The Annals of Applied Statistics
2020, Vol. 14, No. 3, 1409–1431
https://doi.org/10.1214/20-AOAS1356

DOUBLY ROBUST TREATMENT EFFECT ESTIMATION WITH MISSING ATTRIBUTES

BY IMKE MAYER[1], ERIK SVERDRUP[2,*], TOBIAS GAUSS[3,‡], JEAN-DENIS MOYER[3,§], STEFAN WAGER[2,†] AND JULIE JOSSE[4]

[1]*Centre d'Analyse et de Mathématique Sociales, École des Hautes Études en Sciences Sociales, imke.mayer@ehess.fr*
[2]*Graduate School of Business, Stanford University, *erikcs@stanford.edu; †swager@stanford.edu*
[3]*Department of Anesthesia and Intensive Care, Beaujon Hospital, ‡tgauss@protonmail.com; §jean-denis.moyer@aphp.fr*
[4]*Centre de Mathématiques Appliquées École Polytechnique, Institut Polytechnique de Paris, julie.josse@polytechnique.edu*

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 outperforms others. We find, however, that doubly robust modifications of standard methods for average treatment effect estimation with missing data repeatedly perform better than their nondoubly 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 observational 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 nondoubly 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 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; treatment effect estimation on such observational studies can serve as a precursor for future randomized placebo-controlled studies, namely, by helping to define 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 8% of traumatic brain injury patients, and among

Received October 2019; revised May 2020.
Key words and phrases. Missing data, causal inference, potential outcomes, observational data, propensity score estimation, incomplete confounders, major trauma, public health.

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

TABLE 1
Occurrence and frequency table for traumatic brain injury patients (*total number*: 8248)

	Survived	Died
TA not administered	6238 (76%)	1327 (16%)
TA administered	367 (4%)	316 (4%)

all patients 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 (46% vs. 18%). 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 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 (Mayer et al. (2020)).

As discussed further in the following section, the statistics of treatment effect estimation under unconfoundedness are 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 noninformative missingness, for example, medical staff simply forgetting to log some numbers, but in other cases the missingness is clearly informative; 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

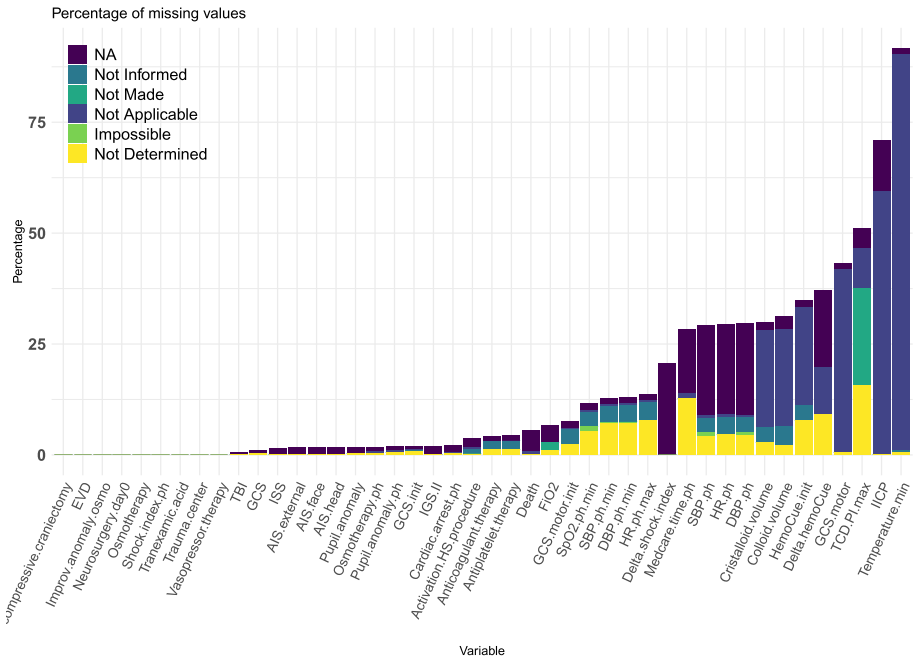


FIG. 1. *Percentage of missing values for a subset of variables relevant for traumatic brain injury. Different encodings of missing values*: *NA* (*not available*), *not informed*, *not made*, *not applicable*, *impossible*.

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.

1.2. *Summary of contributions and outline.* In this paper we consider several popular methods for treatment effect estimation with missing covariates that rely on various unconfoundedness assumptions or assumptions about the missingness mechanism. We then discuss natural doubly robust generalizations of these methods, and compare them in numerical experiments. We find 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 with parametric methods fit via the EM algorithm (Dempster, Laird and Rubin (1977)) are better, whereas other times nonparametric estimators do better; overall, the performance of each method strongly depends on the underlying confounding mechanism. However, we systematically find our doubly robust modifications of standard methods to outperform their baselines.

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 ith 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

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

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

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

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

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

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)

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

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)),

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

where $\mu_{(w)}(x) \triangleq \mathbb{E}[Y \mid X_i = x, W_i = w]$ and $\hat{\mu}_{(w)}(x)$ is an estimate thereof. The AIPW estimator is often referred to as "doubly robust" because $\hat{\tau}_{\mathrm{AIPW}}$ is consistent for τ if either the estimated outcome functions $\hat{\mu}_{(w)}(x)$ or the estimated propensity scores $\hat{e}(x)$ are consistent.

A key fact about doubly robust estimators as in (4) is that $\hat{\tau}_{\mathrm{AIPW}}$ can be $\sqrt{n}$-consistent for τ and asymptotically Gaussian even in a nonparametric setting where $\hat{\mu}_{(w)}(\cdot)$ and $\hat{e}(\cdot)$ are estimated, for instance, using generic machine learning methods, at slower nonparametric rates (Farrell (2015)). In particular, provided use "cross-fitting", that is, we do not use the ith datapoint itself for making the predictions $\hat{\mu}_{(w)}(X_i)$ and $\hat{e}(X_i)$, $\hat{\tau}_{\mathrm{AIPW}}$ using any choice of $\hat{\mu}_{(w)}(X_i)$, and $\hat{e}(X_i)$ attains $\sqrt{n}$ rates of convergence whenever the product of the root-mean squared errors of $\hat{\mu}_{(w)}(X_i)$ and $\hat{e}(X_i)$ decays faster than $1/\sqrt{n}$ (Chernozhukov et al. (2018), van der Laan and Rose (2011)).[2]

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 elementwise product, such that $X_{ij}^* = X_{ij}$ if $R_{ij} = 1$ and $X_{ij}^* = \mathrm{NA}$ if $R_{ij} = \mathrm{NA}$.[3]

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) (D'Agostino and Rubin (2000), Leyrat et al. (2019), Mattei (2009), Rosenbaum and Rubin (1984), Seaman and White (2014)).

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) 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), that is, that (Rosenbaum and Rubin (1984))

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

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

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

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. (2020), Mattei (2009)):

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

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)).

[2]Other methods, including those based on inverse-weighting as in (3), can also sometimes achieve similarly 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]This representation of the incomplete data where the missing values are treated as a special category is chosen in view of the random forest approach handling this type of data.

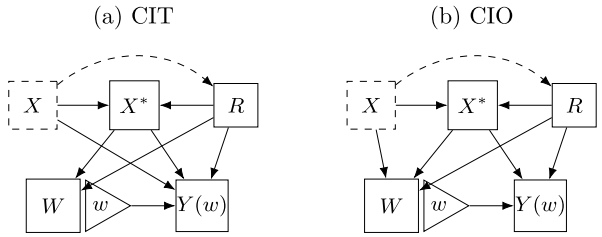


FIG. 2. *Causal graph depicting the assumptions* (7).

We note that fitting (6) may appear difficult from the perspective of classical parametric statistics; for example, 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 (Josse et al. (2019)).

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)), that is, for each possible mask $r \in \{1, \mathrm{NA}\}^p$,

$$\mathbb{P}(R_i = r \mid X_i = x, W_i, Y_i) = \mathbb{P}\big(R_i = r \mid (X_i)_r = x_r, W_i, Y_i\big), \tag{8}$$

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," that is, that the missing attributes are simulated from the correct conditional distribution and correct model specification 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),

$$\mathbb{P}(R_i = r \mid X_i, W_i, Y_i) = \mathbb{P}(R_i = r),$$

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, that is, $X_i = X_i^*$. 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., Candès 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[4] 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 prehospital 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 prehospital 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. But overall—due to the multitude of agents collecting the data in different circumstances and under important time constraints—such statements about the plausibility of MAR are difficult to assess on the whole of the registry.

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

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

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

There are general results about AIPW that immediately guarantee that the above estimator $\hat{\tau}_{\text{AIPW}^*}$ is $\sqrt{n}$-consistent and asymptotically normal around τ given only weak regularity conditions provided the product of the root-mean squared errors of the nuisance component estimates decay as $o(n^{-1/2})$ (Chernozhukov et al. (2018)), and these results extend directly to the case where the X_i may contain missing values. Specifically, in order to get such results for $\hat{\tau}_{\text{AIPW}^*}$, it suffices to assume that

$$\mathbb{E}\left[\left(\frac{1}{\hat{e}^*(X_i^*)(1-\hat{e}^*(X_i^*))} - \frac{1}{e^*(X_i^*)(1-e^*(X_i^*))}\right)^2\right]^{\frac{1}{2}} \times \mathbb{E}\left[\left(\hat{\mu}^*_{(W)}(X_i^*) - \mu^*_{(W)}(X_i^*)\right)^2\right]^{\frac{1}{2}} = o\left(\frac{1}{\sqrt{n}}\right), \tag{10}$$

that is, that $\hat{\mu}^*_w(x^*)$ and $\hat{e}^*(x^*)$ are good approximations to the best predictors we could have using on the partially observed predictors x^*. Below, we instantiate the approach (9) via both a parametric approach, based on logistic regression, and a nonparametric approach, based on random forests.

4.1.1. *Parametric estimation of nuisance components.* For the parametric approach we build on work by Jiang et al. (2020) and Schafer (1997) and logistic and linear forms, respectively, for the generalized propensity score and outcome using the *complete* covariates x. The functions μ^* and e^* that take in incomplete covariates x^* are then estimated via EM (Dempster, Laird and Rubin (1977)). The exact description of this parametric procedure for the AIPW estimator is outlined in Procedure 1; the resulting IPW and AIPW estimators will be denoted $\hat{\tau}_{\text{EM}}$.

A major limitation of this approach is that, in order to justify use of the EM algorithm, one typically needs to make further assumptions on the missing value mechanism; in particular, it is common to make the missing at random assumption (8). In other words, although we did not require the missing at random assumption to identify τ, this assumption is used for consistent parametric estimation of $e^*(x^*)$ and $\mu^*_{(w)}(x^*)$. Below, we describe a nonparametric alternative that only needs the identifying assumption (5) to get consistency for τ.

4.1.2. *Nonparametric estimation of nuisance components.* As an alternative to fitting parametric models via EM as discussed above, one can also directly estimate the functions $e^*(x^*)$ and $\mu^*_{(w)}(x^*)$ nonparametric. This task may appear somewhat unusual, as the features x^* take values in the augmented space $\{\mathbb{R} \cup \text{NA}\}^p$. However, many popular machine learning methods—including decision trees, kernels and neural networks—can be adapted to

Procedure 1: 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 $(Y_{i:W_i=0}, X^*_{i:W_i=0})$, 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.

Procedure 2: 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. Train a causal forest on the potentially incomplete features X^* using MIA splitting.
2. Extract out-of-bag estimates $\hat{\mu}^*_{(w)}(X_i^*)$ and $\hat{e}^*(X_i^*)$ from the causal forest.
3. Combine the predictions as in (9) to obtain a doubly robust estimate $\hat{\tau}$ for τ.

this context, and standard arguments for verifying consistency of these methods still apply (Josse et al. (2019)). Then, once we have estimates of $e^*(x^*)$ and $\mu^*_{(w)}(x^*)$, we can proceed to estimate the treatment effect using the AIPW estimator (9) or the analogous IPW estimator.

In this paper we focus on nonparametric nuisance component estimation via (generalized) random forests (Athey, Tibshirani and Wager (2019a), Breiman (2001)), with missing data handled using the *missing incorporated in attributes* (MIA) method of Twala, Jones and Hand (2008). The main idea of the MIA approach is, give each split additional flexibility, such that missing values may be sent on either side of the split independently of where the split occurred. More specifically, as outlined by Twala, Jones and Hand (2008), consider splitting on the jth attribute and assume that, for some individuals, the value of X_j is missing. MIA treats the missing values as a separate category or code and the considers the following splits:

- $\{i : X_{ij} \leq t$ or X_{ij} is missing$\}$ vs. $\{i : X_{ij} > t\}$
- $\{i : X_{ij} \leq t\}$ vs. $\{i : X_{ij} > t$ or X_{ij} is missing$\}$
- $\{X_{ij}$ is missing$\}$ vs. $\{X_{ij}$ is observed$\}$,

for some threshold t. The MIA approach does not seek to model why some features are unobserved; instead, it simply tries to use information about missingness to make the best possible splits for modeling the desired outcome. Thus, the MIA strategy work with arbitrary missingness mechanisms and does not require the missing data to be MAR.[5]

In order to estimate the average treatment effect, we use the estimator (9) with nuisance components extracted from a variant of the causal forests of Athey, Tibshirani and Wager (2019a) that use MIA splitting to handle missing values.[6] To do so, we have added the MIA splitting rule to the `causal_forest` function in `grf` (Tibshirani et al. (2020)), and our proposed estimator can be computed by simply calling the function `average_treatment_effect` on a trained causal forest.

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), that is, 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 $\hat{\tau}^{(m)}_{\text{AIPW}}$ given

[5]We conjecture that consistency proofs for random forests following, for example, Scornet, Biau and Vert (2015) or Wager and Walther (2015) extend to the case of MIA splitting and missing covariates. However, formal results of this type are not currently available.

[6]We refer to Section 2.1 of Athey and Wager (2019) for a detailed discussion of how the doubly robust scores used in (9) can be extracted from a causal forest.

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. Impute the initial data X^* using an M times with the chosen imputation method to obtain M complete data matrices $(X^{(1)}, \ldots, X^{(M)})$.
2. For every imputed data matrix $X^{(m)}$, $m \in \{1, \ldots, M\}$:

   Option 1 Nonparametric regression:
   a) Train a causal forest on the imputed features $X^{(m)}$.
   b) Extract out-of-bag estimates $\hat{\mu}_{(w)}(X_i^{(m)})$ and $\hat{e}(X_i^{(m)})$ from the causal forest.
   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 $e(X_i^{(m)})$.
   b) Fit two separate linear models on $(Y_{i:W_i=1}, X_{i:W_i=1}^{(m)})$ and on $(Y_{i:W_i=0}, X_{i:W_i=0}^{(m)})$ respectively to obtain predictions for $\mu_{(1)}(X_i^{(m)})$ and $\mu_{(0)}(X_i^{(m)})$, respectively.
   c) Combine the predictions following (4) to obtain a doubly robust estimation $\hat{\tau}^{(m)}$ for τ.

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

in (4) instead of the IPW estimate $\hat{\tau}_{\text{IPW}}^{(m)}$ given in (3). This approach is outlined in Procedure 3. We note that this method relies on the performance of the multiple imputation strategy; for instance, in the case of FCE the method requires correct specification of the conditional models which can be hard to assess in practice. We refer to Carpenter and Kenward (2013) for a discussion on imputation strategies.

Another recent solution is based on matrix factorization (Kallus, Mao and Udell (2018)) as outlined in Procedure 4 in the Supplementary Material (Mayer et al. (2020)). Note that, unlike with multiple imputation, we only impute each datapoint once, and consistency guarantees are only given under MCAR.

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 and 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, that is, we do not study the (parametric) estimators' performances in case of model misspecification. 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 (2020)).[7]

[7]The code for reproducing the experiments presented in this work is given in the Code material (Mayer et al. (2020)).

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*)

	Confounders & Covariates		Missingness		Unconfoun-dedness		Models for (W, Y)	
	Multivariate normal	General	M(C)AR	General	(1)	(5)	Logistic-linear	Non-param.
saem	✓	✗	✓	✗	✗	✓	✓	✗
grf	✓	✓	✓	✓	✗	✓	✓	✓
mice	✓	✓	✓	✗	✓	✓	✓	(✗)
mf	✓	✗	✓	✗	✓ (on U)	✗	✓	(✗)
mean.loglin	✗	✗	✗	✗	✗	✗	✗	✗

5.1. *Methods overview.* We compare our approaches $\hat{\tau}_{\mathrm{EM}}$ and $\hat{\tau}_{\mathrm{MIA}}$, denoted *saem* and *grf* in the experiments,[8] to the following methods, where we summarize their assumptions in Table 2:

- *mice*: Procedure 3 (and its IPW analogue detailed in the Supplementary Material (Mayer et al. (2020))) 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 (Mayer et al. (2020))) with Option 2; we adapt the implementation[9] 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.

For the parametric $\hat{\tau}_{\mathrm{EM}}$ we use the R package `misaem` (Jiang (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 concatenate the initial or imputed data matrix X and the binary mask R 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).

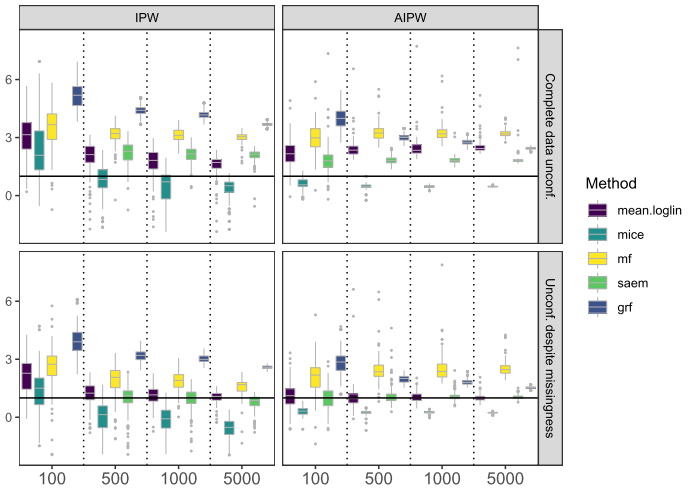
5.2. *Data generation.* We define different models for the generation of the confounders, covariates, missing values, treatment assignment and outcome.

5.2.1. *Confounders and covariates.*

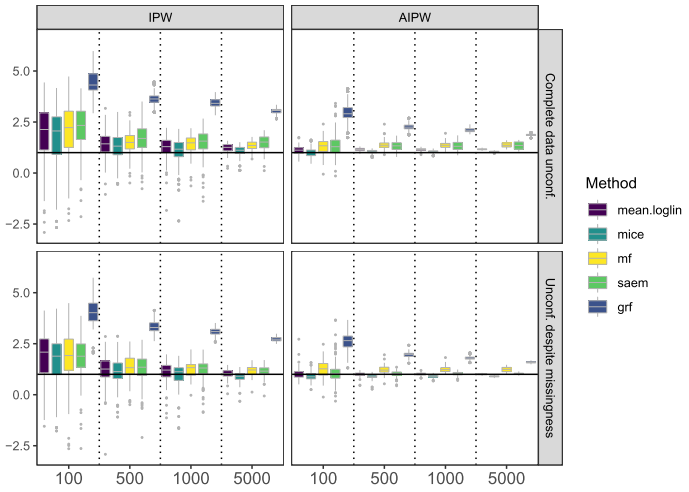
Model 1: *Multivariate normally distributed confounders.* We generate normally distributed confounders $X_{i\cdot} = [X_{i1} \dots X_{ip}]^T \sim \mathcal{N}(\mathbf{1}, \Sigma)$, $i \in \{1, \dots, n\}$, for $p = 10$, where $\Sigma = I - 0.6 \times (I - 1)$, $\mathbf{X} = [X_{1\cdot} \dots X_{p\cdot}]^T \in \mathbb{R}^{n \times p}$. Results for this model are reported in Figure 3.

[8]These abbreviations refer to the algorithms used for the estimation of the nuisance parameters in the presence of missing values. For instance, *saem* stands for (stochastic approximation) EM algorithm.

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



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



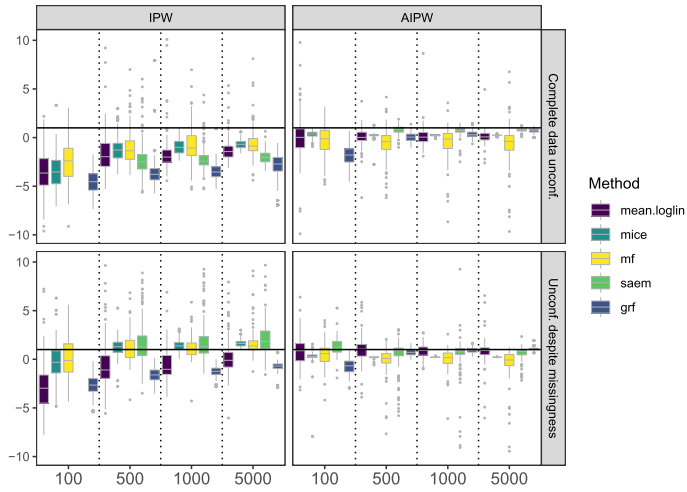
(b) Informative missing values (with 30% missing values in $X_{\cdot,1:5}$)

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* $\in \{\cdot$ *despite missingness*, *complete data* $\cdot\}$. *Results are displayed for* 100 *runs of every setting.*

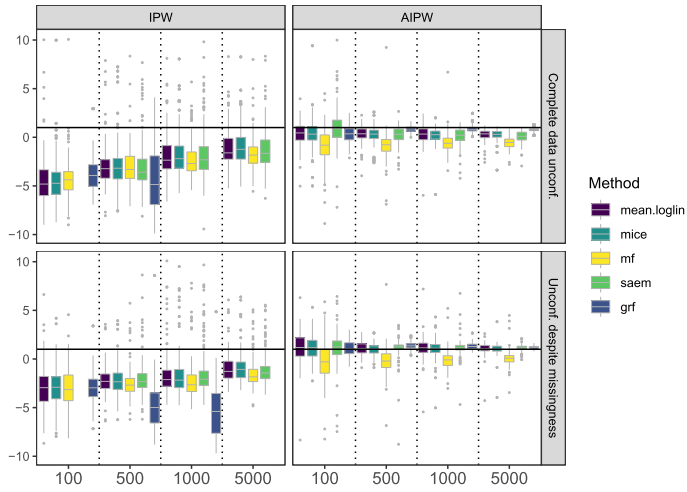
Model 2: *Latent classes model.* We consider a Gaussian mixture model, that is, we first generate class labels C from a multinomial distribution with three categories. Then, the confounders of observation i, $X_{i\cdot}$, are sampled from the corresponding class distribution, that is, $X_{i\cdot} \sim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

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 us to add an additional interaction between treatment and the latent class. Analogously, the outcome is defined as $Y_i \sim \mathcal{N}((\beta(C_i))^T X_{i\cdot}^* + \tau W_i, \sigma^2)$. The corresponding results are reported in Figure 4.

Model 3: *Low rank matrix factorization.* We adapt the simulation framework from Kallus, Mao and Udell (2018) by generating $U_{i\cdot} = [U_{i1} \dots U_{id}]^T \sim \mathcal{N}(0, I_d)$ and defining $X = UV^T$ for some fixed matrix $V \in \mathbb{R}^{p \times d}$, with $d = 3$. Results for this model are reported in Figure 5.



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



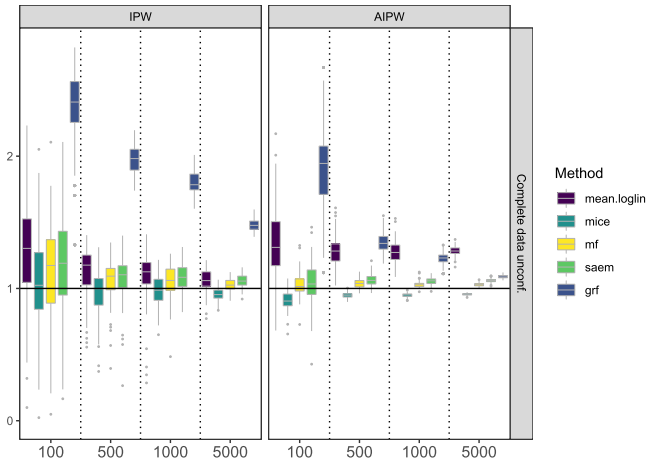
(b) Informative missing values (with 30% missing values in $X_{\cdot,1:5}$)

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* $\in \{\cdot$ *despite missingness*, *complete data* $\cdot\}$. *Results are displayed for* 100 *runs of every setting.*

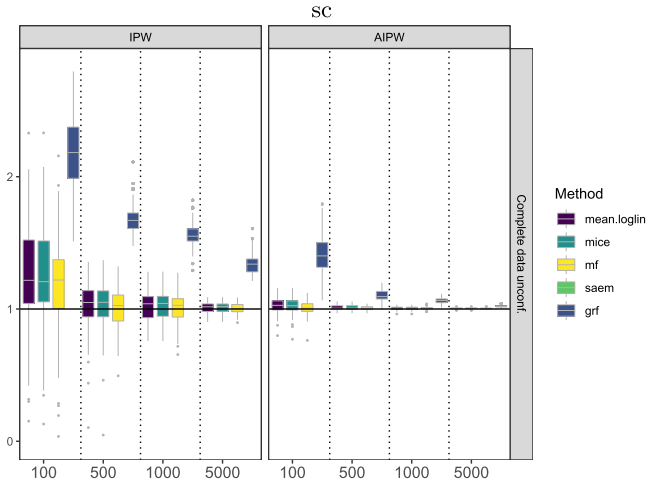
Model 4: *Hierarchical data-generating model.* An alternative to defining a Gaussian mixture model is to use a simplified shallow version of a *deep latent variable model* (DLVM, Kingma and Welling (2014)): the codes C are sampled from a normal distribution $\mathcal{N}_d(0, 1)$. Covariates X_i are then sampled from $\mathcal{N}_p(\mu(c), \Sigma(c))|C_i = c$, where

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

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 $d = 3$. Results for this model are reported in the Supplementary Material (Mayer et al. (2020)).



(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* $n \in \{100, 500, 1000, 5000\}$ *and missing values mechanism* $\in \{MCAR, general\}$. *Results are displayed for* 100 *runs of every setting.*

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[10] 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. We define $\text{logit}(e^*(X_i^*)) = \alpha_0 + \alpha^T X_i^*$. Analogously, the outcome is defined as $Y_i \sim \mathcal{N}(\beta_0 + \beta^T X_i^* + \tau W_i, \sigma^2)$.

[10]By informative we designate all nonignorable missingness mechanisms, where the probability of observing missing values depends on the missing values.

Complete data 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 (Mayer et al. (2020)) 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 splitting rule (*grf*) generally performs at least as well as multiple imputation but yields unbiased results if "unconfoundedness despite missingness" (5) holds.
- Mean imputation coupled with concatenation of the imputed data with the mask and parametric estimation empirically performs well, provided that (5) holds. However, the concatenation of the mask R appears necessary, since 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.

In conclusion, the type of unconfoundedness assumption is important for the choice of the estimation strategy. Once the type is fixed, the choices between simple and doubly robust and between parametric and nonparametric estimation depend on the a priori on the data generating processes. However, in general, we recommend privileging the doubly robust strategy.

For a more detailed discussion of the simulation results, we refer to the Supplementary Material (Mayer et al. (2020)).

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*. For our analysis we used 20,037 currently available validated patient records, validated by the medical expert team after a first pre-treatment. The pretreatment 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.).[11] Out of these 20,047 patients, 8269 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[12] greater or equal 2). Additionally, we excluded a total of 21 patients among this group coming from Trauma centers with too few observations, having joined the registry group several years after the majority of all other Trauma centers.

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 a 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 Blake et al. (2020), Lederer et al. (2019). The causal graph in Figure 6 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 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 average treatment effect of the treatment on mortality among traumatic brain injury patients. When adjusting for confounding using the identified confounders (nodes with two outgoing arcs on the graph in Figure 6), using additional predictors for the outcome model (nodes with one outgoing arc pointing to the outcome node on the graph in Figure 6), we obtain the following estimations in Figure 7 of the direct causal effect of TA on in-ICU mortality among traumatic brain injury patients.

[11]The code for pretreatment and for estimating the treatment effect on this data are available in the Code material (Mayer et al. (2020)).

[12]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 (Mayer et al. (2020)) or https://www.aaam.org/abbreviated-injury-scale-ais/ for more information.

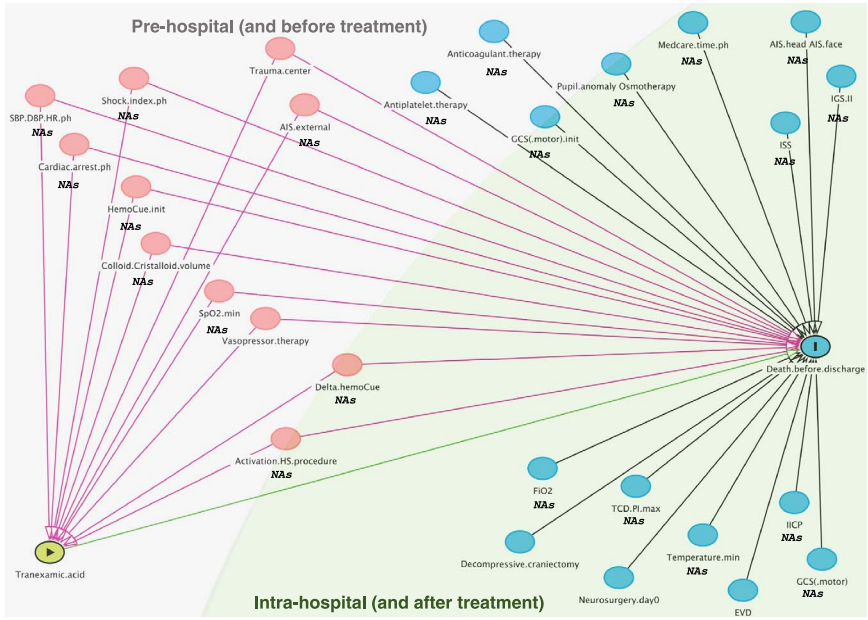


FIG. 6. *Causal graph representing treatment, outcome, confounders and other predictors of outcome* (*Figure generated using DAGitty* (*Textor, Hardt and Knüppel* (2011)); `NAs` *indicates variables that still have missing values after pretreatment*).

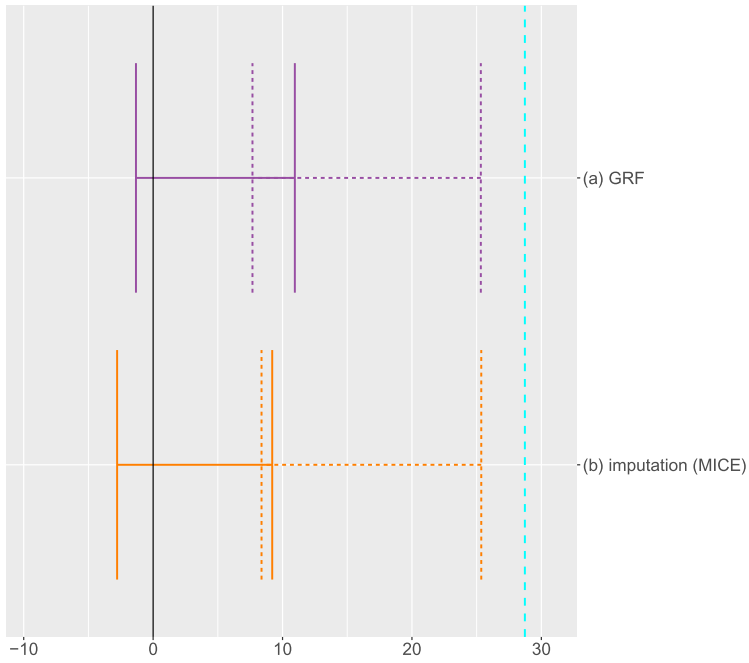


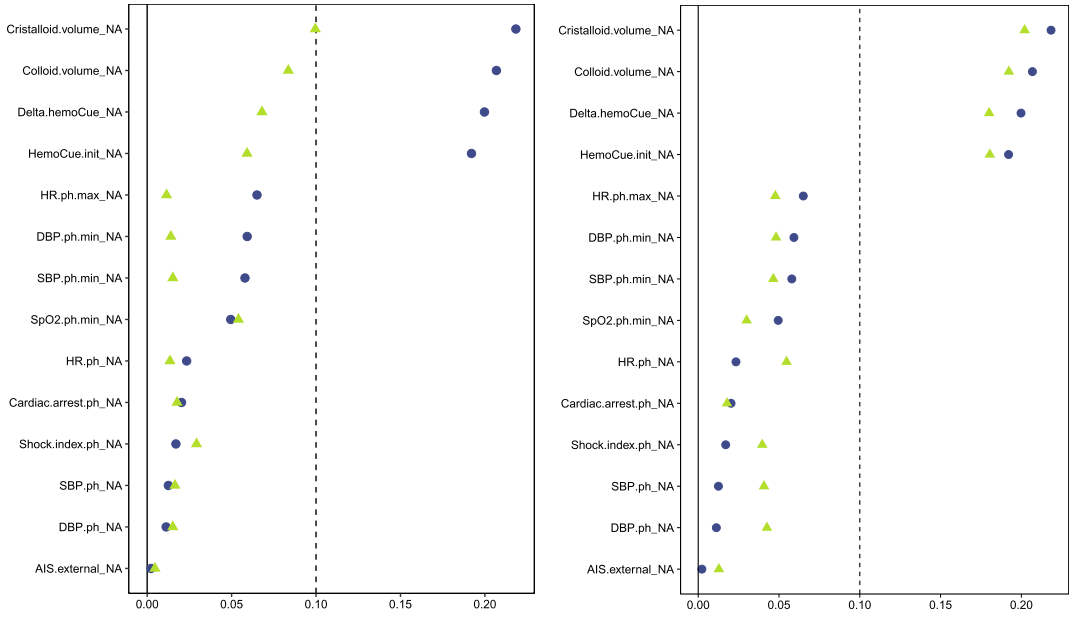
FIG. 7. *ATE estimations on Traumabase data* (*solid*: *doubly robust estimates*; *dotted*: *IPW estimates*; *dashed vertical line*: *without adjustment*; *x-axis*: $\hat{\tau}$ *and bootstrap confidence intervals*[13]). *Note: Positive ATE* $\equiv$ *increase of mortality.*

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

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 with MIA splitting rule does not require any pre-processing of the data and is therefore straightforward when using the *grf* package.

Here, we only consider three pairs of methods: *grf* and *mice*. We do not test *saem* and *mf* since currently both these methods have not been derived for heterogeneous data.[14] A first observation on the results reported in Figure 7 is the concordance of the two 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 *grf* estimations—are most suited for our specific application. When comparing covariate balance for both methods in terms of standardized mean differences, we note that both methods achieve similar balance on the observed values (see results reported in the Supplementary Material (Mayer et al. (2020))), but, as expected, only GRF additionally achieves balance on the response pattern (Figure 8). Since there is consensus by the medical experts that certain missing values are not missing at random, achieving balance on the response pattern is a relevant feature for interpreting the estimation results. A remaining issue might consist in the overlap assumption which is generally difficult to



(a) GRF (b) MICE

FIG. 8. *Absolute difference in proportion for observed and missing values*; *circles*: *before adjustment*, *triangles*: *after adjustment*.

[14]Concatening the mask with the data matrix does not lead to major changes in the estimations, therefore we only report results obtained when including the mask.

assess in most medical applications and which might be slightly violated due in part to the heterogeneity of patient profiles, and it could be argued that, for certain patients, the probability of receiving the treatment is zero. However, the lack of a standardized protocol for tranexamic acid administration favors the overlap assumption even for this group of patients. A solution to handle 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 (Mayer et al. (2020)).

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 beneficial effects of TA for some of their TBI patients in practice, and a previous clinical trial, focussing on a slightly different patient group, has also exhibited a potential benefit from the drug for patients with TBI; moreover, the results of the clinical trial studying the effect of the drug on all TBI patients indicate that on average there is neither benefit nor harm in prescribing the drug (Cap (2019)); 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 (see Figure 6). Finally, the compared estimates have similar standard errors and asymptotic confidence intervals which are also close to the estimated bootstrap confidence intervals (the latter are not reported in Figure 7).

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. (2020)—we advocate our tree-based estimator $\hat{\tau}_{\mathrm{MIA}}$ 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 $\hat{\tau}_{\mathrm{EM}}$ 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 $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$, for several reasons. For instance, one might be interested in estimating how treatment effects vary across subpopulations 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 (Athey and Wager (2017)).

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. *Weighted treatment effects*. Throughout this paper we have focused on cases with overlap (2), that is, where all units have a realistic chance of being randomized to both treatment and control. In some cases, however, there may be subjects who are quasi-deterministically assigned to one of the two treatment arms—in which case the methods developed here may be unstable and/or have very large variance. When this happens, it is common to shift focus away from the average treatment effect and towards alternative weighted estimands that are more robust to lack of overlap. For example, if some units are quasi-deterministically assigned to control (but no units are quasi-deterministically assigned to treatment, i.e., propensity scores are uniformly bounded below 1), then estimating the average treatment effect on the treatment is a popular way to avoid overlap problems (Imbens (2004)). Crump et al. (2009) and Li, Morgan and Zaslavsky (2018) discuss other weighted estimands that can be used when overlap problems get more severe and propensity scores may get arbitrarily close to both 0 and 1.

Although we do not discuss it here, the arguments developed in this paper can be applied directly to estimators of other weighted treatment effects. We implement extensions of the random forest based estimator described in 4.1.2 for estimating both the average treatment effect and the overlap-weighted treatment effect of Li, Morgan and Zaslavsky (2018) in the `R` package `grf` (Tibshirani et al. (2020)).

7.4. *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 (2019) consider a setting with outcome-independent missingness, $Y_i \perp\!\!\!\perp R_i \mid \{X_i, W_i\}$, 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.

Acknowledgments. We thank Jean-Pierre Nadal for fruitful discussion, Helen Blake and Julie Tibshirani for their suggestions 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 (DOI: 10.1214/20-AOAS1356SUPPA; .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.

Code: Functions to replicate simulations (DOI: 10.1214/20-AOAS1356SUPPB; .zip). In this material we provide R code (R Core Team (2020)) with functions allowing to replicate the presented simulation results. Previous and potential future versions extending this code will be available at https://github.com/imkemayer/causal-inference-missing.

REFERENCES

ABADIE, A. and IMBENS, G. W. (2016). Matching on the estimated propensity score. *Econometrica* **84** 781–807. MR3481379 https://doi.org/10.3982/ECTA11293

ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 597–623. MR3849336 https://doi.org/10.1111/rssb.12268

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019a). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

ATHEY, S. and WAGER, S. (2017). Efficient policy learning. ArXiv preprint. Available at arXiv:1702.02896.

ATHEY, S. and WAGER, S. (2019). Estimating treatment effects with causal forests: An application. Observational Studies 5.

BLAKE, H. A., LEYRAT, C., MANSFIELD, K. E., TOMLINSON, L. A., CARPENTER, J. and WILLIAMSON, E. J. (2020). Estimating treatment effects with partially observed covariates using outcome regression with missing indicators. *Biom. J.* **62** 428–443. https://doi.org/10.1002/bimj.201900041

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

CANDÈS, E. J. and PLAN, Y. (2010). Matrix completion with noise. *Proc. IEEE* **98** 925–936.

CAP, A. P. (2019). CRASH-3: A win for patients with traumatic brain injury. *Lancet*.

CARPENTER, J. and KENWARD, M. (2013). *Multiple Imputation and Its Application*. Wiley, Chichester, West Sussex, UK.

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. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144 https://doi.org/10.1093/biomet/asn055

D'AGOSTINO, R. B. JR and RUBIN, D. B. (2000). Estimating and using propensity scores with partially missing data. *J. Amer. Statist. Assoc.* **95** 749–759.

DALKEY, N. and HELMER, O. (1963). An experimental application of the Delphi method to the use of experts. *Manage. Sci.* **9** 458–467.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

FARRELL, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *J. Econometrics* **189** 1–23. MR3397349 https://doi.org/10.1016/j.jeconom.2015.06.017

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. 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. *Lancet* **390** 1260–1344.

HIJAZI, N., FANNE, R. A., ABRAMOVITCH, R., YAROVOI, S., HIGAZI, M., ABDEEN, S., BASHEER, M., MARAGA, E., CINES, D. B. et al. (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. MR1995826 https://doi.org/10.1111/1468-0262.00442

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge Univ. Press, Cambridge.

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

JIANG, W., JOSSE, J., LAVIELLE, M. and GROUP, T. (2020). Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Comput. Statist. Data Anal.* **145** 106907. MR4053706 https://doi.org/10.1016/j.csda.2019.106907

JONES, J. and HUNTER, D. (1995). Consensus methods for medical and health services research. *BMJ* **311** 376–380. https://doi.org/10.1136/bmj.311.7001.376

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. MR3783340 https://doi.org/10.3982/ECTA13288

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. 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. *Ann. Amer. Thorac. Soc.* **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? *Stat. Methods Med. Res.* **28** 3–19. MR3894510 https://doi.org/10.1177/0962280217713032

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. MR3803473 https://doi.org/10.1080/01621459.2016.1260466

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1925014 https://doi.org/10.1002/9781119013563

LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384

MATTEI, A. (2009). Estimating and using propensity score in presence of missing background data: An application to assess the impact of childbearing on wellbeing. *Stat. Methods Appl.* **18** 257–273. MR2515482 https://doi.org/10.1007/s10260-007-0086-0

MAYER, I., SVERDRUP, E., GAUSS, T., MOYER, J.-D, WAGER, S. and JOSSE, J. (2020). Supplement to "Doubly robust treatment effect estimation with missing attributes." https://doi.org/10.1214/20-AOAS1356SUPPA, https://doi.org/10.1214/20-AOAS1356SUPPB

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710. MR1380809 https://doi.org/10.1093/biomet/82.4.669

QU, Y. and LIPKOVICH, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Stat. Med.* **28** 1402–1414. MR2667534 https://doi.org/10.1002/sim.3549

R CORE TEAM (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

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, Univ. Washington.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROBINS, J. M. and WANG, N. (2000). Inference for imputation estimators. *Biometrika* **87** 113–124. MR1766832 https://doi.org/10.1093/biomet/87.1.113

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. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

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. Amer. Statist. Assoc..

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0899519 https://doi.org/10.1002/9780470316696

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Classics Library*. Wiley Interscience, Hoboken, NJ. MR2117498

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. *Monographs on Statistics and Applied Probability* **72**. CRC Press, London. MR1692799 https://doi.org/10.1201/9781439821862

SCORNET, E., BIAU, G. and VERT, J.-P. (2015). Consistency of random forests. *Ann. Statist.* **43** 1716–1741. MR3357876 https://doi.org/10.1214/15-AOS1321

SEAMAN, S. and WHITE, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Comm. Statist. Theory Methods* **43** 3499–3515. MR3239323 https://doi.org/10.1080/03610926.2012.700371

SHAKUR, H., ROBERTS, I., BAUTISTA, R., CABALLERO, J., COATS, T., DEWAN, Y., EL-SAYED, H., GOGICHAISHVILI, T., GUPTA, S. 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.

TEXTOR, J., HARDT, J. and KNÜPPEL, S. (2011). DAGitty: A graphical tool for analyzing causal diagrams. *Epidemiology* **22** 745. https://doi.org/10.1097/EDE.0b013e318225c2be

TIBSHIRANI, J., ATHEY, S., FRIEDBERG, R., HADAD, V., HIRSHBERG, D., MINER, L., SVERDRUP, E., WAGER, S. and WRIGHT, M. (2020). grf: Generalized Random Forests. R package version 1.1.0.

TWALA, B., JONES, M. and HAND, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recogn. Lett.* **29** 950–956.

VAN BUUREN, S. (2018). *Flexible Imputation of Missing Data*, 2nd ed. CRC Press/CRC, Boca Raton, FL.

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning*: *Causal Inference for Observational and Experimental Data*. *Springer Series in Statistics*. Springer, New York. MR2867111 https://doi.org/10.1007/978-1-4419-9782-1

WAGER, S. and WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests. ArXiv preprint. Available at arXiv:1503.06388.

YANG, S., WANG, L. and DING, P. (2019). Causal inference with confounders missing not at random. *Biometrika* **106** 875–888. MR4031203 https://doi.org/10.1093/biomet/asz048

ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. MR3010898 https://doi.org/10.1080/01621459.2012.695674

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107** 1360–1371. MR3036400 https://doi.org/10.1080/01621459.2012.703874