

---

# To Impute or not to Impute? Missing Data in Treatment Effect Estimation

---

**Jeroen Berrevoets**  
University of Cambridge

**Fergus Imrie**  
UCLA

**Trent Kyono**  
Meta

**James Jordon**  
Alan Turing Institute

**Mihaela van der Schaar**  
University of Cambridge  
Alan Turing Institute

## Abstract

Missing data is a systemic problem in practical scenarios that causes noise and bias when estimating treatment effects. This makes treatment effect estimation from data with missingness a particularly tricky endeavour. A key reason for this is that standard assumptions on missingness are rendered insufficient due to the presence of an additional variable, treatment, besides the input (e.g. an individual) and the label (e.g. an outcome). The treatment variable introduces additional complexity with respect to *why* some variables are missing that is not fully explored by previous work. In our work we introduce *mixed confounded missingness* (MCM), a new missingness mechanism where some missingness *determines* treatment selection and other missingness *is determined by* treatment selection. Given MCM, we show that naively imputing all data leads to poor performing treatment effects models, as the act of imputation effectively *removes* information necessary to provide unbiased estimates. However, no imputation at all also leads to biased estimates, as missingness determined by treatment introduces bias in covariates. Our solution is *selective* imputation, where we use insights from MCM to inform precisely which variables should be imputed and which should not. We empirically demonstrate how various learners benefit from selective imputation compared to other solutions for missing data. We highlight that our experiments encompass both average treatment effects *and* conditional average treatment effects.

## 1 Introduction

Treatment effects are arguably the most important estimand in causal inference [44, 46, 55], with a treatment *effect* lying at the heart of a causal question, and have been adopted in a wide range of fields such as medicine [2, 8–10, 45], marketing [7, 16, 17], and even human resources [51]. Using causal inference, we try to more explicitly attribute effect to the treatment by carefully disentangling the role of the environment. We define *effect* as the difference in outcome when applying the treatment versus not applying the treatment (or any alternative treatment for that matter).

Literature on inferring (or predicting) treatment effects is largely concerned with handling *selection bias*. That is, we identify a possible difference between the treated and non-treated subpopulations of the data, since treatment is rarely distributed uniformly random across the population. For example, cancer patients are assigned different oncology therapies based on their individual disease and patient characteristics. If not accounted for, selection bias leads to biased estimates. As such, many works focus on novel strategies to handle this bias [3, 22, 29, 30, 55, 61, 72].

However, these methods almost exclusively assume that data is complete. From a practitioner’s standpoint, this may not always be the case and, in reality, *data is often incomplete* [12, 35]. Work on learning from incomplete data has focused on non-causal prediction problems and, in particular, on methods for imputing missing values [43, 48, 64, 69, 71]. One strategy adopted in the treatment effects literature is to use imputation methods to infer the missing variables before we train a learner on the data [31, 36, 57, 59].

Missing data is more challenging in treatment effects compared to prediction problems due to the presence of a treatment variable. First, the fact that certain variables are missing may contribute to which treatment is administered, contributing to selection bias. For example, being deprived of information, a clinician may opt for a less risky treat-

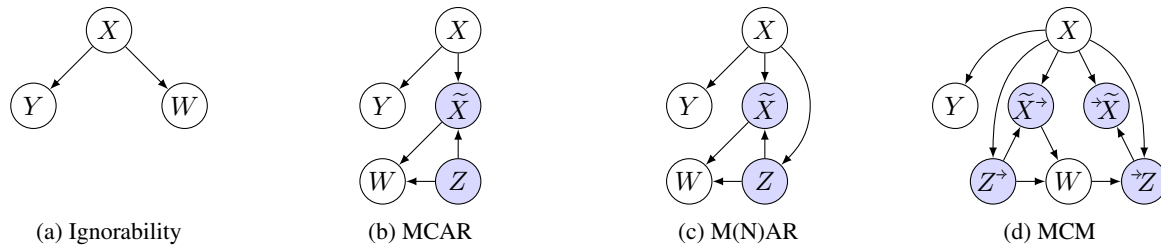


Figure 1: **(a) Ignorability as a graphical model.** From Richardson and Robins [49], we express ignorability as a DAG. For brevity, we drop the parentheses of  $Y(w)$ , as well as its accepted “single world intervention path” from  $w$  to  $Y(w)$ ; **(b-d) DGP for missingness mechanisms.** Shaded nodes indicate nodes that relate to missing variables, white nodes relate to treatment effects. In fig. 1b and fig. 1c we illustrate the standard causal extensions to MCAR and M(N)AR, respectively, including the treatment variable  $W$ . In fig. 1d we illustrate MCM. Unique to MCM is to allow treatment to both *cause* missingness (through  $\tilde{X}^{\rightarrow}$  and  $\rightarrow Z$ ) and *be caused by* missingness (corresponding to  $\tilde{X}^{\leftarrow}$  and  $Z^{\leftarrow}$ ).

ment, just to be safe. Thus, imputation effectively *removes information*— i.e. given that missingness may cause treatment selection, it is much harder to handle selection bias when the information about which variables were missing is removed. This observation has led to many *not imputing* missing values and instead considering a missing variable as another value and using it directly [15, 38, 39, 52].

Conversely, it may also be the case that the treatment decision will impact which variables are measured, i.e. missingness may be *caused by the treatment*. We believe this is a ubiquitous but overlooked scenario. For example, in medicine, many drugs require baseline blood tests to be performed before treatment commences [63]. However, these blood tests typically *would not* be performed if a different treatment decision was made. Missingness as a result of the treatment choice increases the difference between treatment subpopulations, even when there was no difference to begin with, had all patient covariates been observed. Adjusting on these missing variables therefore *introduces bias* to the model. This effect has *not* previously been considered in existing missingness mechanisms. However, in scenarios such as healthcare, this is almost always the case. Thus, to impute or not to impute?

**Example: knee pain.** Let us consider a specific example of MCM in practice (we provide an additional example in appendix E).

Around one in six general practice consultations in England are for arthritis and musculoskeletal issues [23], of which knee pain is one of the most common complaints. Treatment plans often consist of pharmaceutical-based pain management, physiotherapy, or both [5]. However, surgery is considered depending on patient characteristics (e.g. young, active) and severity of their condition. Patients who are treated with pain relief medication or physiotherapy will often not have additional tests performed (such as

radiography or magnetic resonance imaging). However, these tests will almost always be performed before surgery, together with other routine preoperative tests, affecting which patient covariates are collected.

Furthermore, one common surgical intervention is knee arthroscopy [60], a gold standard for simultaneous diagnosis and treatment of knee disorders [53]. As a surgical procedure allowing doctors to examine the knee, knee arthroscopy provides additional information that is not captured if a different treatment plan is chosen. While specific to one indication, our two medical examples clearly illustrate the complexity of missingness patterns in real-world observational medical data, motivating and reinforcing the need for, and relevance of, MCM.

**Contribution.** We introduce (and motivate adoption of) a formal description of missingness in data used to estimate treatment effects [42, 44, 55? ]. In particular, we find that previous attempts— dating as far back as the 1980s [52, Appendix B] —at formalising missingness in treatment effects allow for inaccurate descriptions of missingness and its impact. We illustrate why these descriptions are insufficient and provide an alternative termed *mixed confounded missingness* (MCM). We argue that MCM is a general-purpose missingness mechanism, distinct from well-known missingness mechanisms such as *missing (completely) at random* [36], and a refinement of *conditional independence of treatment* [52] that should be adopted to describe missingness when estimating treatment effects.

Based on the insights provided by MCM, we propose a strategy for handling missing data in treatment effects, termed selective imputation. Our approach is theoretically motivated and we provide empirical evidence of how methods benefit from this approach and demonstrate the harm when missingness is not correctly dealt with.

## 2 Preliminaries

Estimating causal effects requires us to answer a *counterfactual* question. In particular, when we observe the outcome after applying a treatment on an individual, it is impossible to also observe that individual’s outcome under alternative treatment [28]. As a treatment effect is defined as the difference between both outcomes, we are tasked with inferring an estimand which is *never observed*, which is crucially different from standard supervised learning.

One can estimate causal effects by conducting randomised controlled trials (RCTs) [19, 44]. However, RCTs are often very expensive, and are sometimes considered unethical in a clinical setting [18, 24]. However, the alternative we consider in our work—estimating effects from observational data—comes with its own challenges. In particular, comparing each treatment’s subpopulation in an *observational* dataset will result in biased estimates. If treatment is not assigned randomly, but instead based on an individual’s characteristics, with different outcomes, then these individuals are more represented in each subpopulation as a result, and estimating outcomes become biased. This phenomenon is often termed *selection bias*.

**Notation.** Let  $X \in \mathcal{X} \subseteq \mathbb{R}^d$  be the covariates of an individual; let the individual be treated with  $W \in \{0, 1\}$ ; and let  $Y \in \mathcal{Y} \subseteq \mathbb{R}$  be their observed outcome. Practically,  $X$  could be a patient with lung cancer;  $W = 1$  could be chemotherapy (and  $W = 0$  radiotherapy); and  $Y$  their tumour size after treatment. We use a subscript,  $X_i$  to indicate the  $i^{\text{th}}$  element in  $X$ , which means that  $X_i \in \mathbb{R}$ .

**Assumptions in causal inference.** Estimating unbiased treatment effects from observational data has received a lot of attention in recent years. One of the more popular avenues in the literature is the potential outcomes (POs) framework of causality [44, 55]. We define the PO of a treatment  $w \in \{0, 1\}$  as  $Y(w)$ , where  $Y(w)$  corresponds to the outcome an individual would have experienced had they been assigned treatment  $W = w$ . While the standard consistency assumption (see Assum. 1 below) allows us to interpret the observed outcome as the potential outcome of the observed treatment, i.e.  $Y = Y(W)$ ; selection bias makes estimating the counterfactual outcome,  $Y(-W)$ , more involved.

**Goal.** Ultimately, our goal is to correctly identify the treatment effect, for which we need to account for bias in the data. Countering selection bias is achieved by correctly adjusting for the confounders. To do so, we make the following assumptions, which are widely considered as standard in the PO-framework:

**Assumption 1 (Consistency).** The observed outcome  $Y_i = Y(W_i) = Y(w)$  if  $W_i = w$ , for  $w \in \{0, 1\}$  and  $i = 1, 2, \dots, N$ ,<sup>1</sup> i.e. outcomes in the data correspond to one

of the potential outcomes.

**Assumption 2 (Ignorability).** The joint distribution  $p(X, W, Y)$  satisfies strong ignorability:  $Y(0), Y(1) \perp\!\!\!\perp W|X$ , i.e. the potential outcomes are independent of the treatment, conditioned on  $X$ , implying that there are no *additional* (unobserved) confounders beyond the variables in  $X$ .

**Assumption 3 (Overlap).** The distribution  $p(X, W, Y)$  satisfies overlap:  $\exists \delta \in (0, 1)$  s.t.  $\delta < p(W|X = x) < 1 - \delta, \forall x \in \mathcal{X}$ , i.e. each individual has a non-zero probability to receive either treatment.

**Graphical models and causality.** Alternatively to POs, we can express causal relationships as a graphical model. In particular, a causal relationship is depicted as a directed edge in a directed acyclic graph (DAG) [46], where a parent node is the cause and the child node is the effect. The ignorability assumption in Assum. 2 is sometimes illustrated in such a graphical model [49]. Specifically, the ignorability assumption can be expressed as the DAG shown in fig. 1a. Typically the influence of the treatment on the outcome is expressed as a *single world intervention graph* (SWIG):  $(W \circlearrowleft) \rightarrow (Y)$  [49]. We have removed this SWIG-path from our figures in order to focus our discussion on the path(s) between  $X$  and  $W$ .

Note that the set of DAGs (i.e. the Markov equivalence class) that satisfy Assum. 2 encompasses more than just the DAG in fig. 1a. Other DAGs in this equivalence class can just as easily respect Assum. 1 to 3, but they would make no sense. For example, reversing the arrow between  $X$  and  $Y$  would imply that outcome causes the covariates in the individuals, while still respecting ignorability. Instead, fig. 1a is motivated through logical reasoning, where treatment and outcome are caused by the covariates. Throughout the remainder of this paper, we will build (and extend) heavily on fig. 1a, employing similar logical reasoning.

**Causal estimands.** We now arrive at our two estimands of interest: the *average treatment effect* (ATE), and the *conditional average treatment effect* (CATE). Given the notation above, we can define each estimand as follows:

**Definition 2.1 (ATE).** The ATE is defined as the population-wide difference between a treatment’s potential outcomes. Mathematically, we can define the ATE as follows:  $\bar{\tau}(\mathcal{X}) := \mathbb{E}_{\mathcal{X}}[Y(1) - Y(0)]$ .

**Definition 2.2 (CATE).** The CATE is defined as a conditional difference between a treatment’s potential outcomes. Mathematically, we can define the CATE as follows:  $\tau(x) := \mathbb{E}[Y(1) - Y(0)|X = x]$ .

**Missingness.** In practice, a sample  $X$  may be incomplete. For example, a clinician responding to an urgent trauma (SUTVA) assumes both no interference and consistency [58]. The equation in our consistency assumption also implies no interference.

<sup>1</sup>The well-known *stable unit treatment value assumption*

case may have to select treatment based on incomplete information. In this scenario, the incomplete variables are considered *missing*. To learn from these data we could consider *completing* this sample through imputation, but missingness in itself may be informative. Perhaps the clinician’s decision would have been different if they had had complete information. Typically, we define three *mechanisms* to describe how a variable ended up to be missing: a first is *missing completely at random* (MCAR), where missingness in one variable is independent of the other variables as well as its own value, a second is *missing at random* (MAR), where the missingness in one variable may depend on the other (observed) variables, and the third is *missing not at random* (MNAR), where missingness is typically assumed to be caused by variables outside the observed covariates [36, 56, 68]. Generally, MCAR is attributed to noise in data collection.

We indicate the missing data in  $X$  with a variable  $Z \in \{1, \star\}^d$ , where  $Z_i = \star$  if  $X_i$  is missing, and  $Z_i = 1$  if  $X_i$  is observed. Having  $Z \perp\!\!\!\perp X, Y(w)$ , corresponds with MCAR; and  $Z \not\perp\!\!\!\perp X$ , but  $Z \perp\!\!\!\perp Y(w)|X$ , corresponds with M(N)AR. In our paper, we denote the complete (but unobserved) sample as  $X$ , and the incomplete (but observed) sample as  $\tilde{X} := Z \odot X \in \{\star, \mathbb{R}\}^d$  where we define  $\odot$  as the element-wise product and we take  $X_i \times \star = \star$ , i.e. if  $X_i$  is unobserved it equals  $\star$  in  $\tilde{X}$ .

Note that none of the existing missingness mechanisms take the treatment  $W$  into account as standard, since they span a broader literature beyond treatment effects. We have illustrated MCAR and M(N)AR as DAGs in fig. 1b and fig. 1c, respectively, where we have included arrows from  $\tilde{X}$  and  $Z$  to  $W$ . This corresponds to the situation where treatment is decided based on what is actually observed; the alternative where  $X$  is causing  $W$  would lead to a confounded setting, which is commonly assumed not to be the case in the literature. Note that these DAGs respect the conditional independent statements assumed in their respective missingness mechanisms. As we have noted in our introduction, we could either impute a missing value  $X_{ij}$  with an estimate thereof, denoted  $\hat{X}_{ij}$ , such that we can predict a treatment effect from data with imputed samples (i.e. learn  $\bar{\tau}(\hat{X})$  or  $\tau(\hat{X})$ ); or we could predict treatment effects from data with missingness directly (i.e. learn  $\bar{\tau}(\tilde{X})$  or  $\tau(\tilde{X})$ ).

### 3 Mixed Confounded Missingness (MCM)

Some missingness may be *caused by* treatment, while other missingness may *cause* treatment—a situation that cannot be modelled by the MCAR nor the M(N)AR missingness mechanisms. This has important consequences as missingness is now a mixture of confounding and non-confounding elements, leading us to term our proposal: *mixed confounded missingness* (MCM). We have illustrated MCM as a DAG in fig. 1d.

In this section, we explain why there are no other arrows included in MCM, i.e. motivate that it is complete and general. Then, we compare MCM to previous proposals for missingness in CATE.

Like MCAR and M(N)AR, MCM describes the interactions between  $Z$  and the remaining variables,  $X, W, Y(w)$ , (and  $\tilde{X}$ ). However, contrasting MCAR and M(N)AR, MCM specifically models various possible interactions with missingness by splitting  $Z$  into two distinct factors,  $Z^\rightarrow$  and  ${}^\rightarrow Z$ . The former captures missingness causing treatment, and the latter missingness caused by treatment. Acknowledging the possibility of missingness that causes or is caused by treatment, leads to the following assumption:

*Assumption 4* (Missingness factors in MCM). We assume there exists a partition of  $Z = \{Z^\rightarrow, {}^\rightarrow Z\}$  s.t.  $Z^\rightarrow \perp\!\!\!\perp {}^\rightarrow Z | W, X$ , further implying that  $\tilde{X}^\rightarrow \perp\!\!\!\perp {}^\rightarrow \tilde{X} | W, X$  as  $Z^\rightarrow \perp\!\!\!\perp {}^\rightarrow \tilde{X} | X, W$  and  $\tilde{X}^\rightarrow \perp\!\!\!\perp {}^\rightarrow Z | X, W$ , where  ${}^\rightarrow \tilde{X}$  and  $\tilde{X}^\rightarrow$  are the covariate sets with missingness corresponding to  ${}^\rightarrow Z$  and  $Z^\rightarrow$ , respectively.

**MCM is exhaustive.** Being exhaustive is important, as through exhaustiveness we can argue that MCM covers *all* possible scenarios that can lead to missing data with treatments. For our discussion, exhaustiveness thus means that each potential arrow is either included in MCM, or its absence in MCM has a compelling argument.

Five variables ( $X, {}^\rightarrow \tilde{X}, \tilde{X}^\rightarrow, W$ , and  $Y$ ) that are fully connected to  ${}^\rightarrow Z$  and  $Z^\rightarrow$  should result in 20 distinct paths (four for each of the five variables). However, fig. 1d connects only 6 paths to either  ${}^\rightarrow Z$  or  $Z^\rightarrow$ , leaving 14 paths unaccounted for. With assum. 4, we can exclude paths across missingness factors (i.e. the paths that directly connect  $Z^\rightarrow$  and  ${}^\rightarrow \tilde{X}$ , and those that directly connect  ${}^\rightarrow Z$  to  $\tilde{X}^\rightarrow$ ), reducing the number of unaccounted for paths to 10.

We will now argue why these 10 paths (in fig. 2) are not included in MCM. First, we discuss fig. 2a. Having  $(Z^\rightarrow) \rightarrow (X) \leftarrow ({}^\rightarrow Z)$  allows covariates to change *depending on* what other variables are missing, implying that  $X$  is dependent on whatever dataset an individual is included in. Having this dependence would imply that  $X$ , representing the fully observed–*true*–set of covariates, would be different across datasets, despite representing the same individual. Furthermore, including the paths in fig. 2a introduces a cycle:  $(X) \rightarrow ({}^\rightarrow Z) \rightarrow (Z^\rightarrow) \rightarrow (X)$  (and similarly for  $Z^\rightarrow$ ), which violates the DAG structure.

In fig. 2b we immediately observe cycles, meaning that only two of the four presented arrows can exist simultaneously. Let us first consider the paths where  $Y(w)$  is causing missingness, i.e.  $(Z^\rightarrow) \leftarrow (Y) \rightarrow ({}^\rightarrow Z)$ . In the potential outcomes setting,  $Y(w)$  is topologically last in the DAG; take fig. 1a, where inclusion upon of the single world intervention path,  $(W^{(w)}) \rightarrow (Y)$ , we clearly see that no variable is caused by  $Y(w)$ . Having  $Y(w)$  as the final observation



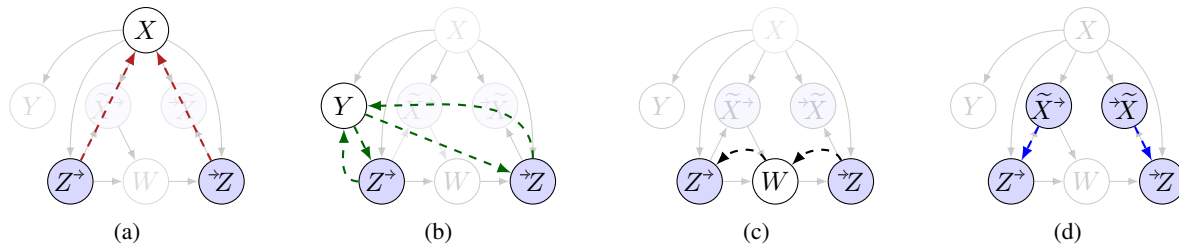


Figure 2: **Arrows that are not included in MCM.** Considering each possible direct arrow from, and to, the missingness variables ( $\tilde{Z}$  and  $Z^\rightarrow$ ), should result in 20 arrows with five remaining variables. Excluding the paths that *are* included, and the paths *across* missingness indicators, we are left with 10 paths that are *seemingly* missing from our definition. From figs. 2a to 2d, we depict each missing arrow in function of one variable:  $X$ , then  $Y(w)$ ,  $W$ , and then  $\tilde{X}^\rightarrow$ ,  $\tilde{\tilde{X}}$ , respectively.

makes sense; once we observe the outcome from a treatment, the covariate observations are left untouched. The absence of the remaining two arrows,  $(Z^\rightarrow \rightarrow Y) \leftarrow (\tilde{Z} \rightarrow Y)$ , is similarly argued as the absence of the arrows in fig. 2a. Namely, if  $Z$  were to directly influence outcome, then an individual represented in two different datasets— with different missing variables —would have conflicting outcomes. Clearly, one person can only have one outcome [28], i.e. the arrows in fig. 2b cannot exist.

Next, fig. 2c, illustrating the existing arrows, *reversed*. Besides these arrows resulting in cycles— specifically,  $(Z^\rightarrow \rightarrow W) \leftarrow (\tilde{Z} \rightarrow W)$ —they also violate our definition for  $\tilde{Z}$  and  $Z^\rightarrow$ . In particular, the *reason* for these distinct factors is precisely their respective directed paths from and to  $W$ .

Lastly, we consider fig. 2d depicting the reverse arrows from  $\tilde{\tilde{X}}$  to  $\tilde{Z}$ , and similarly from  $\tilde{X}^\rightarrow$  to  $Z^\rightarrow$ . Here too, besides the obvious cycles,  $(\tilde{Z} \rightarrow \tilde{\tilde{X}}) \leftarrow (\tilde{\tilde{X}} \rightarrow \tilde{Z})$  and similarly for  $\tilde{\tilde{X}}^\rightarrow$  and  $Z^\rightarrow$ ,  $\tilde{\tilde{X}}$  is deterministically defined an element-wise product between  $Z$  and  $X$ . This relationship is unambiguous. If indeed  $\tilde{Z}$  and  $Z^\rightarrow$  are caused by the fully observed covariates  $X$ , it seems almost silly to consider them to be also caused by the partially observed covariates  $\tilde{\tilde{X}}$ , which represents almost *exactly* the same entity. In fact, their only difference is completely captured in  $Z$ . With this, we can safely remove the arrows in fig. 2, resulting in MCM (with a total of 6 arrows in  $Z$ ) being exhaustive.

**Missing data and existing assumptions.** Having argued each arrow in fig. 1d, we will now relate MCM to previous descriptions of missingness in the treatment effects setting. Introduced in Rosenbaum and Rubin [52] and further investigated in Mattei [37], Mayer et al. [39] and Blake et al. [11], we find that typically the ignorability assumption in the treatment effects literature (that is, assum. 2) is extended to the setting with missingness by simply replacing the condition in assum. 2 by,

$$Y(0), Y(1) \perp\!\!\!\perp W | \tilde{\tilde{X}}, Z, \quad (1)$$

which translates to “*unconfoundedness despite missing-*

*ness*”<sup>2</sup>. To verify eq. (1), typically the above unconfoundedness assumption is combined with one of the two assumptions below:

$$\text{CIT: } W \perp\!\!\!\perp X | \tilde{\tilde{X}}, Z \quad (2)$$

or

$$\text{CIO: } Y(0), Y(1) \perp\!\!\!\perp X | \tilde{\tilde{X}}, Z, \quad (3)$$

where CIT in eq. (2) stands for *conditional independence of treatment*, and CIO in eq. (3) stands for *conditional independence of outcome* [39]. Essentially, CIT and CIO assume there is no additional information in the fully observed  $X$  compared to the observed  $\tilde{\tilde{X}}$  and  $Z$  to predict treatment and outcome, respectively— i.e. adjustment for  $\tilde{\tilde{X}}$  (and  $Z$ ) still warrants ignorability.

The assumptions depicted in eqs. (2) and (3) can be considered a *logical consequence* of the assumption in eq. (1). In particular, for the potential outcomes to be ignorable from treatment, as is assumed through eq. (1), there cannot exist direct arrows beyond those in  $\tilde{\tilde{X}}$  from  $X$  to either  $W$  (which is implied by eq. (2)), or  $Y(w)$  (which is implied by eq. (3)). Violating CIT and CIO would mean that there *has to be* such a variable (i.e. dimension) in  $X$ , that is not represented in the covariate space of  $\tilde{\tilde{X}}$ , that has a direct link to either  $W$ , or  $Y(w)$ , or both. Such a link is in direct conflict with eq. (1).

At first glance, CIT and CIO may seem acceptable assumptions. However, we identify two major issues concerning these assumptions: (i) while further specifying the consequences of ignorability despite missingness (cfr. eq. (1)), CIT and CIO are too general, allowing too many different (mostly completely unrealistic) missingness mechanisms without violating eq. (1) nor CIT or CIO; and (ii) they do

<sup>2</sup>Note that, we do not *have to* explicitly include  $Z$  in the condition (see for example Mayer et al. [39]) as  $\tilde{\tilde{X}}$  is related in a deterministic way to  $Z$ . However, we chose to include  $Z$  in these conditions, as they are related to the DAG presented in fig. 1d, where we have also included  $Z$ . Later in our paper, we will remove  $Z$  from our DAGs and subsequently our conditions, in order to focus on the relationships with  $\tilde{\tilde{X}}$  directly.

not allow for treatments to cause missingness, which we have argued above is an important consideration in treatment effects. Furthermore, problem (ii) hints at a larger underlying problem regarding missingness in treatment effects: creating an additional source of bias, which we shall discuss in Section 4.

Given the variables in eq. (1) and CIT/CIO ( $X$ ,  $\tilde{X}$ ,  $Z$ ,  $Y(w)$ ), while keeping  $(X \rightarrow Z \rightarrow \tilde{X})$  fixed, we can generate 42 DAGs which all respect these assumptions. We have included these DAGs in Appendix F. Note that each of these 42 DAGs respects the independence statements in eq. (1) and those in either eq. (2) or eq. (3). Evaluating these DAGs using the same criteria as we have for MCM leads to only 1 realistic DAG, which turns out to be a special case of MCM. Specifically, when removing the consideration of  $\rightarrow Z$ —treatment causing missingness—we arrive at the one DAG that is realistic. Without explicit consideration of treatment choices that influence missingness (i.e. splitting  $Z$  in  $Z^\rightarrow$  and  $\rightarrow Z$ ), we learn that MCM fits nicely within CIT (and consequentially eq. (1)). Furthermore, this one DAG corresponds exactly to how we introduced M(N)AR in fig. 1c.

## 4 Selective imputation

Having introduced MCM, we now explain why missing data in treatment effects should not be brushed over lightly. We discuss why the two common approaches to handle missing data (recall that typically missing data is either imputed, or kept as is) are not equipped to deal appropriately with missing data that interacts with treatment. Here we offer an alternative to these two approaches.

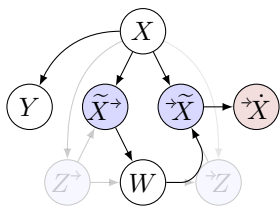


Figure 3: **Selective imputation.** Above DAG depicts MCM where we included  $Z^\rightarrow$  and  $\rightarrow Z$  in the factors  $\tilde{X}^\rightarrow$  and  $\rightarrow \tilde{X}$ , respectively; clearly showing that  $\rightarrow \tilde{X}$  is a collider. This structure motivates selective imputation: we should only impute  $\rightarrow \tilde{X}$ , where the imputed variables are depicted as  $\rightarrow \tilde{X}$ .

**Why naive approaches do not work.** First, we have to discuss what can go wrong when naively dealing with missing data. To aid our discussion, we simplified MCM in fig. 3, merging  $Z^\rightarrow$  into  $\tilde{X}^\rightarrow$ , and  $\rightarrow Z$  into  $\rightarrow \tilde{X}$ . Note that this is equivalent to fig. 1d as the link between the missingness indicator and the observed covariates is completely deterministic. One could, with complete certainty, derive the missingness indicator from the observed covariates. In

fact, the only way of obtaining  $Z$  is to do exactly that.

Imputing all data, i.e. the first naive approach, has the objective of recovering  $X$  from  $\tilde{X}$ , as accurately as possible. Given that  $Y(w)$ , is a direct function of the fully observed set of covariates,  $X$ , and not the partially observed set,  $\tilde{X}$ , regressing  $Y(w)$  on  $\tilde{X}$  should result in better estimates than  $\tilde{X}$ . Contrasting supervised learning, however, the treatment effects literature is not *only* concerned with better test-set estimates. In particular, the target in treatment effects literature goes beyond the *observed* outcomes in the dataset. To accurately estimate treatment effects, one has to account for selection bias [61], which is expressed in the path from  $X$  to  $W$  (see for example, fig. 1a). In our setting, this is more complicated as there exists no direct path. The path from  $X$  to  $W$  is *mediated* by  $\tilde{X}^\rightarrow$ . As explained in Section 3, treatment is decided on an individual’s covariates, but also the missingness in them. As such, imputing away these missing data will result in information loss with respect to selection bias, making it all the more challenging to counteract the bias when predicting a treatment effect.

Should we then impute no data, i.e. the second naive approach? Not imputing data certainly solves the problem of information loss. However, it introduces another problem. We learn from Section 3, that in order to estimate treatment effects from data, one has to assume ignorability despite missingness (see eq. (1)). However, from fig. 3 we learn that  $Y(w) \not\perp W | \tilde{X}^\rightarrow, \rightarrow \tilde{X}$ , i.e. the potential outcomes are not conditionally independent from treatment. We find the reason in  $\rightarrow \tilde{X}$ , which—due to being determined by treatment—composes a collider structure. A property of a collider structure is that it does not bias treatment selection, but when conditioned upon, *it does*. If imputing data causes problems, and not imputing data causes problems, what then should we do?

**Selective imputation.** We argue that one *should* impute  $\rightarrow \tilde{X}$ , but leave  $\tilde{X}^\rightarrow$  as is. While solutions may vary across datasets, we argue that *selective imputation* may offer a general strategy to handle missing data. Consider again fig. 3, where we have included  $\rightarrow \tilde{X}$  in the MCM-DAG. From evaluating fig. 3 with this inclusion we learn that,

$$Y(0), Y(1) \perp W | \tilde{X}^\rightarrow, \rightarrow \tilde{X}, \quad (4)$$

arriving once more at a conditional independence between  $W$  and  $Y(w)$ . Notice that, the independence statement in eq. (1) conditions on the same covariate information (i.e. all is included from  $X$ ), the only thing changed, is that we have to “forget” that some data in  $\rightarrow \tilde{X}$  may be missing as the missingness associated with  $\rightarrow \tilde{X}$  is determined by the treatment—a property we need to account for.

**The role of imputation.** From the above discussion, we learn that missingness is what differentiates treatment-subpopulations in  $\rightarrow \tilde{X}$ . The following theorem shows us what exactly imputation needs to accomplish, in order to

consider eq. (4) to be true (proof of theorem 4.1 can be found in Appendix A).

**Theorem 4.1.** *Suppose the graph structure in fig. 3, and  $W$  and  $\tilde{X}$  are independent, then ignorability defined as  $Y(0), Y(1) \perp W | \tilde{X}$ ,  $\tilde{X}$  holds.*

From theorem 4.1 we learn that, in this context, a good imputation strategy should aim to make  $\tilde{X}$  independent of the treatment, such that ignorability may hold. The intuition behind this is that through the act of (proper) imputation we effectively *balance*  $\tilde{X}$ , making the treatment populations indistinguishable from each other [61]. The better the imputation is, the closer to  $X$  the covariates in  $\tilde{X}$  will become, reducing influence from  $W$ . If our imputation is of poor quality, a model may recover information about the originally missing variables, i.e. information on treatment is retained as  $\tilde{X}$  still correlates with  $W$ , which allows bias to creep into our models.

## 5 Related work

We discuss two major perspectives in missingness in a causal setting: *m-graphs*, and ignorability when there is missingness. A more extensive overview is included in Appendices B and F.

***m-graphs.*** An *m-graph* is a graphical model [32, 46] which explicitly includes missingness indicators as variables [41]. Naturally, this is related as the way in which we describe MCM is through an *m-graph*. Using an *m-graph*, one determines whether or not an effect is identifiable [46]. While this has received some attention in graphical causality [40, 42, 62] which aims to recover identifiability in MAR and MNAR settings or perform structure learning despite missingness [20, 34], *m-graphs* remain relatively underexplored in a potential outcomes setting, as we have. Note that none of the aforementioned works consider partial imputation, specifically to correctly identify a causal effect. A more in-depth discussion can be found in Appendix B.

**Conditional independence.** Assum. 2 is needed to identify a treatment effect. However, through an *m-graph*, we find that assum. 2 may become unrealistic when there is missingness as it requires the fully observed  $X$  which we do not have access to. To our knowledge, Mayer et al. [39] was the first to use *m-graphs* depicting the PO setting (although no mention of *m-graphs* were made in their paper), and we know of none who has done so since. As such, Mayer et al. [39] is by far the most related work to ours, and we compare with them extensively in Appendix F. Mayer et al. [39] propose two *m-graphs* for CIT and CIO (eqs. (2) and (3)), included in figs. 5a and 6a, respectively. In Appendix F, we argue that neither of these *m-graphs* can exist in practice. Furthermore, figs. 5a and 6a only take into account missingness that causes treatment, but not the reverse, making

CIT and CIO much more restrictive, as discussed above.

Recent work by Zhao and Ding [73] also concerns treatment effects with missingness. While their method makes strict assumptions (such as randomised treatments), they do make an interesting observation: by specifically violating their assumption that treatment does not cause missingness, there seems to be an additional source of bias [73, Appendix S1] (despite random treatment assignment). However, they do not offer a remedy as we do, but only acknowledge that this may happen.

## 6 Experiments

We turn now to empirically validating that selective imputation, i.e. imputing only *some* parts of the covariate space, results in better treatment effect predictions.

**Data.** Common in the treatment effects literature is using synthetic datasets. The reason why we have to rely on synthetic data lies at the core of the problem: in any real-world setting, *the counterfactual is unobserved*. As such, with synthetic data we can simulate both potential outcomes, effectively observing the counterfactual for evaluation.

Additionally, simulations allow us to control the missingness mechanism. This allows us to directly test whether treatment effect models are affected by MCM and benefit from Selective Imputation. Our synthetic setup is described in detail in Appendix D.

For each experiment, we sample 10 different simulated datasets, from which we sample 10 random train and test sets, for each treatment effects method. As our finding holds both for ATE as well as CATE (in finite settings [1]), we test on both scenarios. Each simulated dataset contains 10k samples. The datasets span 20 dimensions (with factors of equal size), and a missingness rate of 0.3. We have performed multiple ablations on these values in Appendix C. In Appendix C we have also conducted an experiment on the well-known Twins dataset [4], where we manually include some missingness.

**Imputation.** We define four different scenarios: either we impute missing values across (i) all the variables (indicated as “All”); (ii) none of the variables (“Nothing”); (iii) only the variables in  $\tilde{X}$  (“Selective imputation”); or (iv) only the variables in  $\tilde{X}^c$  (“Sel. Complement”). Following Section 4, scenario (iii) (imputing only  $\tilde{X}$ ) should yield the best results given that bias is removed, while information with respect to treatment selection is retained. Imputation is performed using MICE [69].

**Models.** In our experiments, we evaluate the performance of three classes of learners in the imputation scenarios described above: T-learner, Doubly robust (DR) learner, and X-learner. We pair each treatment effects method with XGBoost due to its ability to naturally handle missing val-

Table 1: **Results.** We evaluate using 10 different train and test sets using 10 *different* simulations and averaged the results, std is in brackets. **ATE (upper).** For each setting we report an MSE between the predicted ATE and the ground truth. Imputing only  $\tilde{X}$  (“Selective imputation”) consistently performs best across different ATE prediction methodologies and treatments. **CATE (lower).** Imputing only  $\tilde{X}$ , while keeping  $\tilde{X}^\rightarrow$  as is (marked as “Selective” below), consistently performs best, in terms of PEHE, across learners and treatments. In all cases, **lower is better**, our proposal is shaded .

		Impute description	covariates	T-Learner	Doubly Rob.	X-Learner
<i>Results on ATE-estimation</i>						
<b>MSE</b>	All		$\tilde{X}^\rightarrow \cup \tilde{X}$	0.0951 (.010)	0.0651 (.008)	0.0472 (.009)
	Nothing		$\emptyset$	0.0642 (.027)	0.0902 (.018)	0.0726 (.024)
	Selective imputation		$\tilde{X}$	<b>0.0403</b> (.014)	<b>0.0381</b> (.009)	<b>0.0309</b> (.014)
	Sel. Complement		$\tilde{X}^\rightarrow$	0.0931 (.026)	0.0902 (.019)	0.0984 (.040)
<i>Results on CATE-estimation</i>						
<b>PEHE</b>	All		$\tilde{X}^\rightarrow \cup \tilde{X}$	0.7603 (0.051)	1.3674 (1.731)	0.6149 (0.063)
	Nothing		$\emptyset$	0.6906 (0.072)	0.9409 (1.943)	0.3027 (0.085)
	Selective imputation		$\tilde{X}$	<b>0.4605</b> (0.045)	<b>0.2042</b> (0.224)	<b>0.2116</b> (0.032)
	Sel. Complement		$\tilde{X}^\rightarrow$	0.9158 (0.064)	4.3657 (8.823)	0.4912 (0.109)
<b>PEHE<sub>w=0</sub></b>	All		$\tilde{X}^\rightarrow \cup \tilde{X}$	0.7371 (0.081)	1.2083 (1.610)	0.6272 (0.070)
	Nothing		$\emptyset$	0.7015 (0.100)	0.8130 (1.287)	0.2907 (0.107)
	Selective imputation		$\tilde{X}$	<b>0.5720</b> (0.079)	<b>0.1787</b> (0.202)	<b>0.2556</b> (0.062)
	Sel. Complement		$\tilde{X}^\rightarrow$	0.9351 (0.120)	4.2306 (9.056)	0.5198 (0.156)
<b>PEHE<sub>w=1</sub></b>	All		$\tilde{X}^\rightarrow \cup \tilde{X}$	0.7726 (0.055)	1.4419 (1.802)	0.6097 (0.068)
	Nothing		$\emptyset$	0.6881 (0.090)	0.9973 (2.317)	0.3091 (0.098)
	Selective imputation		$\tilde{X}$	<b>0.4097</b> (0.045)	<b>0.2169</b> (0.236)	<b>0.1915</b> (0.036)
	Sel. Complement		$\tilde{X}^\rightarrow$	0.9183 (0.075)	4.4322 (8.749)	0.4803 (0.130)

ues. We refer to Künzel et al. [33] or Curth and van der Schaar [14] for an overview of various learners.

### 6.1 Average treatment effects (ATE)

**Objective.** A first estimand to consider is the average treatment effect (defined in def. 2.1). While the ATE has been considered since the 1970’s [55], it is still an important causal estimand today. It can be argued that ATEs suffer more from selection bias than the conditional ATE, as the average is computed over the *entire* dataset, unlike the CATE [1]. In ATE, adjustment plays an important role and is usually achieved by placing non-uniform weights on each element in the dataset when computing an average. These weights often take the form of *inverse propensity weights* (IPW),  $(p(W = 1|X)^{-1})$ , used by, for example, a DR-learner and X-learner.

**Results.** Consider table 1 where we reported the mean squared error (MSE) between an estimated ATE and the ground-truth ATE across ten folds, for ten differently sampled simulations (thus spanning 100 trained learners of each type). Given three popular ATE estimation strategies, we find that imputing  $\tilde{X}$  *significantly* performs better

across all methods, confirming the insights in Section 4.

### 6.2 Conditional average treatment effects (CATE)

**Objective.** Similar to Section 6.1, we will now evaluate how CATE-learners react to MCM. Specifically, we subject the same treatment effects learners to the same scenarios as we have in Section 6.1, and evaluate their CATE-predictions using the PEHE metric described above. As was the case for our ATE experiments, the chosen learners represent a wide range of different methodologies.

CATE estimates are evaluated using the *precision in estimating heterogeneous treatment effects* (PEHE), defined as,  $\mathbb{E}_{\mathcal{X}}[(\tau(X) - \hat{\tau}(X))^2]$  in Hill [27], where  $\hat{\tau}$  is a model’s prediction. Naturally, there is a parallel with the MSE we used in Section 6.1, where PEHE essentially corresponds to the MSE over the predicted vector of CATEs in a test set.

**Results.** table 1 depicts results for various CATE-learners across ten differently sampled simulations, each evaluated with ten different train and test sets— as we have for our ATE in Section 6.1. Given these results, we empirically confirm that one should impute cautiously as imputing all



data, no data, or wrong data, consistently performs worse than what we suggest: impute only  $\hat{X}$ . We confirm our findings with additional configurations in Appendix C.

## 7 Discussion

Estimating treatment effects is becoming more important in many practical settings. The adoption of these methods is largely the result of great academic effort to further push the boundaries of these methods' abilities. While practical adoption is indeed a point in favour of causal methods, it comes with a significant downside: unlike in academia, datasets used in practice are often victim to many imperfections. In our paper, we investigated one imperfection in particular: missing variables.

Missing variables in treatment effect settings behave differently from other settings; *they should be treated differently as a result*. In our paper, we argue that some missingness can be informative of the treatment, i.e. just like certain characteristics of an individual may determine their treatment, so too can the absence of certain measurements determine treatment. If one should impute these variables, that information is lost and can no longer be used to counter any resulting selection bias.

We argue further that we should not leave *all* variables unimputed. When there are variables that are only missing *because* an individual was given a particular treatment, then these missing variables are informative of the treatment and result in a covariate shift between treatment groups.

We believe the key finding of our paper is summarised as follows: *more care and thought should be put into imputing missing data when estimating treatment effects*. We confirm intuitively, theoretically, and empirically that selectively imputing missing variables can improve our ability to estimate treatment effect, while imputing the wrong variables *can lead to poorly modelled treatment effects*.

## Acknowledgements

This work was supported by the W.D. Armstrong Trust, the National Science Foundation (NSF) under grant number 1722516, the Office of Naval Research (ONR), and The Alan Turing Institute (ATI). We thank all reviewers for their invaluable comments and suggestions.

## References

- [1] Alaa, A. and van der Schaar, M. (2018). Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 129–138. PMLR.
- [2] Alaa, A. M., Gurdasani, D., Harris, A. L., Rashbass, J., and van der Schaar, M. (2021). Machine learning to guide the use of adjuvant therapies for breast cancer. *Nature Machine Intelligence*, 3(8):716–726.
- [3] Alaa, A. M. and van der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- [4] Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083.
- [5] Alshami, A. M. (2014). Knee osteoarthritis related pain: a narrative review of diagnosis and treatment. *International journal of health sciences*, 8(1):85–104.
- [6] Andersen, P. K., Syriopoulou, E., and Parner, E. T. (2017). Causal inference in survival analysis using pseudo-observations. *Statistics in medicine*, 36(17):2669–2681.
- [7] Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1):80–98.
- [8] Berrevoets, J., Alaa, A., Qian, Z., Jordon, J., Gimson, A. E., and Van Der Schaar, M. (2021). Learning queueing policies for organ transplantation allocation using interpretable counterfactual survival analysis. In *International Conference on Machine Learning*, pages 792–802. PMLR.
- [9] Berrevoets, J., Jordon, J., Bica, I., Gimson, A., and van der Schaar, M. (2020). OrganITE: Optimal transplant donor organ offering using an individual treatment effect. In *Advances in Neural Information Processing Systems*, volume 33, pages 20037–20050. Curran Associates, Inc.
- [10] Bica, I., Alaa, A. M., Lambert, C., and Van Der Schaar, M. (2021). From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1):87–100.
- [11] 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. *Biometrical Journal*, 62(2):428–443.
- [12] Burton, A. and Altman, D. (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British journal of cancer*, 91(1):4–8.
- [13] Curth, A., Lee, C., and van der Schaar, M. (2021). Survite: Learning heterogeneous treatment effects from time-to-event data. *Advances in Neural Information Processing Systems*, 34.

- [14] Curth, A. and van der Schaar, M. (2021). Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1810–1818. PMLR.
- [15] D’Agostino Jr, R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95(451):749–759.
- [16] Debaere, S., Devriendt, F., Brunner, J., Verbeke, W., De Ruyck, T., and Coussemont, K. (2019). Reducing inferior member community participation using uplift modeling: Evidence from a field experiment. *Decision Support Systems*, 123:113077.
- [17] Devriendt, F., Moldovan, D., and Verbeke, W. (2018). A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big data*, 6(1):13–41.
- [18] Edwards, S. J., Braunholtz, D. A., Lilford, R. J., and Stevens, A. J. (1999). Ethical issues in the design and conduct of cluster randomised controlled trials. *BMJ*, 318(7195):1407–1409.
- [19] Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd, 1<sup>st</sup> edition.
- [20] Gain, A. and Shpitser, I. (2018). Structure learning under missing data. In *International Conference on Probabilistic Graphical Models*, pages 121–132. PMLR.
- [21] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- [22] Hatt, T., Berrevoets, J., Curth, A., Feuerriegel, S., and van der Schaar, M. (2022). Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv preprint arXiv:2202.12891*.
- [23] Helliwell, T., Mallen, C., Peat, G., and Hay, E. (2014). Research into practice: improving musculoskeletal care in general practice. *British Journal of General Practice*, 64(624):372–374.
- [24] Hellman, S. and Hellman, D. S. (1991). Of mice but not men. *New England Journal of Medicine*, 324(22):1585–1589.
- [25] Hernán, M. A., Brumback, B. A., and Robins, J. M. (2002). Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures. *Statistics in medicine*, 21(12):1689–1709.
- [26] Hernán, M. A. and Robins, J. M. (2010). *Causal inference*. CRC Boca Raton, FL.
- [27] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.
- [28] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.
- [29] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- [30] Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning Research*, volume 48 of *Proceedings of Machine Learning Research*, pages 3020–3029, New York, New York, USA. PMLR.
- [31] Kallus, N., Mao, X., and Udell, M. (2018). Causal inference with noisy and missing covariates via matrix factorization. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- [32] Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.
- [33] Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.
- [34] Kyono, T., Zhang, Y., Bellot, A., and van der Schaar, M. (2021). Miracle: Causally-aware imputation via learning missing data mechanisms. *Advances in Neural Information Processing Systems*, 34.
- [35] Little, R. J., D’Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360.
- [36] Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.
- [37] Mattei, A. (2009). Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing. *Statistical Methods and Applications*, 18(2):257–273.
- [38] Mayer, I., Josse, J., Raimundo, F., and Vert, J.-P. (2020a). MissDeepCausal: Causal inference from incomplete data using deep latent variable models. *arXiv preprint arXiv:2002.10837*.

- [39] Mayer, I., Sverdrup, E., Gauss, T., Moyer, J.-D., Wager, S., and Josse, J. (2020b). Doubly robust treatment effect estimation with missing attributes. *The Annals of Applied Statistics*, 14(3):1409–1431.
- [40] Mohan, K. and Pearl, J. (2014). Graphical models for recovering probabilistic and causal queries from missing data. *Advances in Neural Information Processing Systems*, 27:1520–1528.
- [41] Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, pages 1–16.
- [42] Mohan, K., Pearl, J., and Jin, T. (2013). Missing data as a causal inference problem. In *Proceedings of the neural information processing systems conference (nips)*.
- [43] Nazábal, A., Olmos, P. M., Ghahramani, Z., and Valera, I. (2020). Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501.
- [44] Neyman, J. (1923). Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51.
- [45] Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216.
- [46] Pearl, J. (2009). *Causality*. Cambridge university press.
- [47] Prague, M., Wang, R., Stephens, A., Tchetgen Tchetgen, E., and DeGruttola, V. (2016). Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes. *Biometrics*, 72(4):1066–1077.
- [48] Raghunathan, T., Lepkowski, J., Hoewyk, J., and Solenberger, P. (2000). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27.
- [49] Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013.
- [50] Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Halloran, M. E. and Berry, D., editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 1–94, New York, NY. Springer New York.
- [51] Rombaut, E. and Guerry, M.-A. (2020). The effectiveness of employee retention through an uplift modeling approach. *International Journal of Manpower*.
- [52] Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524.
- [53] Roßbach, B. P., Pietschmann, M. F., Gülecüyüz, M. F., Niethammer, T. R., Ficklscherer, A., Wild, S., Jansson, V., and Müller, P. E. (2014). Indications requiring preoperative magnetic resonance imaging before knee arthroscopy. *Archives of medical science : AMS*, 10(6):1147–1152.
- [54] Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, 93(444):1321–1339.
- [55] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.
- [56] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- [57] Rubin, D. B. (1978). Multiple imputations in sample surveys—a phenomenological bayesian approach to non-response. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association.
- [58] Rubin, D. B. (1980). Comment on "randomization analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association*, 75(371):591.
- [59] Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.
- [60] Shah, N. V., Solow, M., Kelly, J. J., Aylyarov, A., Doran, J. P., Bloom, L. R., Akil, S., Siddiqui, B., Newman, J. M., Chatterjee, D., Pancholi, N., Dixit, A., Kavousi, B., Barbash, S. E., Urban, W. P., and Neuman, D. T. (2018). Demographics and rates of surgical arthroscopy and postoperative rehabilitative preferences of arthroscopists from the arthroscopy association of north america (aana). *Journal of orthopaedics*, 15(2):591–595.
- [61] Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085. PMLR.
- [62] Shpitser, I., Mohan, K., and Pearl, J. (2015). Missing data as a causal and probabilistic problem. Technical report, CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE.
- [63] Specialist Pharmacy Services, N. (2022). Medicines monitoring. Accessed: 2022-05-15.

- [64] Stekhoven, D. J. and Bühlmann, P. (2012). Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28 1:112–8.
- [65] Stitelman, O. M., De Gruttola, V., and van der Laan, M. J. (2012). A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. *The international journal of biostatistics*, 8(1).
- [66] Tabib, S. and Larocque, D. (2020). Non-parametric individual treatment effect estimation for survival data with random forests. *Bioinformatics*, 36(2):629–636.
- [67] Tolles, J. and Lewis, R. J. (2019). *Time-to-Event Analysis*. McGraw-Hill Education, New York, NY.
- [68] van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press LLC.
- [69] van Buuren, S. and Groothuis-Oudshoorn, C. (2011). MICE: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45.
- [70] Van der Laan, M. J., Laan, M., and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.
- [71] Yoon, J., Jordon, J., and van der Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5689–5698. PMLR.
- [72] Zhang, Y., Berrevoets, J., and Van Der Schaar, M. (2022). Identifiable energy-based representations: an application to estimating heterogeneous causal effects. In *International Conference on Artificial Intelligence and Statistics*, pages 4158–4177. PMLR.
- [73] Zhao, A. and Ding, P. (2021). To adjust or not to adjust? estimating the average treatment effect in randomized experiments with missing covariates. *arXiv preprint arXiv:2108.00152*.



# Appendix: MCM

## Table of Contents

---

<b>A</b>	<b>Proof of Theorem 4.1</b>	<b>13</b>
<b>B</b>	<b>Related work</b>	<b>14</b>
<b>C</b>	<b>Additional experiments</b>	<b>15</b>
<b>D</b>	<b>Data and reproducibility</b>	<b>17</b>
<b>E</b>	<b>MCM examples</b>	<b>18</b>
<b>F</b>	<b>Exhaustive DAG-search over CIT and CIO</b>	<b>19</b>

---

### A Proof of Theorem 4.1

**Theorem 4.1.** *Suppose the graph structure in fig. 3, and  $W$  and  $\overset{\rightarrow}{X}$  are independent, then ignorability defined as  $Y(0), Y(1) \perp\!\!\!\perp W | \tilde{X}^{\rightarrow}, \overset{\rightarrow}{X}$  holds.*

*Proof.* Consider  $p(Y = y, W = w | \tilde{X}^{\rightarrow} = \tilde{x}, \overset{\rightarrow}{X} = \dot{x})$ , which we write as  $p(y, w | \tilde{x}, \dot{x})$ . Then by introducing  $X = x$  and using the definition of the conditional density, we have

$$\begin{aligned} p(y, w | \tilde{x}, \dot{x}) &= \int_x p(y, w, x | \tilde{x}, \dot{x}) \\ &= \int_x \frac{p(y, w, x, \tilde{x}, \dot{x})}{p(\tilde{x}, \dot{x})} \end{aligned} \quad (5)$$

We can rewrite  $p(y, w, x, \tilde{x}, \dot{x})$  using the graph structure as

$$p(y, w, x, \tilde{x}, \dot{x}) = p(y|x)p(\tilde{x}|x)p(w|\tilde{x})p(\dot{x}|x, w)p(x). \quad (6)$$

Noting that  $X$  is *not* a collider of  $W$  and  $\overset{\rightarrow}{X}$ , and by independence of  $W$  and  $\overset{\rightarrow}{X}$  we have that

$$p(\dot{x}|x, w) = p(\dot{x}|x) \quad (7)$$

By Lemma A.1,  $\tilde{X}^{\rightarrow}$  and  $\overset{\rightarrow}{X}$  are conditionally independent given  $X$ . Therefore  $p(\tilde{x}|x)p(\dot{x}|x) = p(\tilde{x}, \dot{x}|x)$ . Similarly,  $X$  d-separates  $Y$  and the pair  $(\tilde{X}^{\rightarrow}, \overset{\rightarrow}{X})$  and thus,

$$p(y|x)p(\tilde{x}|x)p(\dot{x}|x) = p(y, \tilde{x}, \dot{x}|x). \quad (8)$$

Substituting (7) and (8) into (6), gives us

$$p(y, w, x, \tilde{x}, \dot{x}) = p(y, \tilde{x}, \dot{x}|x)p(w|\tilde{x})p(x). \quad (9)$$

We then note that  $\tilde{X}^{\rightarrow}$  is *not* a collider of  $W$  and  $\overset{\rightarrow}{X}$  so that,

$$p(w|\tilde{x}) = p(w|\tilde{x}, \dot{x}). \quad (10)$$

Substituting this all into (5) gives us

$$\begin{aligned} p(y, w | \tilde{x}, \dot{x}) &= \int_x \frac{p(y, \tilde{x}, \dot{x}|x)p(w|\tilde{x}, \dot{x})p(x)}{p(\tilde{x}, \dot{x})} \\ &= \frac{p(y, \tilde{x}, \dot{x})p(w|\tilde{x}, \dot{x})}{p(\tilde{x}, \dot{x})} \\ &= p(y|\tilde{x}, \dot{x})p(w|\tilde{x}, \dot{x}) \end{aligned} \quad (11)$$

which proves the result. □

**Lemma A.1.** *Suppose the setting of Theorem 4.1. Then  $\tilde{X}^\rightarrow$  and  $\dot{X}$  are conditionally independent given  $X$ .*

*Proof.* By appealing to the DAG to write the full joint distribution of  $p(x, \tilde{x}^\rightarrow, w, \dot{x}, \dot{x})$  we have

$$\begin{aligned}
 p(\tilde{x}^\rightarrow, \dot{x}|x) &= \int_{\tilde{x}, w} p(\tilde{x}^\rightarrow|x)p(w|\tilde{x}^\rightarrow, x)p(\dot{x}|x, w)p(\dot{x}|\tilde{x}, x) \\
 &= \int_w p(\tilde{x}^\rightarrow|x)p(w|\tilde{x}^\rightarrow)p(\dot{x}|x) \\
 &= p(\tilde{x}^\rightarrow|x)p(\dot{x}|x) \int_w p(w|\tilde{x}^\rightarrow) \\
 &= p(\tilde{x}^\rightarrow|x)p(\dot{x}|x)
 \end{aligned}
 \tag{12}$$

□

## B Related work

Ideas of dealing with missing variables have been around for some time [52], leading to many relevant works. However, let us first address works that may *seem* relevant, but are not. In particular, we are not concerned with missingness in either treatment or outcome variables [50, 54], only in the confounding variables  $X$ .

**Censoring.** The case of missing outcomes, or non-response [54], is well known in survival analysis [25, 26, 70, Chapter 17]. While survival analysis is indeed concerned with a particular case of missingness (that of outcomes), it is only relevant in a setting over time<sup>3</sup>. For example, if one wishes to predict the number of years a patient will live, when given a drug, we will only know the answer when they die. If the patient hasn't died yet at the time of analysis, we actually don't know the answer and have to consider the outcome missing (censored). Some works that aim to connect both fields include Andersen et al. [6], Curth et al. [13], Prague et al. [47], Stitelman et al. [65], Tabib and Larocque [66].

**$m$ -graphs.** Another line of research is that of *missingness graphs* or  $m$ -graphs [41]. In short, an  $m$ -graph is a probabilistic graphical model [32, 46] which explicitly includes missingness indicators as variables. Naturally, this is related as the way in which we describe MCM is through an  $m$ -graph. Using an  $m$ -graph, one can determine whether or not an effect is identifiable [46]. While this has received some attention in graphical causality [40, 42, 62] which aim to recover identifiability in MAR and MNAR settings or perform structure learning despite missingness [20, 34], never before were these  $m$ -graphs considered in a potential outcomes setting, as we have. Note that none of the aforementioned works consider partial imputation, specifically to correctly identify a causal effect.

**Conditional independence.** Identifying the causal effect in a potential outcomes setting, requires conditional independence:  $W \perp\!\!\!\perp Y(w)|X$ , for all  $w$  and  $X$ . If this is not the case, then we cannot interpret the estimated effect as *causal*. While much of the potential outcomes literature considers the setting where this is simply the case, it seems that these assumptions are simply carried over to the case that includes missingness. Having only three variables, the conditional independence assumptions merely amount to assuming that variables beyond  $X$  are *ignorable*, i.e. all relevant variables are measured. While this may very well be the case for three variables, this assumption may be “stretching it” when we include missingness indicators. This is especially clear when we portray the situation as an  $m$ -graph.

We are not the first to portray missingness in the potential outcomes setting as an  $m$ -graph. To our knowledge, Mayer et al. [39] was the first to do so (although no mention of  $m$ -graphs was made in their paper), and we know of none who has done so since. As such, Mayer et al. [39] is by far the most related work to ours, and we compare with them extensively in Appendix F. In particular, Mayer et al. [39] proposes two  $m$ -graphs for CIT and CIO (eqs. (2) and (3)), which we have included in figs. 5a and 6a, respectively. In Appendix F, we argue that neither of these  $m$ -graphs can exist in practice. Furthermore figs. 5a and 6a, only take into account missingness that causes treatment, but not the reverse, making CIT and CIO much more restrictive, as we have indicated in our main text.

<sup>3</sup>In fact, survival analysis is sometimes termed “time-to-event analysis” [67].

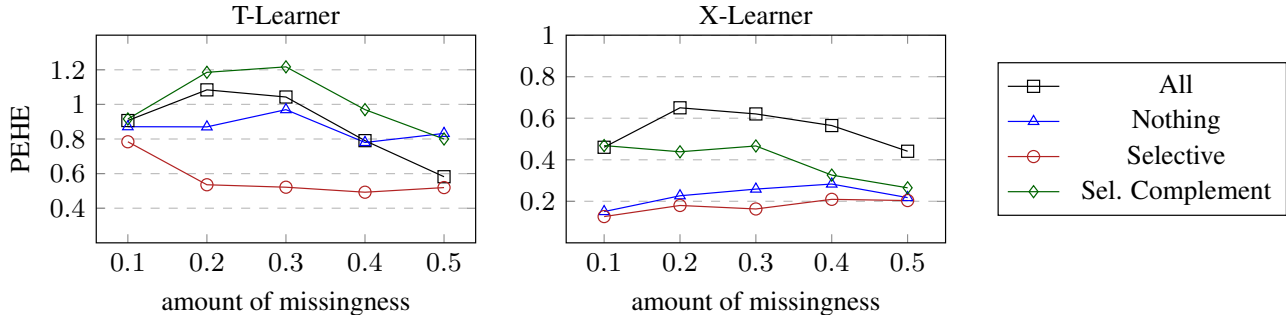


Figure 4: **Sensitivity analysis on the amount of missingness: CATE.** We report for two learners their performance (PEHE, y-axis) as a function of the amount of missingness (x-axis). For each setting of the amount of missingness, we sample 50 different train and test sets to calculate an average and std (in line with what is reported in table 1 and documented in our provided code-files).

### C Additional experiments

Here we present the following additional experiments to those presented in our main text: the same experimental setup as in table 1 with an alternative imputation scheme, a sensitivity analysis on the amount of missingness in a dataset, and the same experimental setup as in table 1, but using a different (real) dataset.

**Imputation with representation learning.** Consider table 2, where we have repeated the experiment in table 1 with GAIN [71]; a deep learning imputation method based on GANs [21]. While notoriously difficult to train, we find the combination of GAIN and MCM to be an interesting one. Specifically, the discriminator is tasked with classifying imputed samples from real samples, making it harder to predict which samples were imputed. The way the discriminator does this, is by predicting a binary mask of sorts, corresponding to what we have defined as  $Z$ . Having 50/50 prediction (resulting in the worst cross-entropy for binary prediction), we say that  $Z$  is independent of the imputed result, which then corresponds with theorem 4.1. Note that in theorem 4.1 we pose independence of  $W$ . Seeing that  $\tilde{Z}$  is a mediator variable between  $W$  and  $\tilde{X}$ , we have  $\tilde{X} \perp\!\!\!\perp W|X$  given that  $\tilde{X} \perp\!\!\!\perp Z|X$ .

**Sensitivity.** As a second additional experiment, consider fig. 4, where we have repeated the same experiment in table 1, but with different levels of missingness. Not only do these results confirm that selective imputation is best across the board, they also bring to light an interesting phenomenon: the trade-off between prediction gain from imputation versus bias reduction. As  $Y(w)$  is a function of the fully observed  $X$ , it makes sense that predicting  $Y(w)$  from  $X$  directly would yield the best results. This insight is the motivation why imputation is so common in standard supervised learning. From our discussion in Section 4, we learn that imputation in the treatment effects setting introduces bias, but perhaps there is only very little bias. In some cases, it might be better to impute, *despite bias*.

What we see in fig. 4 is the combination of the balance trade-off described above, with another realisation: if too much data is missing, bias might actually reduce, as the associated patterns in the data become less apparent. Take the most extreme case where all data in  $\tilde{X}$  is missing. Of course, then there is no bias anymore, as there is no more data to bias. With less bias in the data, it might be best for a model to impute. This is clearly visible in the leftmost plot (T-Learner), where not imputing does not seem to improve with more missingness; only the “Sel. Complement” and “All”— which introduce bias —seem to improve as more data is missing, likely since there is less bias available to introduce by imputing. Overall, we find that “Selective” outperforms the rest.

**Semi-real data.** In order to not only rely on synthetic data, we also test our findings in the well-known and used Twins dataset [4]. Naturally, it is semi-real, as we need to simulate the missingness mechanism such that we know the exact factors of  $\tilde{Z}$  and  $Z^*$ . In order to not let any other bias creep into the missingness mechanisms, we removed all rows that already contained missing variables.

Consider table 3 which presents results using the same setup as tables 1 and 2. Again, we find no surprises and report consistent performance improvements when selectively imputing the  $\tilde{Z}$  factor of the data. For this, we again relied on MICE for imputing the data [69].

Table 2: **CATE results with GAIN.** We repeat our experiment from the main text. For each setting, we report a PEHE on 100 *differently sampled* simulations and averaged the results, standard deviation is reported in brackets. Imputing only  $\tilde{X}$ , while keeping  $\tilde{X}^\rightarrow$  as is, performs best across learners and treatments. **Lower is better**, our proposal is shaded .

		Impute		T-learner	Doubly rob.	X-learner
		description	covariates			
<b>PEHE</b>	All	$\tilde{X}^\rightarrow \cup \tilde{X}$		0.8720 (0.072)	1.4276 (1.410)	0.5388 (0.112)
	Nothing	$\emptyset$		0.8233 (0.064)	0.2421 (0.579)	0.4724 (0.072)
	Selective	$\tilde{X}$		<b>0.5260</b> (0.105)	<b>0.1497</b> (0.210)	<b>0.2633</b> (0.069)
	Sel. Complement	$\tilde{X}^\rightarrow$		0.8855 (0.056)	0.4726 (0.971)	0.6924 (0.152)
<b>PEHE<sub>w=0</sub></b>	All	$\tilde{X}^\rightarrow \cup \tilde{X}$		0.8918 (0.107)	1.6743 (1.218)	0.5684 (0.138)
	Nothing	$\emptyset$		0.7079 (0.087)	0.2719 (0.340)	0.4829 (0.116)
	Selective	$\tilde{X}$		<b>0.5754</b> (0.092)	<b>0.1539</b> (0.209)	<b>0.2699</b> (0.064)
	Sel. Complement	$\tilde{X}^\rightarrow$		0.8342 (0.105)	0.3531 (0.604)	0.7313 (0.200)
<b>PEHE<sub>w=1</sub></b>	All	$\tilde{X}^\rightarrow \cup \tilde{X}$		0.8650 (0.074)	2.0708 (0.709)	0.5169 (0.107)
	Nothing	$\emptyset$		0.8844 (0.080)	0.2798 (0.706)	0.4650 (0.070)
	Selective	$\tilde{X}$		<b>0.5035</b> (0.132)	<b>0.1482</b> (0.213)	<b>0.2590</b> (0.075)
	Sel. Complement	$\tilde{X}^\rightarrow$		0.9149 (0.068)	0.5332 (1.197)	0.6664 (0.157)

Table 3: **Results on Twins data.** We repeat our experiments in table 1, presented in our main text, but on a different (real world) dataset [4]. Again, we sampled 100 different subsets of the data, with a different outcome and missing variables, and averaged the results with standard deviation in brackets. Different from our reported results in the main text, we normalised the results in order to better compare across methods as we found the performance discrepancy to increase when compared to the fully synthetic setup. Again, **lower is better**, and our proposal is shaded .

		Impute		T-learner	Doubly rob.	X-learner
		description	covariates			
<i>Results on ATE-estimation</i>						
<b>MSE</b>	All	$\tilde{X}^\rightarrow \cup \tilde{X}$		5.7554 (2.535)	8.9752 (3.047)	30.792 (8.459)
	Nothing	$\emptyset$		6.3382 (3.156)	10.310 (4.182)	8.9802 (5.661)
	Selective	$\tilde{X}$		<b>3.8482</b> (2.017)	<b>5.7543</b> (2.403)	<b>8.1534</b> (4.562)
	Sel. Complement	$\tilde{X}^\rightarrow$		6.5303 (3.150)	9.2979 (3.739)	10.369 (6.097)
<i>Results on CATE-estimation</i>						
<b>n-PEHE</b>	All	$\tilde{X}^\rightarrow \cup \tilde{X}$		0.7572 (0.101)	0.0699 (0.122)	0.8971 (0.065)
	Nothing	$\emptyset$		0.7393 (0.295)	0.0015 (0.003)	0.9191 (0.114)
	Selective	$\tilde{X}$		<b>0.7289</b> (0.302)	<b>0.0013</b> (0.015)	<b>0.8868</b> (0.119)
	Sel. Complement	$\tilde{X}^\rightarrow$		0.7591 (0.095)	0.1441 (0.376)	0.9067 (0.062)
<b>n-PEHE<sub>w=0</sub></b>	All	$\tilde{X}^\rightarrow \cup \tilde{X}$		0.9590 (0.064)	0.0792 (0.116)	0.9191 (0.071)
	Nothing	$\emptyset$		0.7451 (0.079)	0.0026 (0.004)	0.7557 (0.066)
	Selective	$\tilde{X}$		<b>0.7407</b> (0.076)	<b>0.0028</b> (0.028)	<b>0.7406</b> (0.064)
	Sel. Complement	$\tilde{X}^\rightarrow$		0.9570 (0.066)	0.1522 (0.376)	0.9139 (0.068)
<b>n-PEHE<sub>w=1</sub></b>	All	$\tilde{X}^\rightarrow \cup \tilde{X}$		0.7152 (0.109)	0.0695 (0.122)	0.8803 (0.066)
	Nothing	$\emptyset$		0.7188 (0.319)	0.0015 (0.002)	0.9090 (0.120)
	Selective	$\tilde{X}$		<b>0.7086</b> (0.327)	<b>0.0001</b> (0.015)	<b>0.8700</b> (0.126)
	Sel. Complement	$\tilde{X}^\rightarrow$		0.7175 (0.104)	0.1438 (0.363)	0.8810 (0.064)



## D Data and reproducibility

Standard in treatment effects literature, proper empirical validation requires synthetic data. Here we describe our synthetic setup, and how we sample from it precisely. We do this by including Python code, here in the appendix. For further clarifications (beyond this appendix), both on data generation as well as selective imputation, we refer the reader to our online code repository:

<https://github.com/jeroenbe/mcm>

We also like to point the interested reader to a wide collection of related code repositories to this work, which includes the code repository of this paper:

<https://github.com/vanderschaarlab/mlforhealthlabpub>

**[Step 1] Sample  $X$ .** In order for imputation to make sense, there has to exist some correlation between variables in  $X$ , as such sampling from a standard normal with the identity matrix as a covariance matrix, will not suffice. As such, we sample from a normal distribution with a random (positive semidefinite) covariance matrix, spanning 20 dimensions:

```

1 import numpy as np
2
3 def _generate_covariates(d, n):
4     assert 0 < d
5     assert 0 < n
6
7     A = np.random.rand(d, d)
8     cov = np.dot(A, A.transpose())
9
10    X = np.random.multivariate_normal(np.zeros(d), cov, size=n)
11    X /= (X.max() - X.min())
12
13    return X

```

In the above code, we have  $d=20$  the dimension count, and  $n=10000$  the sample size. In our experiments, we sample 10 different train and test sets from the data to calculate per simulation descriptives.

**[Step 2] Sample  $Z^\rightarrow$ .** From  $X$  we generate  $Z^\rightarrow$  as follows:

```

14    highest_border = X[:, :z_dim].argsort(axis=1)[:,
15        -int(np.max((int(np.round(amount_of_missingness * z_dim)), 1)))]
16    Z_out = list(x >= x[highest_border[i]] for i, x in enumerate(X[:, :z_dim]))
17    Z_out = np.array(Z_out).astype(int)
18    Z_out = np.abs(Z_out-1)

```

In the above code, we have two main parameters:  $z\_dim=10$  indicating the number of variables in  $Z^\rightarrow$ , and  $amount\_of\_missingness=0.3$  indicating the fraction of the data that is missing. In ln 14 above, we calculate a threshold, when the value of  $X_i$  above this threshold, that variable is missing. This threshold corresponds to the number of dimensions that need to be missing in order to respect  $amount\_of\_missingness$ . Notice that  $Z^\rightarrow$  now corresponds to  $X[:, :z\_dim]$ .

**[Step 3] Sample  $W$ .** We make very explicit that treatment choices depend on  $Z^\rightarrow$ .

```

19    W = []
20    for z_d in Z_out:
21        if 0 == z_d[-1]:
22            w = 0
23        elif 0 in z_d[:int(np.floor(z_dim/2))]:
24            w = 1
25        else:
26            w = np.random.binomial(1, .5)
27        W.append(w)
28    W = np.array(W)

```

[Step 4] **Sample  $\vec{Z}$ .** Sampling  $\vec{Z}$ , which depends on  $W$  in such a way that the arrow is identifiable, requires interaction between  $X$  and  $W$ <sup>4</sup>. For this, we sample two random vectors (`theta_z_in_0`, and `theta_z_in_1`) and let those interact with  $X$  to decide  $\vec{Z}$ . As was the case with  $Z^\rightarrow$ , we also calculate the number of variables that should be affected (calculated in ln 29–32). Note that the `dim_count` in this setting only corresponds *approximately* to the `amount_of_missingness`.

```

29 import scipy
30
31 dim_count = np.round(amount_of_missingness * (d - z_dim) * 2)
32 dim_count = np.max((dim_count, 1))
33 dim_count = np.min((dim_count, int((d - z_dim) / 2)))
34 dim_count = int(dim_count)
35
36 theta_z_in_0 = np.full(dim_count, scipy.stats.norm.ppf(1 - amount_of_missingness))
37 theta_z_in_1 = np.full(dim_count, scipy.stats.norm.ppf(1 - amount_of_missingness))
38
39 Z_in = np.zeros((n, d - z_dim))
40 for i, z in enumerate(Z_in):
41     x = X[i, z_dim:z_dim+dim_count]
42     if W[i]:
43         Z_in[i, -dim_count:] = (x - X[:, z_dim:z_dim+dim_count].mean(axis=0)) > (
theta_z_in_1 * x.std(axis=0))
44     else:
45         Z_in[i, :dim_count] = (x - X[:, z_dim:z_dim+dim_count].mean(axis=0)) > (
theta_z_in_0 * x.std(axis=0))
46 Z_in = np.abs(Z_in-1)

```

[Step 5] **Sample  $Y(w)$ .** Generating outcomes is done simply by sampling two linear functions.

```

47 def _generate_outcomes(X, W):
48     theta = np.random.randn(X.shape[1]) / 10
49     theta_y0 = np.ones(X.shape[1]) + theta
50     theta_y1 = np.ones(X.shape[1]) * -1 + theta
51
52     Y0 = np.sum(X * theta_y0, 1)
53     Y1 = np.sum(X * theta_y1, 1)
54
55     Y = np.array([Y0[i] if w == 0 else Y1[i] for i, w in enumerate(W)])
56     Y += np.random.randn(X.shape[0]) * .1
57
58     CATE = Y1 - Y0
59
60     return Y0, Y1, CATE, Y

```

[Step 6] **Identifiability.** All the above are simple functions, which are made identifiable (such that they respect the DAG) through a non-linearity. Then,  $\tilde{X}$  is generated by combining  $Z$  into  $X$ .

```

61 X = np.abs(X)
62
63 X_tilde = X.copy()
64 X_tilde[:, z_dim:] [Z_out==0] = missing_value
65 X_tilde[:, :z_dim] [Z_in==0] = missing_value

```

## E MCM examples

In the main manuscript, we provide an example (knee pain) of how MCM arises in practice, and in particular how missingness may be caused by the treatment. In addition to our example, we provide an additional (non-medical) example below.

**Job-training program.** We now present a quite different example of how MCM may arise in practice from outside of medicine.

<sup>4</sup>Note that in the non-binary setting, this interaction may not be necessary.

*Imagine a job-training program to boost employment. Before sponsoring such a program, a legislative body may want to estimate its effect before widespread adoption. Using past data on the program, the body has to rely on causal methods to infer the effect of the training program before formal adoption.*

Naturally, one requires data to learn a treatment effects model. But as is motivated throughout the literature, these data may be biased and cannot directly be used to learn a good treatment effects model.

*In the past, the program was not offered to everyone. Those with a job were less likely to be considered. As were those that had been unemployed for a long time. Perhaps age played a role; being less likely to switch careers, older people would benefit less from job training.*

While there exists “natural” bias, for example age may influence whether someone may or may not receive the treatment. In MCM, we argue that the fact that a variable is missing, may further bias the data.

*Without knowledge of someone’s age, it was deemed better to not offer the program. This is understandable, as the high price of the program requires some level of certainty.*

Bias does not stop there. Even when treatment is assigned, data may be biased further.

*Once an applicant has accepted their offer for job training, the organisers require some additional information. Imagine, the program’s registration process asking the applicant’s current address, the job of their spouse, the number of children they have, etc. Had they not accepted to participate in the program, they would be less likely to provide this additional information.*

Understanding where bias may come from, what could go wrong when these sources are ignored, or brushed over too lightly? One way this bias may sneak into our models is by using an estimator that knows how to handle missing data (such as an xgboost model) and choose to not impute anything.

*The organisers found that they had vastly overestimated the effect of job training. Those willing to provide additional information were also very motivated to advance their careers, and conversely, those that were reluctant to provide information were not motivated at all. By not imputing, the organisers’ model was identifying the effect of motivation, not of job training.*

However, when naively imputing everything, bias still creeps into our model.

*However, when they did decide to impute, their model seemed to underestimate the treatment effect, as the imputation strategy filled in absent age close to the mean age, however, it was mostly older people who did not provide their age. When adjusting for “younger” people, their model was actually confusing its estimates with data from older people, where their job-training program yielded less effect.*

This final point is subtle. Through the act of debiasing, the model ignored vital information. Indeed, there exists some bias from  $X$  to  $W$ , through the missing variables: namely, older people are more likely to have missing age. This is a clear example of removing a placeholder (the missingness indicator) for information that contributed to bias. Crucially, age plays an important role in outcome as well, propagating this bias into the effect estimates.

## F Exhaustive DAG-search over CIT and CIO

CIT and CIO are assumptions over  $X$ ,  $\tilde{X}$ ,  $Z$ ,  $Y(w)$ , and  $W$ . While keeping  $(X) \rightarrow (Z) \rightarrow (\tilde{X})$  fixed, we discuss each other DAG that respects eq. (2) or eq. (3).

In particular, for a DAG to be a valid description for missingness in CATE, they *should contain*:

$(X) \rightarrow (Y)$ , which is a standard arrow in CATE (cfr. fig. 1a). Essentially, the potential outcome  $Y(w)$  is the result of a natural process involving the fully observed  $X$ .

$(\tilde{X}) \rightarrow (W)$ , when a clinician determines treatment, they have to make do with what is given to them, i.e. the observed covariate set  $\tilde{X}$ , which may be partially unobserved due to missingness. This proxy to  $X$  is the best a clinician has available to them as the fully observed  $X$  is not always available to them.

$(Z) \rightarrow (W)$ , throughout our work we consider  $Z$  to be different from  $\tilde{X}$ , where  $\tilde{X}$  entails the actual *values* of the covariate set,  $Z$  indicates their presence. If or not a variable is present for a clinician to base their treatment decision on, can have an effect on their eventual decision. Say that it is too risky for a particular treatment when a patient’s blood pressure is not observed, then a different treatment option will be chosen (or the variable will be measured before a decision is made). In

this setting, the absence of a value has determined treatment, leading  $Z$  into  $W$ .

Continuing our discussion, for a DAG to be valid as a missingness description, they *should not contain* the following:

$(X) \rightarrow (W)$ , as, again,  $X$  is simply not available for a clinician to base their treatment-decision on.

$(\tilde{X}) \rightarrow (Y)$ , as  $Y(w)$  is the result of a natural process, it should depend on  $X$ , not  $\tilde{X}$ .

$(Z) \rightarrow (Y)$ , similarly,  $Y(w)$ , should not depend on the missingness  $Z$  as different datasets on the same person would register different outcomes when outcomes depend directly on  $Z$ , which cannot happen.

**Permutations on CIT.** For CIT, there cannot be a direct arrow from  $X$  to  $W$  (a feat we agree with), as it automatically violates eq. (2). As long as there exists a path between  $W$  and  $Y$ , excluding the connecting SWIG-path [49], we consider the DAG valid. When excluding paths *going up* from  $W$  or  $Y(w)$ , the total amount of valid DAGs amounts to  $(C_3^1 + C_3^2 + C_3^3) \cdot (C_2^1 + C_2^2) = 21$ , where  $C_3$  is coming from three potential paths into  $Y$  (having three variables, other than  $W$  and  $Y(w)$ ), and  $C_2$  is coming from two potential paths into  $W$  (three variables, excluding  $X$ ). All the CIT DAGs, including the one presented in Mayer et al. [39] are illustrated in fig. 5 (the DAGs presented in [39] correspond to figs. 5a and 6a for CIT and CIO, respectively).

Figure 5 is organised as follows: each two rows contain 7 graphs (corresponding to  $(C_3^1 + C_3^2 + C_3^3)$ ) where we vary the arrows going into  $Y(w)$ ; there are three sets of graphs (where a set is two rows), for each set we vary the arrows going into  $W$  (where three then corresponds to  $(C_2^1 + C_2^2)$ ).

Handy with table 4, we find that fig. 5g is the only valid DAG from all CIT (and later we see all CIO) compatible DAGs. Inspecting fig. 5g we notice that it corresponds exactly with fig. 1d, without the bidirectional arrow between  $W$  and  $Z$ , which is a result of informative versus uninformative missingness. Splitting  $Z$  in  $Z^\rightarrow$  and  $^\rightarrow Z$  results automatically in MCM.

**Permutations on CIO.** For CIO, the definition does not allow an arrow from  $X$  to  $Y(w)$ , similarly to the definition of CIT not allowing an arrow from  $X$  to  $W$ . Contrasting CIT, however, we find this restriction not sensible. As we have already argued above,  $X$  is the *only* variable “justified” to influence the potential outcomes,  $Y(w)$ , directly. Any other variable influencing  $Y(w)$  would result in contrasting outcomes over different datasets, which would imply, for example, different tumour sizes for the same person across different datasets. Having CIO and accompanying eq. (3), results in 21 (applying the same calculation as for CIT) non-sensible DAGs, including the DAG presented in Mayer et al. [39]. All these DAGs are listed in fig. 6, and evaluated (like for CIT) in table 4.



Table 4: **Validity of CIT and CIO.** There are six criteria we argue missingness in CATE should follow. These criteria are indicated in the column headers as directed dependencies missingness should, or should not, include. Assuming that there are no arrows *going up* from  $W$  and  $Y(w)$  (for example, the potential outcomes  $Y(w)$ , cannot influence the covariates  $X$ ), we have 42 DAGs that respect eq. (1) and one of eq. (2) or eq. (3). Only one of these DAGs (indicated in **green**), respects each criteria. This DAG (fig. 5g) is a permutation of CIT, and a version of MCM that does not assume factors,  $Z^{\rightarrow}$  and  $\rightarrow Z$  in  $Z$ . In the below table, “✓” means presence and “✗” means absence, when these icons are black absence or presence is positive, when they are gray they are negative.

		Does not contain			Does contain		
		$(X) \rightarrow (W)$	$(\tilde{X}) \rightarrow (Y)$	$(Z) \rightarrow (Y)$	$(X) \rightarrow (Y)$	$(\tilde{X}) \rightarrow (W)$	$(Z) \rightarrow (W)$
CIT	fig. 5a	assum.	✓	✓	✓	✓	✓
	fig. 5b	assum.	✗	✓	✓	✓	✓
	fig. 5c	assum.	✓	✗	✓	✓	✓
	fig. 5d	assum.	✓	✓	✗	✓	✓
	fig. 5e	assum.	✓	✗	✗	✓	✓
	fig. 5f	assum.	✗	✓	✗	✓	✓
	fig. 5g	assum.	✗	✗	✓	✓	✓
	fig. 5h	assum.	✓	✓	✓	✗	✓
	fig. 5i	assum.	✗	✓	✓	✗	✓
	fig. 5j	assum.	✓	✗	✓	✗	✓
	fig. 5k	assum.	✓	✓	✗	✗	✓
	fig. 5l	assum.	✓	✗	✗	✗	✓
	fig. 5m	assum.	✗	✓	✗	✗	✓
	fig. 5n	assum.	✗	✗	✓	✗	✓
	fig. 5o	assum.	✓	✓	✓	✓	✗
	fig. 5p	assum.	✗	✓	✓	✓	✗
	fig. 5q	assum.	✓	✗	✓	✓	✗
	fig. 5r	assum.	✓	✓	✗	✓	✗
	fig. 5s	assum.	✓	✗	✗	✓	✗
	fig. 5t	assum.	✗	✓	✗	✓	✗
fig. 5u	assum.	✗	✗	✓	✓	✗	
CIO	fig. 6a	✓	✓	✓	✓	✓	✓
	fig. 6b	✓	✓	✓	✓	✗	✓
	fig. 6c	✓	✓	✓	✓	✓	✗
	fig. 6d	✓	✓	✓	✗	✓	✓
	fig. 6e	✓	✓	✓	✗	✓	✗
	fig. 6f	✓	✓	✓	✗	✗	✓
	fig. 6g	✓	✓	✓	✓	✗	✗
	fig. 6h	✓	✗	✓	✓	✓	✓
	fig. 6i	✓	✗	✓	✓	✗	✓
	fig. 6j	✓	✗	✓	✓	✓	✗
	fig. 6k	✓	✗	✓	✗	✓	✓
	fig. 6l	✓	✗	✓	✗	✓	✗
	fig. 6m	✓	✗	✓	✗	✗	✓
	fig. 6n	✓	✗	✓	✓	✗	✗
	fig. 6o	✓	✓	✗	✓	✓	✓
	fig. 6p	✓	✓	✗	✓	✗	✓
	fig. 6q	✓	✓	✗	✓	✓	✗
	fig. 6r	✓	✓	✗	✗	✓	✓
	fig. 6s	✓	✓	✗	✗	✓	✗
	fig. 6t	✓	✓	✗	✗	✗	✓
fig. 6u	✓	✓	✗	✓	✗	✗	

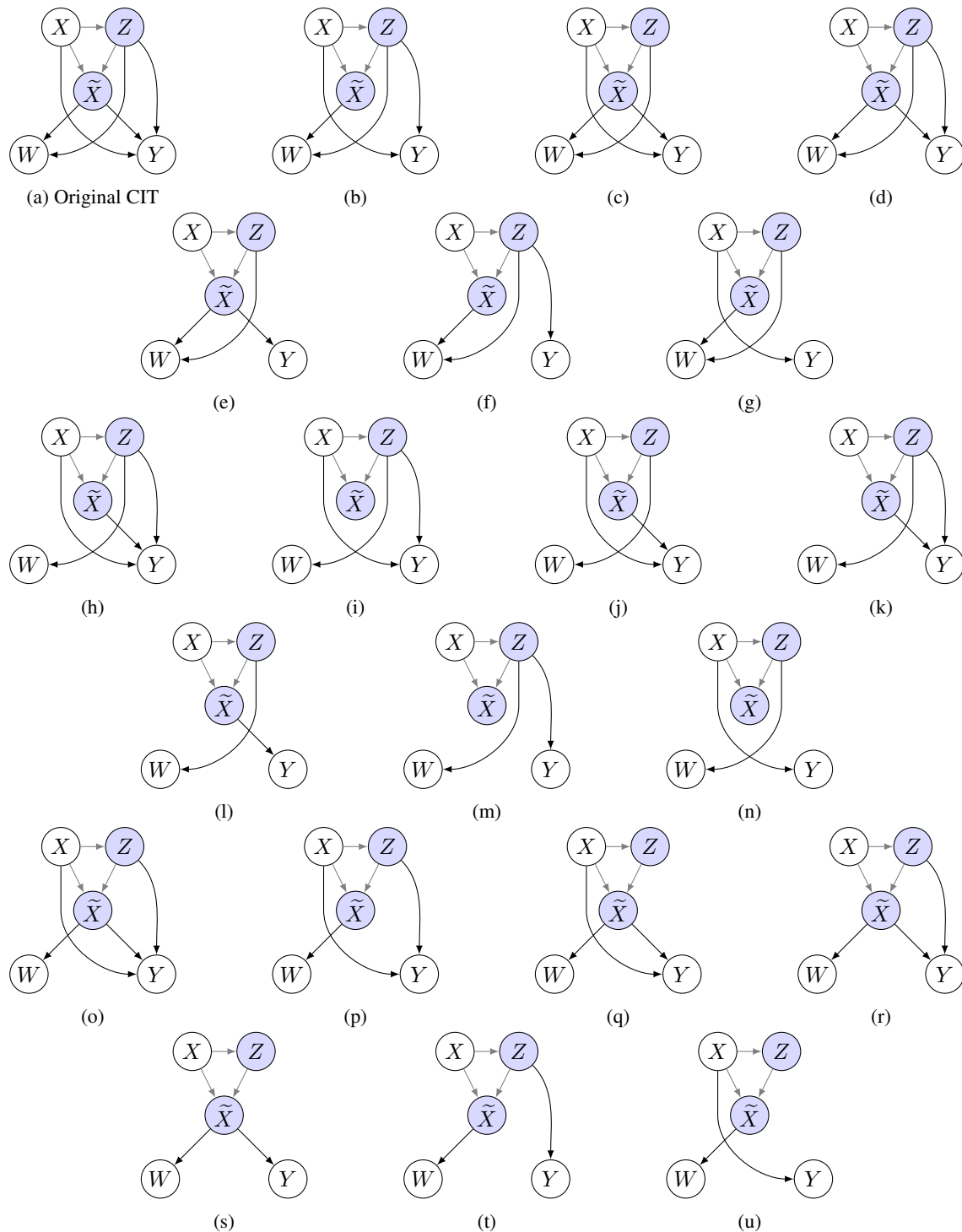


Figure 5: **Permutations on CIT.** There are a total of 21 DAGs that respect eq. (1) and eq. (2). From these DAGs, only one — Figure 5g — is acceptable. Add only the assumption that some elements in  $Z$  are informative, and some elements in  $Z$  are uninformative, and we automatically arrive at MCM. Note that we exclude edges *going up* from  $W$  or  $Y(w)$ , similarly to Richardson and Robins [49].

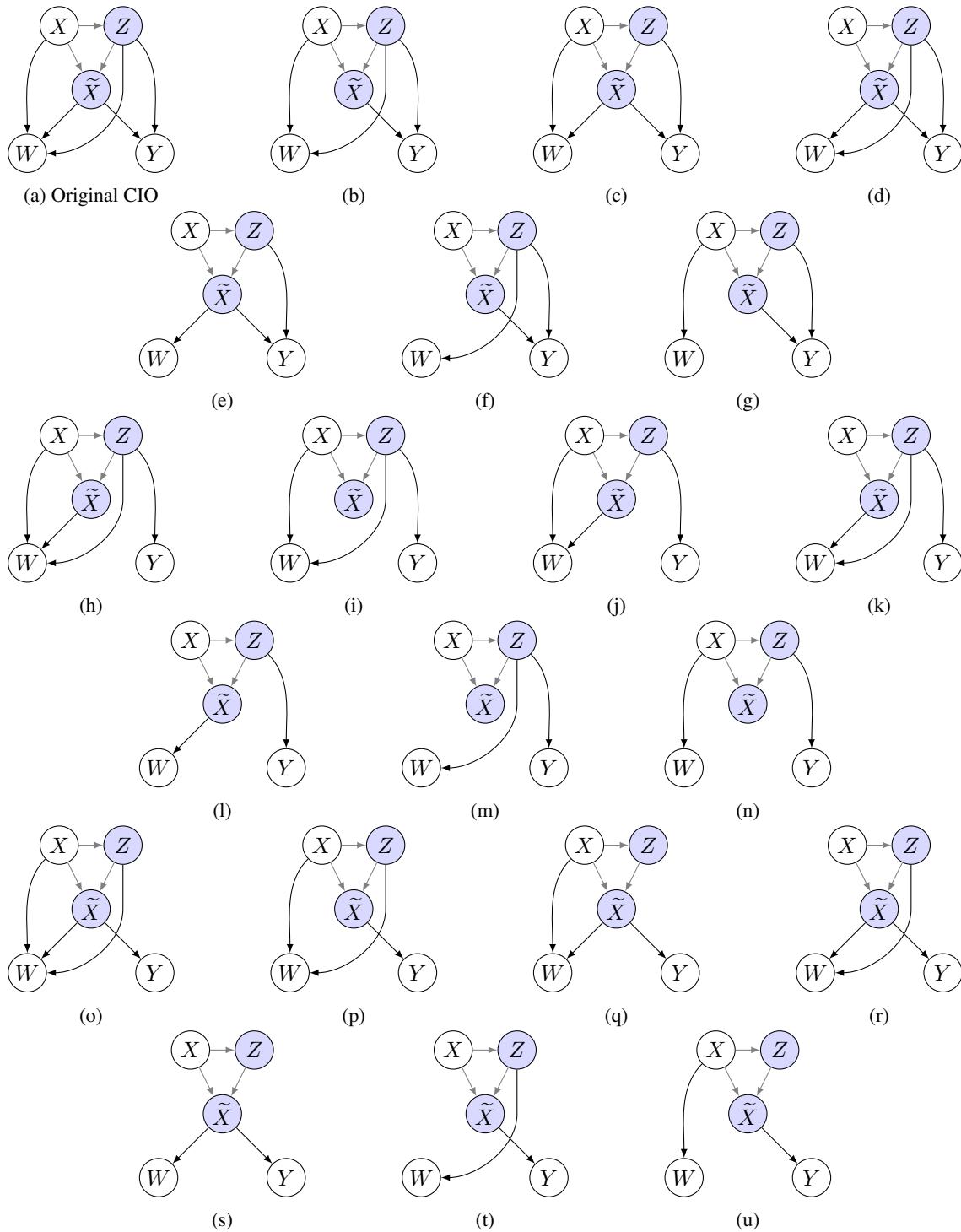


Figure 6: **Permutations on CIO.** There are 21 possible DAGs that respect eq. (3). None of them is acceptable as they, by definition, cannot include a direct edge from  $X$  to  $Y(w)$ . Having a direct edge between  $X$  and  $Y(w)$  encodes dependence, despite conditioning on  $\tilde{X}$  and  $Z$ . Note that we have excluded edges *going up* from  $W$  or  $Y(w)$ , in similar fashion to Richardson and Robins [49].