and Mahesh K B Parmar (Dec. 2013). "Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome". en. In: *BMC Med. Res. Methodol.* 13.1, p. 152.

Rubin, Donald B. (Oct. 1974). "Estimating causal effects of treatments in randomized and nonrandomized studies.". en. In: *Journal of Educational Psychology* 66.5, pp. 688–701.

Stensrud, Mats J et al. (2019). "Limitations of hazard ratios in clinical trials". In: *European heart journal* 40.17, pp. 1378–1383.

Willems, SJW et al. (2018). "Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator". In: *Statistical Methods in Medical Research* 27.2. PMID: 26988930, pp. 323–335.

Appendix

A desirable property: Collapsibility

Collapsibility: The population effect is equal to a weighted sum of local effects (conditional effects).

Direct collapsibility - weights are equal to the population's proportions

$$\tau = \mathbb{E}[\tau(X)]$$

Risk Difference is directly collapsible

	τ_{RD}	τ_{RR}	τ_{SR}	τ_{NNT}	τ_{OR}
All (P_R)	-0.0452	0.6	1.05	22	0.57
$X = 1$	-0.006	0.6	1.01	167	0.6
$X = 0$	-0.08	0.6	1.1	13	0.545

$$\tau_R^{RD} = p_R(X = 1) \cdot \tau_R^{RD}(X = 1) + p_R(X = 0) \cdot \tau_R^{RD}(X = 0)$$

A desirable property: Collapsibility

Collapsibility: The population effect is equal to a weighted sum of local effects (conditional effects).

Collapsibility: weights depend on the baseline distribution $Y(0)$

$$\mathbb{E}[w(X, P(X, Y(0)))\tau(X)] = \tau$$

with $w \geqslant 0$, $\mathbb{E}[w(X, P(X, Y(0)))] = 1$.

Risk Ratio is collapsible:

$$\mathbb{E}\left[\tau_{RR}(X)\frac{\mathbb{E}[Y(0)|X]}{\mathbb{E}[Y(0)]}\right] = \tau_{RR}$$

Summary of Causal Measure Properties

Direct collapsibility

$$\mathbb{E}[\tau(X)] = \tau$$

Collapsibility: weights depend on the baseline distribution $Y(0)$

$$\mathbb{E}[w(X, P(X, Y(0)))\tau(X)] = \tau$$

with $w > 0$, $\mathbb{E}[w(X, P(X, Y(0)))] = 1$.

Measure	Directly Collapsible	Collapsible
Risk Difference	Yes	Yes
Number Needed to Treat	No	Yes
Risk Ratio	No	Yes
Survival Ratio	No	Yes
Odds Ratio	No	No
Hazard Ratio	No	No

Dependent censoring introduces selection bias

If we do not adjust for dependent censoring, the estimator is biased in this context (ex: severe people are more likely to be censored → under-represented group):

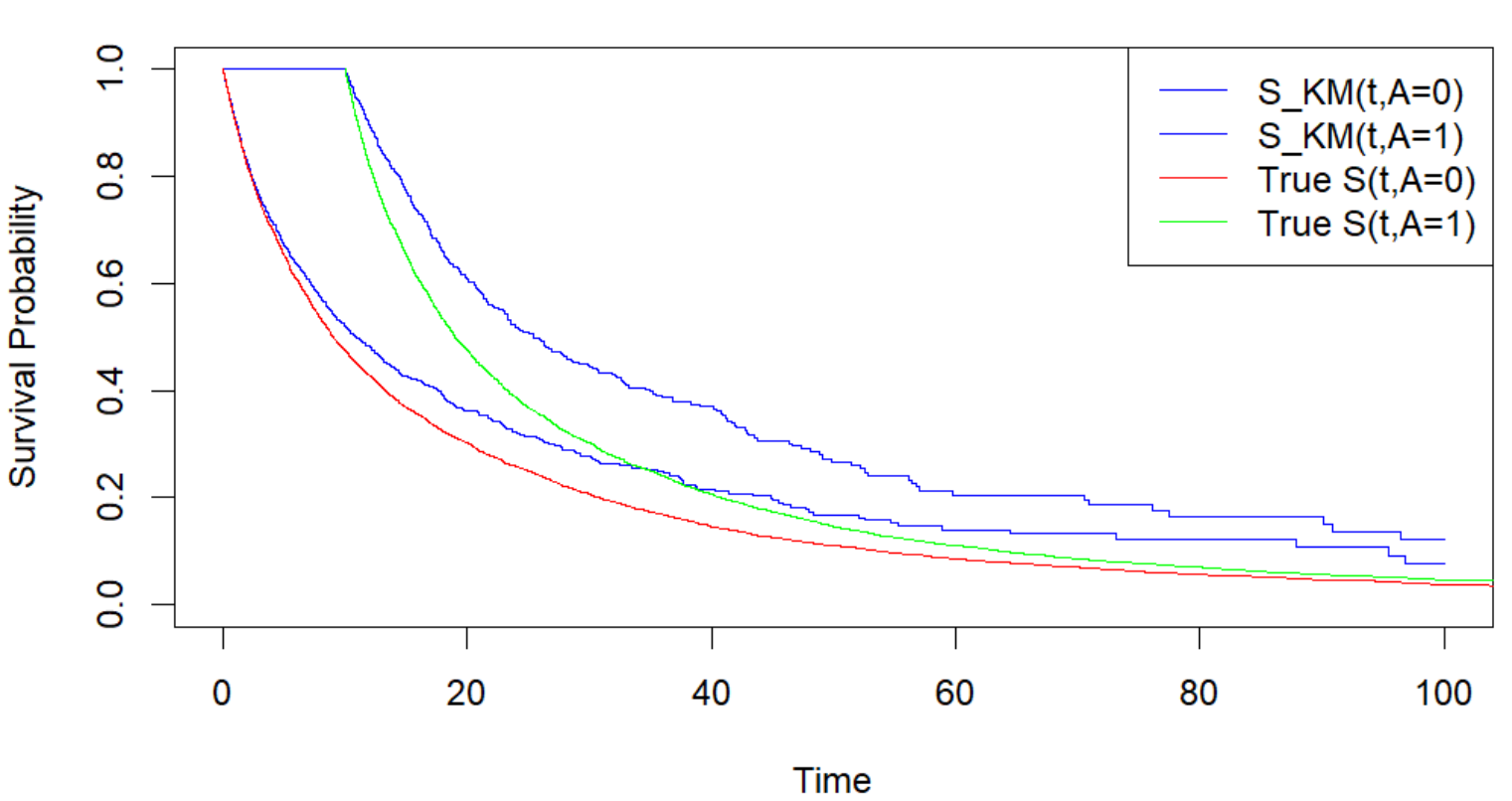


Figure 4: Plot of stratified kaplan meier survival function (A=1 and A=0) and the true survival function

⇒ The probability of survival is no longer consistent in using Non adjusted KM [Willems et al. 2018].

Identifiability assumptions in each context

S.T.U.V.A. $T = AT(1) + (1 - A)T(0)$

RCT & Independent censoring

- **Random treatment assignment**
  $A \perp\!\!\!\perp (T(0), T(1), C, X)$
- **Independent censoring**
  $C \perp\!\!\!\perp T(0), T(1), X, A$

Obs & Independent censoring

- **Unconfoundedness** $A \perp\!\!\!\perp (T(0), T(1))|X$
- **Positivity for treatment**
  $1 > P(A = a \mid X = x) > 0$
- **Independent censoring**
  $C \perp\!\!\!\perp T(0), T(1), X, A$

RCT & Dependent censoring

- **Random treatment assignment**
  $A \perp\!\!\!\perp (T(0), T(1), C, X)$
- **Conditionally independent censoring**
  $C \perp\!\!\!\perp T(0), T(1)|X, A$
- **Positivity for censoring**
  $0 < P(C > t \mid X = x, A = a) < 1$

Obs & Dependent censoring

- **Unconfoundedness** $A \perp\!\!\!\perp (T(0), T(1))|X$
- **Positivity for treatment**
  $1 > P(A = a \mid X = x) > 0$
- **Conditionally independent censoring**
  $C \perp\!\!\!\perp T(0), T(1)|X, A$
- **Positivity for censoring**
  $0 < P(C > t \mid X = x, A = a) < 1$

Simulation of observational study ((semi)-parametric DGP)

For the simulation, n samples $(X_i, A_i, C, T_i(0), T_i(1))$ are generated in the following way:

- $X \sim \mathcal{N}\left(\mu = [1, 1, -1, 1]^\top, \Sigma = I_4\right)$
- $logit(e(X)) = \beta_A^T X$ where $\beta_A = (1, 1, 2.5, 1)$ and $logit(p) = log(p/(1-p))$ the logistic function.
- Then $A_i \sim \text{Bernoulli}(e_i), \quad \forall i \in \{1, \ldots, n\}$
- $\lambda^{(0)}(t|X) = 0.01 \cdot \exp\{0.5X_1 + 0.5X_2 - 0.5X_3 + 0.5X_4\}$ hazard for the event time $T(0)$
- $T(1) = T(0) + 10$
- The hazard for the censoring time C:
  - Scenario 1: $\lambda_c = 0.03$.
  - Scenario 2: $\lambda_c(X) = 0.03 \cdot \exp\{0.7X_1 + 0.7X_2 - 0.25X_3 - 0.1X_4 - 0.2A\}$.
- The threshold time τ is set to 25.

Simulation of observational study for mispecification

We generate n samples $(X_i, A_i, C, T_i(0), T_i(1))$ as follows:

- $X \sim \mathcal{N}(\mu, \Sigma)$ and $\mu = (0.5, 0.5, 0.7, 0.5)$, $\Sigma = \mathrm{Id}_4$.
- The hazard function of $T(0)$ is given by

  $$\lambda^{(0)}(t|X) = \exp\{\beta_0^\top Y\} \quad \text{where} \quad \beta_0 = (0.2, 0.3, 0.1, 0.1, 1, 0, 0, 0, 0, 1),$$
  $$\text{and} \quad Y = (X_1^2, X_2^2, X_3^2, X_4^2, X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2X_4, X_3X_4).$$

- The distribution of $T(1)$ is the one of $T(0)$ but shifted: $T(1) = T(0) + 1$.
- The hazard function of C is given by

  $$\lambda_C(t|X) = \exp\{\beta_C^\top Y\} \quad \text{where} \quad \beta_C = (0.05, 0.05, -0.1, 0.1, 0, 1, 0, -1, 0, 0).$$

- The propensity score is

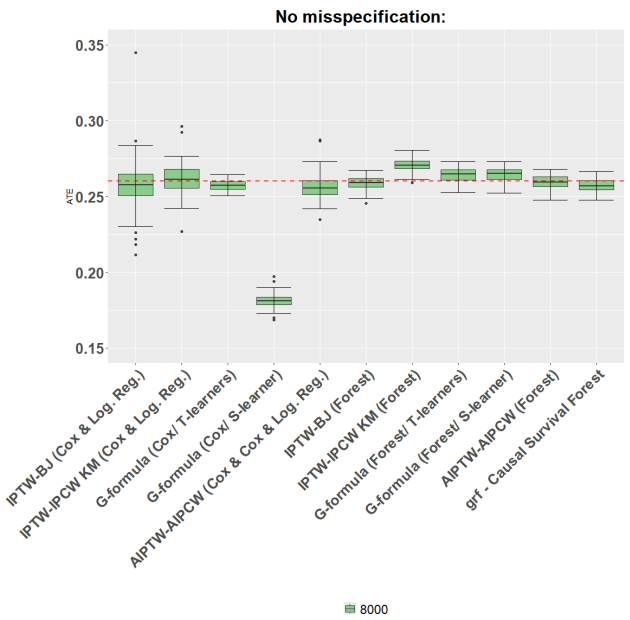
  $$\mathrm{logit}(e(x)) = \beta_A^\top Y \quad \text{where} \quad \beta_A = (0.05, -0.1, 0.5, -0.1, 1, 0, 1, 0, 0, 0).$$

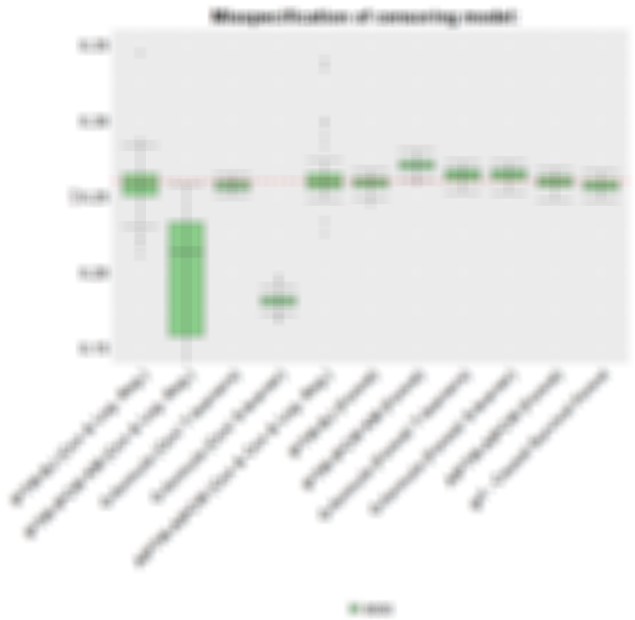
- The threshold time τ is set to 0.5.

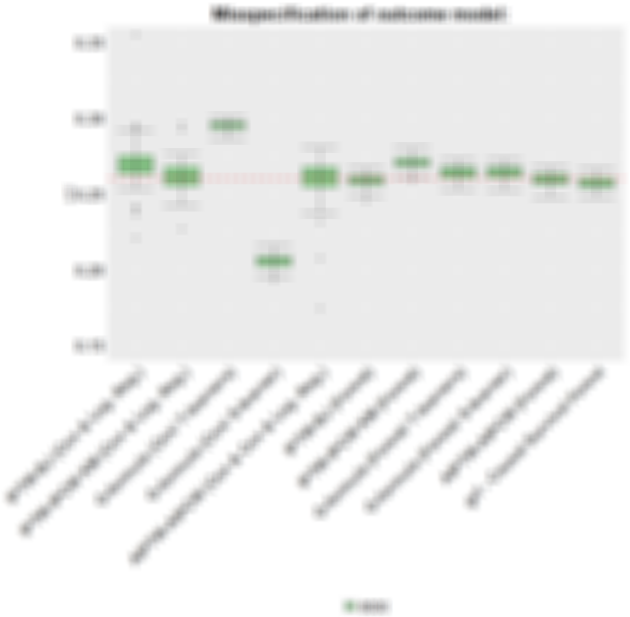
When the model is not well-specified, only the first half of Y corresponding to $(X_1^2, X_2^2, X_3^2, X_4^2)$ is given as an input.

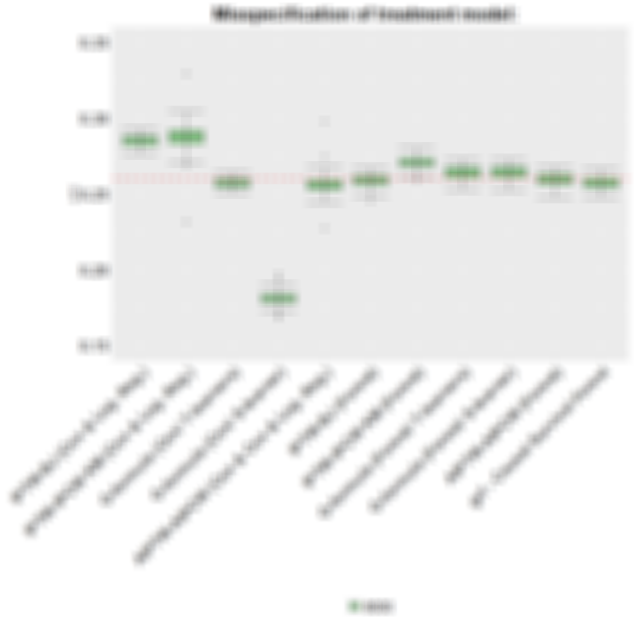
Misspecification: Obs & Dependent Censoring

100 studies at sample size 8,000 are generated.







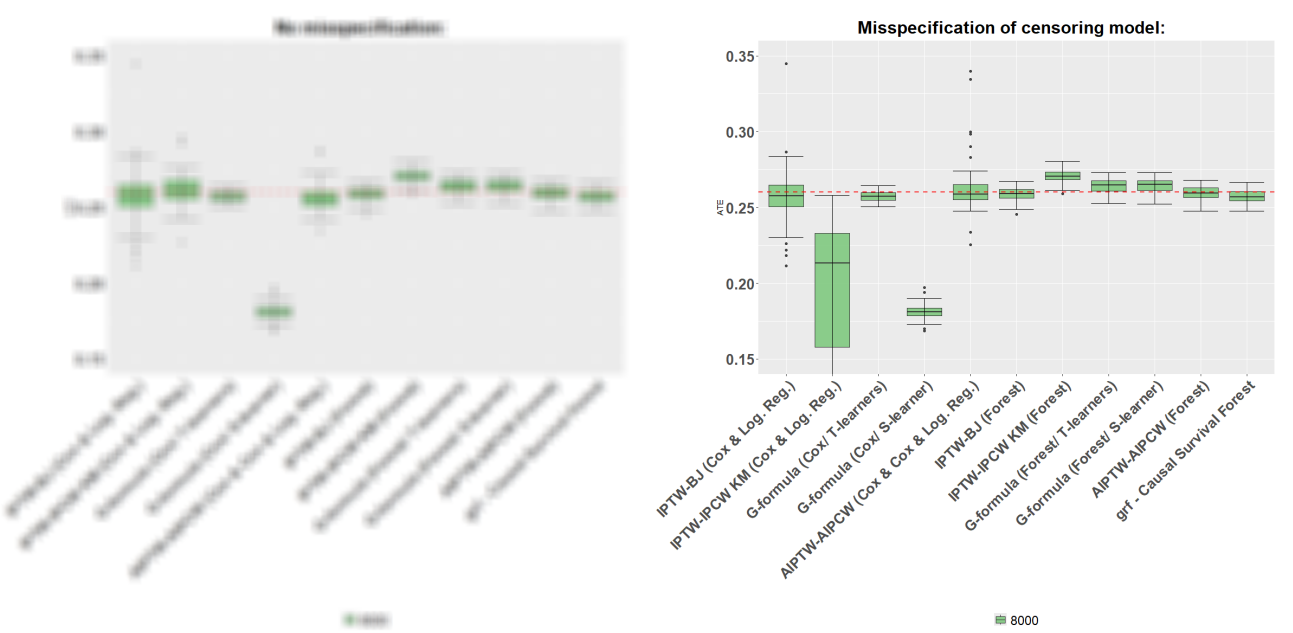


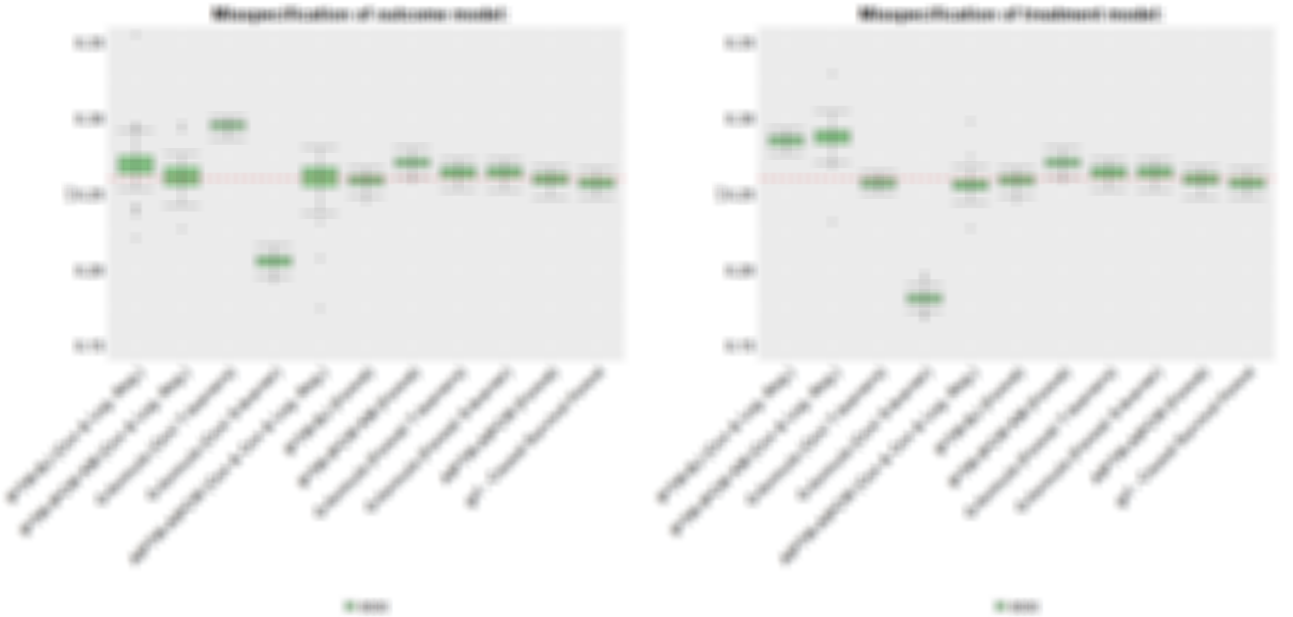
For parametric and semi-parametric models, all relevant interaction terms are explicitly included. In contrast, for the Forest model, interaction terms are not provided manually, as the model automatically captures them during training.

1. Forest are performing at least as good as classical regressions at high sample size (8,000). In general, these methods need more data to converge.
2. With increasing misspecification, the added value of robust estimators (AIPTW-AIPCW) and Forest becomes evident. All the estimators that use misspecified working models are biased.

Misspecification: Obs & Dependent Censoring

100 studies at sample size 8,000 are generated.



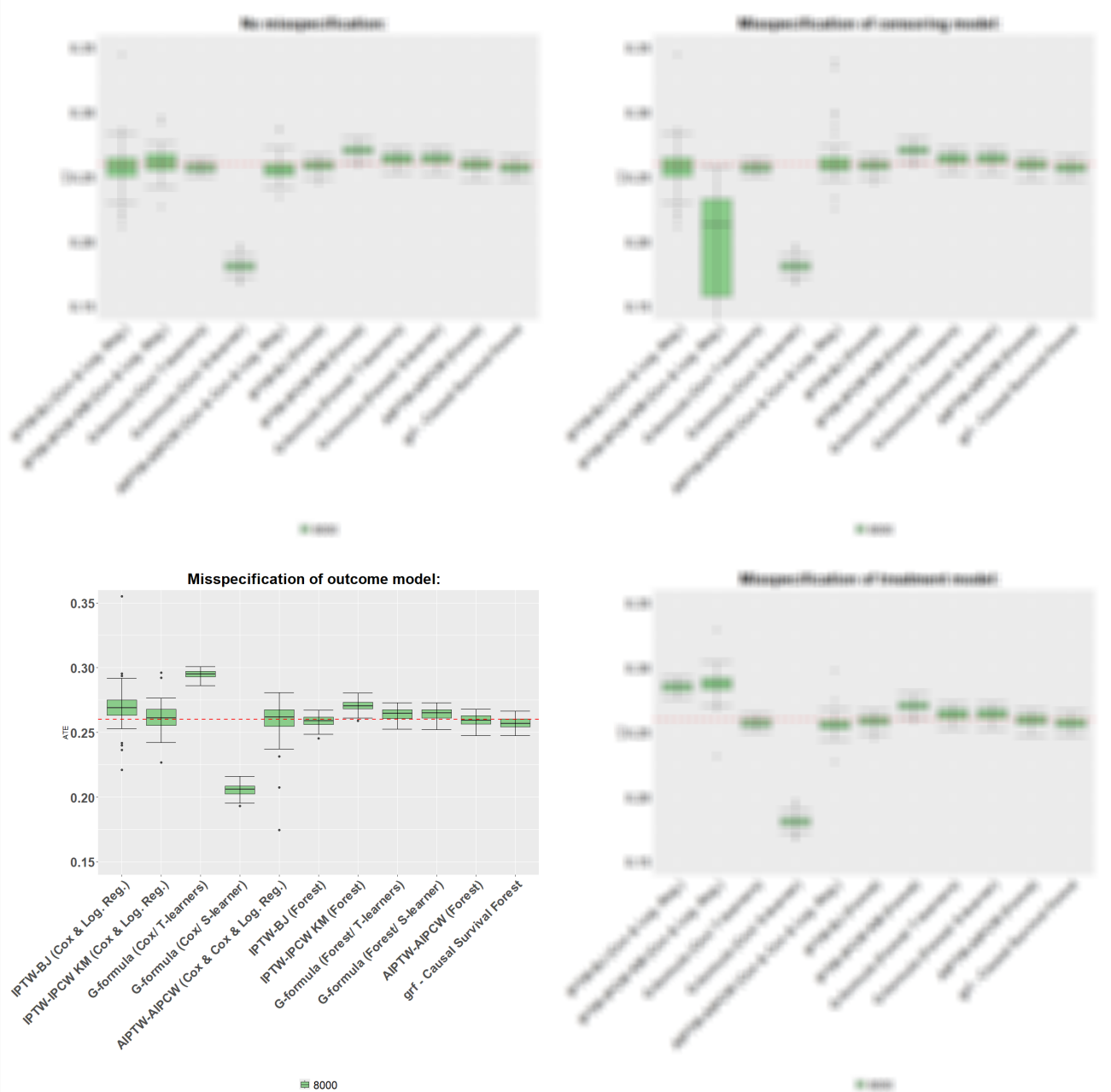


For parametric and semi-parametric models, all relevant interaction terms are explicitly included. In contrast, for the Forest model, interaction terms are not provided manually, as the model automatically captures them during training.

1. Forest are performing at least as good as classical regressions at high sample size (8,000). In general, these methods need more data to converge.
2. With increasing misspecification, the added value of robust estimators (AIPTW-AIPCW) and Forest becomes evident. All the estimators that use misspecified working models are biased.

Misspecification: Obs & Dependent Censoring

100 studies at sample size 8,000 are generated.



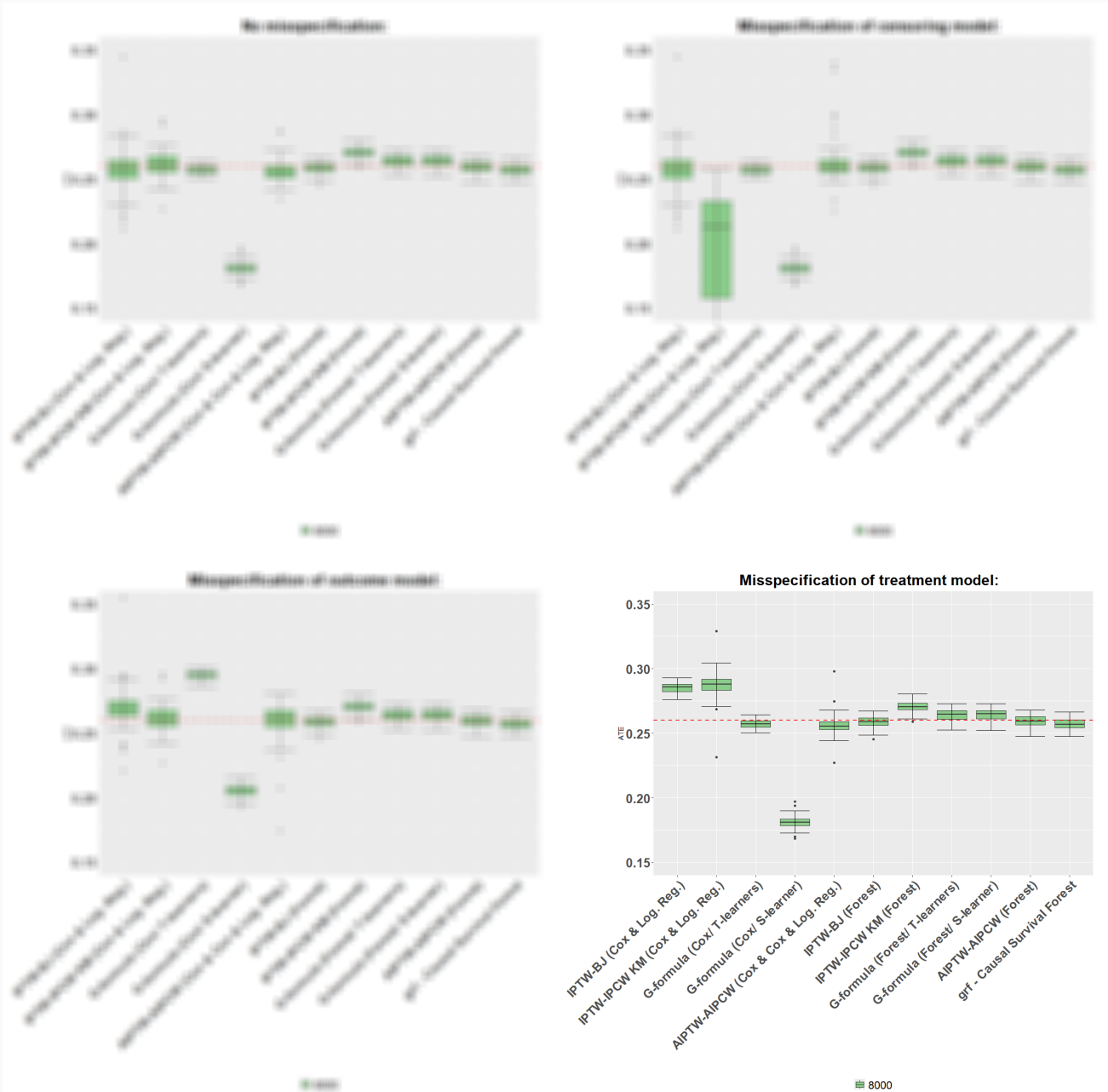
For parametric and semi-parametric models, all relevant interaction terms are explicitly included. In contrast, for the Forest model, interaction terms are not provided manually, as the model automatically captures them during training.

1. Forest are performing at least as good as classical regressions at high sample size (8,000). In general, these methods need more data to converge.
2. With increasing misspecification, the added value of robust estimators (AIPTW-AIPCW) and Forest becomes evident. All the estimators that use misspecified working models are biased.

Forest remains more **flexible** than traditional regression models when there is **uncertainty about the model form (or missing interaction)**

Misspecification: Obs & Dependent Censoring

100 studies at sample size 8,000 are generated.



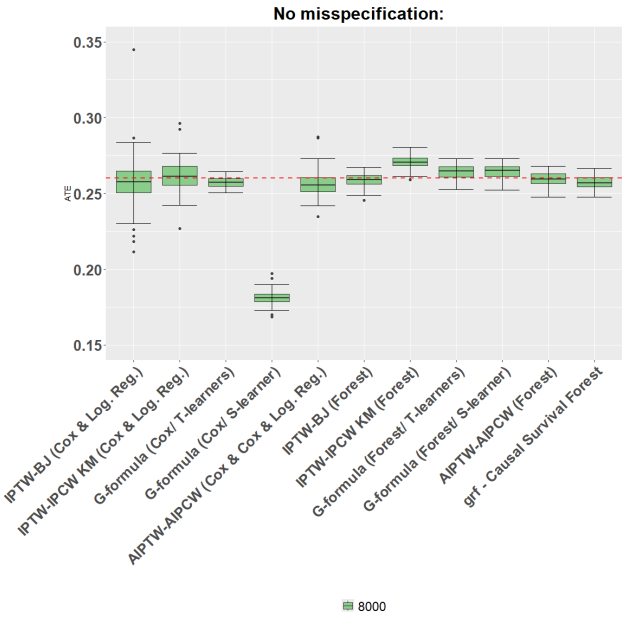
For parametric and semi-parametric models, all relevant interaction terms are explicitly included. In contrast, for the Forest model, interaction terms are not provided manually, as the model automatically captures them during training.

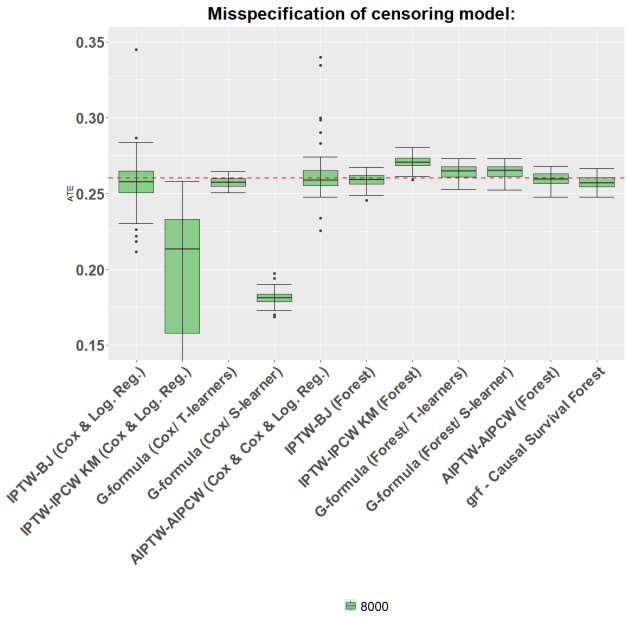
1. Forest are performing at least as good as classical regressions at high sample size (8,000). In general, these methods need more data to converge.
2. With increasing misspecification, the added value of robust estimators (AIPTW-AIPCW) and Forest becomes evident. All the estimators that use misspecified working models are biased.

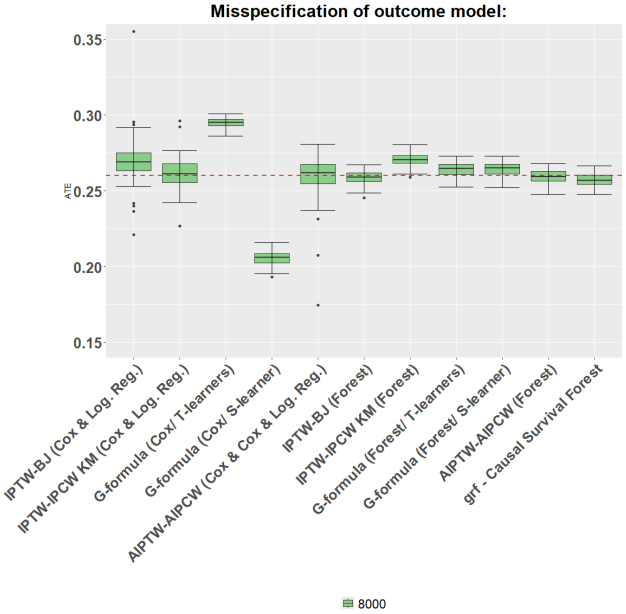
Forest remains more **flexible** than traditional regression models when there is **uncertainty about the model form (or missing interaction)**

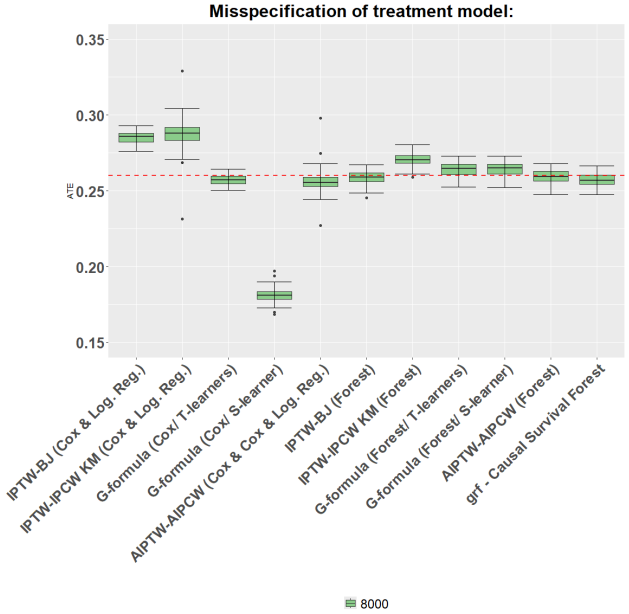
Misspecification: Obs & Dependent Censoring

100 studies at sample size 8,000 are generated.









For parametric and semi-parametric models, all relevant interaction terms are explicitly included. In contrast, for the Forest model, interaction terms are not provided manually, as the model automatically captures them during training.

1. Forest are performing at least as good as classical regressions at high sample size (8,000). In general, these methods need more data to converge.
2. With increasing misspecification, the added value of robust estimators (AIPTW-AIPCW) and Forest becomes evident. All the estimators that use misspecified working models are biased.

Forest remains more **flexible** than traditional regression models when there is **uncertainty about the model form (or missing interaction)**

WARNING:

In this context, misspecification refers to **omitting interaction terms**. However, many other forms of misspecification exist, such as:

- **Omitting important variables**
- **Incorrect assumptions** about the model structure (e.g., assuming proportional hazards when the true model is non-parametric or assuming linearity instead of a nonlinear relationship)