SUPPLEMENTARY MATERIAL FOR *DOUBLY ROBUST TREATMENT EFFECT ESTIMATION WITH INCOMPLETE ATTRIBUTES*

By Imke Mayer[¶], Stefan Wager[††] Tobias Gauss[‡‡] Jean-Denis Moyer[‡‡] and Julie Josse[**]

*École des Hautes Études en Sciences Sociales[¶], École Polytechnique[¶**], INRIA Saclay[**], Stanford University[††] and Beaujon Hospital[‡‡]*

1. Proofs.

1.1. *Consistency of AIPW.*

Proof. In order to show the double robustness of $\hat{\tau}_{AIPW}$, as given in (**??**), let us rewrite it by rearranging the terms:

$$\begin{aligned}\hat{\tau}_{AIPW} &= \frac{1}{n}\sum_{i=1}^{n}\frac{W_iY_i}{\hat{e}(X_i)} - \frac{W_i-\hat{e}(X_i)}{\hat{e}(X_i)}\hat{\mu}_1(X_i) - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-W_i)Y_i}{1-\hat{e}(X_i)} + \frac{W_i-\hat{e}(X_i)}{1-\hat{e}(X_i)}\hat{\mu}_0(X_i)\\ &=: \hat{\mu}_{1,AIPW} - \hat{\mu}_{0,AIPW}.\end{aligned}$$

First note that by the law of large numbers, $\hat{\mu}_{1,AIPW}$ and $\hat{\mu}_{0,AIPW}$ respectively estimate $\mathbb{E}[Y_i(1)] + \eta_1$ and $\mathbb{E}[Y_i(0)] + \eta_0$ where η_1 and η_0 are given by

$$\eta_1 \triangleq \mathbb{E}\left[\frac{W_i-e(X_i)}{e(X_i)}(Y_i(1)-\mu_1(X_i))\right],\ \eta_0 \triangleq \mathbb{E}\left[\frac{W_i-e(X_i)}{1-e(X_i)}(Y_i(0)-\mu_0(X_i))\right].$$

Indeed we have that

$$\begin{aligned}\mathbb{E}\left[\frac{W_iY_i}{e(X_i)} - \frac{W_i-e(X_i)}{e(X_i)}\mu_1(X_i)\right] &= \mathbb{E}\left[\frac{W_iY_i(1)}{e(X_i)} - \frac{W_i-e(X_i)}{e(X_i)}\mu_1(X_i)\right]\\ &= \mathbb{E}[Y_i(1)] + \mathbb{E}\left[\frac{W_i-e(X_i)}{e(X_i)}(Y_i(1)-\mu_1(X_i))\right],\end{aligned}$$

where the first equality results from SUTVA: $W_iY_i = W_i(W_iY_i(1) + (1-W_i)Y_i(0)) = W_iY_i(1) + W_i(1-W_i)Y_i(0)$. And similar for the derivation of η_0.

[¶]E-mail: imke.mayer@ehess.fr
[††]E-mail: swager@stanford.edu
[**]E-mail: julie.josse@polytechnique.edu

Now, the double robustness can easily be shown by considering these two terms:

- If the propensity model $e(x)$ is correctly specified but the outcome model $(\mu_0(x), \mu_1(x))$ is mis-specified we have

$$\begin{aligned}\eta_1 &= \mathbb{E}\left[\mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}(Y_i(1) - \mu_1(X_i))\right] \mid Y_i(1), X_i\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[W_i \mid Y_i(1), X_i] - e(X_i)}{e(X_i)}(Y_i(1) - \mu_1(X_i))\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[W_i \mid X_i] - e(X_i)}{e(X_i)}(Y_i(1) - \mu_1(X_i))\right] = 0.\end{aligned}$$

  We use the unconfoundedness assumption to go from the second to the third line and the definition of the propensity score for the last equality.
- If the propensity model $e(x)$ is mis-specified but the outcome model $(\mu_0(x), \mu_1(x))$ is correctly specified we have

$$\begin{aligned}\eta_1 &= \mathbb{E}\left[\mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}(Y_i(1) - \mu_1(X_i))\right] \mid W_i, X_i\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}\left(Y_i(1) - \mathbb{E}[Y_i \mid W_i = 1, X_i]\right) \mid W_i, X_i\right]\right] \\ &= \mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}\left(\mathbb{E}[Y_i(1) \mid W_i, X_i] - \mathbb{E}[Y_i \mid W_i = 1, X_i]\right)\right] \\ &= \mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}\left(\mathbb{E}[Y_i(1) \mid X_i] - \mathbb{E}[Y_i(1) \mid X_i]\right)\right] = 0,\end{aligned}$$

  where we use SUTVA and unconfoundedness to go from the third to the fourth line.

Analogously we obtain in both cases of mis-specification that $\eta_0 = 0$, proving the double robustness of $\hat{\tau}_{AIPW}$. □

1.2. *Treatment Effect Estimation with Missing Attributes.* Here we prove the balancing property of the generalized propensity score (**??**).

Proof. We note that the distribution of W is fully specified by its mean. Therefore we need to prove that:

$$\mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, X_i^*] = \mathbb{E}[W_i|X_i^*] \Rightarrow \mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, e^*(X_i^*)] = \mathbb{E}[W_i|e^*(X_i^*)]$$

a) By the law of total expectation we have:

$$\mathbb{E}[W_i \,|\, e^*(X_i^*)] = \mathbb{E}[\mathbb{E}[W_i|X_i^*, e^*(X_i^*)] \,|\, e^*(X_i^*)] = \mathbb{E}[\mathbb{E}[W_i|X_i^*] \,|\, e^*(X_i^*)] = e^*(X_i^*)$$

b) And again using the law of total expectation we have the following:

$$\begin{aligned}
&\mathbb{E}[W_i \,|\, \{Y_i(0), Y_i(1)\}, e^*(X_i^*)] \\
&\quad = \mathbb{E}[\mathbb{E}[W_i \,|\, \{Y_i(0), Y_i(1)\}, X_i^*, e^*(X_i^*)] \,|\, \{Y_i(0), Y_i(1)\}, e^*(X_i^*)] \\
&\quad = \mathbb{E}[\mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, X_i^*] \,|\, \{Y_i(0), Y_i(1)\}, e^*(X_i^*)] \\
&\quad = \mathbb{E}[\mathbb{E}[W_i \,|\, X_i^*] \,|\, \{Y_i(0), Y_i(1)\}, e^*(X_i^*)] \qquad \text{(assuming (??))} \\
&\quad = \mathbb{E}[e^*(X_i^*) \,|\, \{Y_i(0), Y_i(1)\}, e^*(X_i^*)] = e^*(X_i^*)
\end{aligned}$$

□

2. Procedures. In this section we give the details of all procedures omitted in the main article. The IPW counterparts to the Procedures presented in the main article and to Procedure 4 are obtained by simply dropping the regressions of Y on the (proxies for the) confounders and by estimating τ using expression (**??**) and its generalized extension

$$\hat{\tau}_{IPW^*} \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}^*(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}^*(X_i)}\right), \tag{1}$$

instead of expressions (**??**) and (**??**).

Procedure 4: AIPW with matrix factorization pre-processing.
This algorithm provides an estimation for the average treatment effect τ using a matrix factorization pre-processing, given observed covariates with missing attributes, observed treatment assignment and outcome. We assume a low rank matrix factorization model for X and unconfoundedness (**??**) given the latent factors U as detailed in Section **??** and MCAR.

1. Estimate the latent factors U using SVD decomposition of X, choose the number of latent factors by cross-validation.
2. **Option 1** Nonparametric regression.
   (a) Train a causal forest on $(\hat{U}, W, Y)$.
   (b) Take the average over the out-of-bag predictions of conditional average treatment effects $\tau(\hat{U}_i) = \mathbb{E}[Y_i(1) - Y_i(0)|\hat{U}_i]$ using the trained causal forest to obtain an estimation $\hat{\tau}$ for τ as in (**??**).

   Option 2 Parametric regression (we additionally assume logistic-linear model specification for $(e, \mu_{(0)}, \mu_{(1)})$).
   (a) Fit a logistic model to obtain predictions for the propensity score $e(\hat{U}_i)$
   (b) Fit two separate linear models on $(Y_{i:W_i=1}, \hat{U}_{i:W_i=1})$ and on $(Y_{i:W_i=1}, \hat{U}_{i:W_i=1})$ respectively to obtain predictions for $\mu_{(1)}(\hat{U}_i)$ and $\mu_{(0)}(\hat{U}_i)$ respectively.
   (c) Combine the predictions following (**??**) to obtain a doubly robust estimation $\hat{\tau}$ for τ.

3. Simulation study on synthetic data.

3.1. *Interpretation and discussion of the results from Section* **??**. Figure **??** shows that if the data is MCAR and satisfies (**??**), *saem* works well as expected, i.e. it converges to the true value τ. Note however that the EM-based estimators fail in the small sample case ($(n, p) = (100, 10)$). This is likely due to the strong correlation in the covariates, leading to numerically singular variance-covariance estimates for low sample sizes. Note that *mia.grf* also converges but very slowly which is expected due to the smoothness of e^* and $\mu^*_{(w)}$ and as it does not use the strong parametric assumptions which are met in these simulations. The method *mean.grf* gives similar results than *mia.grf*, which is expected according to the results from Josse et al. (2019). We observe that *mean.loglin* performs similarly to *saem*, in terms of convergence and behavior w.r.t. the unconfoundedness assumptions. Figure **??**

shows as well that *mice* works under both unconfoundedness assumptions as expected[1]. In particular, when only (**??**) holds and (**??**) is violated, then all methods but multiple imputation give biased results.

In the general missingness case, Figure **??**, we only expect *mia.grf* and *mean.grf* to perform well as explained in Section **??**. However their convergence seems to be very slow which again can be explained with the strong parametric and smooth models we defined with the attributes X and that are hard to estimate with random forests. The good performance of the others estimators in this general case can only be observed when the mask R is used in the estimation, otherwise these methods fail in this setting, as expected but not shown in Figure **??**.

Under Model 3, Figures **??** show that, as expected, if (**??**) is satisfied, our estimator *mia.grf* converges quickly to the true value τ while the other methods remain biased. With the exception of *mice*, all other methods fail if the "unconfoundedness despite missingness" assumption is violated, independently from the missingness mechanism. However *mia.grf* and *mean.grf* in AIPW-form seem to cope well even under the standard unconfoundedness (**??**).

Figures **??** and **??** show that under Model 4, *mia.grf* converges to the true value τ in all cases but rather slowly, provided assumption (**??**) is met. Even in the "simplest" MCAR case, the parametric observed-likelihood based approach, namely *saem*, fails under DLVM for small sample sizes ($n \in \{100, 500\}$). Indeed, while satisfying the necessary normality assumption, the observations X_i are not i.i.d. due to their (nonlinear) dependence on the (latent) codes C_i. This behavior of *mia.grf* and *saem* is again in accordance with Section **??**. The multiple imputation method yields some biased estimations in the MCAR case but performs well in the general case (with the mask). Note that the poor performance of the estimator based on low-rank matrix factorization (*mf*) is not surprising since the latency structure arises in the covariate generating process, but the confounders themselves are defined as the observed X rather than the latent factors (C or $\mu(C)$).

For model 4, where treatment and outcome are unconfounded given some latent factors U, we observe on Figure **??** that the estimator based on low-rank matrix factorization in the MCAR performs well. This result is expected, since we assume confoundedness on to the latent factors U and not the partially observed covariates X. Hence the crucial point for recovering the treatment effect is the recovery of these latent factors U, as pointed out by Kallus, Mao and Udell (2018). Interestingly, all methods—except *saem*

[1]Note that the small remaining bias with multiple imputation is likely to vanish as the number of imputations increases.

which fails in the case of informative missingness—empirically perform well in this scenario scenario. This again, is only observed as long as the mask is used for estimation. Furthermore, our *mia.grf* and *mean.grf* seem to converge to the true value of τ despite the "wrong" unconfoundedness assumption.

3.2. *Simulation results for a variant of Model 4.* The hierarchical data-generating model used in Section **??** can be modified in order to allow for correlation between covariates by defining the code-depending Gaussian parameters as

$$(\mu(c), \Sigma(c)) = (U(V \tanh(Wc + a) + b), U \exp(\gamma^T (Wc + a) + \delta) I_p U^T),$$

for some randomly generated orthonormal matrix U.

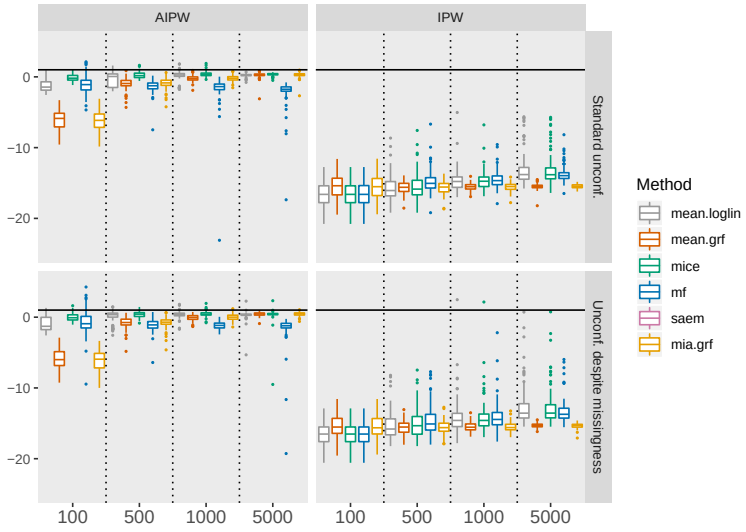
The difference in terms of bias and variability between the AIPW-type estimators and their IPW-type equivalent is clear in this scenario. However the difference in terms of bias w.r.t. the different unconfoundedness assumptions is less apparent. More precisely, *mia.grf* and *mean.grf* seem to approximate the true treatment effect τ for large sample sizes ($n \geq 500$) similarly in both scenarios (first and second line in Figure 1a and 1b). These observations require further investigations in the future.

4. Details on the medical application (Traumabase).

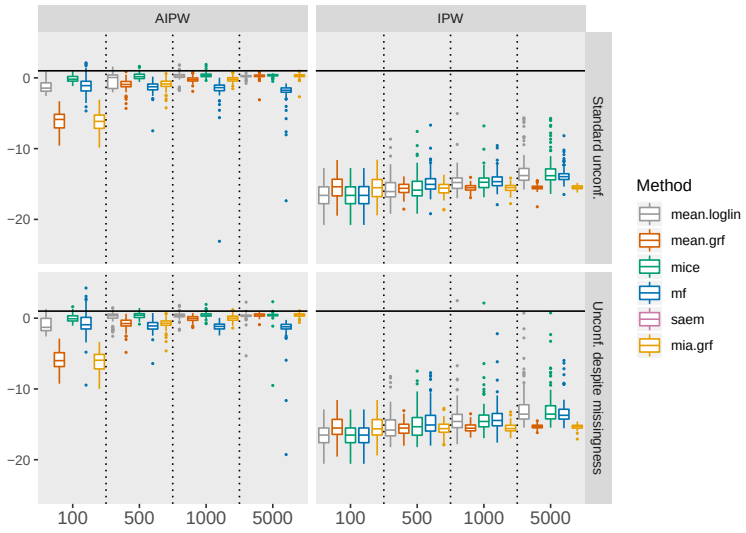
4.1. *Definition of the variables of the Traumabase used in the analysis.* Here we provide the names and short descriptions of the variables we use in our causal analysis. The moment at which the variable is first available is given in parentheses (ph = pre-hospital phase, h = hospital phase).

List of confounders:.

- *Trauma.center* (categorical): name of the trauma center. (ph/h)
- *SBP.ph*, *DBP.ph*, *HR.ph* (continuous): systolic and diastolic arterial pressure and heart rate during pre-hospital phase (SBP.ph = min(SBP.min, SBP.MICU), etc.); MICU = mobile intensive care unit. (ph)
- *Cardiac.arrest.ph* (categorical): cardiac arrest during pre-hospital phase. (ph)
- *HemoCue.init* (continuous): prehospital capillary hemoglobin concentration (the lower, the more the patient is probably bleeding and in shock); hemoglobin is an oxygen carrier molecule in the blood. (ph)
- *SpO2.min* (continuous): peripheral oxygen saturation, measured by pulse oxymetry, to estimate oxygen content in the blood ($95 - 100\%$:



(a) MCAR (with 30% missing values in $X_{\cdot,1:10}$)



(b) MNAR (with 30% missing values in $X_{\cdot,1:5}$)

Fig 1: Estimated average treatment effect $\hat{\tau}$. **Hierarchical data-generating model with dense covariance matrices** for confounders.

considered normal; $< 90\%$ critical and associated with considerable trauma, danger and mortality). (ph)
- *Vasopressor.therapy* (continuous): treatment with catecholamines in case of physical or emotional stress increasing heart rate, blood pressure, breathing rate, muscle strength and mental alertness. (ph)
- *Cristalloid.volume* (continuous): total amount of prehospital administered cristalloid fluid resuscitation (volume expansion). (ph)
- *Colloid.volume* (continuous): total amount of prehospital administered colloid fluid resuscitation (volume expansion). (ph)
- *Shock.index.ph* (continuous): ratio of heart rate and systolic arterial pressure during pre-hospital phase. (ph)
- *AIS.external* (discrete, range: $[0, 6]$): Abbreviated Injury Score for external injuries, here it is assumed to be a proxy of information available/visible during pre-hospital phase. (ph/h)
- *Delta.shock.index* (continuous): Difference of shock index between arrival at the hospital and arrival on the scene. (h)
- *Delta.hemoCue* (continuous): Difference of hemoglobin level between arrival at the hospital and arrival on the scene. (h)

List of predictors of mortality and that are not associated with treatment assignment.

- *Anticoagulant.therapy* (categorical): oral anticoagulant therapy before the accident. (ph)
- *Antiplatelet.therapy* (categorical): anti-platelet therapy before the accident. (ph)
- *GCS(.init)* (discrete, range: $[3, 15]$): Initial Glasgow Coma Scale (GCS) on arrival on scene of enhanced care team and on arrival at the hospital ($GCS = 3$: deep coma; $GCS = 15$: conscious and alert). (ph & h)
- *GCS.motor(.init)* (discrete, range: $[1, 6]$): Initial Glasgow Coma Scale motor score ($GCS.motor = 1$: no response; $GCS.motor = 6$: obeys command/purposeful movement). (ph & h)
- *Pupil.anomaly* (categorical): pupil dilation indicating brain herniation. (ph & h)
- *Osmotherapy* (categorical): administration of osmotherapy to alleviate compression of the brain (either Mannitol or hypertonic saline solution). (ph & h)
- *Improv.anomaly.osmo* (categorical): change of pupil anomaly after administration of osmotherapy. (ph)
- *Medcare.time.ph* (continuous): total duration of prehospital care team engaged (arrival on scene to arrival at hospital). (h)

- *FiO2* (discrete, range: $[0, 5]$): inspired concentration of oxygen on ventilatory support (the higher the more critical; $Ventilation = 0$: no ventilatory support). (h)
- *Temperature.min* (continuous): Minimal body temperature. (h)
- *TCD.PI.max* (continuous): pulsatility index (PI) measured by echodoppler sonographic examen of blood velocity in cerebral arteries ($PI > 1.2$: indicates altered blood flow maybe due to traumatic brain injury). (h)
- *IICP* (categorical): at least one episode of increased intracranial pressure; mainly in traumatic brain injury; usually associated with worse prognosis. (h)
- *EVD* (categorical): external ventricular drainage (EVD); mean to drain cerebrospinal fluid to reduce intracranial pressure. (h)
- *Decompressive.craniectomy* (categorical): surgical intervention to reduce intracranial hypertension. (h)
- *Neurosurgery.day0* (categorical): neurosurgical intervention performed on day of admission. (h)
- *AIS.head, AIS.face* (discrete, range: $[0, 6]$): Abbreviated Injury Score, describing and quantifying facial and head injuries ($AIS = 0$: no injury; the higher the more critical).(h)
- *ISS* (discrete, range: $[0, 108]$): Injury Severity Score, sum of squares of top three AIS scores. (h)
- *IGS.II* (continuous): Simplified Acute Physiology Score. (h)

4.2. *ATE estimation on the Traumabase using overlap weights.* An often raised concern with many medical observational data sets is the potential violation of the overlap assumption. For instance some patients might never get the treatment due to infrastructural circumstances or due to recommendations followed strictly by the entire medical staff. The overlap assumption however is needed for consistency of the treatment effect estimations and states that every patient has a non-zero probability of being in either treated or control group. Another way of describing this assumption is that the treatment groups are sufficiently comparable, otherwise the attempt of drawing causal inferences is doomed to failure from the beginning.

Given the important level of heterogeneity among trauma patients, especially among patients with traumatic brain injury, and the multi-level and multi-actor nature of the data, it cannot be ruled out that the treatment groups have only small overlap. When considering standardized mean differences of the confounding variables between treatment and control groups in Figure 2 it appears indeed that certain features such as the hemoglobin level differ considerably between the two groups. As detailed in Section **??**, a pos-

sible solution to deal with this potential situation is the use of overlap weights instead of the inverse propensity weights (Li, Morgan and Zaslavsky, 2018). However, in our case, when using the corresponding modified estimands and estimators, i.e. the average treatment effect on the overlap population, the results reported in Figure 3 are very similar to those from the normal average treatment effect estimation on the entire population (Figure **??**) and lead to the same conclusion about the treatment effect.

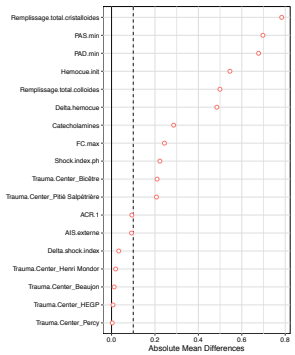


Fig 2: Unadjusted standardized mean differences of the confounding variables.

References.

Josse, J., Prost, N., Scornet, E. and Varoquaux, G. (2019). On the consistency of supervised learning with missing values. *arXiv preprint*.

Kallus, N., Mao, X. and Udell, M. (2018). Causal Inference with Noisy and Missing Covariates via Matrix Factorization. In *Advances in Neural Information Processing Systems* 6921–6932.

Li, F., Morgan, K. L. and Zaslavsky, A. M. (2018). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association* **113** 390–400.

I. Mayer
Centre d'Analyse et de Mathématique Sociales
EHESS
75006 Paris, France
E-mail: imke.mayer@ehess.fr

J. Josse
Centre de Mathématiques Appliquées
École Polytechnique
91128 Palaiseau Cedex, France
E-mail: julie.josse@polytechnique.edu

S. Wager
Graduate School of Business
Stanford University
CA 94305, USA
E-mail: swager@stanford.edu

T. Gauss and J.-D. Moyer
Department of anesthesia and intensive care
Beaujon hospital, AP–HP
92110 Clichy, France
E-mail: tgauss@protonmail.com
E-mail: jean-denis.moyer@aphp.fr

[2]Values on the x-axis are multiplied by 100 for better readability.

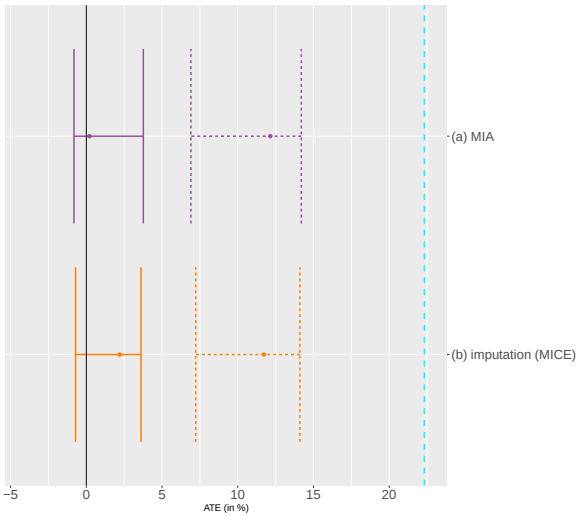


Fig 3: ATE estimations on overlap population on Traumabase data (solid: doubly robust estimates; dotted: IPW estimates; dashed vertical line: without adjustment; x-axis: $\hat{\tau}$ and bootstrap confidence intervals[2]).