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[‡‡]*

The problem of missing values in causal inference has long been ignored and only recently gained some attention due to the non-negligible impacts in terms of bias induced by complete case analyses and misspecified imputation models. We discuss different conditions under which causal inference can be possible despite missing attributes, we review existing solutions and propose a new approach to handle missing attributes in treatment effect estimation. We propose two average treatment effect (ATE) estimators, each in an inverse propensity weighting and a doubly robust form, which directly account for the missing values and show their consistency. The first is built on logistic-linear specification and observed likelihood, appropriate for data *missing at random*, while the second uses semi-parametric estimation based on random forests with the great advantage of handling data *missing not at random*. We compare these two estimators to different methods available in an extensive simulation study. We apply the estimators on a large prospective database counting about over 20,000 severely traumatized patients in France to study the effect on mortality of tranexamic acid administration among patients with traumatic brain injury in the context of critical care management.

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

[¶]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: missing data, causal inference, potential outcomes, observational data, propensity score estimation, causal forest, 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).

randomized placebo-controlled study (Shakur et al., 2010). Our interest in 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 39 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 covari-

[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).

ates to incorporate in our study by soliciting feedback from a number experts 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 semiparametrically 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

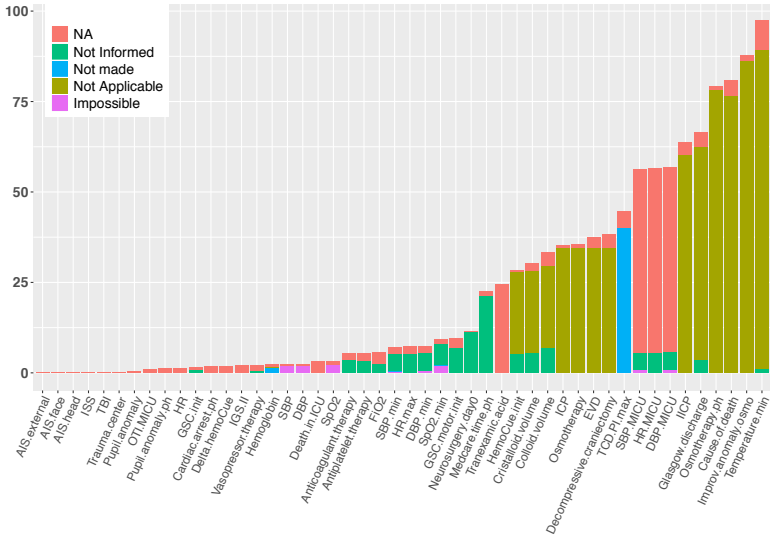


Fig 1: Percentage of missing values for a subset of variables relevant for traumatic brain injury. Different encodings of missing values: *NA* (not available), *not informed*, *not made*, *not applicable*, *impossible*.

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 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. Thus, it appears that the non-doubly-robust baselines did not succeed in using covariates to eliminate the confounding bias described 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),

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

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 at slower non-parametric rates (Farrell, 2015). Methods based on inverse-weighting as in (3) can also sometimes achieve this property, but these results are generally more fragile and require considerably stronger regularity conditions than corresponding AIPW results (Hirano, Imbens and Ridder, 2003). In particular, AIPW methods can achieve $\sqrt{n}$-consistency under considerable generality even when $\hat{\mu}_{(w)}(\cdot)$ and $\hat{e}(\cdot)$ are estimated using generic machine learning methods, provided both estimators achieve reasonable rates of convergence in mean-squared error and 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).

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" $R_i \in \{1, \text{NA}\}^p$ such that the analyst observes $X_i^* \triangleq R_i \odot X_i \in \{\mathbb{R} \cup \text{NA}\}^p$. Here, $\odot$ denotes an element-wise product, such that $X_{ij}^* = X_{ij}$ if $R_{ij} = 1$ and $X_{ij}^* = \text{NA}$ if $R_{ij} = \text{NA}$.

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). However, as AIPW-form methods have been found to often out-perform IPW-form

estimators with complete data, it is natural to ask whether AIPW-form estimators are also applicable with missing data. The main contributions of this paper consist in (1) a dyadic classification of possible approaches to treatment effect estimation with missing attributes, (2) in the proposal of two new estimators in one 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 other class and (4) an extensive comparison of these estimators, with a focus on the AIPW vs. baseline performance. 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 show (5) is to pair (1) together with one of the two assumptions below (Blake et al., 2019; Mattei, 2009)

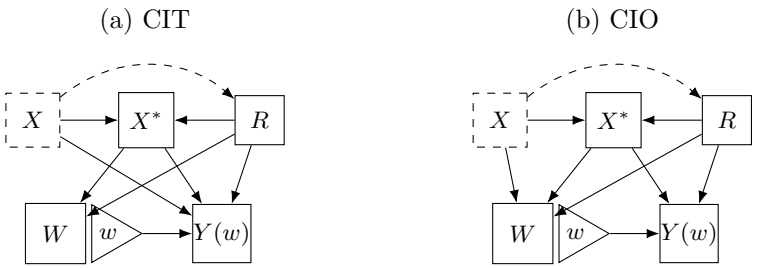
$$\begin{cases} \text{CIT:} & W_i \perp\!\!\!\perp X_i \,|\, X_i^*, R_i \\ \textbf{or} & \\ \text{CIO:} & Y_i(w) \perp\!\!\!\perp X_i \,|\, X_i^*, R_i \quad \text{for } w \in \{0,1\}, \end{cases} \tag{7}$$

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).

Note however that (6) requires fitting many models (one model per pattern). We demonstrate in this work, that (6) can be efficiently estimated using random forests incorporating the missingness information.

3.2. *Missing values mechanisms.* Another choice is to make assumptions about the missingness mechanism R_i. The most popular approach is to take

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



the missingness mechanism to be random (MAR) (Little and Rubin, 2002; Rubin, 1976), i.e., for each possible mask $r \in \{1, \text{NA}\}^p$,

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

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.

Finally, the third and most challenging case in the missing values taxonomy introduced by Rubin (1976) consists in considering the missingness mechanism to be non random (MNAR), more formally this includes all cases that are not MAR; hence when referring to the "general" or "informative missingness" case in the remainder of this article we consider all cases that are not MAR. The general case has not been addressed yet in the context of causal inference with the sole exception of a recent proposal of Yang, Wang and Ding (2017). They consider a setting where $Y_i \perp\!\!\!\perp R_i \,|\, \{X_i, W_i\}$ – which, in particular, allows to drop the MAR assumption – and find that τ can be identified via a set of integral equation[2].

3.3. *Discussion: The Traumabase study.* Before we turn to the main contribution of this work – the proposal of two doubly robust estimators to estimate the treatment effect and an extensive empirical study comparing these estimators to previously proposed methods –, we briefly return to our application, the study of the effect of TA on in-ICU mortality based on the Traumabase data set. 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[3] 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

[2]This work covers many of the approaches to deal with missing attributes in treatment effect estimation that are currently most popular in applications.

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

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

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.

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 estimation, using the EM algorithm that allows to do valid inference on the observed values (Dempster, Laird and Rubin, 1977). A limitation of this approach, as opposed to the previous one, is the additional 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.

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). A shortcoming of this approach however is the need to include the outcome Y in the imputation model. Intuitively, this is necessary as it allows to estimate the joint data distribution over (X, W, Y) and therefore also to estimate τ which is a functional of this joint distribution.

Another recent solution is based on matrix factorization (Kallus, Mao and Udell, 2018) as outlined in Section 3.2. Note that, unlike with multiple imputation, we only impute each datapoint once and consistency guarantees are only given under MCAR.

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

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, the (generalized) propensity score is defined as a logistic function and $\mu_{(0)}$ and $\mu_{(1)}$ are linear functions of the covariates (or of the confounders in Section 5.1).

We compare our approaches relying on assumption (5), $\hat{\tau}_{EM}$ and $\hat{\tau}_{MIA}$, denoted *saem* and *mia.grf* in the experiments, to the following methods:

- *mice*: Multiple imputation by conditional equations (MICE) (van Buuren, 2018) that imputes using conditional models, followed by logistic and linear regressions for the propensity and the outcome models respectively, the final estimates $\{\hat{\tau}^{(m)}\}_{m\in\{1,\dots,M\}}$, where $M = 10$, are aggregated to obtain a final estimate $\hat{\tau}_{MI}$. More specifically,

$$\hat{\tau}_{MI} = \frac{1}{M}\sum_{m=1}^{M} \hat{\tau}^{(m)},$$

  where $\hat{\tau}_m$ is obtained by applying either (3) or (4) to the m-th imputed data matrix, to obtain the baseline and the doubly robust estimator respectively.
- *mf*: Estimation of latent confounders (Kallus, Mao and Udell, 2018) using low-rank matrix factorization, again followed by logistic and linear regressions on the estimated latent confounders. Therefore we define $\hat{\tau}_{MF}$ via (4) (or (3) for the baseline) by replacing the covariates X by the estimated latent factors $\hat{U}$ as outlined in Section 3.2.
- *mean.loglin/mean.grf*: Imputation by the mean for the missing values and using either logistic-linear regression or random forests.

Finally, 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.

All simulations are implemented in R (R Core Team, 2018), 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 (the MIA approach being implemented by replacing each missing value by two "surrogates", $-\infty$ and $+\infty$); then, the estimator (9) is implemented in the command `average_treatment_effect`. For multiple imputation we use the R package `mice` (van Buuren and Groothuis-Oudshoorn, 2011) and default options. Finally for the matrix factorization approach we adapt the implementation[4] of Kallus, Mao and Udell (2018) based on the R package `softImpute` (Hastie and Mazumder, 2015).[5]

All figures in this section are generated from 100 simulations for sample sizes $n \in \{100, 500, 1000, 5000\}$, we fix the proportion of missing values at

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

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

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).

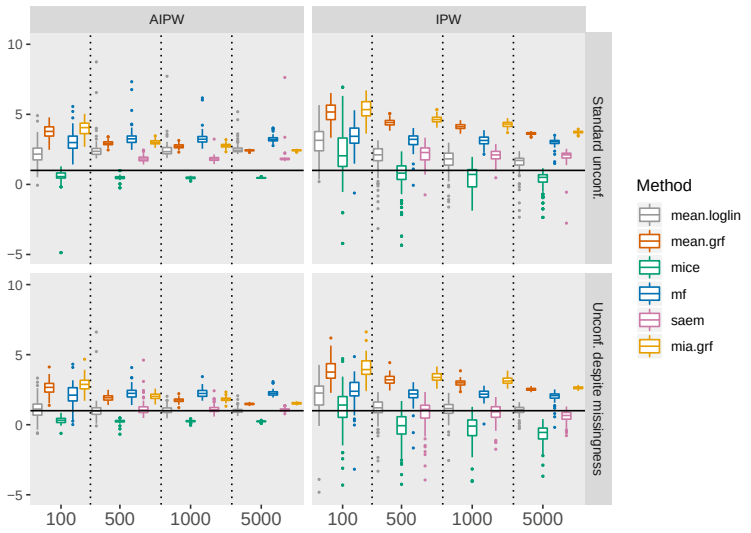
5.1. *Parametric double robustness.* We illustrate the importance of assumption (5) for the parametric AIPW estimator $\hat{\tau}_{EM}$ using the following setting: we generate normally distributed confounders $X_{i\cdot} = [X_{i1} \ \dots \ X_{ip}]^T \sim \mathcal{N}(\mathbf{1}, \Sigma)$, $i \in \{1, \dots, n\}$, for $p = 10$, where $\Sigma = I - 0.6 \times (I - 1)$, $\mathbf{X} = [X_{1\cdot} \ \dots \ X_{p\cdot}]^T \in \mathbb{R}^{n \times p}$. 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 $\eta = 0.3$. We refer to the Supplementary material for details on how to simulate treatment and outcome under assumption (5) (or rather (1) and (7)); in all cases, they are simulated under a logistic and linear regression model respectively.

Figure 3a shows that if the data is MCAR and satisfies (5), *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 3a shows as well that the *mice* works under both unconfoundedness assumptions as expected[6]. In particular, when only (1) holds and (7) is violated, then all methods but multiple imputation give biased results.

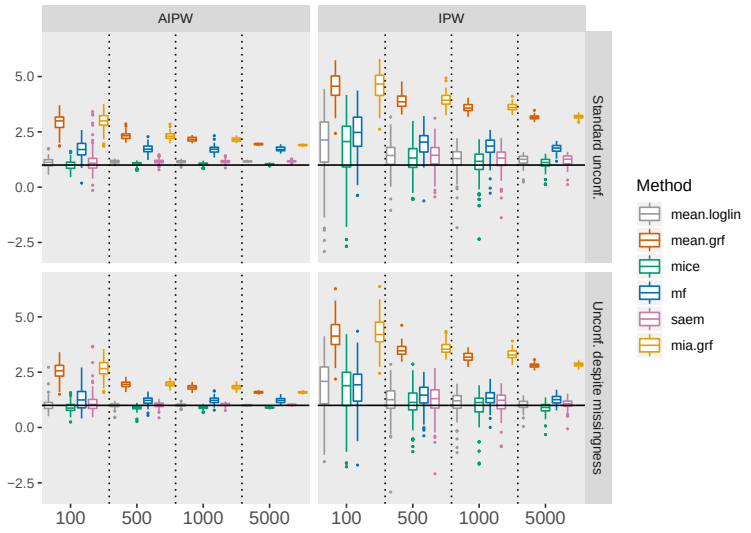
In the general missingness case, Figure 3b, we only expect *mia.grf* and *mean.grf* to perform well as explained in Section 4.1.1. However their convergence seems to be very slow which again can be explained with the strong parametric and smooth models we defined with the attributes X and that are hard 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

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

expected but not shown in Figure 3b.



(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: Estimated average treatment effect $\hat{\tau}$ for **strongly correlated attributes**.

Simulation results under latent confounders model. We reproduce the latent confounders case studied for instance in Kallus, Mao and Udell (2018) and briefly discussed in Section 3.2.

Indeed on Figure 4 we observe good performance of the estimator based on low-rank matrix factorization in the MCAR. 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* 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.

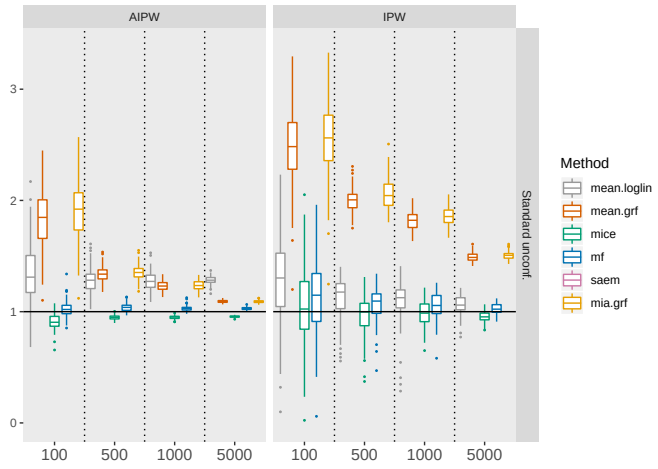
5.2. *Nonparametric double robustness.* Next we assess the performance of the second estimator, $\hat{\tau}_{MIA}$ and its sensitivity to the "unconfoundedness despite missingness" assumption (5). We consider two data generating models: a latent class model and a standard hierarchical data-generating model.

Latent class 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., $X_{i\cdot} \sim \mathcal{N}(\mu(c_i), \Sigma(c_i)) \,|\, C_i = c_i$. Missing values are generated either under MCAR or as informative missing values (as in the previous setting). Treatment and outcome are defined using the logistic-linear model but the following way: $\text{logit}(e(X_{i\cdot})) = (\alpha(C_i))^T X_{i\cdot}$. This allows to add an additional interaction between treatment and the latent class. Analogously, the outcome is defined as $Y_i \sim \mathcal{N}((\beta(C_i))^T X_{i\cdot} + \tau W_i, \sigma^2)$.

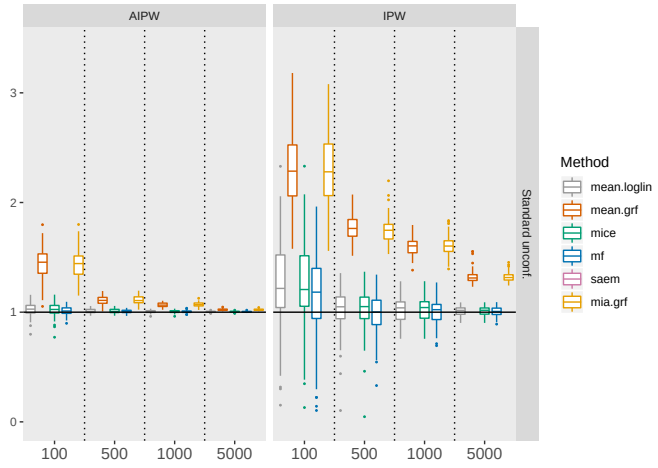
Figures 5 show that, as expected, if (5) is satisfied, our estimator *mia.grf* converges quickly to the true value τ while the other methods remain biased. With the exception of *mice*, all other methods fail if the "unconfoundedness despite missingness" assumption is violated, independently from the missingness mechanism. However *mia.grf* and *mean.grf* in AIPW-form seem to cope well even under the standard unconfoundedness (1).

Note that especially in the general case, the IPW-type estimators perform poorly: even for the largest sample size ($n = 5000$), the estimators are still far from the true τ.

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



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



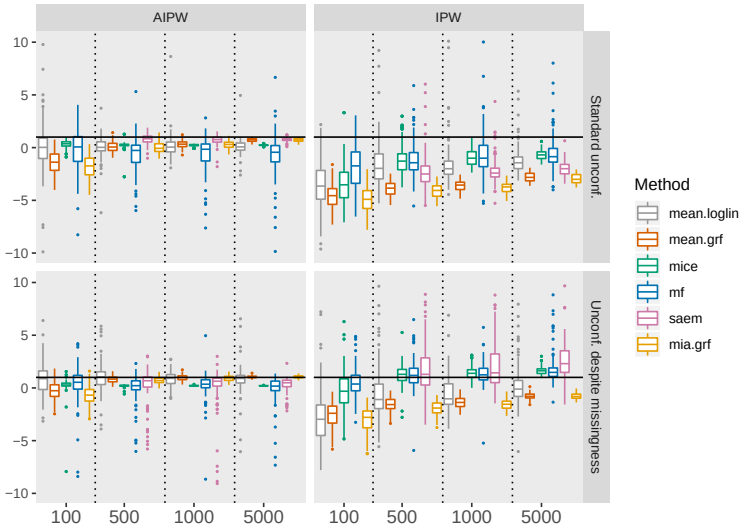
(b) Informative missing values (with 30% missing values in $X_{\cdot,1:5}$)

Fig 4: Estimated average treatment effect $\hat{\tau}$. **Latent confounders**.

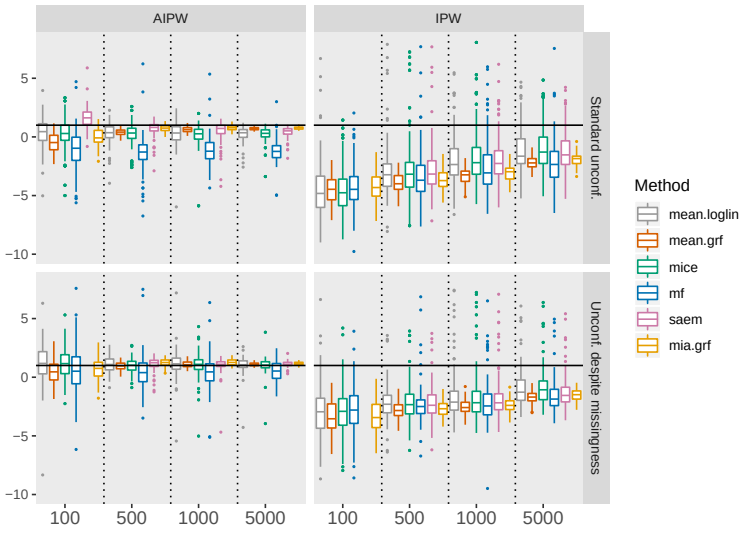
$\mathcal{N}_p(\mu(c), \Sigma(c)) \,|\, C_i = c$, where

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

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). In our simulations we fix $d = 3$. Missing values are generated as



(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: Estimated average treatment effect $\hat{\tau}$. **Latent classes model for confounders**.

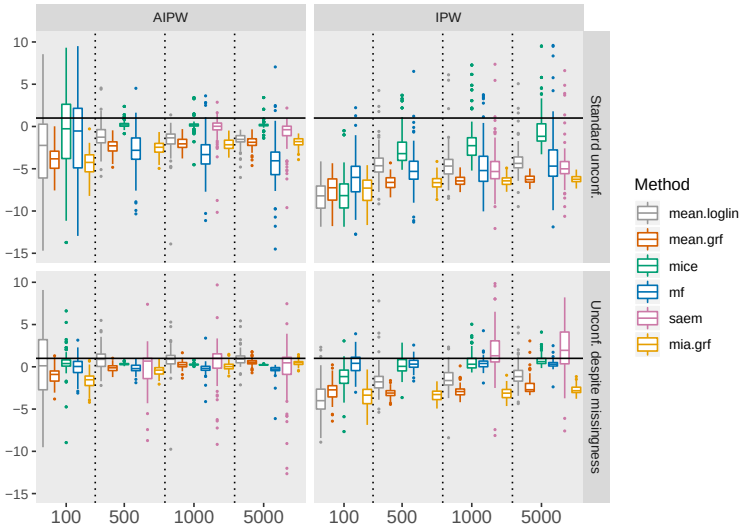
for the latent classes model. Treatment and outcome are again defined via the logistic-linear model in the following way: $\text{logit}(e(X_{i\cdot})) = \alpha^T X_{i\cdot}$ and $Y_i \sim \mathcal{N}((\beta^T X_{i\cdot} + \tau W_i, \sigma^2)$, i.e., without an additional class interaction in the treatment or outcome model.

Figures 6a and 6b show that *mia.grf* converges to the true value τ in all cases but rather slowly, provided assumption (5) 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 4. 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 $\mu(C)$).

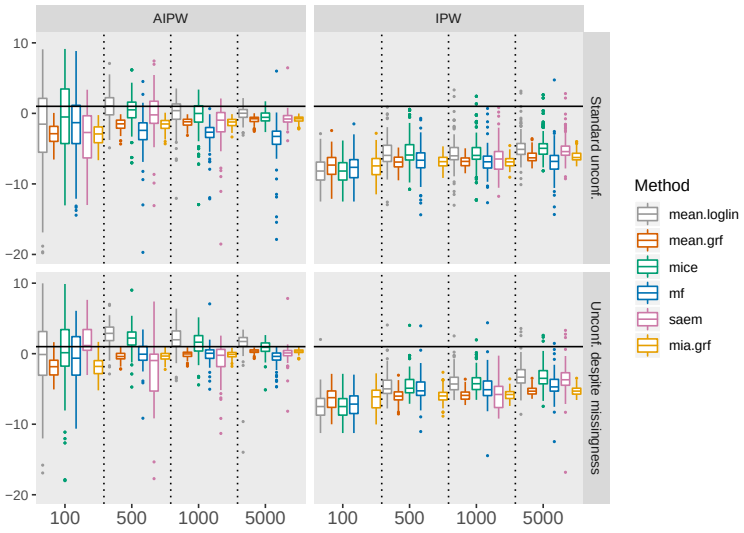
Further experiments with similar results but under a different generative data model, the standard hierarchical data-generating model with non-diagonal covariance structure (Kingma and Welling, 2014) can be found in the Supplementary material.

5.3. *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.
- Multiple imputation (*mice*) using all available information (X^*, W, Y) generally performs well for the standard unconfoundedness and the "unconfoundedness despite missingness". However in most scenarios there is a small remaining bias, even for large sample sizes. 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 at least for linear-logistic models.
- 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



(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: Estimated average treatment effect $\hat{\tau}$. **Hierarchical data-generating model** for confounders.

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.

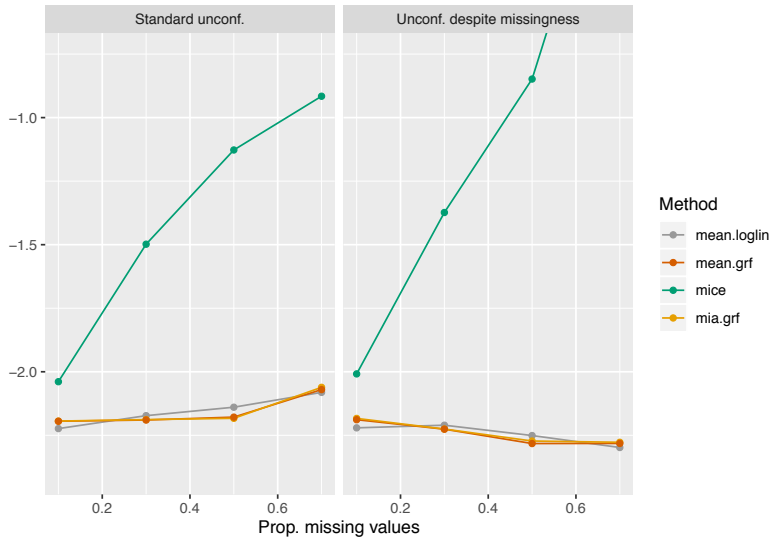
- The EM-based estimator (*saem*) performs well under correct specification (multivariate Gaussian confounders, logistic treatment assignment, linear outcome, M(C)AR missing data mechanism, (5) satisfied) and adding the mask to the initial data matrix yields unbiased estimates even if the missing data mechanism is not ignorable. It fails however in the cases where the data is not i.i.d. Gaussian.

5.4. *Simulation study on IHDP data.* Before we turn to the open medical question presented in the introduction we first apply and compare our estimation methods on a benchmark data set from the Infant Health and Development Program (IHDP)[7]. We follow Hill (2011) for the emulation of an observational data set from the original experimental data[8]. The experimental data set is composed of six quantitative and 19 binary variables, recorded for 985 individuals. In order to "transform" the experimental data into observational data, Hill (2011) proposed to select a nonrandom subset among the treated, stratified along an ethnicity variable, which leads to two unbalanced treatment groups. In total there are 139 treated and 608 control observations in the new data set. For our comparison study we only focus on the quantitative variables and simulate the response surfaces μ_w and potential outcomes $Y(w) \sim \mathcal{N}(\mu_w, 1)$, $w \in \{0, 1\}$ according to scenario A in Hill (2011), but using only these six quantitative variables. As in the previous simulation study we modify this simulation step to obtain two scenarios: one with standard unconfoundedness (1) and the other with (5).

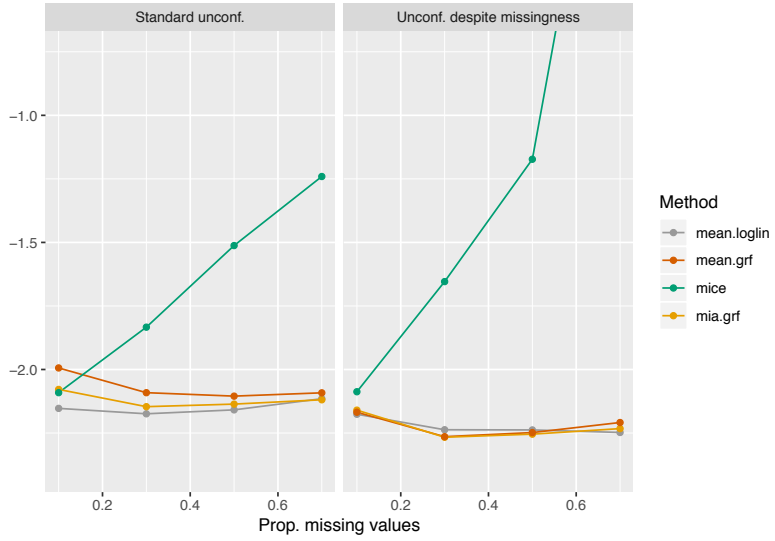
When comparing the different methods w.r.t. the unconfoundedness assumptions in Figure 7, it consistently appears that, as expected, all methods – except for *mice* – perform better under assumption (5), i.e., the potential outcomes only depend on the observed variables and the mask. The difference in performance increases as the amount of missing values increases. This is true even in the MCAR case, provided that the proportion of missing values is sufficiently large (more than 50% of missing values). Interestingly, the tree-based methods (*mia.grf* and *mean.grf*) slightly improve as the number

[7] https://github.com/vdorie/npci/tree/master/examples/ihdp_sim/data

[8] We use and adapt the corresponding code from V. Dorie: https://github.com/vdorie/npci/.



(a) MCAR



(b) Informative missing values

Fig 7: RMSE of estimated average treatment effect $\hat{\tau}_{AIPW}$ on IHDP dataset; (200 simulations for varying proportion of missing values; y-axis in log-scale).

of missing values increases, with one exception: in the MCAR setting under

standard unconfoundedness assumption. Note that in all these simulations the multiple imputation performs poorly even for small amounts of missing values and that results for *saem* are not reported as this method numerically fails on this data set. Similar behavior is observed for the IPW estimators, see the Supplementary material for more details.

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; or the measured reaction to a specific treatment is evidently missing if the treatment has not been administrated, etc.)[9]. 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[10] 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. After a large clinical trial demonstrating the effectiveness of this treatment in limiting the risk of death due to hemorrhagic shock

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

[10]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.

or excessive bleeding (Shakur et al., 2010), it is now recommended to administer this drug to patients at risk of developing an hemorrhagic shock. In the pre-treated set of patients with major trauma (counting 7,240 observations) 1,158 patients receive the treatment, corresponding to 16%. In Table 1 we read that among the patients with traumatic brain injury, 602 receive the treatment, corresponding to 19%.

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 8 is the result of a two-step Delphi procedure (Dalkey and Helmer, 1963) consisting first in a selection of covariates related to either treatment or outcome or both and second in a classification of these covariates into confounders and predictors of only treatment or outcome. The Delphi procedure consists in consulting a panel of experts, in this case six anesthetists and resuscitators specialized in critical care, to identify a stable set of quantities relevant for this study. 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 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 8), using additional predictors for the outcome

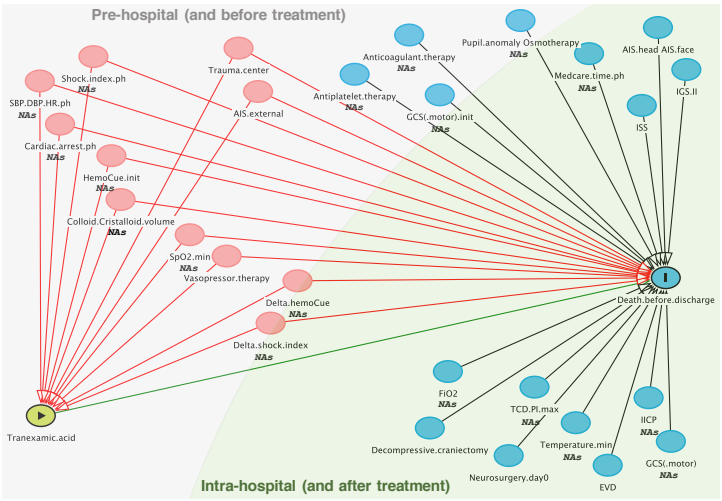


Fig 8: Causal graph representing treatment, outcome, confounders and other predictors of outcome (Figure generated using DAGitty (Textor, Hardt and Knüppel, 2011); `NAs` indicates variables that still have missing values after pre-treatment).

model (blue nodes on the graph in Figure 8), we obtain the following estimations in Figure 9 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

[11] 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.

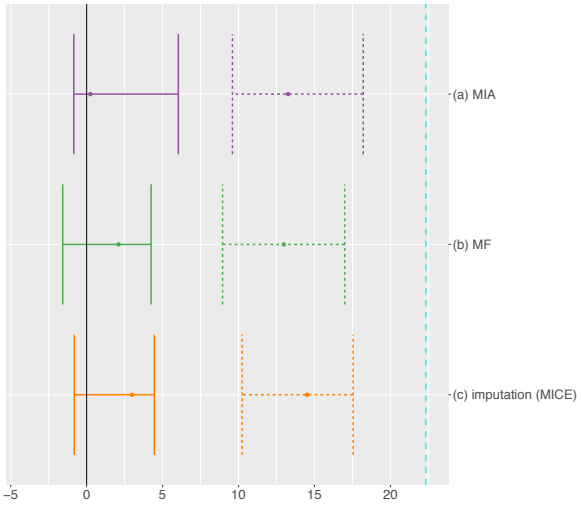


Fig 9: ATE estimations on Traumabase data (solid: doubly robust estimates; dotted: IPW estimates; dashed vertical line: without adjustment; x-axis: $\hat{\tau}$ and bootstrap confidence intervals[11]).

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.

Unlike the simulation study, we compare only three methods in this part: *mia.grf*, *mf* and *mice*. We drop *mean.grf* since it is equivalent to *mia.grf*. We also do not test *saem* since currently the EM for logistic regression with missing attributes has only been derived for continuous variables, and in this application there are many important categorical attributes to consider. A first observation on the results reported in Figure 9 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 8). Finally, all compared methods have similar empirical variances as can be observed on the reported bootstrap confidence intervals in Figure 9. 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, but its limited applicability in practice, namely the exclusion of informative missing data, 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. The main contribution of this work is therefore the proposition of several estimators belonging to this second class that we have implemented and tested on simulated and real data sets.

7.2. *AIPW-type estimation with missing attributes.* The estimators that we propose, $\hat{\tau}_{MIA}$ and $\hat{\tau}_{EM}$, can be used with an IPW or an AIPW formulation. Our proposed AIPW-type methods in the second class differ from previous solutions in several ways:

- they allow for doubly robust estimation whereas previous solutions only consider regression adjustment (Kallus, Mao and Udell, 2018) or IPW (Seaman and White, 2014),
- they are based on consistency results that allow for theoretical guarantees of our methods in a wide range of scenarios,
- they can easily be implemented and allow a simple handling of informative missing data in practice as opposed to a recent treatment effect estimation for confounders with informative missing values based on integral equations (Yang, Wang and Ding, 2017).

For better comparability with the most popular method from the first class, namely multiple imputation, we have also proposed an AIPW formulation for this approach.

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.

It remains to study how consistency rates and variance results from Wager and Athey (2018) extend to the incomplete case, namely random forests with the MIA splitting rule. Furthermore, simulation studies on mixed confounders are necessary to corroborate the effectiveness of the proposed estimators in scenarios with this additional – yet in practice often encountered – aspect.

7.3. *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.

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://www.imkemayer.com/papers/2019-09-08_DR-TreatmentEffect-WithMissingAttributes_supp.p
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.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20** 217–240.

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