Generalizing treatment effects with incomplete covariates

Imke Mayer, Julie Josse, Traumabase Group

Generalizing treatment effects with incomplete covariates

Imke Mayer* Julie Josse† Traumabase Group‡

September 6, 2021

Abstract

We focus on the problem of generalizing a causal effect estimated on a randomized controlled trial (RCT) to a target population described by a set of covariates from observational data. Available methods such as inverse propensity weighting are not designed to handle missing values, which are however common in both data sources. In addition to coupling the assumptions for causal effect identifiability and for the missing values mechanism and to defining appropriate estimation strategies, one difficulty is to consider the specific structure of the data with two sources and treatment and outcome only available in the RCT. We propose and compare three multiple imputation strategies (separate imputation, joint imputation with fixed effect, joint imputation without source information), as well as a technique that uses estimators that can handle missing values directly without imputing them. These methods are assessed in an extensive simulation study, showing the empirical superiority of fixed effect multiple imputation followed with any complete data generalizing estimators. This work is motivated by the analysis of a large registry of over 20,000 major trauma patients and a RCT studying the effect of tranexamic acid administration on mortality. The analysis illustrates how the missing values handling can impact the conclusion about the effect generalized from the RCT to the target population.

Keywords: Causal effect transportability; missing values; external validity; data integration; multiple imputation; evidence-based medicine.

*Centre d'Analyse et de Mathématique Sociales, EHESS, PSL University, CNRS, Paris, France (email: imke.mayer@ehess.fr). Corresponding author.

†Inria Sophia-Antipolis, Montpellier, France (email: julie.josse@inria.fr).

‡Department of anesthesia and intensive care, Beaujon hospital, AP–HP, Clichy, France.

1 Introduction

Observational and clinical trial data can provide different perspectives when evaluating an intervention or a medical treatment. Combining the information gathered from experimental and observational data is a promising avenue for medical research, because the knowledge that can be acquired from integrative analyses would not be possible from any single-source analysis alone. Such integrative analyses can be used for example to predict a treatment effect on a specific target population using the one estimated from the RCT, to validate methods, especially applied to observational data by emulating a trial, to better estimate heterogeneous effects (which generally cannot be estimated from experimental data due to underpowered studies). Here, we are interested in the former case, where the experimental data or randomized controlled trial (RCT) is considered as a biased sample of a target population and we would like to estimate the treatment effect on the target population represented by an observational study. More precisely, the effect is estimated on a RCT composed of covariates, treatment and outcome while the observational study is composed only of covariates. There exists a multitude of methods to generalize a treatment effect, for detailed reviews we refer to Colnet et al. (2020) and Degtiar and Rose (2021). But all these methods do not consider the problem of missing data which is ubiquitous in data analysis practice (Josse and Reiter, 2018). We emphasize here that we focus on incomplete covariates in both studies, while the outcome and treatment (available in the RCT) are assumed to be fully observed.

In certain cases, naive approaches such as complete-case analysis can yield unbiased treatment effect estimates (Bartlett et al., 2015); however, in many settings, especially for observational data, the estimations are known to be biased since the complete-case observations are generally not a representative subsample of the population of interest (Little and Rubin, 2014). Furthermore, this approach is sometimes not even possible since in high-dimensional settings the probability of having complete observations decreases rapidly Zhu et al. (2019). There exists a multitude of methods to handle missing values (Little and Rubin, 2014; van Buuren, 2018; Mayer et al., 2019), such as maximum likelihood estimation or multiple imputations with the aim of estimating as well as possible a parameter and its variance. These methods make assumptions on the mechanism that generated the missing values. More recent works also consider the question of supervised learning with missing values which is a different problem from statistical inference of model parameters (Josse et al., 2019; Le Morvan et al., 2020). In the context of causal inference there exist several recent works that address missing data (Mattei and Mealli, 2009; Seaman and White, 2014; Yang et al., 2019; Kallus et al., 2018; Mayer et al., 2020). One difficulty in the context of causal inference, is that one has to couple identifiability assumptions for the causal parameter with assumptions on the missing values mechanisms to define an appropriate strategy. According to Mayer et al. (2020), identifiability of the causal effect with missing values is ensured by adapting the causal inference assumptions to the missing values setting with an *unconfoundedness despite missing values* (UDM) assumption (Rosenbaum and Rubin, 1984). However, these works only consider the case of a single dataset—or potentially multiple datasets with the same data distribution, i.e., sampled from the same population of interest—and do not treat the case of transporting or generalizing a treatment effect from an RCT to a target distribution defined through an observational dataset; the RCT representing a "distorted" population due to sampling or selection bias (generally defined via eligibility criteria). In practice, the observed distributions between the observational data and the RCT do not only differ due to the selection bias but may also differ in terms of missing values patterns. Another field that explicitly studies data integration and missing values therein is meta-analysis. Burgess et al. (2013) study a multiple imputation approach in the context of meta-analysis where missing values can occur in different data sources.

In this paper we propose to revisit the standard identifiability assumptions and estimators for generalizing a treatment effect from one to the other under the perspective of identifying the impact of missing values and suggesting appropriate strategies to handle missing values according to the different assumptions, such as the missing values mechanisms. Indeed, to the best of our knowledge, there exists no directly applicable method for this setting nor guidelines for a correct handling of missing values in either of or both data sources. Our main contributions are:

- we define several multiple imputation strategies adapted for the aforementioned data integration problem in Section 3;
- we propose alternative identifiability assumptions which can be seen as an an extension of the *unconfoundedness despite missingness* (UDM) assumption from the observational data case (Mayer et al., 2020) and suggest adapted estimators in Section 4;
- we assess the performance of the proposed estimators and naive complete case estimators in an extensive simulation study in Section 5;
- we present the results of the real-world data analysis which has motivated this work in Section 6.

For simplicity, we will assume that the relevant covariates that allow for adjustment of the sampling bias are observed both in the RCT and the observational dataset, and we focus on the case of different missing values mechanisms for these covariates.
Before diving into the statistical aspects of the problem of treatment effect generalization, we briefly introduce a medical question about major trauma patients that has motivated this work. Major trauma denotes injuries that endanger the life or the functional integrity of a person. The World Health Organization (WHO) has recently shown that major trauma including road-traffic accidents, interpersonal violence, falls, etc. remains a world-wide public health challenge and major source of mortality and handicap (Leigh et al., 2018). We focus on trauma patients suffering from a traumatic brain injury (TBI). TBI is a sudden damage to the brain caused by a blow or jolt to the head and can lead to intracranial bleeding that can be observed on a computed tomography (CT) scan. Ongoing intracranial bleeding can lead to raised intracranial pressure, brain herniation, and death. Over 10 million people pass away or are hospitalized worldwide because of TBI each year (Dewan et al., 2012). Tranexamic acid (TXA) is an antifibrinolytic agent that limits excessive bleeding, commonly given to surgical patients. Previous clinical trials showed that TXA decreases mortality in patients with traumatic *extracranial* bleeding (Shakur-Still et al., 2009). Such a result raises the possibility that it might also be effective in TBI, because *intracranial* hemorrhage is common in TBI patients. Therefore the question here is to assess the potential decrease of mortality in patients with intracranial bleeding when using TXA. To answer this question, we have at disposal two data sources:"CRASH-2", a multi-center international RCT, and "Traumabase", an observational national registry. The details about these data are provided in Section 6, as well as the analysis results for generalizing the treatment effect from the CRASH-2 study to the Traumabase registry.

2 Background and notations

We first introduce the notations, standard assumptions and estimators in the complete case, following the line of Colnet et al. (2020).

2.1 Notations

The general case consists in assuming that each patient from the two populations is fully characterized by a random tuple $(X, Y(0), Y(1), A, S)$, where $X \in \mathcal{X}$ is a p-dimensional vector of covariates, A denotes the binary treatment assignment, $Y(a)$ is the potential outcome for treatment level a, $a \in \{0, 1\}$ and S indicates RCT participation[1]. The data considered is composed of $n + m$ independent random tuples: $(X_i, Y_i(0), Y_i(1), A_i, S_i)_{i=1,\dots,n+m}$. For simplicity of exposition, we introduce an additional indicator variable Q that allows to distinguish between observations from the RCT (indexed by the set $Set_{\mathcal{R}} \triangleq \{1, \dots, n\}$) and from the observational cohort (indexed by $Set_{\mathcal{O}} \triangleq \{n+1, \dots, n+m\}$). Thus $Q_i \triangleq \begin{cases} \mathcal{R} & \text{if } i \in Set_{\mathcal{R}} \\ \mathcal{O} & \text{if } i \in Set_{\mathcal{O}} \end{cases}$. Note that (i) $P(Q = \mathcal{R}|S = 1) = 1$, and we assume that conditionally on $S = 0$, Q is independent of all other variables considered in this setting. In this paper, we consider the case where for each RCT sample i such that $Q_i = \mathcal{R}$, we observe $(X_i, A_i, Y_i, S_i = 1)$, while for observational data i such that $Q_i = \mathcal{O}$, we consider that we only observe the covariates X_i, more precisely:

- the RCT samples $i = 1, \dots, n$ are identically distributed according to $\mathcal{P}(X, Y(0), Y(1), A, S \mid S = 1)$,
- and the observational data samples $i = n+1, \dots, n+m$ are identically distributed following $\mathcal{P}(X, S)$.

An illustration of the data we consider in this work is provided in Figure 1.

	Q	Covariates			Treatment	Outcome under A=0	Outcome under A=1
		X_1	X_2	X_3	A	$Y(0)$	$Y(1)$
1	$\mathcal{R}$	1.1	20	5.4	1	23.4	24.1
...	$\mathcal{R}$		...		...	...	...
$n-1$	$\mathcal{R}$	-6	45	8.3	0	26.3	27.6
n	$\mathcal{R}$	0	15	6.2	1	28.1	23.5
$n+1$	$\mathcal{O}$	-2	52	7.1	NA	NA	NA
$n+2$	$\mathcal{O}$	-1	35	2.4	NA	NA	NA
...	$\mathcal{O}$		...		NA	NA	NA
$n+m$	$\mathcal{O}$	-2	22	3.4	NA	NA	NA

	Q	Covariates			Treatment	Outcome under A
		X_1	X_2	X_3	A	Y
1	$\mathcal{R}$	1.1	20	5.4	1	24.1
...	$\mathcal{R}$		...		...	...
$n-1$	$\mathcal{R}$	-6	45	8.3	0	26.3
n	$\mathcal{R}$	0	15	6.2	1	23.5
$n+1$	$\mathcal{O}$	-2	52	7.1	NA	NA
$n+2$	$\mathcal{O}$	-1	35	2.4	NA	NA
...	$\mathcal{O}$		...		NA	NA
$n+m$	$\mathcal{O}$	-2	22	3.4	NA	NA

Figure 1: Example of data structure in the full data problem setting. Left: complete underlying data with potential outcomes in the RCT. Right: observed data with factual outcomes.

We define the conditional average treatment effect (CATE):

$$\forall x \in \mathcal{X}\,, \quad \tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]\,, \tag{1}$$

and the population average treatment effect (ATE):

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

while we define the RCT (or sample) average treatment effect as

$$\tau_1 = \mathbb{E}[Y(1) - Y(0) \mid S = 1].$$

We denote by $\mu_a(x)$ and $\mu_{a,1}(x)$ the conditional response surfaces under treatment $a \in \{0, 1\}$ in the general and in the RCT population, respectively:

$$\mu_a(x) = \mathbb{E}\left[Y(a) \mid X = x\right]\,, \quad \mu_{a,1}(x) = \mathbb{E}\left[Y(a) \mid X = x, S = 1\right]\,,$$

[1]This indicator S comprises several components: eligibility, selection into the trial and willingness to participate.

and by $\pi_S(x)$ the selection score:

$$\pi_S(x) = P(S = 1 \mid X = x)\,.$$

Note that $\pi_S(x)$ is the probability of being eligible for selection in the RCT and of being willing to participate given covariate values x. It is different from the probability that an individual with covariates x, known to be in the study (RCT or observational population), is selected in the RCT:

$$\pi_S(x) \neq \pi_{\mathcal{R}}(x), \quad \pi_{\mathcal{R}}(x) = P(\exists i, Q_i = \mathcal{R}, X_i = x \mid \exists i \in Set_{\mathcal{R}} \cup Set_{\mathcal{O}}, X_i = x)\,.$$

We similarly note

$$\pi_{\mathcal{O}}(x) = P(\exists i, Q_i = \mathcal{O}, X_i = x \mid \exists i \in Set_{\mathcal{R}} \cup Set_{\mathcal{O}}, X_i = x) = 1 - \pi_{\mathcal{R}}(x)\,.$$

Finally, we denote by $\alpha(x)$ the conditional odds that an individual with covariates x is in the RCT or in the observational cohort:[2]

$$\alpha(x) = \frac{P(i \in \mathcal{R} \mid \exists i \in Set_{\mathcal{R}} \cup Set_{\mathcal{O}}, X_i = x)}{P(i \in \mathcal{O} \mid \exists i \in Set_{\mathcal{R}} \cup Set_{\mathcal{O}}, X_i = x)} = \frac{\pi_{\mathcal{R}}(x)}{\pi_{\mathcal{O}}(x)} = \frac{\pi_{\mathcal{R}}(x)}{1 - \pi_{\mathcal{R}}(x)}\,.$$

This quantity arises in several approaches that have been proposed to generalize a treatment effect from an RCT to a target population, see for example Westreich et al. (2017). As we will see in the following, this conditional odds is identifiable under certain assumptions and can be used instead of the selection score π_S to generalize a treatment effect. Indeed the latter is only identifiable in the case of a nested trial design (Dahabreh et al., 2021) which we do not cover in this work.

2.2 Assumptions for identifiability of the ATE on the target population in the full data case

The main identifiability assumptions that allow for generalizing an ATE from the RCT onto a target population are as follows:

- Internal validity of the RCT:
  - Consistency of potential outcomes $Y = A\,Y(1) + (1 - A)\,Y(0)$
  - Treatment randomization $Y(a) \perp\!\!\!\perp (A, X) \mid S = 1$ for all X and $a = 0, 1$.
- Generalizability of the RCT to the target population:
  - Ignorability on trial participation

$$\{Y(0), Y(1)\} \perp\!\!\!\perp S \mid X. \tag{2}$$

  - Positivity of trial participation

$$\begin{aligned}&\exists\, c \text{ such that for all } x,\ P(\pi_S(x) \geq c > 0) = 1,\\ &\text{and } 0 < P(A = a \mid X = x, S = 1) < 1 \text{ for all } a \text{ and for all } x \text{ such that } P(S = 1 \mid X = x) > 0.\end{aligned} \tag{3}$$

[2] In the statistical and econometric literature, this term is sometimes also referred to as *conditional odds ratio* even though, by definition, it is an odds and not a ratio of odds.

Assumption (2) implies that we have measured all variables related to the trial eligibility indicator S that are treatment effect modifiers. In other words, participation in the RCT is randomized within levels of X. This is analogous to the *ignorability assumption* on treatment assignment in causal inference with observational data. With Assumption (3) we require adequate overlap of the covariate distribution between the trial sample and the target population, as well as between the treatment groups in the trial.

2.3 Estimators in the full data case

The covariate distribution of the RCT sample is generally different from that of the target population; therefore, τ_1 is different from τ, and an estimator based solely on the RCT is biased for the ATE of interest τ. Under the previous identifiability assumptions, different estimators are available to estimate the ATE τ: one proposes to reweight the RCT sample so that it "resembles" the target population with respect to the observed shifted covariates and treatment effect modifiers. Another proposes to model the conditional outcomes with and without treatment in the RCT, and then to apply the model to the target population of interest. Doubly robust approaches are combinations of the former two, improving the robustness and efficacy. For simplicity, we assume a standard RCT with a constant propensity score of 0.5 for all individuals.

Inverse probability of sampling weighting (IPSW). This estimator is defined as the weighted difference of average outcomes between the treated and control groups in the trial. The observations are weighted by the inverse odds $1/\alpha(x) = \pi_{\mathcal{O}}(x)/\pi_{\mathcal{R}}(x)$ to account for the shift of the covariate distribution from the RCT sample to the target population. The IPSW estimator can be written as follows:

$$\widehat{\tau}^{IPSW} = \frac{2}{m}\sum_{i=1}^{n} \frac{Y_i\,(2A_i - 1)}{\widehat{\alpha}(X_i)}\,, \tag{4}$$

where $\widehat{\alpha}$ is an estimate of α. The IPSW estimator is consistent as soon as α is consistently estimated by $\widehat{\alpha}$.

Conditional outcome-based estimation. It fits models of the conditional response surfaces among trial participants. Applying these models to the covariates of the observational data, gives the corresponding expected outcome (Robins, 1986). This outcome-model-based estimator, also called g-formula estimator, is then defined as:

$$\widehat{\tau}^{CO} = \frac{1}{m}\sum_{i=n+1}^{n+m} \left(\widehat{\mu}_{1,1}(X_i) - \widehat{\mu}_{0,1}(X_i)\right), \tag{5}$$

where $\widehat{\mu}_{a,1}(X_i)$ is an estimator of $\mu_{a,1}(X_i)$. If the model is correctly specified, then the estimator is consistent.

Doubly robust estimators The selection score and outcome models used in the first two estimators can be combined to form an augmented IPSW estimator (AIPSW):

$$\begin{aligned}\widehat{\tau}^{\text{AIPSW}} = \frac{2}{m}\sum_{i=1}^{n} \frac{1}{\widehat{\alpha}(X_i)} \left[A_i\left\{Y_i - \widehat{\mu}_{1,1}(X_i)\right\} - (1-A_i)\left\{Y_i - \widehat{\mu}_{0,1}(X_i)\right\}\right] \\ + \frac{1}{m}\sum_{i=n+1}^{m+n} \left\{\widehat{\mu}_{1,1}(X_i) - \widehat{\mu}_{0,1}(X_i)\right\}.\end{aligned}$$

It is doubly robust, i.e., consistent and asymptotically normal when either one of the two models for $\widehat{\pi}_{\mathcal{R}}$ and $\widehat{\mu}_{a,1}(X)$ $(a = 0, 1)$ is consistent.

Calibration weighting. It is well known that IPSW is likely to be unstable as soon as some of the estimated odds are very small. To resolve the instability of IPSW calibration weighting (Dong et al., 2020) have been proposed. They calibrate, i.e. weight subjects in the RCT sample in such a way that afterwards, the covariates are balanced between the RCT sample and the target population: usually the balance is enforced on the first and second moments of the covariates $X_1, \ldots, X_p$ such that the weighting mean and variance for each variable in the RCT match the ones in the observational data. More precisely, in order to calibrate, they assign an entropy-balancing weight ω_i to each subject i in the RCT sample obtained by solving an optimization problem:

$$\begin{aligned} &\min_{\omega_1,\ldots,\omega_n} \sum_{i=1}^{n} \omega_i \log \omega_i, \\ \text{subject to} \quad &\omega_i \geq 0, \text{ for all } i, \\ &\sum_{i=1}^{n} \omega_i = 1, \sum_{i=1}^{n} \omega_i \, \mathbf{g}(X_i) = \widetilde{\mathbf{g}}, \text{ (the balancing constraint)} \end{aligned} \tag{6}$$

where $\widetilde{\mathbf{g}} = m^{-1} \sum_{i=n+1}^{m+n} \mathbf{g}(X_i)$ is a consistent estimator of $\mathbb{E}[\mathbf{g}(X)]$ from the observational sample. The balancing constraint calibrates the covariate distribution of the RCT sample to the target population in terms of $\mathbf{g}(X)$. The objective function in (6) is the negative entropy of the calibration weights; thus, minimizing this criterion ensures that the empirical distribution of calibration weights are not too far away from the uniform, such that it minimizes the variability due to heterogeneous weights. Based on the calibration weights, the CW estimator is then defined as

$$\widehat{\tau}^{\mathrm{CW}} = 2 \sum_{i=1}^{n} \widehat{\omega}_i Y_i (2A_i - 1). \tag{7}$$

This estimator is doubly robust in that it is a consistent estimator for τ if either the selection score follows a log-linear model, or if the CATE (1) is linear in the calibration constraint.

2.4 Missing values mechanisms

In the taxonomy proposed by Rubin (1976), an ignorable missingness mechanism is either *missing completely at random* (MCAR) or *missing at random* (MAR). The former means that the missingness mechanism is independent of the data, whereas the latter states that the missingness only depends on the observed values. In a nutshell, ignorable missingness means that the missing data mechanism can be "ignored" when doing inference for the parameter in the data likelihood function since the two are separable in the full data likelihood function, whereas non-ignorable missingness complicates analyses more significantly.
More formally, we denote the response pattern of the i-th sample as $R_i \in \{0, 1\}^p$ such that $R_{ij} = 1$ if X_{ij} is observed and $R_{ij} = 0$ otherwise. We model $1 - R_i$ as a random vector and its (conditional) distribution is known as the missing values mechanism. Additionally, for our problem, for all response patterns r and $X = (X_{obs(r)}, X_{mis(r)})$ the partition of the data in realized observed and missing values given a specific realization of the pattern,

$$(MCAR) \quad \forall\, r,\ P(R = r|X, S, Q, A, Y) = P(R = r) \tag{8}$$

$$(MAR) \quad \forall\, r,\ P(R = r|X, S, Q, A, Y) = P(R = r|X_{obs(r)}, S, Q, A, Y) \tag{9}$$

If the missingness mechanism is non-ignorable, it is qualified as *missing not at random* (MNAR) and it formally states that the mechanism does not satisfy (8) or (9), in other words the missingness is allowed to depend on the missing values themselves.
In Figure 2 we provide an example of observed incomplete data and recall the underlying data for our problem.

	Q	Covariates			Treatment	Outcome under A=0	Outcome under A=1
		X_1	X_2	X_3	A	$Y(0)$	$Y(1)$
1	$\mathcal{R}$	1.1	20	5.4	1	23.4	24.1
...	$\mathcal{R}$		...		...	...	...
$n-1$	$\mathcal{R}$	-6	45	8.3	0	26.3	27.6
n	$\mathcal{R}$	0	15	6.2	1	28.1	23.5
$n+1$	$\mathcal{O}$	-2	52	7.1	NA	NA	NA
$n+2$	$\mathcal{O}$	-1	35	2.4	NA	NA	NA
...	$\mathcal{O}$		...		NA	NA	NA
$n+m$	$\mathcal{O}$	-2	22	3.4	NA	NA	NA

	Q	Covariates			Treatment	Outcome under A
		X_1^*	X_2^*	X_3^*	A	Y
1	$\mathcal{R}$	1.1	20	NA	1	24.1
...	$\mathcal{R}$		...		...	...
$n-1$	$\mathcal{R}$	-6	NA	8.3	0	26.3
n	$\mathcal{R}$	0	15	6.2	1	23.5
$n+1$	$\mathcal{O}$	-2	52	NA	NA	NA
$n+2$	$\mathcal{O}$	-1	NA	2.4	NA	NA
...	$\mathcal{O}$		...		NA	NA
$n+m$	$\mathcal{O}$	NA	NA	3.4	NA	NA

Figure 2: Example of data structure in the incomplete data problem setting. Left: complete underlying data with potential outcomes in the RCT. Right: observed incomplete data with factual outcomes.

3 Multiple imputation

3.1 General concept

Multiple imputation (MI) is one of the most powerful approaches to estimate parameters and their variance from incomplete data (Little and Rubin, 2014; Kim and Shao, 2013; Schafer, 2010). In a nutshell, for a single dataset, it consists in generating M plausible values for each missing entry, which leads to M completed datasets. Then, an analysis is performed on each imputed data set $m = 1, \ldots, M$, to get an estimate for the parameter of interest, say θ as $\widehat{\theta}^m$ and an estimate of its variance $\widehat{V}^m(\widehat{\theta}^m)$ and the results are combined using Rubin (1987)'s rules to get correct inference with missing values, namely confidence intervals with the appropriate coverage.

3.2 Adapted multiple imputation for multiple data sources with different data design

For our problem, there are multiple possibilities to derive a multiple imputation strategy to generalize a treatment effect. This is due to the multi-source structure of the data and the fact that there is an additional complication due to the number of variables are not the same in the RCT and in the observational study. Indeed, we assume that the observational study does not include treatment and outcome but only covariates. We stress again that we assume missing values only occur in the covariates of both data. We suggest and describe three strategies to tackle this problem:

1. Within-study multiple imputation, see also Figure 3a:
   (a) Multiple imputation of the RCT: Impute M times the covariates of the RCT using $(X_i, A_i, Y_i)_{i:\, Q_i=\mathcal{R}}$.
   (b) Multiple imputation of the observational data: Impute M times the covariates of the observational data using only the covariates $(X_i)_{i:\, Q_i=\mathcal{O}}$.

(c) Create $M \times M$ complete tables by concatenating all possible combinations of imputed RCT and observational data. Estimate the treatment effect on every combination using any complete case estimator as in Section 2 and aggregate these estimations using Rubin's rules.

2. Ad-hoc joint covariates multiple imputation, ignoring the group variable, see also Figure 3b:

   (a) Impute M times the joint datasets (the concatenation of the covariates from the RCT and the ones from the observational study) with covariates X and ignore the "source" indicator variable Q during the imputation.

   (b) Concatenate the outcome Y and treatment A for each imputed RCT.

   (c) Compute the M treatment effect estimators using any complete case estimator as in Section 2 and aggregate them using Rubin's rules.

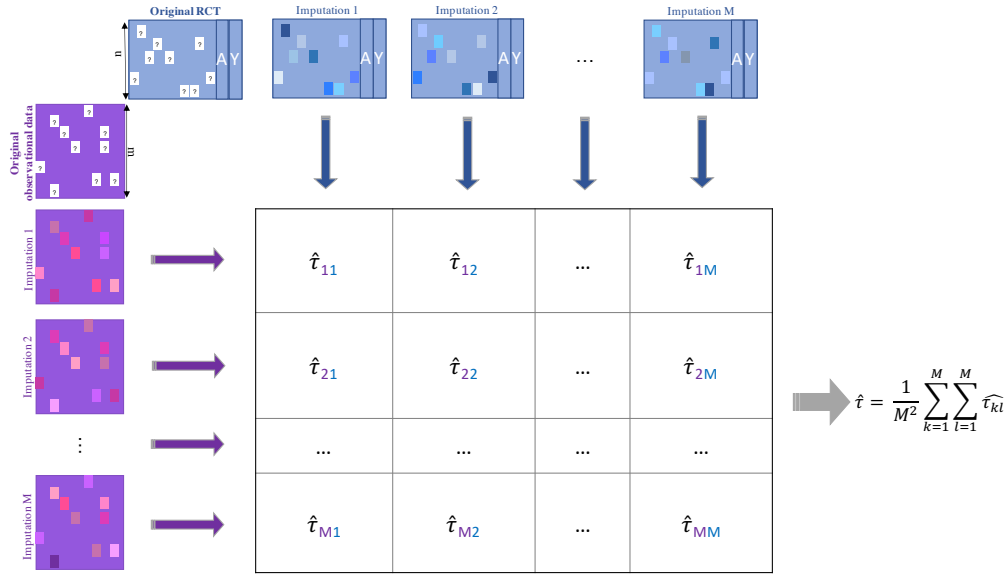
3. Joint covariates multiple imputation, modeling the group variable as a fixed effect, i.e. keep the indicator variable Q indicating the corresponding "group"/"source" during the imputation, see also Figure 3c:

   (a) Impute M times the joint datasets with covariates X and model the source indicator variable Q as a fixed effect during the imputation.

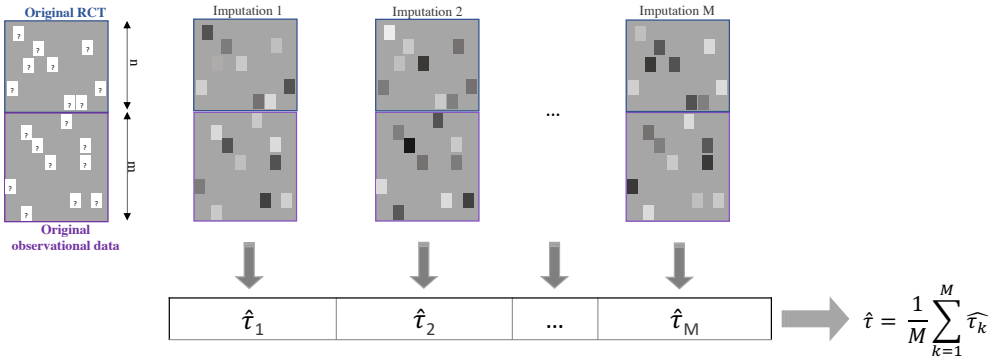
   (b) Concatenate the outcome Y and treatment A for each imputed RCT.

   (c) Compute the M treatment effect estimators using any complete case estimator as in Section 2 and aggregate them using Rubin's rules.

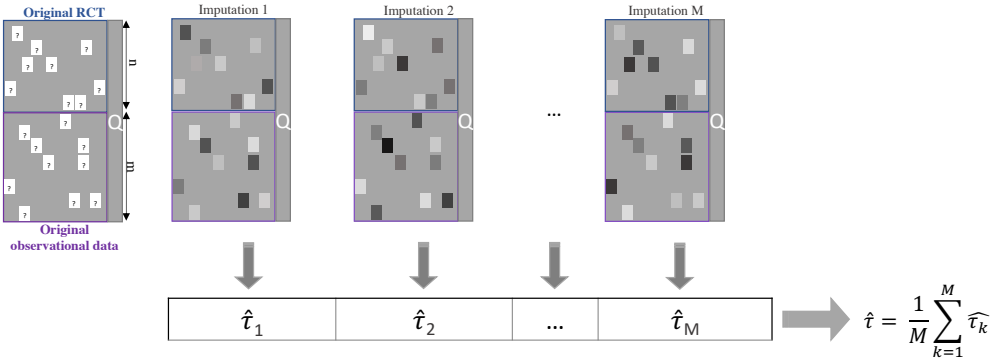
The first strategy has the advantage that it takes into account the outcome Y and treatment A which are dependent variables of the covariates X, when imputing the covariates of the RCT as suggested by Leyrat et al. (2019); Seaman and White (2014); Mattei and Mealli (2009) as it models the entire joint distribution. The other strategies only consider the covariates X. The second strategy solely relies on the relationships between the covariates X and the assumption that these are stable across the data sources, i.e., $Cov(X) = Cov(X|S = 1)$. The third strategy allows to additionally take into differences between both sources when imputing by modeling the source as a fixed effect. Strategy 3 can thus be seen as a fixed effect method, where the variable Q is included as a variable in the imputation model which allows, e.g., in case of multiple imputation with conditional regression models, to impute according to an analysis of covariance model. A drawback of this approach is that it generally inflates the between-group variability (Andridge, 2011).
In the case of several observational data sources, a fourth strategy could be considered, namely a multilevel multiple imputation approach, adding random effects into the model for each (observational) data source. Indeed, multilevel multiple imputation is specifically designed for cases of hierarchical or clustered observations, allows for a random intercept between groups (or, in our case, data sources) (van Buuren, 2018; Burgess et al., 2013; Audigier et al., 2018). It could be also appropriate when the observational data consists of patient records coming from different hospitals. It is indeed useful to encode the clustered structure of these records to account for between-hospital variability in terms of patient population and treatment practices. For a broader overview of multilevel imputation, we refer to Audigier et al. (2018); van Buuren (2018). All three strategies can be implemented easily using the R package mice (van Buuren, 2018) which uses conditional models such as linear and logistic regressions to perform multiple imputation. The sketched fourth strategy could be implemented using the `micemd` R package (Audigier et al., 2018).



(a) Within-study multiple imputation.



(b) Ad-hoc joint covariates multiple imputation.



(c) Joint covariates multiple imputation, modeling the source variable as fixed effect.

Figure 3: Schematic illustrations of different multiple imputation strategies.

Multiple imputation is suited if the missing values are ignorable as described in Section 2.4 and if the identifiability assumptions of the ATE τ in the full data case are met, namely Assumptions (2) and (3).
Note that for the considered imputation strategies, increasing the number of imputations does not improve (significantly) the final result. There exists no clear rule about the number of multiple imputations to achieve good performances, however an accepted rule of thumb is to choose the number of imputations to be similar to the percentage of incomplete cases (Hippel, 2009) or to the average percentage of missing data (van Buuren, 2018).

4 Missing incorporated in attributes under adapted ignorability assumption

Multiple imputation makes classical identifiability assumptions of the causal effect and ignorability assumptions of the missing values mechanism (Seaman and White, 2014; Leyrat et al., 2019). An alternative to handle missing covariates values consists in modifying the identifiability assumptions so that they directly handle missing values but do not necessarily require assumptions on the missing values mechanism. This can be seen as an advantage as it possibly allows for MNAR data, but the new identifiability assumptions may be more difficult to satisfy than in the full data case. More precisely, we extend the work of Mayer et al. (2020) who adapt the unconfoundedness assumption to the incomplete covariates in order to identify the (average) treatment effect in the observational data case. In the following we lay out the modified assumptions for generalizability of the RCT to the target population we make to achieve a similar identifiability in the case of missing values in the RCT and the observational data.[3] We then explain how to generalize treatment effects under these assumptions. First, we introduce an additional notation, required for the following approach: The matrix of observed covariates can be written with $X_i^* \triangleq X_i \odot R_i + \texttt{NA} \odot (\mathbf{1} - R_i)$, with $\odot$ the element-wise multiplication and $\mathbf{1}$ the matrix filled with 1, so that X_i^* takes its value in the half discrete space $\mathcal{X}^* \triangleq \underset{1 \leq j \leq |\mathcal{X}|}{\times} \{\mathcal{X}_j \cup \{\texttt{NA}\}\}$.
And, similar to D'Agostino and Rubin (2000); Mayer et al. (2020), we define the generalized conditional response surfaces μ_a^* and $\mu_{a,1}^*$ as follows:

$$\begin{aligned} \mu_a^*(x^*) &= \mathbb{E}(Y(a) \mid X^* = x^*), \\ \mu_{a,1}^*(x^*) &= \mathbb{E}(Y(a) \mid X^* = x^*, S = 1), \end{aligned} \tag{10}$$

The possibility to infer causal effects and to generalize the effect(s) from the RCT to another (broader) target population in the presence of missing data, i.e., using observations $(X_i^*, A_i, Y_i, S_i = 1)_{i=1,\dots,n}$ and $(X_i^*)_{i=n+1,\dots,n+m}$, depends on the following additional assumptions on the joint law of $(X_i, A_i, Y_i(0), Y_i(1), S_i = 1, R_i)_{i=1,\dots,n}$ and $(X_i, R_i)_{i=n+1,\dots,n+m}$.

- **Ignorability on trial participation, conditionally independent selection (CIS)**

$$\text{Assumption (2) and } \{Y(0), Y(1)\} \perp\!\!\!\perp S \mid X^*. \tag{11}$$

- **Positivity of trial participation**

$$\begin{aligned} &\text{Assumption (3) and } \exists\, c^* \text{ such that for all } x^*,\ P(\pi_S(x^*) \geq c > 0) = 1, \\ &\text{and } 0 < P(A = a \mid X^* = x^*, S = 1) < 1 \text{ for all } a \text{ and for all } x^* \text{ such that } P(S = 1 \mid X^* = x^*) > 0. \end{aligned} \tag{12}$$

[3] Note that the assumptions for internal validity of the RCT are the same as in the full data case.

Assumption (11) means that being eligible to the RCT does not affect the potential outcomes conditionally on the covariates X^*.
The intuition behind these additions is to assume that instead of requiring conditional independence of potential outcomes and trial eligibility conditionally on all covariates, we only require conditional independence conditionally on the *observed* information, meaning the observed values and the pattern of missing values. Taking the example given in Figure 2, for observation 1, only X_1 and X_2 and the fact that X_3 is unobserved are decisive for trial eligibility, while for observation 2, only X_1 and X_3 and the fact that X_2 is missing decide upon trial eligibility, etc. A possible scenario could be the existence of a list of sufficient but not necessary trial eligibility criteria. In other words, one could imagine a "check list" of L conditions and it is necessary to fulfill at least $l < L$ of these to be eligible.
Similar to the UDM assumption in Mayer et al. (2020), Assumption (11) can be replaced by two sufficient assumptions: Assumption (2) and $S \perp\!\!\!\perp X \mid X^*$, thus the term *conditionally independent selection*.

4.1 Generalized parameters and estimators

Since the conditional response surfaces μ_a^* and $\mu_{a,1}^*$ now depend on X^* rather than X and thus depends on the pattern of missing values, the estimators that involve an estimation of these quantities require an adaptation. Analogously to the generalized response surfaces, on can also define the generalized conditional odds α^*, following the same logic.

$$\mu_a^*(x^*) = \mathbb{E}\left[Y(a) \mid X^* = x^*\right], \qquad \mu_{a,1}^*(x^*) = \mathbb{E}\left[Y(a) \mid X^* = x^*, S = 1\right]. \tag{13}$$

The resulting estimators are then formed analogously to the estimators in the full data case, by substituting the corresponding nuisance or intermediary parameters with their generalized counterparts. More explicitly, the outcome-model-based estimator defined by (5) becomes

$$\widehat{\tau}^{CO,*} = \frac{1}{m} \sum_{i=n+1}^{n+m} \left(\widehat{\mu}_{1,1}^*(X_i^*) - \widehat{\mu}_{0,1}^*(X_i^*)\right). \tag{14}$$

Fitting the new nuisance or intermediary parameters (13) is not straightforward, since these require to fit a separate regression model for each possible pattern r of missing values. For example, if we have three incomplete covariates X_1, X_2, X_3, this means we need to fit a separate regression on $\{X_1, X_2, X_3\}$, on $\{X_1, X_2\}$, on $\{X_2, X_3\}$, on $\{X_1\}$, etc. We can see from this example that this is not possible in moderate and high dimensions with classical regression methods. This is why we propose to use random forests with a splitting criterion adapted to missing data. Indeed, as noted already by Athey et al. (2019); Mayer et al. (2020), many modern machine learning methods, including tree ensembles and neural networks, can be adapted to this context and thus readily handle missing data and enable direct fitting of the generalized models above (Josse et al., 2019).

Nonparametric estimation. As an example of such a modern nonparametric estimation approach, we propose to estimate the generalized parameters via random forests (Breiman, 2001; Athey et al., 2019), with missing data handled using the *missing incorporated in attributes* (MIA) method of Twala et al. (2008). The resulting IPSW, CO and AIPSW estimators will be denoted by $\hat{\tau}_{MIA}^{IPSW}$, $\hat{\tau}_{MIA}^{CO}$, and $\hat{\tau}_{MIA}^{AIPSW}$ respectively.
In random trees, the MIA approach extends the classical splitting rules such that missing values are incorporated in the splitting criterion. More specifically, consider splitting on the j-th at-

tribute 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 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 works with arbitrary missingness mechanisms and does not require the missing data to follow a specific mechanism. This MIA approach for (generalized) random forests is implemented in the R package `grf` (Tibshirani et al., 2020) which is also used in the simulation part of this work presented in Section 5.

Parametric alternative. Parametric estimation is however possible in the case of logistic and linear regression models. This is based on work by Jiang et al. (2020) and Schafer (2010) for logistic and linear regressions with missing covariates. The functions μ^* and α^* that take in incomplete covariates x^* are estimated via EM (Dempster et al., 1977). The resulting IPSW, CO and AIPSW estimators will be denoted by $\hat{\tau}_{EM}^{IPSW}$, $\hat{\tau}_{EM}^{CO}$, and $\hat{\tau}_{EM}^{AIPSW}$ respectively.
The details of this approach are given in the Appendix A.
However, a major limitation of this approach is that, in addition to the modified identifiability assumptions (11) and (12), in order to justify the use of the EM algorithm, one typically needs to make further assumptions on the missing value mechanism; in particular, this approach assumes the MAR mechanism (9). 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 the generalized conditional models α^* and $\mu_{a,1}^*$.

5 Simulations

We conduct a detailed simulation study to assess the performance of the previously introduced estimators to handle missing values. This controlled study allows to quantify the impact of different missing values mechanisms and identifiability assumptions on the final estimate for the ATE τ.

5.1 Data generation

5.1.1 Classical assumptions for identifiability in treatment effect generalization and on missing values mechanisms

We consider the selection model generated as a logistic model as follows

$$\text{logit}\left\{\pi_S(X)\right\} = -2.5 - 0.5X_1 - 0.3X_2 - 0.5X_3 - 0.4X_4, \tag{15}$$

where every X is drawn from a multivariate normal distribution with mean 1 and covariance matrix Σ such that $\Sigma_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0.6 & \text{if } i \neq j \end{cases}$ to have correlated covariates. The outcome is generated according to the linear model below such that X_1 is a treatment effect modifier and the true ATE τ is set to 27.4.

$$Y(a) = -100 + 27.4aX_1 + 13.7X_2 + 13.7X_3 + 13.7X_4 + \epsilon \quad \text{with } \epsilon \sim \mathcal{N}(0,1). \tag{16}$$

We do not modify the treatment assignment mechanism since by assumption it is independent of the rest and in standard RCT design it is constant for all individuals. Missing values in the covariates are generated as follows, defining different models for the response indicator R (see Section 2.4):

- Missing values can occur in all four covariates.
- Proportion of missing values in each incomplete covariate: 20%.
- The missing values mechanism can be one out of the following and is implemented in the `produce_NA` function proposed by Mayer et al. (2019):
  - MCAR where the probability to have missing values does not depend on any variable:
    $$P(R_{i\cdot} = r|X_i, S_i, Q_i, Y_i, A_i) = P(R_{i\cdot} = r) \tag{17}$$
    We choose $P(R_{ij} = 0) = 0.2$ for all $i \in \{1, \ldots, n+m\}$, and $j \in \{1, 2, 3, 4\}$.
  - MAR:
    $$P(R_{i\cdot} = r|X_i, S_i, Q_i, Y_i, A_i) = f(X_{obs(r)}), \tag{18}$$
    for some function $f : \mathcal{X} \to [0, 1]$, for all $i \in \{1, \ldots, n+m\}$. For example, missing values in X_1 are introduced for observation i using a logistic model on X_2, X_3, X_4, assuming these three variables are observed for observation i.
  - MNAR:
    $$P(R_{ij} = 0|X_i, S_i, Q_i, Y_i, A_i) = g(X_{ij}), \tag{19}$$
    for some function $g : \mathcal{X}_j \to [0, 1]$, for all j and $i \in \{1, \ldots, n+m\}$. In this study, we use a self-masking MNAR mechanism, i.e., the missingness of a variable depends on its value alone. More precisely, we use an upper quantile censorship approach. The quantile level q is chosen such that when missing values are generated on the q-quantile at random, the requested proportion of missing values is achieved. For more details about this chosen approach, we refer to the documentation of the `produce_NA` function[4].

We assume that the trial selection, randomization and potential outcomes are completely independent from the missing values. The simulation design is summarized by Algorithm 1.

Algorithm 1: Steps for simulation design under the standard assumption

Result: Joint data table X^* of RCT and observational data and additional variables A, Y for the RCT.

1. Sample $N >> n$ observations $X_1, \ldots, X_N$ from the target population $\mathcal{P}(X)$;
2. Sample S according to the logistic model (15) on X;
3. Keep the $\{S = 1\}$ indexed observations $X_{\mathcal{R}} \leftarrow X_{\{i:\, S_i=1\}}$ as the RCT;
4. Sample A according to a Bernoulli distribution $\mathcal{B}(0.5)$ (coin flip);
5. Sample Y according to the linear model (16) on $X_{\mathcal{R}}$;
6. Sample m observations $X_{\mathcal{O}}$ from the target population $\mathcal{P}(X)$ as the observational data;
7. Concatenate the datasets: $X \leftarrow [X_{\mathcal{R}}^T,\ X_{\mathcal{O}}^T]^T$ and append the indicator Q to the data $(X \leftarrow [X\ ,\ Q])$;
8. Sample missing values for the $n+m$ observations according to either (17), (18) or (19);

[4] https://rmisstastic.netlify.app/how-to/generate/missSimul.pdf

For ease of comparison with the alternative simulation scenario, we provide the expression of the joint distribution over (X, R, A, Y, S, Q) factorized according to the underlying generative model:

$$\begin{aligned} p(X, R, A, Y, S, Q) &\propto p(R|S, Q, X, A, Y)p(S, Q, Y|X, A)p(A|X)p(X) \\ &\propto p(R|S, Q, X, A, Y)p(Q|S)p(S|X, A)p(Y|X, A)p(A|X)p(X) \end{aligned} \tag{20}$$

In the MCAR case (8), this factorization simplifies as follows:

$$p(X, R, A, Y, S, Q) \propto p(R)p(Q|S)p(S, Y|X, A)p(A|X)p(X).$$

5.1.2 Modified assumptions on treatment effect generalization

We also generate data according where the CIS assumption (11) is met. In such scenarios, we expect that the methods described in section 4 will work best.
The main difference with the previous setting of simulation, lies in the definition of the selection model, the outcome model remaining unchanged. To relate this setting to the previous one, we begin by giving the expression of the factorization of the joint distribution under these assumptions on the data generating process.

$$\begin{aligned} p(X, R, A, Y, S, Q) &\propto p(S, Q|X, R, A, Y)p(R|X, A, Y)p(Y|X, A)p(A|X)p(X) \\ &\propto p(Q|S)p(S|X, R, A, Y)p(R|X)p(Y|X, A)p(A|X)p(X) \end{aligned} \tag{21}$$

Note that we can see from this factorization, the difference w.r.t. the previous case induced by the modified transportability assumptions, and in particular the modified ignorability (11) which induces a (conditional) dependence between R and $\{S, Q\}$ by assumption.
In order to simulate data under the CIS assumption (11), we need to modify the definition of π_S such that it becomes pattern-dependent.

$$\text{logit}\left\{\pi_S(X)\right\} = -2.5 - 0.5X_1 \odot R_1 - 0.3X_2 \odot R_2 - 0.5X_3 \odot R_3 - 0.4X_4 \odot R_4, \tag{22}$$

The simulation design and the adapted CIS ignorability assumption is summarized in Algorithm 2 and is different from the one described in Algorithm 1.

Algorithm 2: Steps for simulation design under the CIS assumption.

Result: Joint data table X of RCT and observational data and additional variables A, Y for the RCT.

1. Sample $N >> n$ observations $X_1, \ldots, X_N$ from the target population $\mathcal{P}(X)$;
2. Sample missing values for the N observations according to either (17), (18) or (19);
3. Sample S according to pattern-dependent logistic model on X;
4. Keep the $\{S = 1\}$ indexed observations $X_{\mathcal{R}} \leftarrow X_{\{i:\, S_i=1\}}$ as the RCT;
5. Sample A according to a Bernoulli distribution $\mathcal{B}(0.5)$ (coin flip);
6. Sample Y according to the linear model (16) on $X_{\mathcal{R}}$;
7. Sample m observations $X_{\mathcal{O}}$ from the target population as the observational data;
8. Sample missing values for the m observations $X_{\mathcal{O}}$ using the same mechanism as before but possibly with different proportions;
9. Concatenate $X_{\mathcal{R}}$ and $X_{\mathcal{O}}$ ($X \leftarrow [X_{\mathcal{R}}^T\ ,\ X_{\mathcal{O}}^T]^T$), and append the indicator Q to the data ($X \leftarrow [X\ ,\ Q]$);

5.2 Estimation methods

We consider different scenarios of data generating processes by varying the type of missing values (MCAR, MAR, MNAR), ignorability assumption (standard or CIS), as well as the number of observations.
We compare the following methods to handle missing values (the following acronyms are identical to the method labels used in Figures 4 –5:

- Full data: we apply the standard full data estimators from Section 2 on the full data before introducing missing values (this would serve as a reference).
- Complete cases (CC): we apply the standard full data estimators from Section 2 on the complete observations extracted from the incomplete data (by deleting observations with missing values).[5]
- EM (see Section 4.1): we use EM to fit logistic and linear regression models of α^* and $\mu^*_{a,1}$ on the incomplete data using the R package `misaem` (Jiang et al., 2020).
- MIA (see Section 4.1): we use generalized random forests with MIA splitting criterion to estimate the generalized models α^* and $\mu^*_{a,1}$ on the incomplete data, using the R package `grf` (Athey et al., 2019).
- Multiple imputation (MI, Section 3): we apply the standard full data estimators from Section 2 on the imputed data (5-10 imputations obtained using the R package `mice`) where we use either
  - within-study multiple imputation (WI-MI), or
  - ad-hoc multiple imputation (AH-MI), or
  - fixed effect multiple imputation (FE-MI).

We do not assess the random effect multiple imputation (i.e., the joint covariates multiple imputation with a 2-level model accounting for the multiple data sources, sketched Strategy 4 from Section 3) in this simulation study because this does not correspond to our motivating data example from the introduction.
Note that for the EM and MIA approach, we only compute the IPSW, CO and AIPSW estimators (see Section 2.3) since the calibration weighting estimator in its current form is not applicable on incomplete data and future work is required to adapt this estimator to incomplete data.

5.3 Results

Due to the large number of different scenarios we consider in this simulation study, we first provide a summary of the expected behaviors of the different estimators in various cases before we report results of our experiments.

5.3.1 Summary of expected and empirical results

In Table 1 we summarize the expected results in terms of consistency of the different approaches, depending on the mechanisms that generate the data. For example, the MIA approach being suited for the modified ignorability assumption CIS (11), we expect it to perform well under CIS, independently of the missingness mechanism. In contrast, we do not expect it to be consistent

[5]Note that this approach is the most common default option in many implementations.

under the standard ignorability assumption (2), whereas multiple imputation should be consistent under this standard assumption, provided the missingness is either MCAR or MAR.

Table 1: Expected behavior under different assumptions about the data generating process and used estimation approach. Color code: blue=no bias; red=bias.

		Full data	Complete cases	EM	MIA	Multiple imputation
MCAR	CIS (Assmpt. 11)					
	I (Assmpt. 2)					
MAR	CIS (Assmpt. 11))					
	I (Assmpt. 2)					
MNAR	CIS (Assmpt. 11)					
	I (Assmpt. 2)					

This Table 1 is based on the Table 2 that summarizes the required assumptions for each method.

Table 2: Methods for handling incomplete observations in treatment effect generalization 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 good empirical performance.)

	Covariates		Missingness			Identifiability of generalized τ		Models for (S, Y)	
	multivariate normal	other	MCAR	MAR	MNAR	I (≡ (2) & (3))	CIS (≡ (11) & (12))	Generalized linear models	Non-parametric models
CC	✓	✓	✓	✗	✗	✓	✗	✓	✓
EM	✓	✗	✓	✓	✗	✗	✓	✓	✗
MIA	✓	✓	✓	✓	✓	✗	✓	✓	✓
MI	✓	✓	✓	✓	✗	✓	✓	✓	(✗)

Some simulations results are in agreement with what is expected but there are also some gaps between the expected and empirical behavior in the chosen simulation settings. Indeed, in Figures 4 and 5 we report the empirical bias with at 95% Monte Carlo confidence interval (based on a Monte Carlo standard error) of the estimated generalized ATE $\hat{\tau}$ relative to the true value $\tau = 27.4$, using $n_{sim} = 100$ repetitions of each scenario. We define the empirical bias and its Monte Carlo standard error respectively by

$$\widehat{B}_{\hat{\tau}} = \widehat{Bias}(\hat{\tau}) = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\tau}_i - \tau$$

$$\widehat{SE}(\widehat{B}_{\hat{\tau}}) = \sqrt{\frac{1}{n_{sim}(n_{sim}-1)} \sum_{i=1}^{n_{sim}} (\hat{\tau}_i - \overline{\tau})^2},$$

where $\overline{\tau} = \sum_{i=1}^{n_{sim}} \hat{\tau}_i$, following Morris et al. (2019).

- As expected, in Figure 4 we note that the full-data estimations are unbiased in all scenarios under the standard ignorability assumption (2). Under the CIS assumption, only the full-data estimators that (partly) rely on the outcome model, namely CO, AIPSW, and CW

are unbiased, whereas the parametric IPSW estimator fails under CIS. Surprisingly, the nonparametric full-data IPSW estimator recovers the true value in the MAR and MNAR case (see Figure 5).

- The complete case estimations are, unsurprisingly, unbiased only in the MCAR case.
- The behavior of the EM estimations is as expected: all estimators are unbiased under CIS under MCAR and MAR. However the AIPSW estimator performs better than the IPSW and the CO. In the MNAR case, the algorithm fails to converge.
- The MIA estimations overall have either small or no bias under CIS, especially the AIPSW estimator. Furthermore, under the CIS assumption, the AIPSW estimator always performs at least as well as the simple estimators (IPSW and CO). Under the standard ignorability assumption, the behavior is heterogeneous and tends toward biased results for all missingness mechanisms.
- Surprisingly, the within-study MI IPSW estimator is biased in all cases except the CIS+MCAR, and CIS+MNAR cases. This behavior is not expected and it remains to be investigated whether this is due to the simulation design.
- The joint fixed effect MI estimator (FE-MI) comes closer to the expected behavior of the multiple imputation approach than the ad-hoc MI (AH-MI) and the within-study MI (WI-MI) estimator as it has small or no bias under the standard ignorability and ignorable missingness (I+MCAR and I+MAR), but all three fail in the MNAR case (as expected).

Note that for the full data case, the choice of the estimator, namely parametric (in our case, generalized linear models) or non-parametric (here generalized random forests), has an impact on the bias, especially for the single-model estimators IPSW and CO. This is not surprising given the linear specification of the conditional odds and outcome models from (15) and (16) and the rather slow convergence of the chosen non-parametric method, random forest (`grf`), for linear models.[6]

In view of the results, if one has to recommend a method, it is preferable, from the present empirical evidence, to choose the MIA-AIPSW estimator or the joint multiple imputation coupled with the CW estimator. The MIA approach is simple to use in R thanks to the `grf` package which directly handles incomplete variables with the MIA splitting criterion for random forests. However, in terms of computational costs, this approach can be more expensive due to (automatic) parameter tuning. The joint multiple imputation approaches are easy to implement (with the `mice` package) but the relative computational running time is of similar magnitude as the MIA approach.

[6] The random forest approach would require a lot of data to estimate linear regression functions; random forests are however known for their good performance in the presence of non-linearities and high order interaction terms (Breiman, 2001).

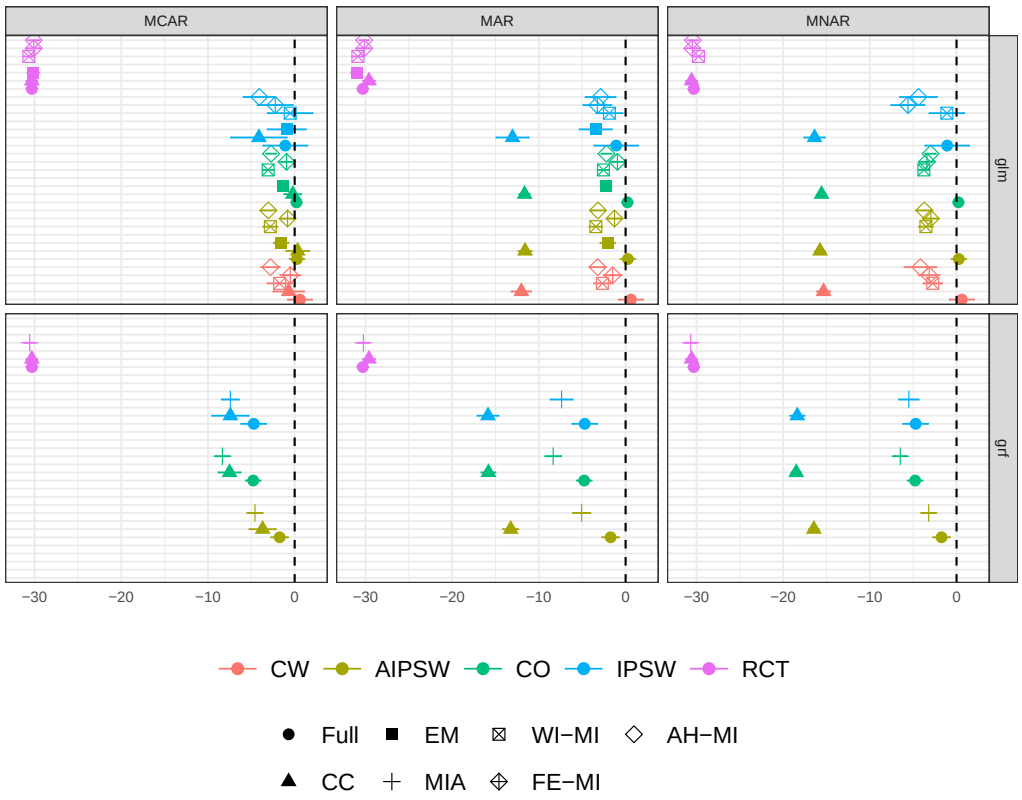


Figure 4: Empirical bias of generalizing ATE estimators under the *standard ignorability assumption*, 95% Monte Carlo confidence intervals, $n = 1000$.

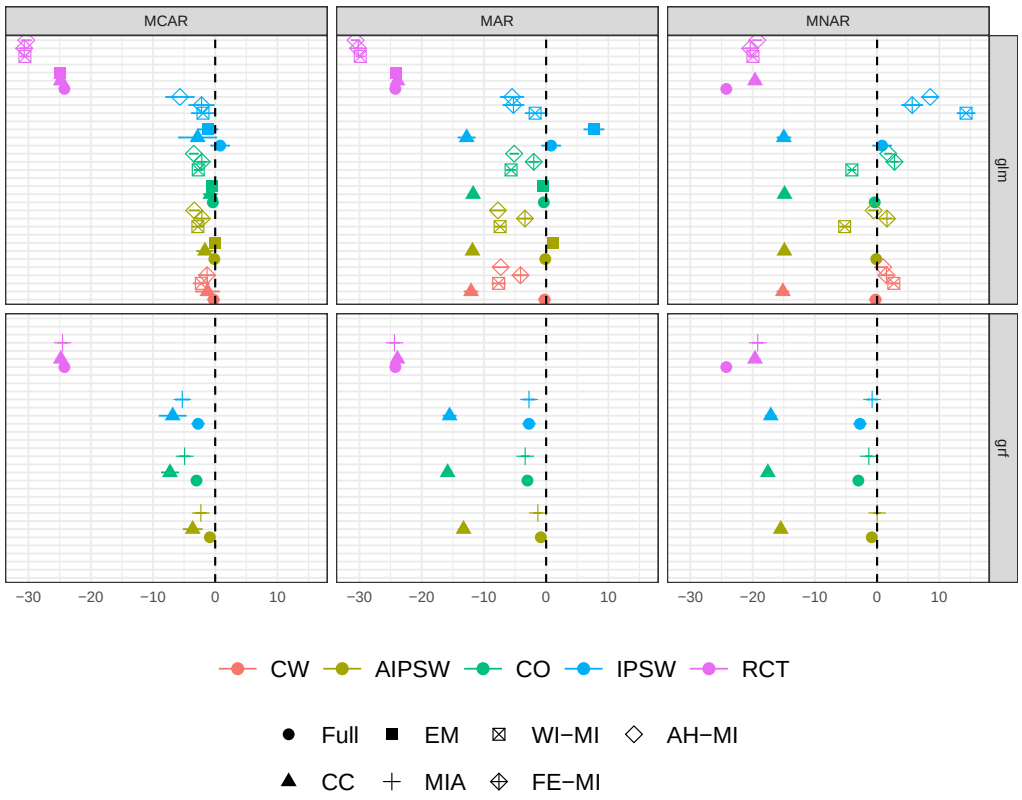


Figure 5: Empirical bias of generalizing ATE estimators under the *conditionally independent ignorability (CIS) assumption*, 95% Monte Carlo confidence intervals, $n = 1000$.

5.3.2 Impact of different proportions of missing values in the RCT and observational data

It is common that the RCT presents significantly less missing values than the observational study due to a more systematic monitoring of the data collection process. This invokes the question of how the above studied methods behave in the case of unbalanced proportions of missing values or different missing values mechanisms in the different data sources.
Extending the previous simulation study by such a case, we summarize in Figure 6 the performance of the different estimators under different scenarios of varying proportions of missing values in the RCT and the observational data when the data is MAR given S (or equivalently, we say it is MCAR in each data set). Note that this implies a slightly different factorization of the joint distribution over all variables than the one in the standard MCAR case (8):

$$\begin{aligned} p(X,R,A,Y,S) &\propto p(R|S,X,A,Y)p(S,Y|X,A)p(A|X)p(X) \\ &\propto p(R|S,X,A,Y)p(S|X,A)p(Y|X,A)p(A|X)p(X) \\ &\propto p(R|S)p(S|X,A)p(Y|X,A)p(A|X)p(X) \end{aligned}$$

As expected, the complete case estimators are unbiased in this special case since conditionally on S, the data is MCAR. The doubly robust multiple imputation estimators can cope with

very different proportions of missing values, either 10%, 50% in RCT and observational data respectively, or 5% and 22% respectively.
These results are supporting our claim that the previous results and methods apply as well to the likely case of different proportions of missing values in the two studies. Indeed the following data analysis in Section 6 is an example of this case.

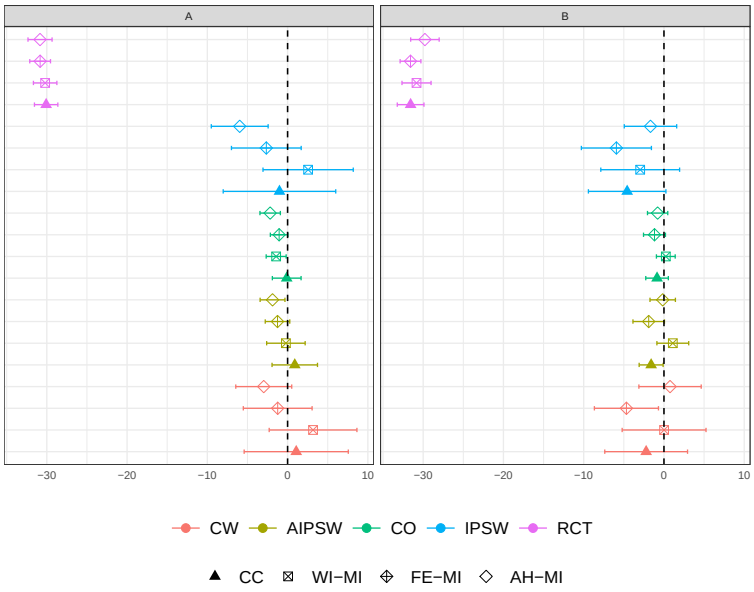


Figure 6: Bias estimates of generalized ATE under *standard ignorability* where missing values are **"study-wise MCAR"** ($\equiv$ MAR given S). For $n \in \{1000, 5000\}$, 20 repetitions. Case A=$\{$m=$10 \times n$, RCT=10% NA, Obs=50% NA$\}$; case B=$\{m = 10 \times n$, RCT=5% NA, Obs=22% NA$\}$.

6 Application on critical care data

In this part, we come back to the medical question introduced in the beginning of this work about the potential effect of tranexamic acid (TXA) on mortality in patients with intracranial bleeding. We recall that, in order to answer this question, we have at disposal two data sources: (1) CRASH-2, a multi-center international RCT, (2) Traumabase, an observational national registry.
A detailed data analysis of the observational registry to address the above medical question has been conducted by Mayer et al. (2020). We thus refer to this previous analysis for a detailed description of the observational registry as well as their findings. In a nutshell, leveraging only the observational registry does not provide evidence towards a beneficial (or detrimental) effect of TXA on trauma patients with TBI in terms of head-injury-related mortality.
We will first recall a summary of the findings of the original CRASH-2 study (Shakur-Still et al., 2009) before turning to focus on how the handling of missing values in the RCT and the observational registry impacts the final estimations of the population average treatment effect.

6.1 Findings of the CRASH-2 RCT

The CRASH-2 (Clinical Randomisation of Antifibrinolytic in Significant Haemorrhage) trial enrolled 20,211 patients in 274 hospitals in 40 countries between May 2005 and 2009 (Shakur-Still et al., 2009).
The aim of this trial was to study the effect of tranexamic acid in adult trauma patients with ongoing significant hemorrhage or at risk of significant hemorrhage, within 8 hours of injury (inclusion criteria), except those for whom antifibrinolytic agents were thought to be clearly indicated or clearly counter-indicated (exclusion criteria).[7].
More precisely, eligible patients were defined as trauma patients within 8 hours of the injury, of age at least 16 years:

- with ongoing significant hemorrhage (systolic blood pressure less than 90 mmHg and/or heart rate more than 110 beats per minute)
- or who are considered to be at risk of significant hemorrhage.

The inclusion criteria and other baseline regressors are summarized in the graph of Figure 7.

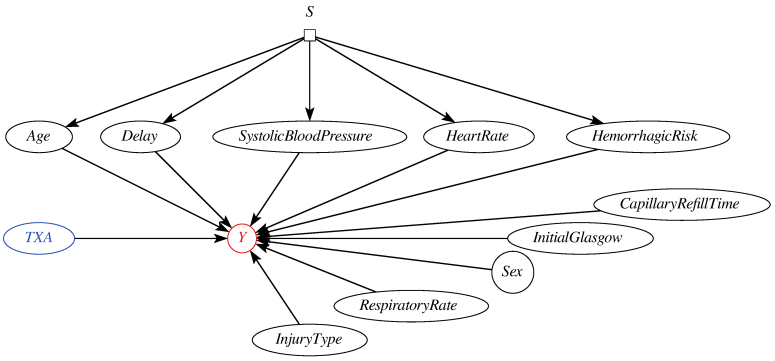


Figure 7: Causal graph of CRASH-2 trial representing treatment, outcome, inclusion criteria with S and other predictors of outcome (Figure generated using the Causal Fusion software by Bareinboim and Pearl (2016)).

The results of the CRASH-2 study are reported in Shakur-Still et al. (2009) and show a beneficial effect of TXA on the trial population for the primary outcome of interest (all-cause 28 day death).

6.2 Integration of the CRASH-2 trial and the Traumabase registry

In the following, we discuss common variables definition, outcome, treatment, and designs in order to leverage both sources of information. We recall the causal question of interest: "What is the effect of the TXA on brain-injury death on patients suffering from TBI?" This part is important for the harmonization of the study protocol.

Treatment exposure. The treatment protocol of CRASH-2 frames the timing and mean of administration precisely (a first dose given by intravenous injection shortly after randomization, i.e., within 8 hours of the accident, and a maintenance dose given afterwards (Shakur-Still et al., 2009)). The Traumabase study being a retrospective analysis, this level of granularity concerning

[7]Extract from the study protocol available at https://www.thelancet.com/protocol-reviews/05PRT-1.

TXA is unfortunately not available. Neither the exact timing, nor the type of administration are specified for patients who received the drug. However, the expert committee agreed that the assumption of treatment within 3 hours of the accident is very likely since this drug is administered in pre-hospital phase or within the first 30 minutes at the hospital (Mayer et al., 2020).

Outcome of interest. The CRASH-2 trial defined primary outcome as any-cause death in hospital within 28 days of injury. This outcome is also available in the Traumabase.

Covariates accounting for trial eligibility. For the CRASH-2 trial, four criteria determined inclusion: age (patients of at least 16 years old were eligible), ongoing or risk of significant hemorrhage (defined as systolic blood pressure below 90 mmHg or heart rate above 110 beats per minute, or clinicians evaluation of a risk), within 8 hours of injury and absence of a clear indication or counter-indication of antifibrinolytic agents. The necessary variables are also available in the Traumabase, either exactly or in form of proxies, which allows the estimation of the trial inclusion model on the combined data.

Additional covariates. Note that other covariates are (partially) available in both data sets, while not responsible of trial inclusion according to CRASH-2 investigators. But as this could still be covariates moderating the outcome and treatment effect, we include them in the outcome models used in the CO and AIPSW estimators (5) and (6). According to the two data sets, we could add three of them: sex (binary), type of injury (categorical, 1 =blunt, 2 = penetrating, 3 = blunt and penetrating), and initial Glasgow coma scale (numeric, integers from 3 to 15). Note that this three covariates are all mentioned in the baseline of CRASH-2 results (Shakur-Still et al., 2009), arguing that they should impact the outcome. The variables *central capillary refill time* and *respiratory rate* are also mentioned but are not available in the Traumabase, we thus omit them from this joint study.

Missing values. First, note that the RCT contain almost no missing values, whereas the variables for determining eligibility in the observational data contains important fractions of missing values, as shown in Table 8, while the sample sizes of the data sets are similar, see Table 3.

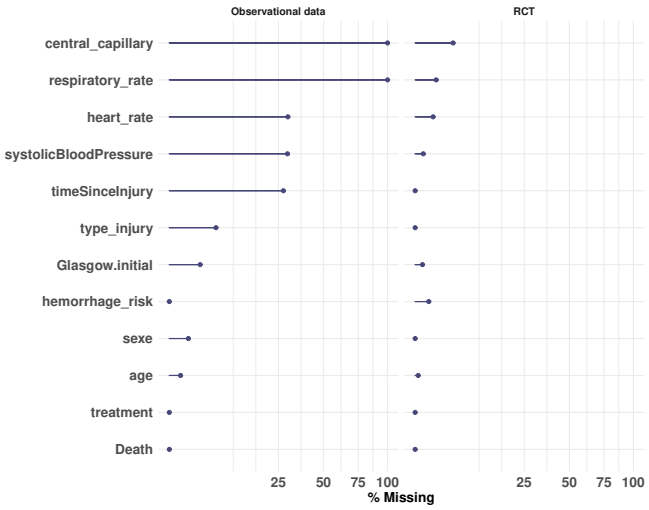


Figure 8: Percentages of missing values in each covariate for the Traumabase and CRASH-2 RCT.

Table 3: Sample sizes for the two studies.

	m	n	#treated	#all-cause 28d death
Traumabase	8248	–	686	1648
CRASH-2	–	3727	1866	1176

While the MCAR assumption is admitted for the RCT (Shakur-Still et al., 2009), the missing values in the observational Traumabase are more complex and, according to the medical experts monitoring the collection process, partly non-ignorable. For example, the pre-hospital systolic blood pressure (SBP) is likely to be missing for patients with severe ongoing bleeding. Since the latter is informed in the *hemorrhage risk* variable, we could admit the missing values in the SBP variable as being MAR. A similar reasoning can be applied for the delay between the accident and treatment administration. However, there remains uncertainty as to whether the observed variables allow to fully explain the missingness in this variable.

Distribution shift. There are different ways of assessing the shift between the distributions of the two studies, e.g., by univariate comparisons. We provide a simplified comparison of the means of the covariates between the treatment groups of the two studies in Figure 9. This graph illustrates again the fundamental difference between the two studies, namely the treatment bias in the observational study and the balanced treatment groups in the RCT, but also a covariate shift between the two studies. To provide an example, the average patient age in the RCT is 7-9 years below the average age in the observational study; and there are only 16% of female patients in the RCT while there are over 20% in the observational study.

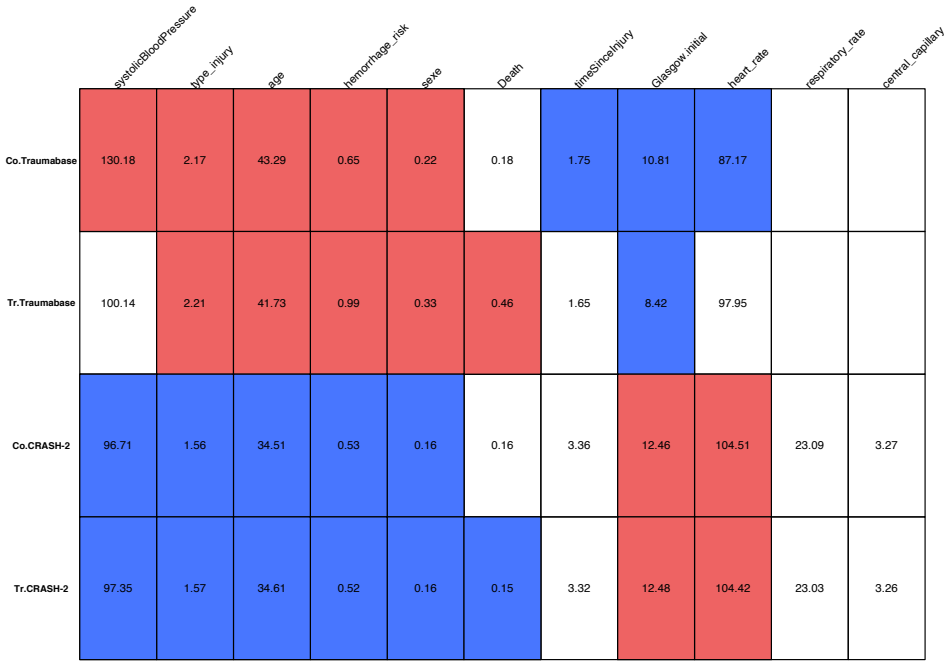


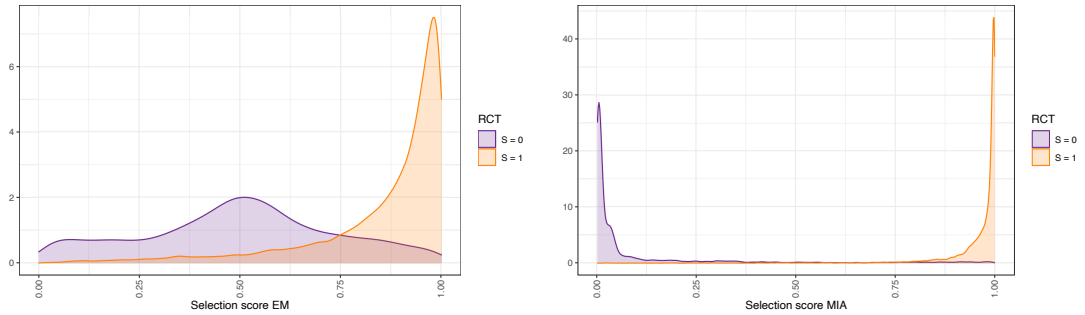
Figure 9: Distributional shift and difference in terms of univariate means of the trial inclusion criteria (red: group mean greater than overall mean, blue: group mean less than overall mean, white: no significant difference with overall mean, numeric values: group mean (resp. proportion for binary variables). Graph obtained with the `catdes` function of the `FactoMineR` package (Lê et al., 2008).

Ignorability. Due to the design of the CRASH-2 study, namely the eligibility criteria which all need to be informed to decide upon trial eligibility, the modified ignorability assumption CIS (11) is less plausible to hold in this case and we rather consider the standard assumptions (2) and (3) to be satisfied by the CRASH-2 and Traumabase studies.[8] The additional assumptions concerning the missing values mechanism(s) have been outlined above and we consider that the assumptions for the multiple imputation strategy to be applicable are sufficiently plausible in this real-world example.

The distribution of the estimated selection scores are given in Figures 10a (logistic regression via EM), 10b (generalized random forest with MIA), and 10c (logistic regression on joint fixed effect multiple imputations, MI). We notice that the scores obtained using EM and MI are similar and suggest that the positivity assumption is satisfied since we observe a good degree of overlap between the distributions of the scores for the two data sets. The scores estimated via MIA

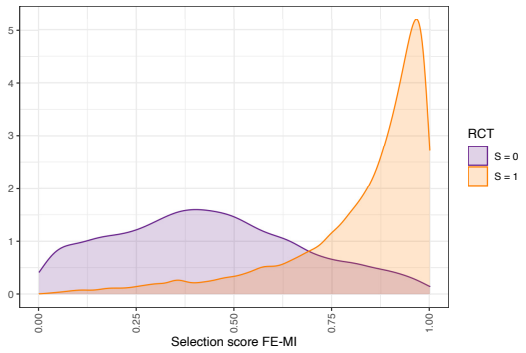
[8]The CIS assumption may be plausible in a context where trial eligibility (and participation) are defined via a set of sufficient conditions that are not all necessary conditions. More specifically, if we consider five covariates $X_1, \ldots, X_5$ and assume that there exist different possibilities to be included in the trial, e.g., a condition on X_1, X_2 and X_3 regardless of X_4 and X_5 and another alternative condition on X_2 and X_5. In the case of such a design, the CIS assumption is potentially a suited assumption to generalize a treatment effect onto another population.

however concentrate around 0 and 1 for the observational and RCT observations respectively, suggesting poor overlap under this model. In Appendix B, we provide further comparisons of the estimated selection scores, pointing towards the multiple imputation strategy as the suited approach in this case.
These results provide an additional argument in favor of the multiple imputation strategy which appears to be more adapted to the handling of missing values in this analysis; in particular we will apply a joint fixed effect multiple imputation strategy since it outperforms the other multiple imputation strategies in the simulation study of Section 5.



(a) EM. Obtained with the `misaem` R package.

(b) MIA. Obtained with the `grf` R package.



(c) Joint fixed effect MI. Obtained with the `mice` R package.

Figure 10: Estimated densities of the fitted selection scores.

6.3 Final results when generalizing the ATE from the CRASH-2 trial to the observational data population

We now apply the estimators presented in this work and implemented first for the simulation study of Section 5. The confidence intervals for the corresponding point estimators are computed via non-parametric stratified bootstrap (Efron and Tibshirani, 1994) using 100 bootstrap samples (using stratified sampling to preserve the study-specific sample sizes).
We additionally report two consistent ATE estimators from the solely CRASH-2 data:

- `Difference_in_mean`: the difference in mean estimator (classical estimator for RCT);
- `Difference_in_condmean_ols` the difference in conditional means where the we assume

linear-logistic outcome models for $Y(1)$ and $Y(0)$ (estimator for RCT with smaller variance).

The former only involves treatment assignment A and outcome Y and thus requires no additional handling of the incomplete covariates; the latter is obtained using an EM algorithm for logistic regression with ignorable missing values in the covariates (Jiang et al., 2020).
And we present the AIPW estimators (Robins et al., 1994) for the observational study applied solely on the Traumabase data. For details about the derivation and properties of these estimators applied on incomplete observations we refer to Mayer et al. (2020):

- Generalized random forest after multiple imputation (`MI_AIPW`),
- Generalized random forest using MIA splitting criterion (`MIA_AIPW`).

Since AIPW combined with either missing incorporated in attributes (`MIA`) or multiple imputation (`MI`) is recommended by Mayer et al. (2020) when analyzing observational data, these are the estimators kept in this analysis.
When summarizing the results from the separate analyses on the RCT and the observational data respectively and the results from the joint analysis of both studies, we observe on Figure 11 a discrepancy between the different results. The only approach with consistent estimations throughout all estimators is the EM approach that points towards a beneficial effect of TXA on all-cause mortality. The joint fixed effect multiple imputation IPSW and AIPSW estimators (`MI_IPSW` and `MI_AIPSW`) also conclude on a beneficial effect of the treatment. However, the calibration weighting estimator (`MI_CW`) does not find a significant effect.
The large confidence intervals could be partly explained by the measurement noise in the administration delay variable in the Traumabase[9]: contrary to the RCT, the Traumabase does not encode the exact delay of treatment administration, but is defined by a noisy proxy (delay between accident and admission to the resuscitation bay). However there exists evidence that administration delay is a treatment modifier for TXA and that only early administration has a beneficial effect (Hijazi et al., 2015; CRASH-2 Collaborators et al., 2011). This remark and the discrepancies between the findings, especially the different conclusions of the joint fixed effect multiple imputation estimators call for additional attempts to further refine the administration delay proxy variable in the Traumabase and potentially for additional analyses with supplementary data such as the CRASH-3 study (Cap, 2019) which describes another slightly different TBI patient population.

[9] Another direction to explore could be to trim the scores such as to avoid extreme weights. This could potentially reduce the size of the confidence intervals but it is known to generally provide a biased result since trimming of the weights induces an implicit change in the definition of the target population (Li et al., 2018)

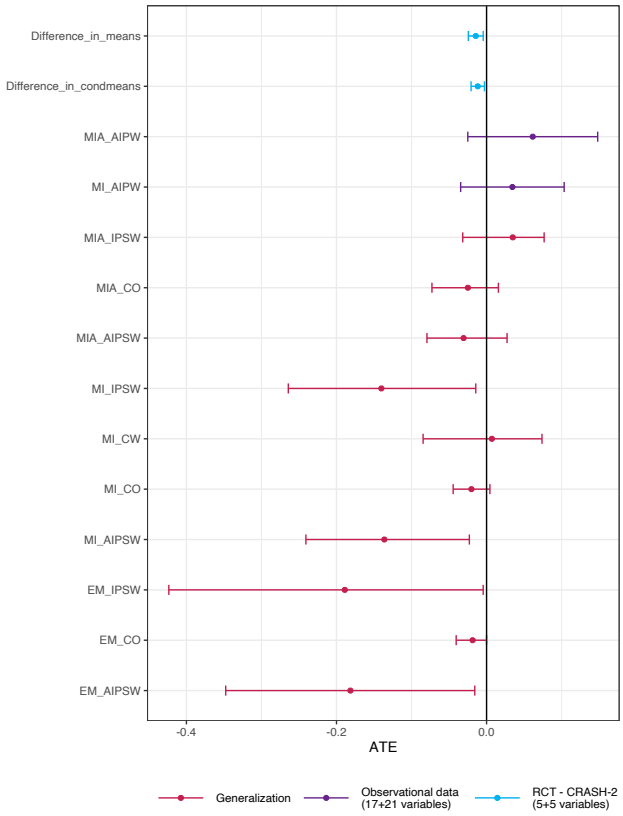


Figure 11: Separate and joint ATE estimators and 95% confidence intervals computed on the Traumabase (observational data set; purple), on the CRASH-2 trial (RCT; cyan), and generalized from CRASH-2 to the Traumabase target population (red). Number of variables used for adjustment in each context is given in the legend. The confidence intervals obtained on the observational data set and on the joint data sets are obtained via nonparametric bootstrap.

7 Conclusion

In this work we have shown that missing values in multiple data sources require additional assumptions either on the mechanism that generated the missing values or on the data generating processes of both data sources to ensure generalizability (or transportability) of the treatment effect. We have proposed several estimators that are suited for generalizing a treatment effect from an RCT to a different target population described by observational covariate data. Which of the proposed methods is preferable depends on the underlying identifiability assumptions for generalizing the treatment effect. If the identifiability assumptions on the full data case are kept and only assumptions on the missingness mechanism are added, we recommend a joint multiple imputation that models the data source as fixed effect as long as the missing values are ignorable. If the identifiability assumptions are altered to account for informative missing values implicated in the selection process, then estimators involving generalized conditional models are suited. These estimators rely on strong assumptions about the form of ignorability that is imposed, and we note that in many common examples in medicine or epidemiology it does not appear

to be the most plausible one. The approach(es) that only imply additional assumptions on the missing values mechanism seem both closer to real application contexts and methodologically more feasible. Indeed, the presented data analysis on a treatment administered in critical care falls into this latter case and the results obtained with the proposed estimators from the two different presented approaches do not come to the same conclusion. This illustrates again the importance of choosing adequate assumptions for identifiability with missing covariate values and corresponding estimation strategies.
On the methodological side, for all considered approaches and simulation settings, the question of varying missing values mechanisms across data sources remains to be addressed in more detail. Note as well that we have focused on missing values in both data sets, but the recommendations extend to the case where there are only missing values in the observational data and not the RCT due to different levels of systematic data collection.
Finally, the problem of incomplete observations addressed here is different from the problem of inconsistent variable sets between an RCT and an observational dataset, e.g., one variable is completely missing in one set, which is a challenging problem of a different kind that is left for future work. We note that this latter problem may cause issues of identifiability and is thus more related to the problem of unobserved confounding and a recent work proposes sensitivity analysis to address this issue (Colnet et al., 2021).

Acknowledgements We would like to thank Shu Yang for fruitful discussions and her valuable feedback on our work. We thank Tobias Gauss, Jean-Denis Moyer and François-Xavier Ageron for their medical insights and interpretation of our data analysis; finally we thank the CRASH-2 trial investigators for sharing the trial data with us.

Funding IM was supported by a EHESS PhD fellowship and a Google PhD fellowship. The funding institutions had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

A Details on the estimation methods with missing values

A.1 Prediction on new incomplete observations with parametric model

As mentioned in Section 4, it is possible to predict the outcome y for new incomplete observations, using the regression model estimated via EM, by marginalizing over the distribution of missing data given the observed. More formally, in the logistic regression case, using a Monte Carlo approach and *maximum a posteriori* estimator, it is possible to predict the response y for a new observation x_i as follows:

1. Sample
$$\left(x_{\text{mis}}^{(s)}, 1 \leq s \leq S\right) \sim \text{p}\left(x_{\text{mis}} \mid x_{\text{obs}}\right)$$

2. Predict the response y by maximum a posteriori

$$\begin{aligned}
\hat{y} = \arg\max_y \text{p}\left(y \mid x_{\text{obs}}\right) &= \arg\max_y \int \text{p}(y \mid x)\text{p}\left(x_{\text{mis}} \mid x_{\text{obs}}\right) dx_{\text{mis}} \\
&= \arg\max_y \mathbb{E}_{p_{z_{\text{mis}}|x_{obs}}} \text{p}(y \mid x) \\
&= \arg\max_y \sum_{s=1}^{S} \text{p}\left(y \mid x_{\text{obs}}, x_{\text{mis}}^{(s)}\right)
\end{aligned}$$

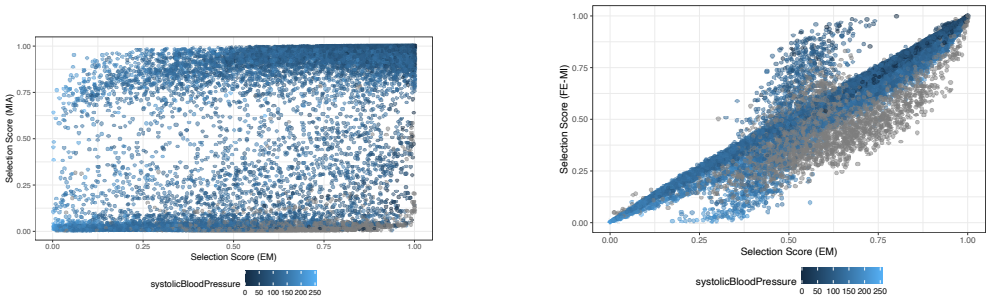
For the linear case, the prediction proceeds in two steps:

1. Imputation of the new observation using the estimated variance-covariance matrix of the covariates $\widehat{\Sigma}$.
$$\hat{x}_{mis}^{new} = -\left[\widehat{\Sigma^{-1}}_{mis,mis}\right]^{-1} \widehat{\Sigma^{-1}}_{mis,obs} x_{obs}^{new}$$

2. Prediction of response y using the imputed observation $[x_{obs}^{new}, \hat{x}_{mis}^{new}]$.

B Details on the critical care management application

We have seen in Section 6 that the scores obtained using EM and MI suggest that the positivity assumption is satisfied, while the scores estimated via MIA however concentrate around 0 and 1 for the observational and RCT observations respectively, suggesting poor overlap under this model. This can be also observed in the scatter plots in Figure 12, where we color the observations according to the values of the systolic blood pressure (SBP, a variable with an important fraction of missing values in the observational data, see Figure 8). We notice that MIA attributes a very low selection score to all observations with missing SBP value and thus selects the response indicator of the SBP variable to partly predict trial eligibility. This method has been studied more extensively in a regression framework and not in a classification framework and here we find that it predicts a class according to the presence of missing values and this is not necessarily what we intend when applying this method.



(a) x-axis: EM; y-axis: MIA.

(b) x-axis: EM; y-axis: joint fixed effect MI.

Figure 12: Scatter plots of different estimated selection scores. The point color is set according to the systolic blood pressure (SBP) covariate values (missing SBP values are indicated by gray points).

References

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical journal 53*(1), 57–74.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Audigier, V., I. R. White, S. Jolani, T. P. Debray, M. Quartagno, J. Carpenter, S. Van Buuren, M. Resche-Rigon, et al. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science 33*(2), 160–183.

Bareinboim, E. and J. Pearl (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences 113*, 7345–7352.

Bartlett, J. W., O. Harel, and J. R. Carpenter (2015). Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American journal of epidemiology 182*(8), 730–736.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Burgess, S., I. R. White, M. Resche-Rigon, and A. M. Wood (2013). Combining multiple imputation and meta-analysis with individual participant data. *Statistics in medicine 32*(26), 4499–4514.

Cap, A. P. (2019). Crash-3: a win for patients with traumatic brain injury. *The Lancet 394*(10210), 1687 – 1688.

Colnet, B., J. Josse, E. Scornet, and G. Varoquaux (2021). Generalizing a causal effect: sensitivity analysis and missing covariates. *arXiv preprint arXiv:2105.06435*.

Colnet, B., I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang (2020). Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*.

CRASH-2 Collaborators et al. (2011). The importance of early treatment with tranexamic acid in bleeding trauma patients: an exploratory analysis of the crash-2 randomised controlled trial. *The Lancet 377*(9771), 1096–1101.

D'Agostino, Jr, R. B. and D. B. Rubin (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association 95*(451), 749–759.

Dahabreh, I. J., S. J. A. Haneuse, J. M. Robins, S. E. Robertson, A. L. Buchanan, E. A. Stuart, and M. A. Hernán (2021). Study designs for extending causal inferences from a randomized trial to a target population. *American Journal of Epidemiology 190*(8), 1632–1642.

Degtiar, I. and S. Rose (2021). A review of generalizability and transportability. *arXiv preprint arXiv:2102.11904*.

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

Dewan, Y., E. Komolafe, J. Mejìa-Mantilla, P. Perel, I. Roberts, and H. Shakur-Still (2012, 06). CRASH-3: Tranexamic acid for the treatment of significant traumatic brain injury: study protocol for an international randomized, double-blind, placebo-controlled trial. *Trials 13*, 87.

Dong, L., S. Yang, X. Wang, D. Zeng, and J. Cai (2020). Integrative analysis of randomized clinical trials with real world evidence studies. *arXiv preprint arXiv:2003.01242*.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Hijazi, N., R. Abu Fanne, R. Abramovitch, S. Yarovoi, M. Higazi, S. Abdeen, M. Basheer, E. Maraga, D. B. Cines, and A. Al-Roof Higazi (2015). Endogenous plasminogen activators mediate progressive intracerebral hemorrhage after traumatic brain injury in mice. *Blood, The Journal of the American Society of Hematology 125*(16), 2558–2567.

Hippel, P. v. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology 39*(1), 265–291.

Jiang, W., J. Josse, M. Lavielle, and T. Group (2020). Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis 145*, 106907.

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

Josse, J. and J. P. Reiter (2018, 05). Introduction to the special section on missing data. *Statist. Sci. 33*(2), 139–141.

Kallus, N., X. Mao, and M. Udell (2018). Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6921–6932.

Kim, J. K. and J. Shao (2013). *Statistical methods for handling incomplete Data*. CRC Press.

Lê, S., J. Josse, and F. Husson (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software 25*(1), 1–18.

Le Morvan, M., J. Josse, T. Moreau, E. Scornet, and G. Varoquaux (2020). Neumiss networks: differential programming for supervised learning with missing values. In *Advances in Neural Information Processing Systems 33*.

Leigh, J., G. Collaborators, Y. Guo, K. Deribe, A. Brazinova, and S. Hostiuc (2018, 11). Global,regional,and national disability- adjusted life years(dalys) for 359 diseases and injuries and healthy life expectancy(hale) for 195 countries and territories,1990-2017: a systematic analysis for the global burden of disease study 2017. *The Lancet 392*, 1859–1922.

Leyrat, C., S. R. Seaman, I. R. White, I. Douglas, L. Smeeth, J. Kim, M. Resche-Rigon, J. R. Carpenter, and E. J. Williamson (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical methods in medical research 28*(1), 3–19.

Li, F., K. L. Morgan, and A. M. Zaslavsky (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association 113*(521), 390–400.

Little, R. J. and D. B. Rubin (2014). *Statistical Analysis with Missing Data*. New York: Wiley.

Mattei, A. and F. Mealli (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*(2), 257–273.

Mayer, I., J. Josse, N. Tierney, and N. Vialaneix (2019). R-miss-tastic: a unified platform for missing values methods and workflows. *arXiv preprint arXiv:1908.04822*.

Mayer, I., E. Sverdrup, T. Gauss, J.-D. Moyer, S. Wager, and J. Josse (2020). Doubly robust treatment effect estimation with missing attributes. *Ann. Appl. Statist. 14*(3), 1409–1431.

Morris, T. P., I. R. White, and M. J. Crowther (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine 38*(11), 2074–2102.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling 7*, 1393–1512.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*(427), 846–866.

Rosenbaum, P. R. and D. B. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association 79*(387), 516–524.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schafer, J. L. (2010). *Analysis of incomplete multivariate data*. London: Chapman and Hall: CRC press.

Seaman, S. and I. White (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods 43*(16), 3499–3515.

Shakur-Still, H., I. Roberts, R. Bautista, J. Caballero, T. Coats, Y. Dewan, H. El-Sayed, G. Tamar, S. Gupta, J. Herrera, B. Hunt, P. Iribhogbe, M. Izurieta, H. Khamis, E. Komolafe, M. Marrero, J. Mejía-Mantilla, J. J. Miranda, C. Uribe, and S. Yutthakasemsunt (2009,

11). 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.

Tibshirani, J., S. Athey, R. Friedberg, V. Hadad, D. Hirshberg, L. Miner, E. Sverdrup, S. Wager, and M. Wright (2020). *grf: Generalized Random Forests*. R package version 1.1.0.

Twala, B., M. Jones, and D. J. Hand (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters 29*(7), 950–956.

van Buuren, S. (2018). *Flexible Imputation of Missing Data. Second Edition*. Boca Raton, FL: Chapman and Hall/CRC.

Westreich, D., J. K. Edwards, C. R. Lesko, E. Stuart, and S. R. Cole (2017). Transportability of trial results using inverse odds of sampling weights. *American journal of epidemiology 186*(8), 1010–1014.

Yang, S., L. Wang, and P. Ding (2019). Causal inference with confounders missing not at random. *Biometrika 106*(4), 875–888.

Zhu, Z., T. Wang, and R. J. Samworth (2019). High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125*.