Submitted to the Annals of Applied Statistics

DOUBLY ROBUST TREATMENT EFFECT ESTIMATION WITH MISSING 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*[‡‡]

Missing attributes are ubiquitous in causal inference, as they are in most applied statistical work. In this paper, we consider various sets of assumptions under which causal inference is possible despite missing attributes and discuss corresponding approaches to average treatment effect estimation, including generalized propensity score methods and multiple imputation. Across an extensive simulation study, we show that no single method systematically out-performs others. We find, however, that doubly robust modifications of standard methods for average treatment effect estimation with missing data repeatedly perform better than their non-doubly robust baselines; for example, doubly robust generalized propensity score methods beat inverse-weighting with the generalized propensity score. This finding is reinforced in an analysis of an observations study on the effect on mortality of tranexamic acid administration among patients with traumatic brain injury in the context of critical care management. Here, doubly robust estimators recover confidence intervals that are consistent with evidence from randomized trials, whereas non-doubly robust estimators do not.

1. Introduction.

1.1. *Hemorrhagic shock and traumatic brain injury in critical care management.* Our work is motivated by a prospective observational study of the causal effect of tranexamic acid (TA), an antifibrinolytic agent that limits excessive bleeding, on mortality among traumatic brain injury patients during their stay at the hospital (from admission to ICU and regular care units). The beneficial effect of TA on mortality has been shown in a large randomized placebo-controlled study (Shakur et al., 2010). Our interest in

[¶]E-mail: imke.mayer@ehess.fr
[††]E-mail: swager@stanford.edu
[**]E-mail: julie.josse@polytechnique.edu
MSC 2010 subject classifications: Primary 93C41, 62G35, 62F35; secondary 62P10
Keywords and phrases: causal inference, potential outcomes, observational data, propensity score estimation, incomplete confounders, major trauma, public health

deceased	no	yes
TA not administered	2,167 (68%)	399 (13%)
TA administered	374 (12%)	228 (7%)

TABLE 1

Occurrence and frequency table for traumatic brain injury patients (total number: 3,168).

developing observational study methods for assessing the effect of TA is twofold: In the long run, observational studies will be able to incorporate data on a larger and more diverse set of patients, thus allowing us to get a better understanding of when and for whom TA works; and treatment effect estimation on such observational studies can serve as a precursor for future randomized placebo-controlled studies, namely by helping defining the most interesting or promising target population beforehand and the associated inclusion rules.

Our study is built on top of the Traumabase® database, which currently indexes around 20,000 major trauma patients.[1] For each patient, 244 measurements are collected both before and during the hospital stay, including both quantitative and categorical variables. As shown in Table 1, TA was administered to roughly 19% of traumatic brain injury patients, and 20% died before the end of their hospital stay. We also see that mortality was much higher among patients who received TA than those who did not (38% vs. 16%). This apparent reversal of the expected causal effect is a standard example of confounding bias (also known as Simpson's paradox): The effect arises because patients who appeared to be in more severe state were more likely to be administered TA and were also more likely to die with or without the treatment.

The goal of our observational study design is to use a subset of 37 auxiliary covariates collected by the Traumabase group to control for confounding and identify the causal effect of TA on mortality. This "unconfoundedness" or "selection on observables" strategy is justified if the treatment of interest (i.e., administration of TA) is as good as random after conditioning on covariates (Imbens and Rubin, 2015; Rosenbaum and Rubin, 1983). In general, such an unconfoundedness assumption cannot be validated from data, and needs to be built into the observational study design.

In order to make unconfoundedness as plausible as possible, the Traumabase group chose which covariates among the total of 244 collected covariates to incorporate in our study by soliciting feedback from a number ex-

[1]Major trauma is defined as any injury that potentially causes prolonged disability or death and it is a public health challenge and a major source of mortality and handicap around the world (Hay et al., 2017).

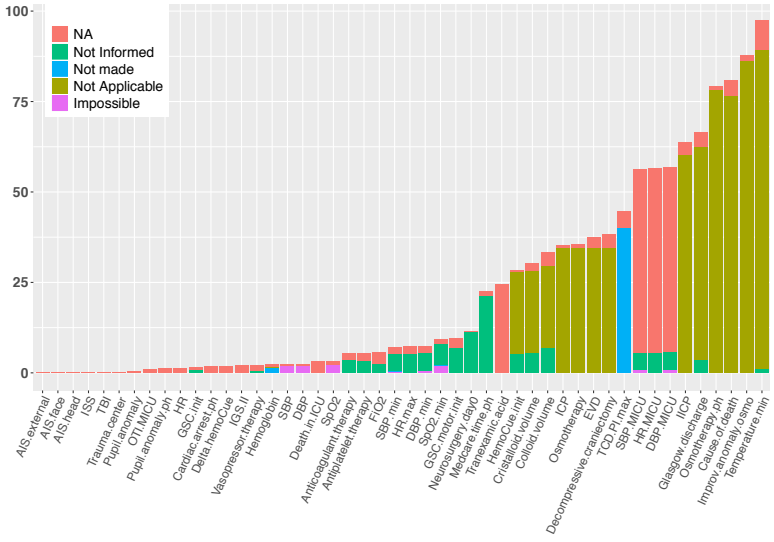


Fig 1: Percentage of missing values for a subset of variables relevant for traumatic brain injury. Different encodings of missing values: *NA* (not available), *not informed*, *not made*, *not applicable*, *impossible*.

perts using the Delphi method (Dalkey and Helmer, 1963; Jones and Hunter, 1995). The focus of the Delphi survey was in understanding which factors were important for understanding health trajectories of major trauma patients. Because the decision whether or not to administer TA was performed by health professionals, it is likely that this same set of variables is also relevant to understanding which patients were more likely than others to be selected for treatment. A detailed list of the confounders and predictors of the outcome, in-ICU mortality, that were chosen via the Delphi method is given in the Supplementary material.

As discussed further in the following section, the statistics of treatment effect estimation under unconfoundedness is by now well understood, with literature covering a range of topics from identification (Imbens and Rubin, 2015; Rosenbaum and Rubin, 1983) and simple weighted estimators (Abadie and Imbens, 2016; Rosenbaum and Rubin, 1984; Zubizarreta, 2012) to semi-parametrically efficient estimation in potentially high-dimensional settings (Athey, Imbens and Wager, 2018; Chernozhukov et al., 2018; Robins, Rotnitzky and Zhao, 1994; Van der Laan and Rose, 2011) and optimal treatment personalization (Athey and Wager, 2017; Kitagawa and Tetenov, 2018; Luedtke and Van Der Laan, 2016; Zhao et al., 2012).

In the case of the Traumabase dataset, however, we have an additional

complication whereby, in Figure 1, many of the variables have missing entries. Some of the missingness is presumably due to non-informative missingness, e.g., medical staff simply forgetting to log some numbers, but in other cases the missingness is clearly informative; and in fact the analysts compiling the dataset used many different phrases to describe missing measurements, ranging from "not made" and "not applicable" to "impossible". The last denomination arises, for example, in the case of blood pressure measurements for patients in cardiac arrest or with dismemberment, as first responders simply cannot measure blood pressure for patients suffering from one of these two conditions. Meanwhile, variables indicating the response to a certain drug, such as the pupil contraction after the administration of a saline solution, systematically take on the value "not applicable" if the treatment has not been administered (the latter is informed in a separate variable).

There are a handful of popular strategies for working with missing values in the context of treatment effect estimation under unconfoundedness, ranging from generalized propensity score methods (D'Agostino and Rubin, 2000; Rosenbaum and Rubin, 1984) to multiple imputation (Little and Rubin, 2002; Rubin, 1976, 1987). However, the methodology for treatment effect estimation with missingness is not as thoroughly fleshed out as corresponding methods without missing data. In particular, although doubly robust and semiparametrically efficient methods have shown considerable promise in cases without missingness (Athey, Imbens and Wager, 2018; Chernozhukov et al., 2018; Robins, Rotnitzky and Zhao, 1994; Van der Laan and Rose, 2011), we are not aware of a study of doubly robust treatment effect methods with missing covariates.

In this paper, we discuss natural doubly robust generalizations of several popular methods for treatment effect estimation with missing covariates, and conduct an extensive simulation comparison. There is considerable variability in which methods perform best in our experiments: Sometimes methods that start from generalized propensity scores do better while other times multiple imputation wins; sometimes parametric methods fit via the EM algorithm (Dempster, Laird and Rubin, 1977) are better whereas other times non-parametric estimators do better. However, we systematically find our doubly robust modifications of standard methods to outperform their baselines.

Finally, in the case of the Traumabase study, all doubly robust estimators give confidence intervals that cover 0, indicating that we need to collect more data before we can use the observational study to guide clinical choices around administration of TA in the context of traumatic brain injury. In

contrast, all baseline methods result in confidence intervals that do not cover 0, and find significantly harmful effects of TA on mortality. It thus appears that using doubly robust estimators is needed to eliminate the selection bias seen in Table 1.

2. Methods for Complete Data. As a preliminary to our discussion on how to estimate causal effects with missing attributes, we first briefly review methods that are widely used in the easier case without missingness. Suppose we observe n independent and identically distributed samples $(X_i, Y_i, W_i) \in \mathbb{R}^p \times \mathbb{R} \times \{0, 1\}$ where X_i is a vector of attributes, Y_i is an outcome of interest, and W_i denotes treatment assignment. We define causal effects via the Neyman-Rubin potential outcomes model under the stable unit treatment value assumption (Imbens and Rubin, 2015). We posit potential outcomes $\{Y_i(0), Y_i(1)\}$ corresponding to the outcome the i-th sample would have experienced had they been assigned treatment $W_i = 0$ or 1 respectively, such that $Y_i = Y_i(W_i)$. The average treatment effect is then defined as

$$\tau \triangleq \mathbb{E}[Y_i(1) - Y_i(0)].$$

In order to identify τ, we further assume unconfoundedness, i.e., that treatment assignment is as good as random conditionally on the attributes X_i (Rosenbaum and Rubin, 1983),

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i, \tag{1}$$

and overlap, i.e., that the propensity score $e(\cdot)$ is bounded away from 0 and 1,

$$e(x) \triangleq \mathbb{P}\left[W_i = 1 \,|\, X_i = x\right], \quad \eta < e(x) < 1 - \eta, \tag{2}$$

for all $x \in \mathbb{R}^p$ and some $\eta > 0$.

In the case without any missingness in the attributes X_i, the problem of average treatment effect estimation in the above setting is well understood. Several popular and consistent approaches to estimating τ are built around the propensity score. The analyst first estimates the propensity score $e(x)$ in (2), and then estimates τ either via inverse-propensity weighting (IPW)

$$\hat{\tau}_{IPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)} \right), \tag{3}$$

or by matching treated and control observations with similar values of the propensity score (Abadie and Imbens, 2016; Rosenbaum and Rubin, 1984; Zubizarreta, 2012).

However, when the propensity score is somewhat difficult to estimate, methods that only rely on the propensity score are in general dominated by bias due to estimation error in $e(\cdot)$, and methods that also model the outcomes Y_i can attain a better sample complexity; see Athey, Imbens and Wager (2018), Chernozhukov et al. (2018) and Van der Laan and Rose (2011) for references and recent results. One particularly successful approach to combining these two approaches to modeling is via augmented inverse-propensity weighting (AIPW) (Robins, Rotnitzky and Zhao, 1994),

$$\hat{\tau}_{AIPW} \triangleq \frac{1}{n}\sum_{i=1}^{n}\Bigg(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + \frac{W_i}{\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(1)}(X_i)\right) - \frac{(1-W_i)}{1-\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(0)}(X_i)\right)\Bigg), \tag{4}$$

where $\mu_{(w)}(x) \triangleq \mathbb{E}\left[Y \,\middle|\, X_i = x,\, W_i = w\right]$ and $\hat{\mu}_{(w)}(x)$ is an estimate thereof.

A key fact about doubly robust estimators as in (4) is that $\hat{\tau}_{AIPW}$ can be $\sqrt{n}$-consistent for τ and asymptotically Gaussian even in a non-parametric setting where $\hat{\mu}_{(w)}(\cdot)$ and $\hat{e}(\cdot)$ are estimated, for instance using generic machine learning methods, at slower non-parametric rates (Farrell, 2015). We use "cross-fitting", whereby we do not use the i-th datapoint itself for making the predictions $\hat{\mu}_{(w)}(X_i)$ and $\hat{e}(X_i)$ (Chernozhukov et al., 2018; Van der Laan and Rose, 2011). Methods based on inverse-weighting as in (3) can also sometimes have good asymptotic performance, but these results are generally more fragile and require considerably stronger regularity conditions than corresponding AIPW results (Hirano, Imbens and Ridder, 2003).

3. Treatment Effect Estimation with Missing Attributes. In this paper, we are interested in a more difficult variant of the above setting where the analyst cannot always observe the full attribute vector. Rather, we assume that there is a "mask" $R_i \in \{1, \mathrm{NA}\}^p$ such that the analyst observes $X_i^* \triangleq R_i \odot X_i \in \{\mathbb{R} \cup \mathrm{NA}\}^p$. Here, $\odot$ denotes an element-wise product, such that $X_{ij}^* = X_{ij}$ if $R_{ij} = 1$ and $X_{ij}^* = \mathrm{NA}$ if $R_{ij} = \mathrm{NA}$.

In current empirical practice, there are several approaches to treatment effect estimation with missing attributes; but the literature studying this problem is rather scarce and most such approaches focus on IPW-form estimators as in (3) (Rosenbaum and Rubin, 1984; D'Agostino and Rubin, 2000; Seaman and White, 2014; Mattei, 2009; Leyrat et al., 2019).

The main contributions of this paper consist in (1) a dyadic classification of possible approaches to treatment effect estimation with missing attributes —the first class relying on a variant of the unconfoundedness assumption while the second uses the classical missing values mechanism taxonomy—(2)

in the proposal of two new estimators in the first class—a parametric and nonparametric estimator, both in an IPW and an AIPW form—(3) the extension of previously introduced IPW estimators to the AIPW form in the second class and (4) an extensive comparison of these estimators.As preliminaries, below we review some paradigms for treatment effect estimation with missing attributes.

3.1. *Unconfoundedness despite missingness.* Perhaps the simplest way to work with missing attributes is to assume that the missingness mechanism does not break unconfoundedness (1), i.e., that (Rosenbaum and Rubin, 1984)

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i^*. \tag{5}$$

In this setting, D'Agostino and Rubin (2000) show that matching on the generalized propensity score

$$e^*(x^*) \triangleq \mathbb{P}\left[W_i = 1 \,|\, X_i^* = x^*\right] \tag{6}$$

is consistent for τ. In general, the simplest way to verify (5) is to pair (1) together with one of the two assumptions below (Blake et al., 2019; Mattei, 2009)

$$\begin{cases} \text{CIT:} & W_i \perp\!\!\!\perp X_i \,|\, X_i^*, R_i \\ \textbf{or} & \\ \text{CIO:} & Y_i(w) \perp\!\!\!\perp X_i \,|\, X_i^*, R_i \quad \text{for } w \in \{0,1\}, \end{cases} \tag{7}$$

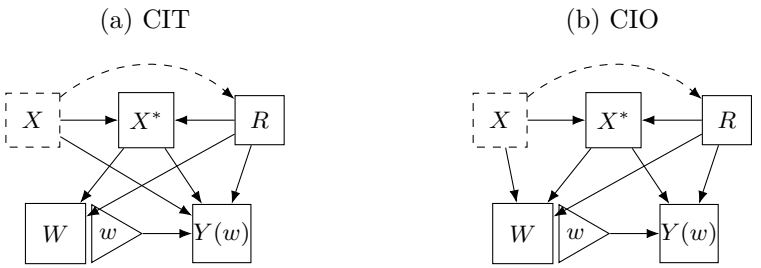
where CIT and CIO stand for *conditional independence of treatment* and *conditional independence of outcome* respectively. Given these assumptions, (5) can be directly derived from the causal graphs shown in Figure 2 (Pearl, 1995; Richardson and Robins, 2013).

We note that fitting (6) may appear difficult from the perspective of classical parametric statistics; e.g., in order to run logistic regression, one needs to fit a separate parameter vector for each mask r. However, many modern machine learning methods, including tree ensembles and neural networks, can readily handle missing data and enable (6) to be fit directly.

3.2. *Missing values mechanisms.* Another choice is to make assumptions about the missingness mechanism R_i. The most popular approach is to take the missingness mechanism to be random (MAR) (Little and Rubin, 2002; Rubin, 1976), i.e., for each possible mask $r \in \{1, \text{NA}\}^p$,

$$pr(R_i = r \,|\, X_i = x, W_i, Y_i) = pr(R_i = r \,|\, (X_i)_r = x_r, W_i, Y_i), \tag{8}$$

Fig 2: Causal graph depicting the assumptions (7).



where X_r is the subset of entries of X indexed by $\{j : r_j = 1\}$. Under these assumptions, multiple imputation (Rubin, 1987; van Buuren, 2018) is a popular approach to treatment effect estimation (Qu and Lipkovich, 2009; Robins and Wang, 2000; Rubin, 1978, 2004; Seaman and White, 2014). Under the condition that this imputation is "proper", i.e., that the missing attributes are simulated from the correct conditional distribution, and a linear-logistic model for the outcome and treatment this method is consistent for IPW estimators (Seaman and White, 2014). Note that multiple imputation does not rely on the assumption (5) or the generalized propensity score, but it only requires the data to be MAR as in (8).

A stronger variant of the missing-at-random assumption (8) is to assume missingness to be completely at random (MCAR),

$$pr(R_i = r \,|\, X_i, W_i, Y_i) = pr(R_i = r),$$

or equivalently

$$R_i \perp\!\!\!\perp \{X_i, Y_i, W_i\}.$$

Under this assumption, further methods become available. First, we can consistently estimate τ using only the subset of the data with no missingness, i.e., $X_i = X_i^*$. Of course, using only a subset of the data results in a loss of efficiency; however, this approach is simple and consistent. We emphasize that complete case analysis is not valid under the weaker assumption (8); in that case, ignoring observations with missingness will result in bias (Little and Rubin, 2002).

Another algorithm that has been studied under the MCAR assumption is based on matrix completion (Kallus, Mao and Udell, 2018). Write X and X^* for the matrices with rows X_i and X_i^* respectively. Then, assuming that X is a potentially noisy realization of a low rank matrix U and that unconfoundedness (1) holds with X_i replaced by U_i, we can approximate U from X^* using methods for low-rank matrix factorization (e.g., Candes

and Plan, 2010), and then apply complete-data methods on the recovered $\hat{U}_i$. In cases where both MCAR and the low-rank assumption hold, matrix factorization may be more efficient than complete case analysis and simpler than multiple imputation.

3.3. *Discussion: The Traumabase study.* In light of the previous discussion on the underlying (additional) assumptions required in the case of missing attributes, we argue that the Traumabase data is more likely to fall under the *unconfoundedness despite missingness* assumption from Section 3.1 than the MAR assumption from Section 3.2. Indeed, the administration of TA in the context of major trauma generally takes place under time pressure—the more blood a patient looses, the more complications can occur—and the medical staff cannot wait too long to collect a lot of information before deciding on the treatment. Therefore, if a value such as the evolution of the shock index level between arrival of the MICU[2] and arrival at the ICU, is not available because at least one measurement is missing—for instance, due to transmission problems –, the decision on the treatment will not depend on this feature. Another example could be information about the pre-hospital hemoglobin level: if the patient is in a severe state and immediate measures (such as resuscitation) are prioritized, then this measurement might not be made, however the consequently missing value is informative in the sense that it is due to the severe state of the patient, which might not necessarily be recorded explicitly in other observed features. These examples point in favor of the *unconfoundedness despite missingness* assumption as they suggest that the missing values are not only missing for the analyst but have already been missing for the physician at the time of treatment administration.

On the contrary, the MAR assumption seems plausible only for a subset of covariates. For instance, if the binary variable *Cardiac.arrest.ph* indicates that the patient needed to be resuscitated, then this can explain the missing values for the blood pressure and heart rate during pre-hospital phase. And there are other incomplete variables such as the total quantity of volume expanders used in pre-hospital phase for which the missing values depend on several other recorded variables describing the need for volume expansion.

4. IPW and augmented IPW with Missing Attributes. The previously discussed assumptions lead to two families of methods for treatment effect estimation with missing attributes. We now propose two IPW and AIPW estimators in the family derived from the *unconfoundedness despite*

[2] *Mobile intensive care unit*, enhanced medical care team that takes care of the patient at the scene of the accident.

missingness assumption (Section 3.1). In the other family that relies on classical assumptions on the *missingness mechanism* (Section 3.2), we extend the existing multiple imputation IPW estimator to a doubly robust AIPW version. For the former family, we only present details for the AIPW estimators, their IPW counterparts can almost directly be read off the AIPW formulation below.

4.1. *Unconfoundedness despite missingness.* Under assumption (5), the generalization to incomplete attributes is direct. First, estimate the generalized propensity score $e^*(x^*)$ from (6) and similarly the generalized outcome model $\mu^*_{(w)}(x^*)$, and then form the AIPW estimator

(9)
$$\hat{\tau}_{AIPW^*} \triangleq \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\mu}^*_{(1)}(X_i^*) - \hat{\mu}^*_{(0)}(X_i^*) + \frac{W_i}{\hat{e}^*(X_i^*)}\Big(Y_i - \hat{\mu}^*_{(1)}(X_i^*)\Big) - \frac{(1-W_i)}{1-\hat{e}^*(X_i^*)}\Big(Y_i - \hat{\mu}^*_{(0)}(X_i^*)\Big)\Big).$$

There are general results about AIPW that immediately guarantee that the above estimator $\hat{\tau}_{AIPW^*}$ is $\sqrt{n}$-consistent and asymptotically normal around τ provided $\hat{e}^*(\cdot)$ and $\hat{\mu}^*_{(w)}(\cdot)$ converge at $o(n^{-1/4})$ rate in root-mean squared error given only weak regularity conditions (Chernozhukov et al., 2018). Below, we consider both a parametric approach based on logistic regression, and a non-parametric approach using random forests.

4.1.1. *Nonparametric approach.* The non-parametric task of learning $e^*(x^*)$ and $\mu^*_{(w)}(x^*)$ is somewhat unusual, since the x^* take values in the augmented space $\{\mathbb{R} \cup \mathrm{NA}\}^p$. However, this problem has received attention in the machine learning literature. For example, random forests (Breiman, 2001) can handle semi-continuous variables therefore allowing for missing values in the data. One solution that takes into account the missingness in tree models is *missing incorporated in attributes* (MIA) (Twala, Jones and Hand, 2008; Josse et al., 2019). It allows optimal splits along the observed variables. Therefore, it selects patterns that are important for predicting the treatment assignment (and also the outcome) instead of adjusting one model per pattern as would be a naive approach to estimate (6). More formally, this procedure estimates the following quantity (Bayes estimate):

$$\mathbb{E}[V|X^*] = \sum_{r\in\{0,1\}^p} \mathbb{E}[V|X^*, R=r]\mathbb{1}_{R=r},$$

where V stands either for the treatment assignment W or for the outcome Y. In the following we will denote by $\hat{\tau}_{MIA}$ the resulting treatment effect estimator, either its IPW or its AIPW formulation. Another, conceptually even simpler approach for prediction with incomplete data is mean imputation which is consistent, provided that one uses a learning algorithm with infinite learning capacity (Josse et al., 2019). Both the MIA and mean imputation strategy are valid for arbitrary missingness mechanisms, provided that (5) holds, i.e., this method does not require the missing data to be MAR; and in many applications it is likely that MAR does not hold, therefore this approach can be a suited alternative if (5) is more likely to hold than MAR. The details of this approach are outlined in Procedure 1[3].[4]

Procedure 1: nonparametric AIPW with generalized propensity score and generalized response surfaces.
This algorithm provides an estimation for the average treatment effect τ via random forests with MIA splitting rule, given incomplete covariates X^*, observed treatment assignment W and outcome Y. We assume unconfoundedness despite missingness (5).

1. Transform the given X^* to $\widetilde{X^*}$:

   Option 1 Duplicate each variable X_j^* with at least one missing observation. Replace the missing values with $-\infty$ and ∞.

   Option 2 For each variable X_j^* with at least one missing observation, impute each missing value with the mean of X_j^*.

2. Train a random forest on $(\widetilde{X^*}, W)$ to get an estimate for e^*.
3. Train two random forest on $(\widetilde{X^*}_{i:W_i=1}, Y_{i:W_i=1})$ and on $(\widetilde{X^*}_{i:W_i=}, Y_{i:W_i=})$ to get estimates for $\mu^*_{(1)}$ and $\mu^*_{(0)}$.
4. Combine the predictions following (9) to obtain a doubly robust estimation $\hat{\tau}$ for τ.

4.1.2. *Parametric approach.* For the parametric approach, we build on work by Jiang, Josse and Lavielle (2018) and Schafer (1997), assuming a logistic and linear model for the generalized propensity score and outcome respectively. These two models are estimated by maximum likelihood esti-

[3]Alternatively, steps 2-4 from Procedure 1 can be replaced with a causal forest approach (Athey, Tibshirani and Wager, 2019).

[4]In this section, all procedures focus on AIPW estimation, the details for the IPW estimators are given in the Supplementary material.

mation, using the EM algorithm that allows to do valid inference on the observed values (Dempster, Laird and Rubin, 1977). The exact description of this parametric procedure for the AIPW estimator is outlined in Procedure 2. A limitation of this approach, as opposed to the previous one, is that additionally to the unconfoundedness despite missingness assumption it makes an assumption on the missingness mechanism, namely (8). The latter is required only for valid estimation of $e^*(x^*)$ and $\mu^*_{(w)}(x^*)$ through the EM algorithm, while it is not necessary for identification of the causal effect τ, as explained above. The resulting IPW and AIPW estimators will be denoted as $\hat{\tau}_{EM}$ in the remainder of this article.

Procedure 2: parametric AIPW with generalized propensity score and generalized response surfaces.
This algorithm provides an estimation for the average treatment effect τ via logistic and linear regressions, given incomplete covariates X^*, observed treatment assignment W and outcome Y. We assume unconfoundedness despite missingness (5) and MAR (8).

1. Fit a logistic model on (W, X^*) using the stochastic approximation EM algorithm to obtain predictions for the generalized propensity score $e^*(X^*_i)$.
2. Fit two separate linear models on $(Y_{i:W_i=1}, X^*_{i:W_i=1})$ and on $(Y_{i:W_i=1}, X^*_{i:W_i=1})$ respectively via an EM algorithm to obtain predictions for $\mu^*_{(1)}(X^*_i)$ and $\mu^*_{(0)}(X^*_i)$ respectively.
3. Combine the predictions following (9) to obtain a doubly robust estimation of τ.

4.2. *Standard unconfoundedness and missingness mechanisms.* As discussed in Section 3.2, multiple imputation is a solution if the missingness mechanism is MAR as defined by (8). We propose to augment the multiple imputation approach to obtain an AIPW estimator: we proceed similarly to Mattei (2009), i.e., we do multiple imputation using fully conditional equation (FCE) where we draw missing values from a joint distribution which is implicitly defined by the set of conditional distributions, proper imputation is ensured using a Bootstrap approach to reflect the sampling variability of the imputation models parameters. Then, on each imputed data set $m \in \{1, \ldots, M\}$, we compute an AIPW estimate $\hat{\tau}^{(m)}_{AIPW}$ given in (4) instead of the IPW estimate $\hat{\tau}^{(m)}_{IPW}$ given in (3). This approach is outlined in Procedure 3.

Another recent solution is based on matrix factorization (Kallus, Mao and Udell, 2018) as outlined in Procedure 4. Note that, unlike with multiple

imputation, we only impute each datapoint once and consistency guarantees are only given under MCAR.

Procedure 3: AIPW with multiple imputation.
This algorithm provides an estimation for the average treatment effect τ using multiple imputation, given incomplete covariates X^*, observed treatment assignment W and outcome Y. We assume unconfoundedness (1) and MAR (8).

1. Choose number of imputations M, for instance $M = 20$. Choose an imputation method, for instance *mice*. Impute the initial data X^* using an M times with the chosen imputation method to obtain M complete data matrices $(X^{(1)}, \dots, X^{(M)})$.
2. For every imputed data matrix $X^{(m)}$, $m \in \{1, \dots, M\}$:

   Option 1 Nonparametric regression.

   (a) Train a random forest on $(X^{(m)}, W)$ to get an estimate for e.

   (b) Train two random forest on $(X^{(m)}_{i:W_i=1}, Y_{i:W_i=1})$ and on $(X^{(m)}_{i:W_i=}, Y_{i:W_i=})$ to get estimates for $\mu_{(1)}$ and $\mu_{(0)}$.

   (c) Combine the predictions following (4) to obtain a doubly robust estimation $\hat{\tau}$ for τ.

   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 $e(X^{(m)}_i)$

   (b) Fit two separate linear models on $(Y_{i:\,W_i=1}, X^{(m)}_{i:\,W_i=1})$ and on $(Y_{i:\,W_i=1}, X^{(m)}_{i:\,W_i=1})$ respectively to obtain predictions for $\mu_{(1)}(X^{(m)}_i)$ and $\mu_{(0)}(X^{(m)}_i)$ respectively.

   (c) Combine the predictions following (4) to obtain a doubly robust estimation $\hat{\tau}^{(m)}$ for τ.

3. Aggregate the M estimations $(\hat{\tau}^{(1)}, \dots, \hat{\tau}^{(M)})$: $\hat{\tau} = \frac{1}{M}\sum_{m=1}^{M} \hat{\tau}^{(m)}$.

5. Simulation study. We assess the performance of the previously introduced treatment effect estimators in different scenarios, modifying the data generating process, the confounders' relationship structure, the unconfoundedness hypothesis, the missingness mechanism, the percentage of missing values, the sample size. The comparisons are twofold: (1) comparisons

between IPW-baseline and AIPW-type estimators, (2) comparisons w.r.t. the assumptions on the underlying unconfoundedness and the missingness mechanism. Note that in all simulations, we only consider the well-specified case, i.e., we do not study the (parametric) estimators' performances in case of model mis-specification. More specifically, $e(x) = \sigma(\alpha_0 + \alpha^T x + \epsilon_e)$ and $\mu_{(w)}(x) = \beta_0 + \beta^T x + w\tau + \epsilon_\mu$, where ϵ_e and ϵ_μ are zero mean and independent noise terms. All simulations are implemented in R (R Core Team, 2018).[5]

5.1. *Methods overview.* We compare our approaches $\hat{\tau}_{EM}$ and $\hat{\tau}_{MIA}$, denoted *saem* and *mia.grf* in the experiments, to the following methods, where we summarize their assumptions in Table 2:

- *mice*: Procedure 3 (and its IPW analogue detailed in the Supplementary material) with Option 2; we use the R package `mice` (van Buuren and Groothuis-Oudshoorn, 2011) and default options.
- *mf*: Procedure 4 (and its IPW analogue detailed in the Supplementary material) with Option 2; we adapt the implementation[6] of Kallus, Mao and Udell (2018) based on the R package `softImpute` (Hastie and Mazumder, 2015).
- *mean.loglin*: Imputation by the mean for the missing values and estimate e with logistic regression on the mean imputed covariates and the two $\mu_{(w)}$ with two separate linear regressions.
- *mean.grf*: Procedure 1 with Option 2.

For the parametric $\hat{\tau}_{EM}$ we use the R package `misaem` (Jiang, 2019); for the nonparametric $\hat{\tau}_{MIA}$ we implemented our approach in the R package `GRF` (Athey, Tibshirani and Wager, 2019). We grow forests with missingness via the the MIA method; then, the estimator (9) is implemented in the command `average_treatment_effect`. Note that it is common to add the binary mask R to the initial or imputed data matrix X for estimation or prediction and it is admitted that this addition can sometimes improve the analysis and generally does not deteriorate the result. Hence, in this work we only report results obtained by adding R.

In all cases, we consider inference using the bootstrap (i.e., we bootstrap the original data and repeat the whole process).

[5]The code for reproducing the experiments presented in this work is available online at https://github.com/imkemayer/causal-inference-missing.

[6]For details on the implementation of this last method, see https://github.com/udellgroup/causal_mf_code.

	Confounders & Covariates		Missingness		Unconfoundedness		Models for (W, Y)	
	multivariate normal	general	M(C)AR	general	(1)	(5)	logistic-linear	non-param.
saem	✓	✗	✓	✗	✗	✓	✓	✗
mia.grf	✓	✓	✓	✓	✗	✓	✓	✓
mice	✓	✓	✓	✗	✓	✓	✓	(✗)
mf	✓	✗	✓	✗	✓ (on U)	✗	✓	(✗)
mean.loglin	✗	✗	✗	✗	✗	✗	✗	✗
mean.grf	✓	✓	✓	✓	✗	✓	✓	✓

TABLE 2

Methods and their assumptions on the underlying data generating process. (✓indicates cases that can be handled by a method, whereas ✗marks cases where a method is not applicable in theory; (✗) indicates cases without theoretical guarantees but with heuristic solutions.)

5.2. *Data generation.* We define different models for the generation of the confounders, covariates, missing values, treatment assignment and outcome.

5.2.1. *Confounders and covariates.*

Model 1: Multivariate normally distributed confounders. We generate normally distributed confounders $X_{i\cdot} = [X_{i1} \quad \dots \quad X_{ip}]^T \sim \mathcal{N}(\mathbf{1}, \Sigma)$, $i \in \{1, \dots, n\}$, for $p = 10$, where $\Sigma = I - 0.6 \times (I - 1)$, $\mathbf{X} = [X_{1\cdot} \dots X_{p\cdot}]^T \in \mathbb{R}^{n \times p}$.

Model 2: Latent classes model. We consider a Gaussian mixture model, i.e., we first generate class labels C from a multinomial distribution with three categories. Then the confounders of observation i, $X_{i\cdot}$, are sampled from the corresponding class distribution, i.e., $X_{i\cdot} \sim \mathcal{N}(\mu(c_i), \Sigma(c_i)) \,|\, C_i = c_i$.

Treatment and outcome are defined using the logistic-linear model in the following way: we define $\text{logit}(e^*(X^*_{i\cdot})) = (\alpha(C_i))^T X^*_{i\cdot}$. This allows to add an additional interaction between treatment and the latent class. Analogously, the outcome is defined as $Y_i \sim \mathcal{N}((\beta(C_i))^T X^*_{i\cdot} + \tau W_i, \sigma^2)$.

Model 3: Hierarchical data-generating model. An alternative to defining a Gaussian mixture model, is to use a simplified shallow version of a *deep latent variable model* (DLVM, Kingma and Welling (2014)): the codes C are sampled from a normal distribution $\mathcal{N}_d(0, 1)$. Covariates X_i are then sampled from $\mathcal{N}_p(\mu(c), \Sigma(c)) \,|\, C_i = c$, where

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

and the weights in $V \in \mathbb{R}^{p\times 5}$ and $W \in \mathbb{R}^{5\times d}$ are respectively sampled from a standard normal and a uniform distribution (and similarly for the offsets a and b). We fix $d = 3$.

Model 4: Low rank matrix factorization. We adapt the simulation framework from Kallus, Mao and Udell (2018) by generating $U_{i\cdot} = [U_{i1} \quad \dots \quad U_{id}]^T \sim \mathcal{N}(0, I_d)$ and defining $X = UV^T$ for some fixed matrix $V \in \mathbb{R}^{p\times d}$, with $d = 3$.

5.2.2. *Missing values.* We generate missing values either under MCAR (i.e., $\mathbb{P}(R_{ij} = 1) = 1 - \mathcal{B}(\eta)$ such that on average we have ηnp missing values) or as informative missing values (missing values in $X_{\cdot,1:5}$ are generated depending on the quantiles of $X_{\cdot,1:5}$ such that there are about $\eta np/2$ missing values). In the results presented here we fix $\eta = 0.3$.

5.2.3. *Treatment assignment and outcome.* For models 1,3 and 4, treatment assignment and outcome are defined under either of the unconfoundedness assumptions.

Unconfoundedness despite missingness. Let $X^* = R \odot X + (1 - R) \odot \mathbf{0}$. We define $\text{logit}(e^*(X^*_{i\cdot})) = \alpha_0 + \alpha^T X^*_{i\cdot}$. Analogously, the outcome is defined as $Y_i \sim \mathcal{N}(\beta_0 + \beta^T X^*_{i\cdot} + \tau W_i, \sigma^2)$.

Standard unconfoundedness. We define $\text{logit}(e(X_{i\cdot})) = \alpha_0 + \alpha^T X_{i\cdot}$. Analogously, the outcome is defined as $Y_i \sim \mathcal{N}(\beta_0 + \beta^T X_{i\cdot} + \tau W_i, \sigma^2)$.

For model 2, treatment assignment and outcome are defined under unconfoundedness on the latent factors U as follows: $\text{logit}(e(U_{i\cdot})) = \alpha_0 + \alpha^T U_{i\cdot}$. Analogously, the outcome is defined as $Y_i \sim \mathcal{N}(\beta_0 + \beta^T U_{i\cdot} + \tau W_i, \sigma^2)$

We refer to the Supplementary material for details on how to simulate treatment and outcome under assumption (5) (or rather (1) and (7)).

5.3. *Results.* We report the estimations for a fixed average treatment effect using the previously described estimation methods. All figures in this study are generated from 100 simulations for sample sizes $n \in \{100, 500, 1000, 5000\}$, we fix the proportion of missing values at 30% throughout all experiments; and the true treatment effect τ is reported as black solid line. The *standard unconfoundedness* setting corresponds to assumption (1), while *unconfoundedness despite missingness* corresponds to (5).

5.4. *Take-home message from the simulation study.* The results from this first simulation study can be summarized in several general observations:

- Augmented IPW outperform their IPW equivalents throughout all scenarios (both in terms of variability and of bias), this behavior is analogous to the behavior in the well understood complete data setting.

- All methods perform well if their assumptions on the underlying data generating process are met (see Table 2).
- For multiple imputation (*mice*) there is a small remaining bias, even for large sample sizes. In some cases, when the assumptions for this method are met, based on the theorem from Seaman and White (2014) on multiple imputation with $M = \infty$ imputations, it is expected that an increase of the number of imputations should decrease this remaining bias in these cases.
- The tree-based estimation, using the MIA criterion (*mia.grf*) or mean imputation (*mean.grf*), generally performs at least as good as multiple imputation but yields unbiased results if "unconfoundedness despite missingness" (5) holds.
- Mean imputation and concatenation of the imputed data with the mask, followed by logistic regression for W and linear regression for Y (*mean.loglin*) leads to unbiased estimates, provided that (5) holds, in many scenario even when the models are misspecified, however this is true only when adding the mask R to the regression models. Otherwise this approach is biased as soon as (5) is violated, and in this case it is outperformed by competing methods.
- The EM-based estimator (*saem*) performs well under correct specification (multivariate Gaussian confounders, logistic treatment assignment, linear outcome, M(C)AR missing data mechanism, (5) satisfied) and adding the mask to the initial data matrix yields unbiased estimates even if the missing data mechanism is not ignorable. It fails however in the cases where the data is not i.i.d. Gaussian.

For a more detailed discussion of the simulation results, we refer to the Supplementary material.

6. Application on observational critical care management data. As announced in the introduction we apply our methods to clinical data from a French observational database on major trauma patients. The medical question we aim to answer is whether administrating the drug TA has an effect on in-ICU mortality for patients with traumatic brain injury.

6.1. *Data and causal DAG.* Out of the 20,000 currently available patient records we consider a subset of 7,240 observations that have been validated by the medical expert team after a first pre-treatment of a subset of 7,495 observations available at the beginning of this study. The pre-treatment consisted in identifying outliers clearly due to erroneous inputs and recoding missing values that are not really missing (for instance the variable informing

previous pregnancies is evidently consistently missing, or ideally set to false, for male patients, etc.)[7]. Out of these 7,240 patients, 3,168 are identified as having a traumatic brain injury (defined by the medical expert team as either the presence of a brain lesion visible on the first computed tomography (CT) scan—which is generally taken within the first three hours after the accident—or as a head AIS score[8] greater or equal 2).

The treatment of interest, TA, is an antifibrinolytic agent limiting excessive bleeding and it is currently used in patients suspected of developing an hemorrhagic shock, a state in which the body is no longer able to provide vital organs with sufficient quantities of dioxygen to sustain them. The average cost of a dose of TA lies below 10€ and the drug is generally available immediately after the arrival of the medical first responders team at the place of the accident. It is now recommended to administer this drug to patients at risk of developing an hemorrhagic shock.

In order to clarify the previously raised causal question given the data, we first establish a causal graph in order to summarize the a priori on existing confounding and to highlight the causal question, as suggested, for instance, by Lederer et al. (2019); Blake et al. (2019). The causal graph in Figure 7 is the result of a two-step Delphi procedure in which six anesthetists and resuscitators specialized in critical care first selected covariates related to either treatment or outcome or both and second classified these covariates into confounders and predictors of only treatment or outcome. The absence of an exact timestamp for the drug administration is compensated by the fact that it is always given within the first three hours from the accident and that the treatment does not have an immediate effect on variables such as blood pressure, hemoglobin level or the Glasgow Coma Scale (GCS) which are measured at various moments within the first three hours.

From this graph it becomes clear as well that a method that incorporates a model of the outcome as a function of the identified potential predictors (red and blue vertices in the graph) might achieve more precise results than a method that uses the observed outcome directly. The large number of predictors of the outcome is due both to the medical complexity of traumatic brain injury and to the ambiguous treatment target: the assignment is made in the context of hemorrhagic shock but recently there is some evidence that

[7]The code for pre-treatment and for estimating the treatment effect on this data are available at https://github.com/imkemayer/causal-inference-missing.

[8]The head Abbreviated Injury Score indicates, on a scale from one to six, the severity of the most severe observed brain lesion. This score is defined in the context of the Abbreviated Injury Scale proposed by the American Association for Automotive Medicine. See the Supplementary material or https://www.aaam.org/abbreviated-injury-scale-ais/ for more information.

there might also be a beneficial effect in the context of traumatic brain injury (Hijazi et al., 2015).

6.2. *Results.* First, we recall the estimand we aim at estimating in this context: we are interested in the effect of the treatment on mortality among traumatic brain injury patients (indicated by the binary variable X_{TBI}), more formally:

$$\tau_{TBI} = \mathbb{E}[Y(1) - Y(0)|X_{TBI} = 1] \tag{10}$$

When adjusting for confounding using the identified confounders (pink nodes on the graph in Figure 7), using additional predictors for the outcome model (blue nodes on the graph in Figure 7), we obtain the following estimations in Figure 8 of the direct causal effect of TA on in-ICU mortality among traumatic brain injury patients.

Unlike the simulations of the previous paragraph, the real-world medical data is more complicated and some concessions have to be made to apply the previously discussed method. For instance, due to an important number of outliers in the variable *Medcare.time.ph* that are related with inconsistent units of the recorded values and with patient transfers from one hospital to another, we chose to drop this variable in our analyses since, according to the practitioners, its predictive power does not outweigh the potential issues related to inconsistent recording of this variable.

Note that apart from the issue with the variable *Medcare.time.ph*, the estimation via random forest and MIA does not require substantial preprocessing of the data and is therefore straightforward, once the MIA recoding and the random forest are implemented. A remaining issue might consist in the overlap assumption which is generally difficult to assess in most medical applications and which might be slightly violated due in part to the heterogeneity of patient profiles. A solution to handle such weak overlap is the use of overlap weights (Li, Morgan and Zaslavsky, 2018) and we give the results using this alternative to inverse propensity weights in the Supplementary material.

Here, we only consider three pairs of methods: *mia.grf* and *mice*. We drop *mean.grf* since it performed similarly to *mia.grf* in the simulations. We also do not test *saem* and *mf* since currently both these methods have not been derived for heterogeneous data. A first observation on the results reported in

[9]Values on the x-axis are multiplied by 100 for better readability. The results can be read as the difference in percentage points between mortality rate in the treated and control groups.

Figure 8 is the concordance of the different estimators: none of the AIPW-type estimation strategies allows to reject the null hypothesis of no treatment effect. As discussed in Section 3.3, it can be argued which family of methods has more plausible underlying assumptions on the Traumabase data, but in our opinion the *unconfoundedness despite missingness*—and therefore the *mia.grf* estimations—are most suited for our specific application.

We notice a large difference between the IPW and the AIPW estimations. The AIPW estimations seem more reasonable for two reasons: first, the medical experts have noticed positive effects of TA for their TBI patients in practice and a previous clinical trial, focussing on a slightly different patient group, has also exhibited a certain benefit from the drug for patients with TBI; second, for the AIPW estimators, we incorporate much more available information, namely all identified features that are strongly related to the outcome Y according to the expert panel (blue nodes on Figure 7). Finally, all compared methods have similar empirical variances as can be observed on the reported bootstrap confidence intervals in Figure 8. Finally, adding the mask to the data matrix does not lead to major changes in the estimations, therefore we only report results obtained when including the mask.

7. Discussion and perspectives.

7.1. *Two families of treatment effect estimators handling missing attributes.* We have stressed the dyadic classification of previously exposed methods that allow treatment effect estimation with missing attributes, both in theory and in practice. The class of methods that relies on assumptions about the missingness mechanisms for treatment effect identifiability is currently often used, in combination with IPW-type estimators. We have also proposed an AIPW formulation for the most popular method from the first class, namely multiple imputation. However, methods of this first class have limited applicability in practice, most importantly they exclude informative missing data; this is a drawback of all developed methods in this class. The second class, relying on the generalized propensity score and a different unconfoundedness assumption, can handle arbitrary missingness mechanisms, in particular the case where MAR does not hold, but to the best of our knowledge, implementable and versatile methods in this class have not been proposed so far.

In practice, if one can exclude smooth regression functions for the treatment assignment and the outcome model, such as logistic and linear models, and if the "unconfoundedness despite missingness" assumption is likely to hold—for more details on this, we refer to Blake et al. (2019)—we advocate our tree-based estimator $\hat{\tau}_{MIA}$ in its AIPW-form and its mean-imputation

variant. If one is willing to make stronger (parametric) assumptions about the structure of X and its relationship with W and Y, then our second estimator $\hat{\tau}_{EM}$ can also be considered as an alternative.

7.2. *Heterogeneous treatment effects and policy learning.* Instead of estimating the average treatment effect τ, one could be interested in the conditional average treatment effect function, defined as $\tau(x) = \mathbb{E}[Y(1) - Y(0) \,|\, X = x]$, for several reasons. For instance one might be interested in estimating how treatment effects vary across sub-populations, or assessing whether there is heterogeneity in the population w.r.t. a given treatment. Such questions anticipate problems of learning decision rules that exploit treatment effect heterogeneity (Wager and Athey, 2018).

In light of our medical application, heterogeneous treatment effect estimation is of particular interest because of the known existing heterogeneity among traumatic brain injury patients in terms of clinical presentation, pathophysiology and outcome. It is even more relevant since to this date there is no general classification of patients with traumatic brain injury. Hence a causal inference approach allowing classification w.r.t. treatment heterogeneity for any given treatment is of interest for practitioners in critical care management.

7.3. *Further identification strategies.* Although the two lines of approaches studied here for identification of average treatment effects with missing attributes are the most prevalent in applied work, they are far from exhaustive. For example, Yang, Wang and Ding (2017) consider a setting with outcome-independent missingness, $Y_i \perp\!\!\!\perp R_i \,|\, \{X_i, W_i\}$, and find that τ can be identified via a set of integral equations. We expect the area of methods development for causal inference with missing attributes to be a fruitful research area for years to come.

Acknowledgement. We thank Jean-Pierre Nadal for fruitful discussion, Helen Blake and Julie Tibshirani for their suggestions for the for the simulation study, and the Delphi expert committee for the medical insight and advice on traumatic brain injury and hemorrhagic shock. We acknowledge funding from the EHESS PhD fellowship.

SUPPLEMENTARY MATERIAL

Supplementary material: Further simulation results and details on the Traumabase
(https://imkemayer.com/papers/2019-10-23_DR-TreatmentEffect-WithMissingAttributes_supp.pdf).
In this material we show additional simulation results, including different

simulation scenarios and estimators. Furthermore we provide a glossary for the Traumabase variables and an additional analysis on this data set.

References.

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica* **84** 781–807.

Athey, S., Imbens, G. W. and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 597–623.

Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized random forests. *The Annals of Statistics* **47** 1148–1178.

Athey, S. and Wager, S. (2017). Efficient policy learning. *arXiv preprint arXiv:1702.02896*.

Blake, H. A., Leyrat, C., Mansfield, K., Seaman, S., Tomlinson, L., Carpenter, J. and Williamson, E. (2019). Propensity scores using missingness pattern information: a practical guide. *arXiv preprint*.

Breiman, L. (2001). Random forests. *Machine learning* **45** 5–32.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE* **98** 925–936.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21** C1–C68.

D'Agostino, R. B. Jr and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* **95** 749–759.

Dalkey, N. and Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management science* **9** 458–467.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39** 1–22.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* **189** 1–23.

Hastie, T. and Mazumder, R. (2015). softImpute: Matrix Completion via Iterative Soft-Thresholded SVD R package version 1.4.

Hay, S. I., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulkader, R. S., Abdulle, A. M., Abebo, T. A., Abera, S. F. et al. (2017). Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet* **390** 1260–1344.

Hijazi, N., Fanne, R. A., Abramovitch, R., Yarovoi, S., Higazi, M., Abdeen, S., Basheer, M., Maraga, E., Cines, D. B. and Higazi, A. A.-R. (2015). Endogenous plasminogen activators mediate progressive intracerebral hemorrhage after traumatic brain injury in mice. *Blood* **125** 2558–2567.

Hirano, K., Imbens, G. W. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jiang, W. (2019). misaem: Logistic Regression with Missing Covariates R package version 0.9.1.

Jiang, W., Josse, J. and Lavielle, M. (2018). Logistic Regression with Missing Covariates–Parameter Estimation, Model Selection and Prediction. *arXiv preprint*.

Jones, J. and Hunter, D. (1995). Consensus methods for medical and health services research. *BMJ: British Medical Journal* **311** 376.

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.

Kingma, D. P. and Welling, M. (2014). Stochastic gradient VB and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica* **86** 591–616.

Lederer, D. J., Bell, S. C., Branson, R. D., Chalmers, J. D., Marshall, R., Maslove, D. M., Ost, D. E., Punjabi, N. M., Schatz, M., Smyth, A. R. et al. (2019). Control of confounding and reporting of results in causal inference studies. Guidance for authors from editors of respiratory, sleep, and critical care journals. *Annals of the American Thoracic Society* **16** 22–28.

Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., Resche-Rigon, M., Carpenter, J. R. and Williamson, E. J. (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical methods in medical research* **28** 3–19.

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.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley.

Luedtke, A. R. and Van Der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics* **44** 713.

Mattei, A. (2009). Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing. *Statistical Methods and Applications* **18** 257–273.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–688.

Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine* **28** 1402–1414.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical Report, Center for Statistics and the Social Sciences, University of Washington.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association* **89** 846–866.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87** 113–124.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79** 516–524.

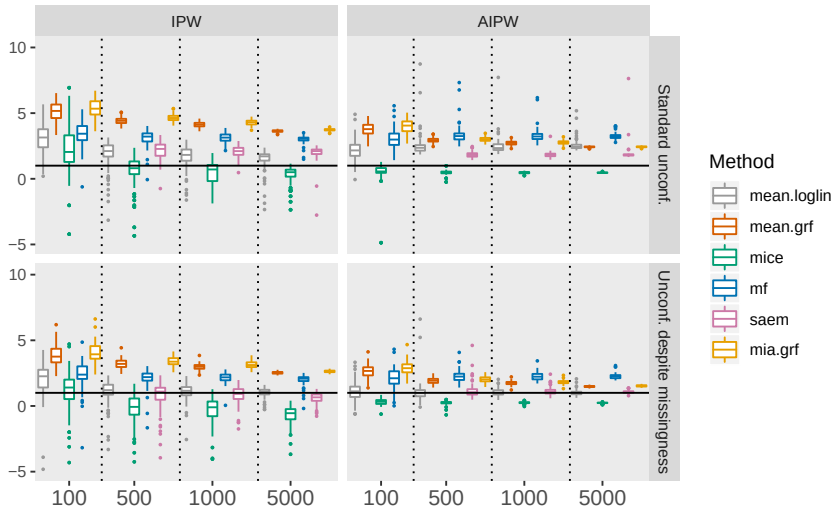
Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.
Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association* **1** 20–34. American Statistical Association.
Rubin, D. B. (1987). *Multlipe Imputation for Nonresponse in Surveys*. Wiley, Hoboken, NJ, USA.
Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* **81**. John Wiley & Sons.
Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data. CRC Monographs on Statistics & Applied Probability*. Chapman and Hall/CRC, Boca Raton, FL, USA.
Seaman, S. and White, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods* **43** 3499–3515.
Shakur, H., Roberts, I., Bautista, R., Caballero, J., Coats, T., Dewan, Y., El-Sayed, H., Gogichaishvili, T., Gupta, S., Herrera, J. et al. (2010). CRASH-2 trial collaborators. Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): a randomised, placebo-controlled trial. *Lancet* **376** 23–32.
R Core Team (2018). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.
Textor, J., Hardt, J. and Knüppel, S. (2011). DAGitty: a graphical tool for analyzing causal diagrams. *Epidemiology* **22** 745.
Twala, B., Jones, M. and Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters* **29** 950–956.
van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, Boca Raton, FL.
van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45** 1-67.
Van der Laan, M. J. and Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.
Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113** 1228–1242.
Yang, S., Wang, L. and Ding, P. (2017). Causal inference with confounders missing not at random. *arXiv preprint arXiv:1702.03951*.
Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107** 1106–1118.
Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association* **107** 1360–1371.

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

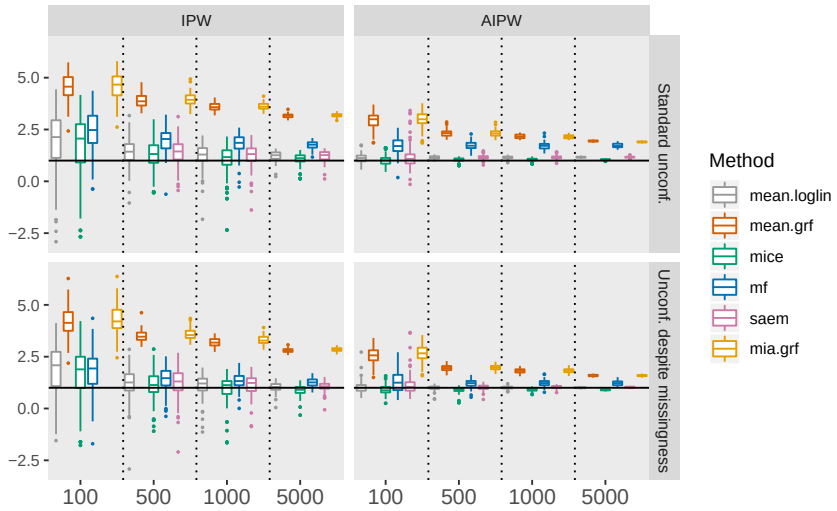
J. Josse
Centre de Mathématiques Appliquées
École Polytechnique
91128 Palaiseau Cedex, France
E-mail: julie.josse@polytechnique.edu

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

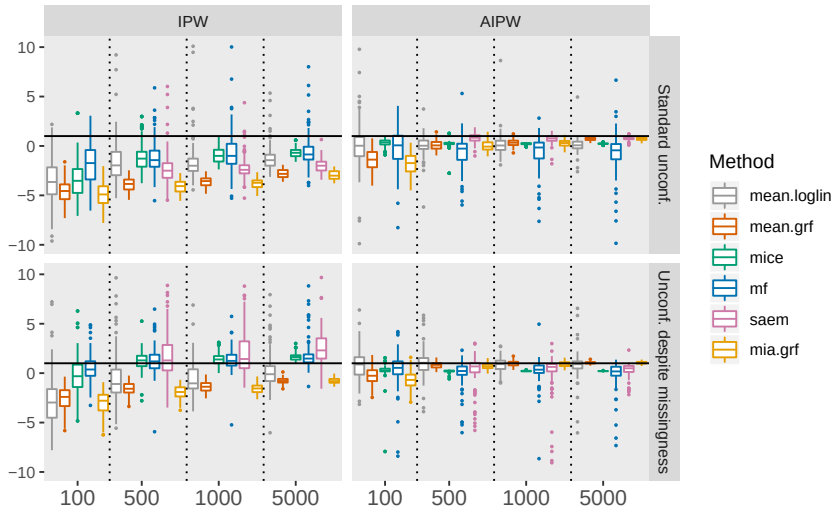


(a) MCAR (with 30% missing values in $X_{\cdot,1:10}$)

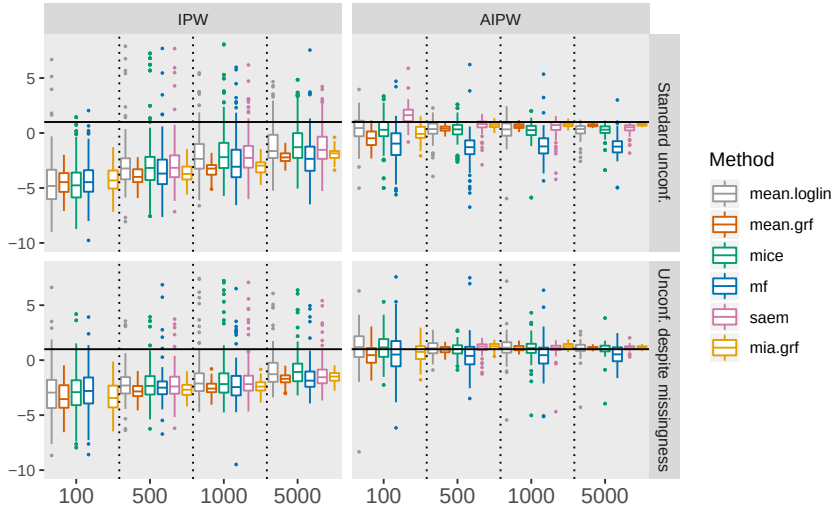


(b) Informative missing values (with 30% missing values in $X_{\cdot,1:5}$)

Fig 3: Model 1. IPW and AIPW estimations across simulation designs described in Section 5.2. We report results for all combinations of $n \in \{100, 500, 1000, 5000\}$, missing values mechanism $\in \{MCAR, general\}$ and unconfoundedness $\in \{unconfoundedness\,despite\,missingness,\ standard\,unconfoundedness\}$. Results are displayed for 100 runs of every setting.

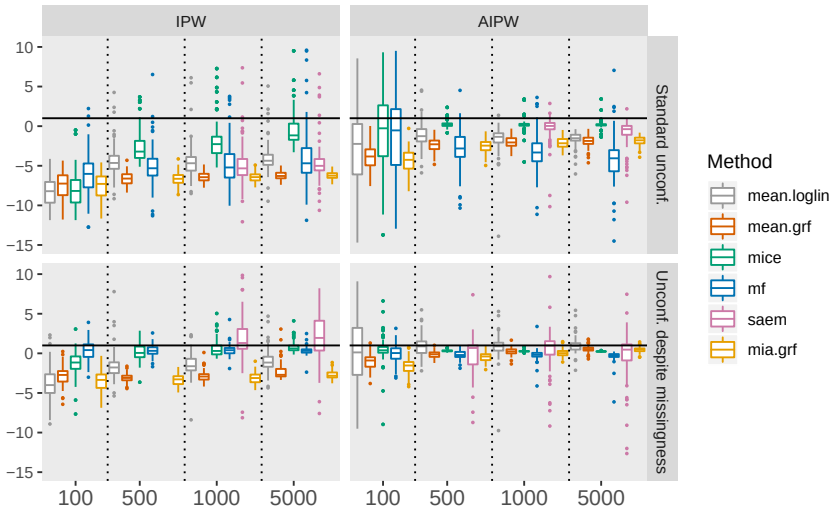


(a) MCAR (with 30% missing values in $X_{\cdot,1:10}$)

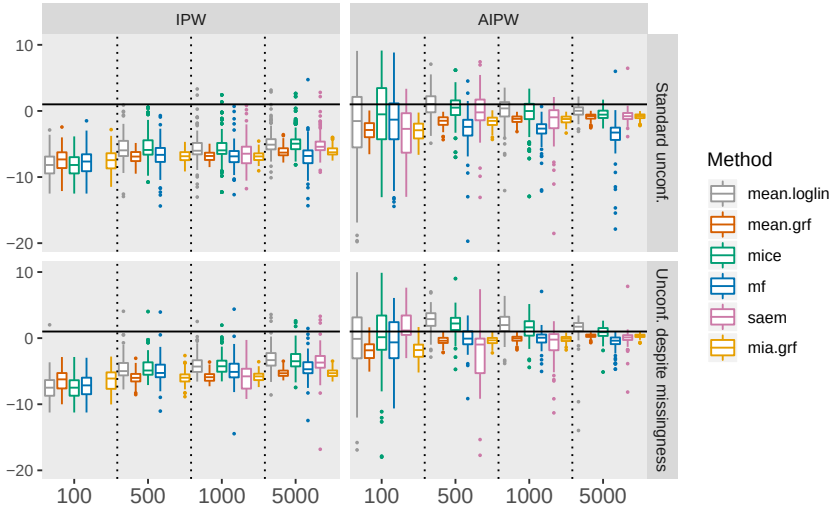


(b) Informative missing values (with 30% missing values in $X_{\cdot,1:5}$)

Fig 4: Model 2. IPW and AIPW estimations across simulation designs described in Section 5.2. We report results for all combinations of $n \in \{100, 500, 1000, 5000\}$, missing values mechanism $\in \{MCAR, general\}$ and unconfoundedness $\in \{unconfoundedness\,despite\,missingness,\ standard\,unconfoundedness\}$. Results are displayed for 100 runs of every setting.

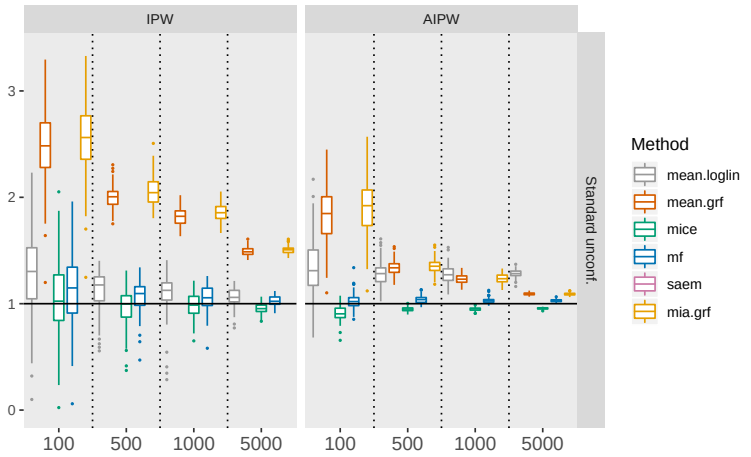


(a) MCAR (with 30% missing values in $X_{\cdot,1:10}$)

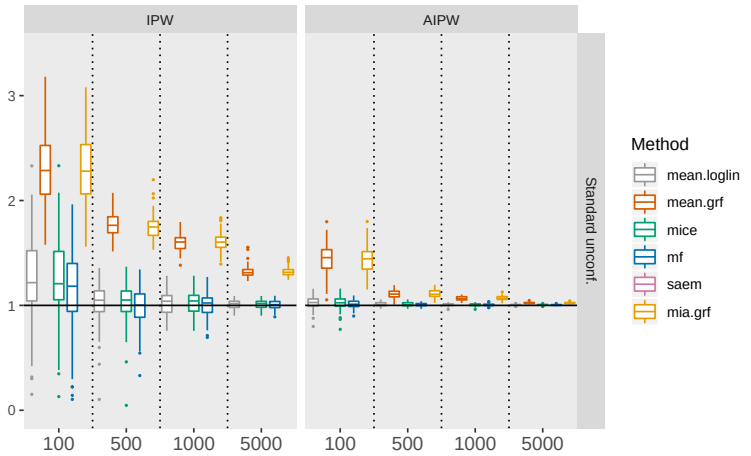


(b) Informative missing values (with 30% missing values in $X_{\cdot,1:5}$)

Fig 5: Model 3. IPW and AIPW estimations across simulation designs described in Section 5.2. We report results for all combinations of $n \in \{100, 500, 1000, 5000\}$, missing values mechanism $\in \{MCAR, general\}$ and unconfoundedness $\in \{unconfoundedness\,despite\,missingness,\ standard\,unconfoundedness\}$. Results are displayed for 100 runs of every setting.



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



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

Fig 6: Model 4. IPW and AIPW estimations across simulation designs described in Section 5.2. We report results for all combinations of $n \in \{100, 500, 1000, 5000\}$ and missing values mechanism $\in \{MCAR, general\}$. Results are displayed for 100 runs of every setting.

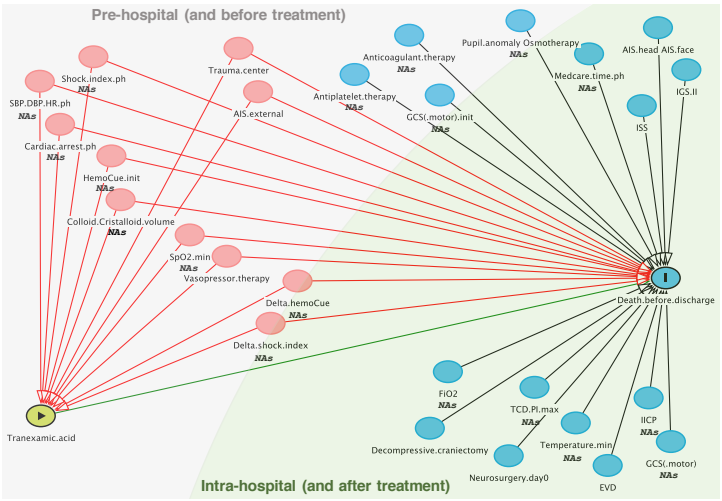


Fig 7: Causal graph representing treatment, outcome, confounders and other predictors of outcome (Figure generated using DAGitty (Textor, Hardt and Knüppel, 2011); `NAs` indicates variables that still have missing values after pre-treatment).

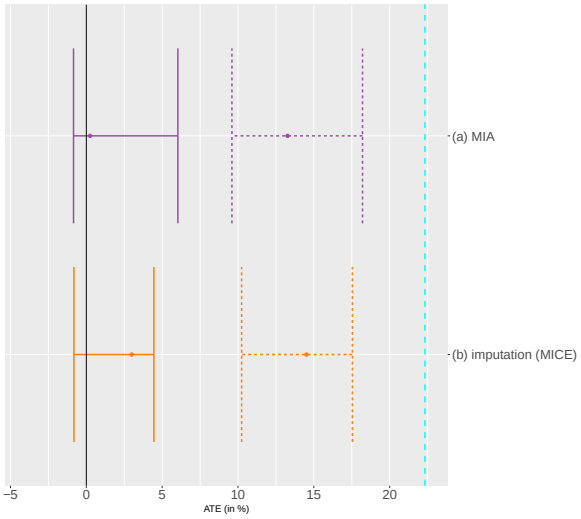


Fig 8: ATE estimations on Traumabase data (solid: doubly robust estimates; dotted: IPW estimates; dashed vertical line: without adjustment; x-axis: $\hat{\tau}$ and bootstrap confidence intervals[9]).