Causal Inference amid Missingness-Specific Independencies and Mechanism Shifts

Johan de Aguas[1,2] **Leonard Henckel**[3] **Johan Pensar**[1,4] **Guido Biele**[2]

[1]Department of Mathematics, University of Oslo
[2]Department of Child Health and Development, Norwegian Institute of Public Health
[3]School of Mathematics and Statistics. University College Dublin
[4]INTEGREAT – The Norwegian Center for Knowledge-Driven Machine Learning

Abstract

The recovery of causal effects in structural models with missing data often relies on m-graphs, which assume that missingness mechanisms do not directly influence substantive variables. Yet, in many real-world settings, missing data can alter decision-making processes, as the absence of key information may affect downstream actions and states. To overcome this limitation, we introduce lm-SCMs and lm-graphs, which extend m-graphs by integrating a label set that represents relevant context-specific independencies (CSI), accounting for mechanism shifts induced by missingness. We define two causal effects within these systems: the *full average treatment effect* (FATE), which reflects the effect in a hypothetical scenario *had no data been missing*, and the *natural average treatment effect* (NATE), which captures the effect under the unaltered CSIs in the system. We propose recovery criteria for these queries and present doubly-robust estimators for a graphical model inspired by a real-world application. Simulations highlight key differences between these estimands and estimation methods. Findings from the application case suggest a small effect of ADHD treatment upon test achievement among Norwegian children, with a slight effect shift due to missing pre-tests scores.

1 INTRODUCTION

In certain causality paradigms, the *fundamental problem of causal inference* —namely, the impossibility to simultaneously observe outcomes under both treatment and no treatment conditions— has traditionally been regarded as a missing data problem (Neyman 1923; Rubin 1974; Ding and Li 2018). Over the past decade, a growing perspective has taken shape in the opposite direction, recasting missing data problems as instances of causal inference. This viewpoint has led to an expanding body of research that combines methodologies from both areas (Mohan et al. 2013; Mohan and Pearl 2014; Saadati and Tian 2019; Bhattacharya et al. 2020; Mohan and Pearl 2021; Nabi et al. 2024).

A central focus of this research has been to determine the conditions under which causal and counterfactual queries can be recovered in the presence of missing data, by combining graphical assumptions about the underlying causal mechanisms with information from potential auxiliary experiments (Mohan et al. 2013; Bhattacharya et al. 2020; Mohan and Pearl 2021; Guo et al. 2023). Common queries include the *average treatment effect* (ATE) of an exposure variable upon an outcome variable (Bia et al. 2024; de Aguas et al. 2025), the *conditional average treatment effect* (CATE) and other functionals conditioned on evidence (Tikka et al. 2021; Kuzmanovic et al. 2023), the interventional distribution (Mohan et al. 2013; Bhattacharya et al. 2020), and the full data distribution (Nabi et al. 2020), among others.

Various techniques have been developed to recover and estimate causal effects under missing data and other forms of endogenous selection of unit information. These techniques include imputation methods (Rubin 1976, 1978; Kyono et al. 2021), covariate adjustment (Correa et al. 2018; Saadati and Tian 2019; Mathur et al. 2023), *inverse probability weighting* (IPW) or other ratios (Huber 2012; Mohan and Pearl 2014), and doubly-robust approaches (Wei et al. 2022; Negi 2024). In certain settings, especially under a semiparametric model and missing outcomes, it has been shown that multiply-robust and efficient estimators for the ATE can incorporate elements from imputation, covariate adjustment, and weighting, thereby providing a unifying framework for estimation (Bia et al. 2024; de Aguas et al. 2025).

In methodological and applied research, special attention has been given to cases of missing outcome data, as these can arise from common phenomena such as attrition or loss to follow-up, which can threaten the validity of causal claims in both observational and experimental studies (Hernán et al.

2004; Lewin et al. 2018; Biele et al. 2019). The problems of recovering causal queries from missing covariate data (Yang et al. 2019; Chang et al. 2023; Levis et al. 2024) and missing exposure data (Kuzmanovic et al. 2023; Shi et al. 2024) have also been studied, though aside from Saadati and Tian (2019), these works generally place less emphasis on graphical criteria. Furthermore, sound recovery algorithms and heuristics have been developed for general queries under multiple missingness and selection mechanisms (Bhattacharya et al. 2020; Tikka et al. 2021).

Several sets of sufficient conditions have been proposed for recovering causal effects from missing data. These typically combine: *(i)* the existence of an admissible adjustment set that, along with relevant missingness indicators, can block all backdoor paths between exposure and outcome in an m-graph —a graphical model encoding both causal relationships and missingness mechanisms; *(ii)* specific d-separation conditions between the missingness indicators and either the exposure or the outcome; *(iii)* a *missing at random* (MAR) assumption for the variables affected by missingness; and *(iv)* the requirement that missingness indicators are not causal descendants of the exposures transmitting the effect to the outcome, among other structural restrictions (Saadati and Tian 2019; Mohan and Pearl 2021). These conditions ensure that causal effect estimands, whether via covariate adjustment or IPW, remain valid despite the presence of missing data. Violations, by contrast, can induce selection bias and undermine causal conclusions for the target population (Hernán et al. 2004).

In most typical setups, if conditions for recovery are satisfied, the interpretation of the recovered effect is framed in an idealized scenario where no data are missing. For instance, a recovered ATE would be understood as the population-level expected difference in outcomes between treated and untreated scenarios, *ceteris paribus*, and *as if no data had been missing* (Nabi et al. 2024). This interpretation allows the problem of inference under missing data to be viewed in a manner akin to a counterfactual scenario.

Motivation Conventional m-graph-based approaches to missing data typically assume that missingness mechanisms affect only each other and do not have a direct causal impact on the substantive variables in the system (Mohan and Pearl 2021; Nabi and Bhattacharya 2023; Nabi et al. 2024). In other words, the missingness status of a variable does not alter the causal mechanisms governing substantive variables. For the problem of missing covariate data, this assumption helps ensuring that the causal mechanisms of the exposure, the outcome, and their relationship remain invariant to the missingness indicators.

However, in certain settings and applications, this constraint may be too restrictive or misaligned with the underlying data-generating process. Consider an observational study where schoolchildren take standardized tests at two consecutive time points, with corresponding scores $Y_0, Y_1 \in \mathbb{R}$. Between the two tests, some children receive stimulant medication, represented by the binary treatment indicator A. The goal is to estimate the ATE of stimulant medication on the second test score. Crucially, some students do not take the first test, indicated by the missingness indicator value $R_{Y_0} = 0$. If Y_0 is a key predictor of treatment assignment, it is plausible that units with missing test scores follow an alternative decision-making process regarding treatment, even after adjusting for other observed covariates. In other words, parents and clinicians of schoolchildren with missing first test scores may rely on different treatment decision rules than those who do observe Y_0. Furthermore, students who took the first test may be more familiar with the testing environment, potentially influencing their performance on the second test. This could result in systematic outcome differences between students who completed the first test and those who did not, even after accounting for other observed factors. In this setting, the missingness indicator directly influences both the exposure and the outcome, violating the conventional assumptions of m-graph models (see figures 1a and 1b) and potentially inducing significant downstream shifts.

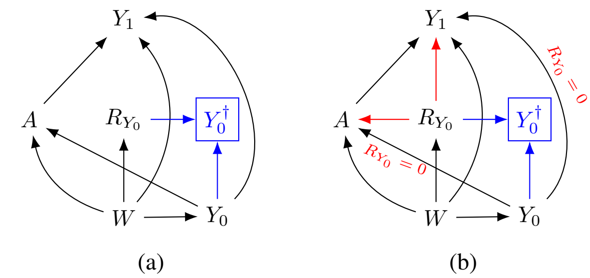


Figure 1: Graphical representations of systems with missing data on covariate Y_0 (a confounder of the casual relationship between A and Y_1): (a) An m-graph including the proxy variable $Y_0^\dagger$, represented in a blue box as a deterministic function of R_{Y_0} and Y_0. (b) An lm-graph illustrating labeled CSIs, where Y_0 is an input for the mechanisms of A and Y_1 when observed ($R_{Y_0} = 1$), but it is not when missing ($R_{Y_0} = 0$). Consequently, R_{Y_0} becomes a causal parent of A and Y_1.

Prior work Early methodological and applied work on missing covariate data highlighted the value of including interaction terms with missingness indicators to enhance inference robustness against potential mechanism shifts (Greenland and Finkle 1995; Jones 1996). This is known as the *missingness indicator method* (MIM), and extensive simulation studies have demonstrated that this approach is *nearly valid* in most common applied scenarios (Song et al. 2021; Chang et al. 2023). However, the MIM is not explicitly grounded in graphical criteria.

Recent work has explored various challenges at the intersection of missing data and mechanism shifts. For instance, Zhou et al. (2023) develop domain adaptation techniques for settings where missingness mechanisms differ across environments (e.g., between source and target populations), while the distributions of substantive variables remain stable. In the context of CATE generalization with missing exposure data, Kuzmanovic et al. (2023) study the impact of covariate shifts between treated and untreated groups, deriving generalization bounds to assess their influence. Addressing a similar generalization problem with missing covariates, Colnet et al. (2022) and Mayer et al. (2023) propose imputation strategies and sensitivity analyses to manage distributional changes. While some of these approaches are inspired by graphical representations of the system, they typically do not employ graphical criteria directly in their solutions.

As part of a broader effort to extend m-graph-based models, Srinivasan et al. (2023) relax standard assumptions by introducing various forms of *entanglement*, which capture classical unit interference as well as cases where one unit's missingness mechanism can influence proxy covariates of other units. However, to the best of our knowledge, no previous work has directly addressed the problem of recovering causal effects in the presence of missingness-specific independencies and mechanism shifts that affect downstream variables; nor has prior work developed semiparametric theory-based estimators tailored to such settings.

Contributions This paper addresses a limitation in current m-graph-based frameworks in causal inference with missing data, which do not allow for substantive descendants of the missingness indicators. As a result, m-graphs fail to capture missingness-induced mechanism shifts, which occur when missing data induce changes in downstream causal mechanisms. Empirical methods, such as the MIM, recommend the explicit modeling of interactions between missingness indicators and other covariates to make inference more robust to such shifts. However, these methods, despite being noted for their practical validity (Song et al. 2021; Chang et al. 2023), lack a solid theoretical justification within the graphical framework. This gap is addressed in this paper by introducing lm-SCMs and lm-graphs, which expand on the existing m-graph models by incorporating a label set with relevant context-specific independencies (CSI). We provide a theoretical justification for these systems as resulting from latent soft interventions, and propose recovery criteria that integrate conditions from both m-graph and labeled graph (l-graph) frameworks. These contributions advance the field of causal inference with missing data, addressing a critical gap in the literature.

2 PRELIMINARIES

Graph operations Consider a *directed acyclic graph* (DAG) $\mathcal{G}$ defined on nodes $\mathcal{V}$, and a subset $X \subseteq \mathcal{V}$. The notations $\mathrm{pa}(X;\mathcal{G})$, $\mathrm{ch}(X;\mathcal{G})$, $\mathrm{an}(X;\mathcal{G})$, and $\mathrm{de}(X;\mathcal{G})$ represent the parents, children, ancestors, and descendants of X in $\mathcal{G}$, respectively. The mutilated graph $\mathcal{G}[\underline{X}]$ is obtained by removing all outgoing edges from the nodes in X, whereas $\mathcal{G}[\overline{X}]$ is formed by eliminating all incoming edges to X.

Causal graphs A *structural causal model* (SCM) is a tuple $\mathcal{M} = (\mathcal{V},\mathcal{U},\mathcal{G},\mathcal{F},P_{\mathcal{U}})$. Here $\mathcal{V}$ represents a finite set of relevant variables; $\mathcal{U}$ is a finite set of noise variables; and $\mathcal{G}$ is a DAG over $\mathcal{V}$. The component $P_{\mathcal{U}}$ specifies a probability distribution for $\mathcal{U}$. The set $\mathcal{F} = \{f_V\}_{V\in\mathcal{V}}$ comprises a collection of measurable functions that describe the direct causal mechanisms: for each $V \in \mathcal{V}$, there exists an associated $U_V \in \mathcal{U}$ and a function $f_V : \operatorname{supp}\mathrm{pa}(V;\mathcal{G}) \times \operatorname{supp} U_V \to \operatorname{supp} V$ such that $V = f_V(\mathrm{pa}(V;\mathcal{G}), U_V)$ almost surely (Pearl 2009). Here, we denote with $\operatorname{supp} X$ the support of random variable X. In *semi-Markovian* models, bidirectional arrows $V_1 \leftrightarrow V_2$ indicate latent confounding $V_1 \leftarrow U \rightarrow V_2$, and $\mathcal{G}$ becomes an *acyclic directed mixed graph* (ADMG) (Evans and Richardson 2014).

Let A and Y denote the *exposure* and the *outcome*, respectively. The *unit-level counterfactual* or *potential outcome*, $Y^a(u)$, represents the value that Y would take if an intervention were performed by setting A to a fixed value $a \in \operatorname{supp} A$ for an individual characterized by $\mathcal{U} = u$ in the SCM $\mathcal{M}$. This intervention propagates through the system, updating descendant variables according to the causal mechanisms $\mathcal{F}$ applied in a topological order dictated by $\mathcal{G}$. Potential outcomes satisfy the consistency axiom: if $A(u) = a$ and $Y(u) = y$, then $Y^a(u) = y$ (Robins 1989; Pearl 2009). The corresponding population-level distribution, known as the *interventional distribution*, is expressed as $p(y \mid \mathrm{do}(A = a)) = \int \mathbb{I}\{u \in \mathcal{U}^a[y]\}\, \mathrm{d}P(u)$, where $\mathcal{U}^a[y] = \{u \in \operatorname{supp}\mathcal{U} : Y^a(u) = y\}$ is the pre-image of $y \in \operatorname{supp} Y$ under $Y^a(\cdot)$ (Bareinboim et al. 2022).

The *average treatment effect* (ATE), ψ, is one of the most commonly studied *causal effects* or *queries*. For a binary point-exposure $A \in \{0, 1\}$ and a continuous outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$, the ATE is defined as:

$$\psi := \Delta_a \mathbb{E}\left[Y \mid \mathrm{do}(A = a)\right], \tag{1}$$

where Δ_a denotes the difference operator relative to a.

A sound and complete set of three rules known as the *do-calculus* aids in the identification of interventional distributions by leveraging the conditional independencies implied by directional separation (*d-separation*) statements embedded in $\mathcal{G}$ and its mutilations (Pearl 1995, 2012).

Soft interventions *Soft / stochastic interventions* extend *do*-actions by enabling surgical modifications of mechanisms within an SCM $\mathcal{M}$ with causal graph $\mathcal{G}$. Instead of fixing variables to specific values, they replace original mechanisms with alternative functions and inputs. For a given $X \in \mathcal{V}$, a soft intervention $\sigma_X = (D, g, \widetilde{U})$ consists of new parents from its original nondescendants, $D \subseteq \mathrm{nde}(X;\mathcal{G})$, a measurable function g, and an independent auxiliary noise $\widetilde{U}$ (which may be empty). This intervention induces a new SCM $\mathcal{M}^{\sigma_X}$, where the original assignment $X \leftarrow f_X(\mathrm{pa}(X;\mathcal{G}), U_X)$ is replaced with $X \leftarrow g(D, U_X, \widetilde{U})$ almost surely. The corresponding *operated graph* $\mathcal{G}^{\sigma_X}$ is obtained by removing edges $\mathrm{pa}(X;\mathcal{G}) \to X$, adding edges $D \to X$, and introducing a *regime arrow* $\sigma_X \to X$. For foundational work on this topic, we refer the reader to Correa and Bareinboim (2019, 2020a,b).

Missingness graphs Let $\mathcal{M}$ be an SCM, and $\mathcal{V}_o \cup \mathcal{V}_m \cup \mathcal{R}$ be a partition of variables $\mathcal{V}$ in $\mathcal{M}$, where $\mathcal{V}_o$ contains fully observed variables, $\mathcal{V}_m$ contains variables affected by missing data, and $\mathcal{R} = \{R_V : V \in \mathcal{V}_m\}$ collects the missingness indicators. That is, if V is observed for an individual unit, then $R_V = 1$ for the same unit; and, if it is missing, then $R_V = 0$. The causal graph $\mathcal{G}$ in $\mathcal{M}$ is called a missingness graph, or *m-graph*, as it involves the indicators $\mathcal{R}$ in addition to the *substantive variables* $\mathcal{V}_o \cup \mathcal{V}_m$. It is typically assumed that $\mathcal{R}$ has no substantive descendants in $\mathcal{G}$ (Mohan and Pearl 2014, 2021; Nabi and Bhattacharya 2023). Let us denote with $V^\dagger$ the *proxy* for $V \in \mathcal{V}_m$, such that $V^\dagger = V$ almost surely when $R_V = 1$, and it takes the empty value $V^\dagger = \emptyset$ otherwise. The distribution $P_\mathcal{V}$ is known as the *full data distribution*, while $P_{\mathcal{V}^\dagger}$, with $\mathcal{V}^\dagger = \mathcal{V}_o \cup \mathcal{V}_m^\dagger \cup \mathcal{R}$, is the *observed data distribution*. In other formulations, V, $V^\dagger$ and $\emptyset$ are denoted respectively with V^1, V and $?/\mathtt{NA}$ (Nabi et al. 2024). The explicit m-graph can incorporate the proxy variables $V^\dagger$, highlighting their deterministic nature given their two parents R_V and V.

A causal query Ψ acting on SCMs, such as the interventional distribution or the ATE, is *(nonparametrically) recoverable* from incomplete observational data if it outputs the same value ψ for all models within the class of SCMs that share the same m-graph $\mathcal{G}$ and the same positive observed data distribution $\mathfrak{M}(\mathcal{G}, P)$. That is, $\forall \mathcal{M} \in \mathfrak{M}(\mathcal{G}, P)$, $\Psi[\mathcal{M}] = \psi$. In other words, the query is uniquely computable as a functional of the graph and the observed data, yielding a recovered statistical *estimand* (Bareinboim et al. 2014; Mohan and Pearl 2021). Graphical criteria can enable the recovery of causal queries even in complicated settings beyond the conventional MAR model (Mohan et al. 2013; Guo et al. 2023). Several sufficient criteria for recovery of causal effects have been proposed, some based on sequential factorizations and covariate adjustment (Mohan and Pearl 2014; Correa et al. 2018; Saadati and Tian 2019; de Aguas et al. 2025), and others based on ratio factorizations and IPW (Horvitz and Thompson 1952; Huber 2012; Mohan and Pearl 2014). More encompassing algorithmic approaches for assessing recoverability include the identification algorithm introduced by Bhattacharya et al. (2020) and the search heuristic by Tikka et al. (2021).

Labeled graphs A *context-specific independence* (CSI) is a form of local statement that extends the concept of conditional independence by holding only in certain specific contexts (Boutilier et al. 1996). For mutually exclusive sets of variables Z, X, B, C, the expression $Z \perp\!\!\!\perp X \mid B, C = c$ represents a CSI, indicating that Z and X are conditionally independent given B and the specific context $C = c$.

A *labeled graph*, l-graph, or LDAG, is a tuple $(\mathcal{G}, \mathcal{L})$ consisting of a DAG $\mathcal{G}$ and a *label set* $\mathcal{L} = \bigcup_{C \in \mathscr{C}} \bigcup_{c \in \nu(C)} \mathcal{L}_C(c)$. Here, $\mathscr{C} \subset \mathcal{V}$ is a set of context variables, $\nu(C) \subseteq \mathrm{supp}\, C$ is a set of context values for $C \in \mathscr{C}$, and $\mathcal{L}_C(c)$ contains a collection of edges $(X, Z) \in \mathrm{edg}(\mathcal{G})$ for which the CSI $Z \perp\!\!\!\perp X \mid \mathrm{pa}(Z;\mathcal{G}) \setminus (X \cup C), C = c$ holds (Pensar et al. 2015; Mokhtarian et al. 2022). If $(X, Z) \in \mathcal{L}_C(c)$, the label "$C = c$" can be graphically depicted on the arrow $X \to Z$.

An l-graph is called *regular-maximal* if *(i)* for all $(X, Z) \in \bigcup_{c \in \nu(C)} \mathcal{L}_C(c)$, one has that $\mathrm{pa}(Z;\mathcal{G}) \cap C \neq \varnothing$, and *(ii)* it is not possible to add an additional edge $(X, Z) \in \mathrm{edg}(\mathcal{G})$ to $\mathcal{L}$ without inducing a new CSI (Mokhtarian et al. 2022).

It is known that the *do*-calculus can fail to identify causal effects in nontrivial l-graphs that are indeed identifiable. Tikka et al. (2019) demonstrated that identifying causal effects from l-graphs is NP-hard and proposed a sound CSI-calculus along with a heuristic approach to obtain identifying formulas. Building on this, Mokhtarian et al. (2022) showed that when all context variables are *control* variables (i.e. having no parents in $\mathcal{G}$) a causal effect is identifiable from a regular-maximal l-graph if it can be identified from a series of causal graphs, each representing a different context world, where each identification problem can be assessed using existing sound and complete procedures.

3 LABELED MISSINGNESS GRAPHS

To formally define the causal effects and their recovery conditions under missingness-specific independencies, we first introduce lm-SCMs and lm-graphs. For notational convenience, we omit curly brackets when referring to singleton sets, allowing us to write X for $\{X\}$.

Definition 1 (**lm-SCM**) *Let $\mathcal{M}$ be a semi-Markovian SCM with m-graph structure $\mathcal{G}^*$. Let the set of variables affected by missingness in $\mathcal{M}$ be partitioned as $\mathcal{V}_m = \mathcal{V}_{\mathrm{no}} \cup \mathcal{V}_{\mathrm{sh}}$, where $\mathcal{V}_{\mathrm{no}}$ are variables whose missingness does not induce mechanism shifts, and $\mathcal{V}_{\mathrm{sh}}$ are those whose missingness may induce mechanism shifts. Let $\mathcal{R} = \mathcal{R}_{\mathrm{no}} \cup \mathcal{R}_{\mathrm{sh}}$ be the corresponding partition for the missingness indicators.*

For each variable $X \in \mathcal{V}_{\mathrm{sh}}$:

- *Let $\mathcal{S}_X \subseteq \mathrm{ch}(X;\mathcal{G}^*) \setminus \mathrm{an}(R_X;\mathcal{G}^*)$ be a nonempty set of selected children of X that are not ancestors of R_X, representing the* ***shifted children*** *of X.*

For any variable $Z \in \mathcal{V}$:

- *Define the set of* ***contextual parents*** *of Z as $\mathcal{T}_Z := \{X \in \mathcal{V}_{\text{sh}} : Z \in \mathcal{S}_X\}$.*
- *For Z with $\mathcal{T}_Z \neq \varnothing$, let σ_Z be a* soft intervention *that modifies the natural assignment $Z \leftarrow f_Z(\text{pa}(Z;\mathcal{G}^*), U_Z)$ governing Z in the original SCM $\mathcal{M}$ to be:*

$$Z \leftarrow \left[\prod_{X \in \mathcal{T}_Z} R_X\right] f_Z(\text{pa}(Z;\mathcal{G}^*), U_Z) + \sum_{\substack{T \subseteq \mathcal{T}_Z \\ T \neq \varnothing}} \left[\prod_{X \in T} (1 - R_X)\right] g_{Z,T}(\text{pa}(Z;\mathcal{G}^*) \setminus T, U_Z),$$

where each $g_{Z,T}$ is a measurable function that defines a shifted mechanism when data from a subset T of contextual parents are missing.

The SCM $\mathcal{M}^\sigma$ induced by joint soft interventions $\sigma = \{\sigma_Z : \mathcal{T}_Z \neq \varnothing\}$ is referred to as a labeled missingness SCM.

An lm-SCM can represent systems such as the one described in the motivation, where some variables Z, directly influenced by variables affected by missingness $\mathcal{T}_Z$, rely on an alternative mechanism that excludes some contextual parent variables when their values are inaccessible.

Definition 2 **(lm-graph)** *A labeled missingness graph, or lm-graph, representing an lm-SCM $\mathcal{M}^\sigma$, is a tuple $(\mathcal{G}, \mathcal{L})$, where these components can be constructed as:*

(i) $\mathcal{G}$ starts with the same structure as the underlying m-graph $\mathcal{G}^*$, including bidirectional arrows, and, for all $X \in \mathcal{V}_{\text{sh}}, Z \in \mathcal{S}_X$, it includes the arrows $R_X \to Z$,

(ii) $\mathcal{L} = \bigcup_{R \in \mathcal{R}_{\text{sh}}} \mathcal{L}_R$, where $\mathcal{L}_{R_X} = \{(X, Z) : Z \in \mathcal{S}_X \wedge X \not\leftrightarrow Z \text{ in } \mathcal{G}\}$,

(iii) For all $(X, Z) \in \mathcal{L}$, the label "$R_X = 0$" can be graphically depicted on the arrow $X \to Z$.

The addition of new arrows to causal graphs carries the risk of creating cycles, which would prevent the system from being represented by a DAG or an ADMG and lead to problematic interpretations of circular causality. That issue is avoided here by requiring that, for any variable X, its children that are ancestors of R_X are never shifted by R_X. This restriction effectively prevents feedback loops. Figure 2 illustrates the construction of an lm-graph from the underlying m-graph and sets of shifted nodes.

An lm-graph combines the graphical expressiveness of both m-graphs and l-graphs, offering a unified framework to represent the problem at hand. Notably, when $\mathcal{V}_{\text{sh}} = \varnothing$, the lm-graph simplifies to an m-graph. Additionally, the following two remarks characterize them as l-graphs.

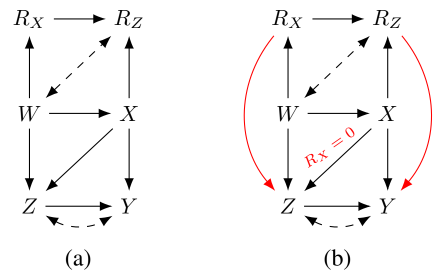


Figure 2: Construction of an lm-graph with $\mathcal{S}_X = Z$ and $\mathcal{S}_Z = Y$: (a) Underlying m-graph $\mathcal{G}^*$; (b) lm-graph $\mathcal{G}$ introducing the arrows $R_X \to Z$ and $R_Z \to Y$. Since there is no latent confounding between Z and X in $\mathcal{G}$, the edge $X \to Z$ is labeled with "$R_X = 0$". In contrast, no label is assigned to the edge $Z \to Y$ because latent confounding is present.

Proposition 1 *Let $(\mathcal{G}, \mathcal{L})$ be an lm-graph representing a semi-Markovian lm-SCM, with $(X, Z) \in \mathcal{L}$, and let $\mathcal{G}_{\setminus(X,Z)}$ be the graph with the edge $X \to Z$ removed from $\mathcal{G}$. If $Z \perp\!\!\!\perp_d X \mid W, R_X$ in $\mathcal{G}_{\setminus(X,Z)}$ then:*

$$Z \perp\!\!\!\perp X \mid W, R_X = 0.$$

Proof: Let $(X, Z) \in \mathcal{L}$, and suppose X and Z are d-separated in $\mathcal{G}_{\setminus(X,Z)}$ when conditioning on R_X and a disjoint set of variables W. Then, this d-separation statement constitutes a CSI-separation. By the soundness of CSI-separation (Koller and Friedman 2009; Corander et al. 2019), it follows that $Z \perp\!\!\!\perp X \mid W, R_X = 0$. Since the label set $\mathcal{L}$ encodes such CSI statements, lm-graphs are l-graphs, in which conditional independence under specific contexts can be assessed via d-separation after removing the labeled edges corresponding to those contexts. □

Proposition 2 *lm-graphs are regular-maximal l-graphs.*

Proof: If $(X, Z) \in \mathcal{L}$, then the edge $R_X \to Z$ is present in $\mathcal{G}$, ensuring regularity as $\text{pa}(Z;\mathcal{G}) \cap R_X \neq \varnothing$. Since each variable in $\mathcal{R}_{\text{sh}}$ is binary, it follows from Corollary 1 in Mokhtarian et al. (2022) that lm-graphs are maximal. That is because it is impossible to add (X, Z) to $\mathcal{L}$ more than once from a different context where $R_X = r \neq 0$. □

Consider the graph in figure 1b. It is an example of an lm-graph with $\mathcal{R}_{\text{sh}} = R_{Y_0}$, $\mathcal{S}_{Y_0} = \{A, Y_1\}$, and $\mathcal{L} = \{(Y_0, A), (Y_0, Y_1)\}$. Building on the motivating case from the Introduction, it represents an lm-SCM in which, when initial test scores Y_0 are missing ($R_{Y_0} = 0$), parents and clinicians of schoolchildren rely on a different rules for prescribing ADHD medication (A) compared to those who observe and leverage the unit's information on Y_0. The assignment mechanism for A is then modified by a latent soft intervention (Correa and Bareinboim 2020a) to be:

$$A \leftarrow R_{Y_0} \cdot f_A(W, Y_0, U_A) + (1 - R_{Y_0}) \cdot g_A(W, U_A),$$

where g_A is a shift function. Clearly, when $R_{Y_0} = 0$, the causal mechanisms for A no longer uses Y_0 as an input. Since the underlying graph contains no latent confounding $A \leftrightarrow Y_0$, this leads to the CSI $A \perp\!\!\!\perp Y_0 \mid W, R_{Y_0} = 0$.

Furthermore, schoolchildren who took the test at the initial time point may have gained more experience in test-taking, potentially affecting their performance even after accounting for all other relevant factors. Thus, a similar argument leads to the other CSI in figure 1b, $Y_1 \perp\!\!\!\perp Y_0 \mid W, A, R_{Y_0} = 0$.

We now introduce causal queries defined within lm-SCMs.

4 CAUSAL EFFECTS ON lm-GRAPHS

In this section, we examine the problem of defining, interpreting, and recovering causal effects from partially observed data generated by an lm-SCM $\mathcal{M}^\sigma$. We work under the assumption that the outcome variable Y has no causal descendants in the system.

Within the lm-SCM framework, and for a binary point exposure $A \in \{0, 1\}$ and a continuous outcome $Y \in \mathbb{R}$, different causal queries analogous to the ATE, as defined in equation (1), can be formulated. To clarify these distinctions, we introduce two specific versions: the FATE and the NATE.

Definition 3 (FATE) *The full average treatment effect (FATE) is the population-level expected difference in outcomes between treated and untreated scenarios, ceteris paribus, had no data been missing, under $\mathcal{M}^\sigma$:*

$$\phi := \Delta_a \mathbb{E}\left[Y \mid \mathrm{do}(A = a, \mathcal{R} = \mathbf{1})\right],$$

where $\mathcal{R} = \mathbf{1}$ is a short notation for $R = 1, \forall R \in \mathcal{R}$.

The FATE corresponds to a hypothetical ATE in a scenario with full information observability and no mechanism shifts. The following proposition outline insights for its recovery.

Proposition 3 *Let $\mathcal{R}_\Phi \subseteq \mathcal{R}_{\text{sh}}$ be the indicators connected to Y via directed (causal) paths that do not intersect $\mathcal{R}_{\text{sh}} \cup A$, as intermediate nodes, in the lm-graph $(\mathcal{G}, \mathcal{L})$, that is:*

$$\mathcal{R}_\Phi := \mathcal{R}_{\text{sh}} \cap \mathrm{an}(Y; \mathcal{G}[\overline{\mathcal{R}_{\text{sh}} \cup A}]).$$

Then, for binary exposure A, the FATE is equivalent to:

$$\phi = \Delta_a \mathbb{E}\left[Y \mid \mathrm{do}(A = a, \mathcal{R}_\Phi = \mathbf{1})\right], \tag{2}$$

and recoverable from $(\mathcal{G}, \mathcal{L})$ if $P(Y \mid \mathrm{do}(A, \mathcal{R}_\Phi))$ is recoverable from $\mathcal{G}$ treated as an unlabelled m-graph with substantive descendants of $\mathcal{R}$.

Proof: The intervention $\mathrm{do}(\mathcal{R}_{\text{sh}} = \mathbf{1})$ makes the labels in $\mathcal{L}$ idle, so the analysis can be conducted using only $\mathcal{G}$. By Rule 3 of *do*-calculus one has that $P(Y \mid \mathrm{do}(A, \mathcal{R})) = P(Y \mid \mathrm{do}(A, \mathcal{R}_{\text{sh}}))$ whenever $Y \perp\!\!\!\perp_d \mathcal{R}_{\text{no}} \mid A, \mathcal{R}_{\text{sh}}$ in $\mathcal{G}[\overline{A, \mathcal{R}}]$ (Pearl 2009). Nodes in $\mathcal{R}_{\text{no}}$ have no parents in $\mathcal{G}[\overline{A, \mathcal{R}}]$ and, since their only possible children belong to $\mathcal{R}$, they have no descendants either. As a result, they are isolated from Y in such graph. Plus, all directed causal paths from nodes $\mathcal{R}_{\text{sh}} \setminus \mathcal{R}_\Phi$ to Y are intersected by $\mathcal{R}_\Phi$ or A, so by Rule 3 $P(Y \mid \mathrm{do}(A, \mathcal{R}_{\text{sh}})) = P(Y \mid \mathrm{do}(A, \mathcal{R}_\Phi))$. $\square$

The set $\mathcal{R}_\Phi$ consists precisely of those missingness indicators that have a causal path to Y under intervention on both A and $\mathcal{R}$; these indicators trigger mechanism shifts and can transmit both their own effects and the effect of the exposure to the outcome Y. This implies that recovering the FATE from observed data, $(\mathcal{V}_o, \mathcal{V}_m^\dagger, \mathcal{R}) \sim P^\sigma$, hinges on the recoverability status of $P(Y \mid \mathrm{do}(A, \mathcal{R}_\Phi))$ from $\mathcal{G}$ as an m-graph (with substantive descendants of $\mathcal{R}$).

Various available methods can be employed to assess this task. For instance, in the scenario depicted in figure 1b, the FATE can be recovered through sequential factorizations Mohan and Pearl (2021), leading to the expressions:

$$\phi = \mathbb{E}_W \mathbb{E}_{Y_0 \mid W, R_{Y_0}=1} \Delta_a Q_1(W, Y_0, a), \text{ with} \tag{3}$$

$$Q_1(W, Y_0, A) = \mathbb{E}\left[Y \mid W, Y_0, A, R_{Y_0} = 1\right]. \tag{4}$$

This result follows from verifying that $\mathcal{R}_\Phi = R_{Y_0}$, and that $P(Y_1 \mid \mathrm{do}(A, R_{Y_0}))$ can be expressed as:

$$\int \mathrm{d}P(W)\, \mathrm{d}P(Y_0 \mid W, R_{Y_0})\, P(Y_1 \mid W, Y_0, A, R_{Y_0}),$$

as the following sequence of graphical conditions hold:

(i) $\{W, Y_0\} \cap \mathrm{de}(A \cup R_{Y_0}; \mathcal{G}) = \varnothing$,
(ii) $Y_1 \perp\!\!\!\perp_d A, R_{Y_0} \mid W, Y_0$ in $\mathcal{G}[\underline{A, R_{Y_0}}]$, and
(iii) $Y_0 \perp\!\!\!\perp_d R_{Y_0} \mid W$ in $\mathcal{G}$.

Although the statistical estimand for the FATE ϕ in equations (3) and (4) matches the expression obtained by recovering the ATE from the underlying m-graph, these do not always coincide. In some cases, the ATE may be recoverable from the underlying m-graph, while the FATE is not from the lm-graph, as illustrated in figure 3a.

For the FATE, a hypothetical intervention makes all variables observed by each unit, thereby avoiding any mechanism shifts that arise from incomplete information. In contrast, a different causal query is the *natural* ATE (NATE[1]). When the system is faithfully described by an lm-graph and units adapt their behavior based on which inputs are missing, the NATE captures the causal effect while accounting for such missingness-driven shifts in the causal mechanism.

[1]The term "NATE" has also been used in recent literature to refer to the *nudge* ATE: the average causal effect among the subgroup of units whose treatment can be manipulated by an instrumental variable (Tchetgen-Tchetgen 2024).

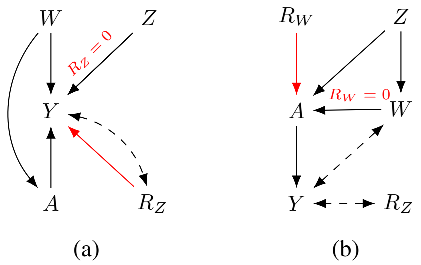


Figure 3: Contrasting recovery results from lm-graphs versus underlying m-graphs (where the red edges and labels are absent): (a) In the m-graph, $P(Y \mid \mathrm{do}(A))$ is recoverable because Z is not needed for adjustment, allowing recovery via identification $\int \mathrm{d}P(W)P(Y \mid W, A)$. In contrast, $P(Y \mid \mathrm{do}(A, R_Z))$ is not recoverable from the lm-graph due to latent confounding $Y \leftrightarrow R_Z$; (b) Here, $P(Y \mid \mathrm{do}(A))$ is not recoverable from the m-graph because Z cannot be used to block the backdoor path $A \leftarrow Z \rightarrow W \leftrightarrow Y$ as R_Z is not d-separable from Y. However, in the lm-graph, it is recoverable by applying Rule 1 of *do*-calculus: $P(Y \mid \mathrm{do}(A)) = P(Y \mid \mathrm{do}(A), R_W)$, and given no confounding when $R_W = 0$, this simplifies to $P(Y \mid A, R_W = 0)$.

Definition 4 (NATE) *The natural average treatment effect (NATE) is the population-level expected difference in outcomes between treated and untreated scenarios, ceteris paribus, under $\mathcal{M}^\sigma$:*

$$\theta := \Delta_a \mathbb{E}\left[Y \mid \mathrm{do}(A = a)\right].$$

Clearly, the NATE is recoverable if $P(Y \mid \mathrm{do}(A))$ is recoverable from the lm-graph. Since lm-graphs are regular-maximal, if all variables in $\mathcal{R}_{\text{sh}}$ were exogenous, recovery could in principle be evaluated using a procedure akin to the one proposed by Mokhtarian et al. (2022). However, because the recovery problem is more complex than the identification problem, such procedure would provide only a sufficient, but not necessary, condition for recovery under missingness-specific independencies. For example, figure 3b illustrates a case in which R_W is exogenous and $P(Y \mid \mathrm{do}(A))$ is recoverable, yet recovery cannot be achieved through a context-by-context sequence. Instead, it requires applying the rules of *do*-calculus prior to selecting a representative context. Notably, in this example, $P(Y \mid \mathrm{do}(A))$ is not recoverable from the underlying m-graph alone, but becomes recoverable from the lm-graph.

In more general settings, recovering the NATE from an lm-graph remains a challenging task. A comprehensive solution would likely require integrating the CSI-calculus introduced by Tikka et al. (2019) with the algorithm for causal effect recovery under missing data proposed by Bhattacharya et al. (2020). However, such combination is not guaranteed to be complete, implying that there may exist recoverable cases that fall outside their scope. Moreover, Tikka et al. (2019) demonstrate that the general identification problem becomes NP-hard when CSI constraints are present.

Recovery of the NATE depends on the recovery of the distribution $P(Y \mid \mathrm{do}(A))$ from the lm-graph $(\mathcal{G}, \mathcal{L})$, where CSIs are non-idle. We now present a sufficient condition for the recovery of this interventional distribution.

Proposition 4 *Let $K, H \subseteq \mathcal{V}_{\text{sh}} \setminus (A \cup Y)$ be disjoint sets of variables whose missingness induce shifts, with $K = \{K_j\}_{j=1}^{\kappa}$ indexed. Let R_K and R_H denote their missingness indicators. For every missingness pattern $r \in \operatorname{supp} R_K \subseteq \{0,1\}^{\kappa}$, let $L_r \subseteq \mathcal{V} \setminus (A \cup Y \cup H)$ be a set such that:*

(i) $L_r \cap \bigcup_{j:r_j=0}^{\kappa} K_j = \varnothing$ and $\bigcup_{j:r_j=1}^{\kappa} K_j \subseteq L_r$, (i.e., L_r excludes variables K_j with $r_j = 0$, and includes all K_j with $r_j = 1$).

Define $M_r := \mathcal{V}_m \cap (A \cup Y \cup L_r \setminus K)$. Let $\mathcal{G}_r$ be the graph obtained by removing from $\mathcal{G}$ all arrows $\bigcup_{j:r_j=0} \mathcal{L}_{R_{K_j}}$ in the label set $\mathcal{L}$, and $\mathcal{H}_r$ be the graph that removes the edges $\mathcal{L}_{R_H}$ from $\mathcal{G}_r$. Let the following conditions hold for every missingness pattern $r \in \operatorname{supp} R_K$:

(ii) $(R_K \cup L_r \cup R_{M_r} \cup R_H) \cap \mathrm{de}(A; \mathcal{G}_r) = \varnothing$,
(iii) $Y \perp\!\!\!\perp_d R_{M_r} \mid A, R_K$ in $\mathcal{G}_r[\overline{A}]$,
(iv) $Y \perp\!\!\!\perp_d R_H \mid A, L_r, R_K, R_{M_r}$ in $\mathcal{G}_r[\overline{A}]$,
(v) $Y \perp\!\!\!\perp_d A \mid L_r, R_K, R_{M_r}, R_H$ in $\mathcal{H}_r[\underline{A}]$.

If K, H and $\{L_r\}_r$ exist and fulfill conditions above then:

$$P(Y \mid \mathrm{do}(A = a)) = \sum_{r \in \operatorname{supp} R_K} \mathbb{P}(R_k = r)\, \Theta_r(a), \tag{5}$$

$$\Theta_r(a) = \int \mathrm{d}P(L_r \mid R_K = r, R_{M_r} = 1) \quad P(Y \mid A = a, L_r, R_k = r, R_{M_r} = 1, R_H = 0) \tag{6}$$

and thus the NATE, θ, is given by:

$$\theta = \sum_{r \in \operatorname{supp} R_K} \mathbb{P}(R_k = r)\, \Delta_a \vartheta_r(a), \text{ where} \tag{7}$$

$$\vartheta_r(a) = \int \mathrm{d}P(L_r \mid R_K = r, R_{M_r} = 1), \quad \mathbb{E}[Y \mid A = a, L_r, R_k = r, R_{M_r} = 1, R_H = 0].$$

Proof: For any set $K \subseteq \mathcal{V}_{\text{sh}} \setminus (A \cup Y)$, one can express $P(Y \mid \mathrm{do}(A)) = \sum_{R_K} \mathbb{P}(R_K \mid \mathrm{do}(A))\, P(Y \mid \mathrm{do}(A), R_K)$. Given a pattern $R_K = r$, terms conditioned in this context can be evaluated within the corresponding graph $\mathcal{G}_r$, obtained by removing the edges associated with the labels "$R_{K_j} = r_j = 0$". By assumption *(ii)*, the first factor simplifies as $\mathbb{P}(R_K)$. By *(iii)* and Rule 1 of *do*-calculus, the second factor is $P(Y \mid \mathrm{do}(A), R_K) = \int \mathrm{d}P(L_r \mid \mathrm{do}(A), R_K, R_{M_r} = 1) P(Y \mid \mathrm{do}(A), L_r, R_K, R_{M_r} = 1)$. Given assumption *(ii)*, the first term on the right-hand side drops the intervention $\mathrm{do}(A)$. By assumption *(iv)* and Rule 1, the second term on the right-hand side becomes

$P(Y \mid \mathrm{do}(A), L_r, R_K, R_{M_r} = 1, R_H = 0)$. Finally, given condition on context $R_K = r, R_H = 0$, analysis of such last term can be done in the graph that removes from $\mathcal{G}_r$ edges labeled with "$R_H = 0$". Assumptions *(ii)* and *(v)*, and Rule 2 of *do*-calculus, applied to the last term lead directly to equations (5) and (6). □

Equation (7) expresses the NATE as a weighted average of context CATEs, denoted $\Delta_a \vartheta_r(a)$, where the weights correspond to the marginal probability of a missingness pattern on R_K occurring.

In the scenario depicted in figure 1b, one can set $K = Y_0$, $H = \varnothing$, $L_0 = W$ and $L_1 = \{W, Y_0\}$. The graph $\mathcal{G}_0 = \mathcal{H}_0$ removes the arrows $\{(Y_0, A), (Y_0, Y_1)\}$. Therefore, in the context $R_{Y_0} = 0$, $P(Y \mid \mathrm{do}(A), R_{Y_0} = 0)$ is recovered from $\mathcal{G}_0$ as:

$$\int \mathrm{d}P(W \mid R_{Y_0} = 0)\, P(Y \mid W, A, R_{Y_0} = 0).$$

In addition, the graph $\mathcal{G}_1 = \mathcal{H}_1$ retains all the arrows. Thus, in the context $R_{Y_0} = 1$, $P(Y \mid \mathrm{do}(A), R_{Y_0} = 1)$ is recovered from $\mathcal{G}_1$ as:

$$\int \mathrm{d}P(W, Y_0 \mid R_{Y_0} = 1)\, P(Y \mid W, Y_0, A, R_{Y_0} = 1).$$

Consequently, the NATE is given by:

$$\begin{aligned} \theta = \mathbb{P}(R_{Y_0} = 0)\, \mathbb{E}_{W \mid R_{Y_0} = 0} \Delta_a Q_0(W, a) \qquad (8) \\ + \mathbb{P}(R_{Y_0} = 1)\, \mathbb{E}_{W, Y_0 \mid R_{Y_0} = 1} \Delta_a Q_1(W, Y_0, a), \\ Q_0(W, A) = \mathbb{E}\left[Y \mid W, A = a, R_{Y_0} = 0\right], \qquad (9) \end{aligned}$$

with $Q_1(W, Y_0, A)$ given in equation (4).

In the scenario in figure 3b, the NATE is recovered under proposition 4 by setting $K = \varnothing$, which makes all L-sets empty by convention; and $H = W$, which highlights the relevant role of such set.

The recovered estimands for the FATE, in equation (3), and the NATE, in equation (8), highlight two key differences between these queries. The FATE relies on a single common response function, Q_1, taking Y_0 as input. This function can be learned from units with observed Y_0, and then imputed for the remaining units. In contrast, the NATE requires two separate response functions, adding Q_0, which does not use Y_0 as input. Each of these estimands can be learned from their respective subpopulation, weighted by their size, and averaged, without any imputation.

The NATE and FATE estimands may be motivated by distinct policy designs. For instance, if a proposed policy requires complete observability of certain variables for all units —such as mandatory pre-testing— then the FATE becomes a more relevant target for pre-evaluating its effectiveness, as it reflects potential outcomes under a scenario of full information. In contrast, the NATE is more appropriate when such observability is not enforced, but missingness remains prevalent. In these cases, the NATE captures the causal effect accounting for mechanism shifts due to missing data and better reflects the anticipated performance of the policy under conditions where not all inputs are observed.

5 ESTIMATION

Building upon the foundational work on doubly-robust estimation of causal effects in settings with missing data (Bang and Robins 2005; Sun et al. 2017), a surge of recent research has lead to the development of misspecification-robust estimation methods, particularly under missing outcome data (Wei et al. 2022, 2023; Negi 2024; Tang et al. 2024; Bia et al. 2024; de Aguas et al. 2025). Mechanism shifts can have significant implications for the efficient estimation of causal queries, especially for robust estimation in a semiparametric model. Doubly-robust estimators for causal parameters remain consistent provided that at least one of their components —either the outcome regression or the treatment assignment model— is correctly specified. However, under CSI either of these components may be affected by shifts.

When missingness-specific independencies are relevant, consistent and robust estimation of the NATE requires tailored parametrization of the nuisance components. Although one could estimate each context CATE separately and compute a weighted average based on context frequencies, this approach may suffer for contexts with sparse observations. In such cases, pooling information from and across more populous contexts can enhance estimation quality.

Building upon the example introduced in the motivation and illustrated in figure 1b, we now present a doubly-robust estimator for the NATE. We parameterize the outcome models Q_1, in equation (4), and Q_0, in equation (9), jointly as:

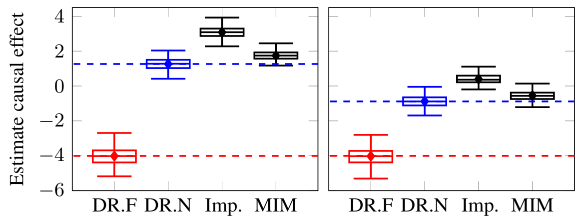
$$Q_r(W, Y_0, A) := Q_0(W, A) + r \cdot q(W, Y_0, A),$$

with q being an auxiliary function. Since this lm-graph implies that the exposure mechanism is also shifted by the missingness of Y_0, this dependence should be reflected in the parametrization of the propensity score, specified as:

$$\log \frac{\pi_r(W, Y_0)}{1 - \pi_r(W, Y_0)} = \log \frac{\pi_0(W)}{1 - \pi_0(W)} + r \cdot \rho(W, Y_0),$$

where $\pi_1(W, Y_0) := \mathbb{P}(A = 1 \mid W, Y_0)$ and $\pi_0(W) := \mathbb{P}(A = 1 \mid W)$ under common positivity constraints, i.e., $0 < \pi_0(W), \pi_1(W, Y_0) < 1$ (almost surely), and employing an auxiliary function ρ.

Then, a *one-step corrected* estimator of the NATE, using



(a) (b)

Figure 4: Boxplots of sampling distribution of estimators in simulation exercise with sample size $n = 5\,000$ and $m = 200$ repetitions, under missingness rates of (a) 50% and (b) 30%. Broken lines indicate the oracle FATE (in red) and NATE (in blue). Abbr.: DR.F = doubly-robust FATE estimator (appendix A), DR.N = doubly-robust NATE estimator (section 5), Imp = Multiple Imputation Method, MIM = Missing Indicator Method.

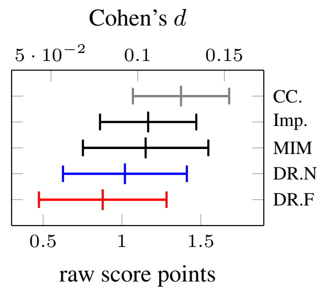


Figure 5: Estimated effects and confidence intervals for different estimators in the application case. Abbr.: CC = Complete cases.

data $\{(W_i, Y^{\dagger}_{0,i}, R_{Y_0,i} = r_i, A_i, Y_{1,i})\}_i^n$, is given by

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \Delta_a \widehat{Q}_{r_i}(W_i, Y_{0,i}, a) + \frac{A_i - \widehat{\pi}_{r_i}(W_i, Y_{0,i})}{\widehat{\pi}_{r_i}(W_i, Y_{0,i})[1 - \widehat{\pi}_{r_i}(W_i, Y_{0,i})]}\left[Y_{1,i} - \widehat{Q}_{r_i}(W_i, Y_{0,i}, A_i)\right] \tag{10}$$

Intuitively, this corresponds to the conventional AIPW estimator (Robins et al. 1995; Robins and Rotnitzky 1997), where the outcome and propensity score models, Q_r and π_r, are parameterized to account for different adjustment sets based on the context $R_{Y_0} = r$. Unfortunately, in more general settings where $|\mathcal{R}_{\text{sh}}|$ is large, a combinatorial explosion of nuisance parameters arises, making it challenging to determine optimal pooling parametrizations or constraints. This, in turn, complicates the computation by hand of the required influence functions and efficient estimators.

6 SIMULATIONS AND APPLICATION

We conduct a simulation study to compare estimators for the FATE and NATE for the case of lm-graph in figure 1b. Data are generated from an lm-SCM with varying parameterizations for two scenarios on the missingness rate of Y_0: (a) 50% and (b) 30%. The FATE is estimated using a doubly-robust method (DR.F) that imputes pseudo-outcomes via one-step corrections from equation (3). The NATE is estimated using the doubly-robust estimator (DR.N) defined in equation (10). We also include a Multiple Imputation estimator (Imp.) and a Missing Indicator Method (MIM) estimator, which uses $(W, A, R_{Y_0} \cdot Y_0)$ as covariates and includes interactions of W, A with R_{Y_0} (Greenland and Finkle 1995; Jones 1996). Implementation details are in appendix A.

Results in figure 4 display the sampling distribution of the estimators with sample size $n = 5\,000$ and $m = 200$ repetitions. They highlight key differences between the FATE and NATE, which may occasionally have opposite signs, though the difference depends on the missingness rate. The DR.N estimator performs well across the two setups. Under data generated from an lm-SCM, both Imp. and MIM estimators target the NATE. The Imp. estimator is biased, as its imputation model fails to fully capture mechanism shifts, while MIM shows less bias, supporting its empirical value (Song et al. 2021; Chang et al. 2023), when justified graphically and implemented with flexible models.

As a motivating application, we study the impact of pharmacological treatment for *attention-deficit/hyperactivity disorder* (ADHD) upon national numeracy test scores among Norwegian 8th-grade students diagnosed with ADHD, using observational data. A key confounder, the grade 5 test score, is missing in 10% of the 8 450 cases due to both exogenous and endogenous factors. Further details are given in appendix B. The results in figure 5 reveal a small effect of medication on scores. One possible explanation for the small positive difference between the NATE and FATE estimates is a compensatory or diminishing returns effect: prior testing-experience also improves performance, and medication may provide greater benefits to children without such experience. Thus, when test-taking in grade 5 is made mandatory, the added impact of medication is slightly reduced. Latent factors could also explain such difference.

7 CONCLUSION AND DISCUSSION

We examined the problem of mechanism shifts that cannot be captured by existing m-graph-based frameworks for causal inference with missing data. To address these limitations, we introduced lm-SCMs and lm-graphs, which retain the expressiveness of m-graphs while incorporating a set of CSIs. We defined relevant causal effects and established recovery criteria for them. Given that causal effect identification in systems with CSI is generally NP-hard, our recovery criteria for the *natural* effects are only declarative and sufficient. Future research will further investigate other causal queries within lm-SCMs with categorial or continuous exposures, broad recovery criteria, and the issue of testability implications and impossibilities under lm-graphs.

Acknowledgements

We thank the organizing committee of the European Causal Inference Meeting (2025) in Ghent for the opportunity to present an early version of this work and for the valuable feedback received. This research was supported by the Research Council of Norway (project number 301081, PI: Guido Biele) and approved by the Regional Committees for Medical and Health Research Ethics (REK) under approval number 96604. Johan Pensar received additional support from the Research Council of Norway through INTEGREAT – The Norwegian Centre for Knowledge-Driven Machine Learning (project number 332645).

REFERENCES

Bang, H. and Robins, J. (Dec. 2005). "Doubly Robust Estimation in Missing Data and Causal Inference Models." In: *Biometrics* 61.4, pp. 962–973. URL: https://doi.org/10.1111/j.1541-0420.2005.00377.x.

Bareinboim, E., Correa, J., Ibeling, D., and Icard, T. (2022). "On Pearl's Hierarchy and the foundations of causal inference." In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. New York, NY, USA: Association for Computing Machinery, pp. 507–556. URL: https://doi.org/10.1145/3501714.3501743.

Bareinboim, E., Tian, J., and Pearl, J. (June 2014). "Recovering from Selection Bias in Causal and Statistical Inference." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 28.1. URL: https://ojs.aaai.org/index.php/AAAI/article/view/9074.

Bhattacharya, R., Nabi, R., Shpitser, I., and Robins, J. (July 2020). "Identification In Missing Data Models Represented By Directed Acyclic Graphs." In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. Ed. by R. P. Adams and V. Gogate. Vol. 115. Proceedings of Machine Learning Research. Pmlr, pp. 1149–1158. URL: https://proceedings.mlr.press/v115/bhattacharya20b.html.

Bia, M., Huber, M., and Lafférs, L. (2024). "Double Machine Learning for Sample Selection Models." In: *Journal of Business and Economic Statistics* 42.3, pp. 958–969. URL: https://doi.org/10.1080/07350015.2023.2271071.

Biele, G. et al. (2019). "Bias from self selection and loss to follow-up in prospective cohort studies." In: *European Journal of Epidemiology* 34.10. Pmid: 31451995, pp. 927–938.

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). "Context-specific independence in Bayesian networks." In: *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*. Uai'96. Portland, OR: Morgan Kaufmann Publishers Inc., pp. 115–123.

Chang, C.-R., Song, Y., Li, F., and Wang, R. (2023). "Covariate adjustment in randomized clinical trials with missing covariate and outcome data." In: *Statistics in Medicine* 42.22, pp. 3919–3935. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9840.

Colnet, B., Josse, J., Varoquaux, G., and Scornet, E. (2022). "Causal effect on a target population: A sensitivity analysis to handle missing covariates." In: *Journal of Causal Inference* 10.1, pp. 372–414. URL: https://doi.org/10.1515/jci-2021-0059.

Corander, J., Hyttinen, A., Kontinen, J., Pensar, J., and Väänänen, J. (2019). "A logical approach to context-specific independence." In: *Annals of Pure and Applied Logic* 170.9. The 23rd Workshop on Logic, Language, Information and Computation, pp. 975–992. URL: https://www.sciencedirect.com/science/article/pii/S0168007219300326.

Correa, J. and Bareinboim, E. (July 2019). "From Statistical Transportability to Estimating the Effect of Stochastic Interventions." In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, pp. 1661–1667. URL: https://doi.org/10.24963/ijcai.2019/230.

Correa, J. and Bareinboim, E. (Apr. 2020a). "A Calculus for Stochastic Interventions: Causal Effect Identification and Surrogate Experiments." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.06, pp. 10093–10100. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6567.

Correa, J. and Bareinboim, E. (2020b). "General Transportability of Soft Interventions: Completeness Results." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 10902–10912. URL: https://papers.nips.cc/paper/2020/hash/7b497aa1b2a83ec63d1777a88676b0c2-Abstract.html.

Correa, J., Tian, J., and Bareinboim, E. (Apr. 2018). "Generalized Adjustment Under Confounding and Selection Biases." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1. URL: https://ojs.aaai.org/index.php/AAAI/article/view/12125.

de Aguas, J., Pensar, J., Varnet Pérez, T., and Biele, G. (2025). "Recovery and inference of causal effects with sequential adjustment for confounding and attrition." In: *Journal of Causal Inference* 13.1, p. 20240009. URL: https://doi.org/10.1515/jci-2024-0009.

Ding, P. and Li, F. (2018). "Causal Inference: A Missing Data Perspective." In: *Statistical Science* 33.2, pp. 214–237. URL: https://doi.org/10.1214/18-STS645.

Evans, R. J. and Richardson, T. S. (2014). "Markovian acyclic directed mixed graphs for discrete data." In: *The Annals of Statistics* 42.4, pp. 1452–1482.

Greenland, S. and Finkle, W. (1995). "A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses." In: *American Journal of Epidemiology* 142.12, pp. 1255–1264.

Guo, A., Zhao, J., and Nabi, R. (2023). "Sufficient identification conditions and semiparametric estimation under missing not at random mechanisms." In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Uai '23. Pittsburgh, PA, USA: JMLR.org. URL: https://proceedings.mlr.press/v216/guo23a/guo23a.pdf.

Hernán, M., Hernández-Díaz, S., and Robins, J. (2004). "A structural approach to selection bias." In: *Epidemiology (Cambridge, Mass.)* 15.5, pp. 615–625.

Hines, O., Dukes, O., Díaz-Ordaz, K., and Vansteelandt, S. (2022). "Demystifying statistical learning based on efficient influence functions." In: *The American Statistician* 76.3, pp. 292–304.

Horvitz, D. and Thompson, D. (1952). "A Generalization of Sampling Without Replacement from a Finite Population." In: *Journal of the American Statistical Association* 47.260, pp. 663–685.

Huber, M. (2012). "Identification of Average Treatment Effects in Social Experiments Under Alternative Forms of Attrition." In: *Journal of Educational and Behavioral Statistics* 37.3, pp. 443–474. URL: https://doi.org/10.3102/1076998611411917.

Jones, M. (1996). "Indicator and stratification methods for missing explanatory variables in multiple linear regression." In: *Journal of the American Statistical Association* 91.433, pp. 222–230.

Kennedy, E. (2023). "Towards optimal doubly robust estimation of heterogeneous causal effects." In: *Electronic Journal of Statistics* 17.2, pp. 3008–3049. URL: https://doi.org/10.1214/23-EJS2157.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Kuzmanovic, M., Hatt, T., and Feuerriegel, S. (2023). "Estimating conditional average treatment effects with missing treatment information." In: *International Conference on Artificial Intelligence and Statistics*. Pmlr, pp. 746–766. URL: https://proceedings.mlr.press/v206/kuzmanovic23a/kuzmanovic23a.pdf.

Kyono, T., Zhang, Y., Bellot, A., and der Schaar, M. van (2021). "MIRACLE: Causally-Aware Imputation via Learning Missing Data Mechanisms." In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34, pp. 23806–23817. URL: https://openreview.net/forum?id=GzeqcAUFGl0.

Levis, A., Mukherjee, R., Wang, R., and Haneuse, S. (2024). "Robust causal inference for point exposures with missing confounders." In: *Canadian Journal of Statistics* n/a.n/a, e11832. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.11832.

Lewin, A., Brondeel, R., Benmarhnia, T., Thomas, F., and Chaix, B. (2018). "Attrition Bias Related to Missing Outcome Data: A Longitudinal Simulation Study." In: *Epidemiology* 29.1.

Mathur, M., Shpitser, I., and VanderWeele, T. (Jan. 2023). *A common-cause principle for eliminating selection bias in causal estimands through covariate adjustment*. OSF Preprints ths4e. Center for Open Science. URL: https://ideas.repec.org/p/osf/osfxxx/ths4e.html.

Mayer, I., Josse, J., and Group, T. (2023). "Generalizing treatment effects with incomplete covariates: Identifying assumptions and multiple imputation algorithms." In: *Biometrical Journal* 65.5, pp. 210–294. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202100294.

Mohan, K. and Pearl, J. (2014). "Graphical Models for Recovering Probabilistic and Causal Queries from Missing Data." In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger. Vol. 27. Curran Associates, Inc. URL: https://papers.nips.cc/paper_files/paper/2014/hash/1835d9d1508eb178b500220a9ddf75a7-Abstract.html.

Mohan, K. and Pearl, J. (2021). "Graphical Models for Processing Missing Data." In: *Journal of the American Statistical Association* 116.534, pp. 1023–1037. URL: https://doi.org/10.1080/01621459.2021.1874961.

Mohan, K., Pearl, J., and Tian, J. (2013). "Graphical Models for Inference with Missing Data." In: *Advances in Neural Information Processing Systems*. Ed. by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/0ff8033cf9437c213ee13937b1c4c455-Paper.pdf.

Mokhtarian, E., Jamshidi, F., Etesami, J., and Kiyavash, N. (Mar. 2022). "Causal Effect Identification with Context-specific Independence Relations of Control Variables." In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by G. Camps-Valls, F. J. R. Ruiz, and I. Valera. Vol. 151. Proceedings of Machine Learning Research. Pmlr, pp. 11237–11246. URL: https://proceedings.mlr.press/v151/mokhtarian22a.html.

Nabi, R. and Bhattacharya, R. (July 2023). "On Testability and Goodness of Fit Tests in Missing Data Models." In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by R. J. Evans and I. Shpitser. Vol. 216. Proceedings of Machine Learning Research. Pmlr, pp. 1467–1477. URL: https://proceedings.mlr.press/v216/nabi23a.html.

Nabi, R., Bhattacharya, R., and Shpitser, I. (2020). "Full law identification in graphical models of missing data: completeness results." In: *Proceedings of the 37th International Conference on Machine Learning*. Icml'20. JMLR.org. URL: http://proceedings.mlr.press/v119/nabi20a/nabi20a.pdf.

Nabi, R., Bhattacharya, R., Shpitser, I., and Robins, J. (2024). *Causal and Counterfactual Views of Missing Data Models*. URL: https://arxiv.org/abs/2210.05558.

Negi, A. (2024). "Doubly weighted M-estimation for nonrandom assignment and missing outcomes." In: *Journal of Causal Inference* 12.1, p. 20230016. URL: https://doi.org/10.1515/jci-2023-0016.

Neyman, J. (1923). "Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes." Trans. by D. M. Dabrowska and T. P. Speed. In: *Statistical Science* 5. Master's Thesis. Excerpts reprinted in English in *Statistical*

Science, Vol. 5, pp. 463–472. Translated by D. M. Dabrowska and T. P. Speed, pp. 463–472.

Pearl, J. (1995). "Causal Diagrams for Empirical Research." In: *Biometrika* 82.4, pp. 669–688. URL: http://www.jstor.org/stable/2337329 (visited on 06/02/2023).

Pearl, J. (2009). *Causality. Models, Reasoning, and Inference*. 2nd ed. Cambridge, UK: Cambridge University Press.

Pearl, J. (2012). "The Do-Calculus Revisited." In: *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*. Uai'12. Catalina Island, CA: AUAI Press, pp. 3–11.

Pensar, J., Nyman, H., Koski, T., and Corander, J. (2015). "Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models." In: *Data Mining and Knowledge Discovery* 29.2, pp. 503–533. URL: https://doi.org/10.1007/s10618-014-0355-0.

Robins, J. (1989). "The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies." In: *Health service research methodology: a focus on AIDS*, pp. 113–159.

Robins, J. and Rotnitzky, A. (1997). "Comment on the Bickel and Kwon article, 'Inference for semiparametric models: Some questions and an answer'." In: *Statistica Sinica* 7.1, pp. 167–199.

Robins, J., Rotnitzky, A., and Zhao, L. (1995). "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data." In: *Journal of the American Statistical Association* 90.429, pp. 106–121.

Rubin, D. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." In: *Journal of Educational Psychology* 66.5, pp. 688–701.

Rubin, D. (1976). "Inference and missing data." In: *Biometrika* 63.3, pp. 581–592.

Rubin, D. (1978). "Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse." In: *Journal of the American Statistical Association* 73.362, pp. 384–395.

Saadati, M. and Tian, J. (2019). "Adjustment criteria for recovering causal effects from missing data." In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 561–577.

Shi, Y., Zhu, Y., and Dubin, J. (2024). *Causal Inference on Missing Exposure via Robust Estimation*. URL: https://arxiv.org/abs/2406.08668.

Song, M., Zhou, X., Pazaris, M., and Spiegelman, D. (2021). *The Missing Covariate Indicator Method is Nearly Valid Almost Always*. URL: https://arxiv.org/abs/2111.00138.

Srinivasan, R., Bhattacharya, R., Nabi, R., Ogburn, E., and Shpitser, I. (2023). *Graphical Models of Entangled Missingness*. URL: https://arxiv.org/abs/2304.01953.

Sun, B. et al. (Nov. 2017). "Inverse-Probability-Weighted Estimation for Monotone and Nonmonotone Missing Data." In: *American Journal of Epidemiology* 187.3, pp. 585–591. URL: https://doi.org/10.1093/aje/kwx350.

Tang, S., Zhan, M., Jiang, Q., and Zhang, T. (2024). "Efficient covariate balancing for the average treatment effect with missing outcome." In: *Economics Letters*, p. 111961. URL: https://www.sciencedirect.com/science/article/pii/S0165176524004452.

Tchetgen-Tchetgen, E. (2024). *The Nudge Average Treatment Effect*. URL: https://doi.org/10.48550/arXiv.2410.23590.

Tikka, S., Hyttinen, A., and Karvanen, J. (2019). "Identifying Causal Effects via Context-specific Independence Relations." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. URL: https://papers.nips.cc/paper_files/paper/2019/hash/d88518acbcc3d08d1f18da62f9bb26ec-Abstract.html.

Tikka, S., Hyttinen, A., and Karvanen, J. (2021). "Causal Effect Identification from Multiple Incomplete Data Sources: A General Search-Based Approach." In: *Journal of Statistical Software* 99, pp. 1–40.

Wei, K., Qin, G., Zhang, J., and Sui, X. (2022). "Doubly robust estimation in causal inference with missing outcomes: With an application to the Aerobics Center Longitudinal Study." In: *Computational Statistics and Data Analysis* 168, p. 107399. URL: https://www.sciencedirect.com/science/article/pii/S0167947321002334.

Wei, K., Qin, G., Zhang, J., and Sui, X. (2023). "Multiply robust estimation of the average treatment effect with missing outcomes." In: *Journal of Statistical Computation and Simulation* 93.10, pp. 1479–1495. URL: https://doi.org/10.1080/00949655.2022.2143501.

Yang, S., Wang, L., and Ding, P. (Sept. 2019). "Causal inference with confounders missing not at random." In: *Biometrika* 106.4, pp. 875–888. URL: https://doi.org/10.1093/biomet/asz048.

Zhou, H., Balakrishnan, S., and Lipton, Z. (2023). "Domain adaptation under missingness shift." In: *International Conference on Artificial Intelligence and Statistics*. Pmlr, pp. 9577–9606. URL: https://proceedings.mlr.press/v206/zhou23b/zhou23b.pdf.

Causal Inference amid Missingness-Specific independencies and Mechanism Shifts (Supplementary Material)

Johan de Aguas[1,2] **Leonard Henckel**[3] **Johan Pensar**[1,4] **Guido Biele**[2]

[1]Department of Mathematics, University of Oslo
[2]Department of Child Health and Development, Norwegian Institute of Public Health
[3]School of Mathematics and Statistics. University College Dublin
[4]INTEGREAT – The Norwegian Center for Knowledge-Driven Machine Learning

A SIMULATION TASK DETAILS

For each sampling batch, we generate 5 000 i.i.d. samples from the following SCM, with associated lm-graph in figure 1b:

$$
\begin{aligned}
W &\sim N(0,1), && (11)\\
Y_0 &= -3 - 2\,W + W^2 + U_0, \qquad U_0 \sim N(0,7), && (12)\\
R_{Y_0} &\sim \mathrm{Ber}_{\mathrm{logit}}(\beta_1 + \beta_2\,W), && (13)\\
A &\sim \mathrm{Ber}_{\mathrm{logit}}\left(R_{Y_0}(W + 0.3\,Y_0) + (1 - R_{Y_0})(-0.5 + 1.5\,W)\right), && (14)\\
Y &= R_{Y_0}(3 + 1.8\,W - 2\,A - 1.5\,Y_0 - 0.8\,A\,W + 4\,A\,Y_0) && (15)\\
&\quad + (1 - R_{Y_0})(4 + 6\,W + 8\,A - 8\,W\,A) + U_1, \qquad U_1 \sim N(0,7). && (16)
\end{aligned}
$$

Here, $N(0,\sigma)$ denotes a Gaussian distribution with mean zero and standard deviation σ, while $\mathrm{Ber}_{\mathrm{logit}}(L)$ represents a Bernoulli distribution parameterized via the logit function, meaning the success probability is given by $p = (1+\exp(-L))^{-1}$. Varying the parameters β_1 and β_2 leads to different missingness rates. Specifically, when $\beta_1 = -0.2$ and $\beta_2 = -1.2$, it is around 50%; and for $\beta_1 = 1.1$ and $\beta_2 = -1.0$, it is approximately 30%.

The doubly-robust estimator for the FATE (DR.F) was computed using *one-step corrected* pseudo-outcomes, akin to the approach propsed by Kennedy (2023) to estimate the CATE. Let the regression surface $\widehat{Q}_1(W, Y_0, A) = \widehat{\mathbb{E}}\left[Y \mid W, Y_0, A, R_{Y_0} = 1\right]$ be learned from units with $R_{Y_0} = 1$. The pseudo-outcomes $\widetilde{\delta}(W, Y_0, A)$ are given by:

$$
\widetilde{\delta}(W, Y_0, A) := \Delta_a \widehat{Q}_1(W, Y_0, a) + \frac{A - \widehat{\pi}_1(W, Y_0)}{\widehat{\pi}_1(W, Y_0)[1 - \widehat{\pi}_{r_i}(W, Y_0)]} \left[Y_1 - \widehat{Q}_1(W, Y_0, A)\right].
$$

And $\widehat{\tau}(W)$ are the predictions from the meta-regression model $\widehat{\tau}(W) = \widehat{\mathbb{E}}\left[\widetilde{\delta}(W, Y_0, A) \mid W, R_{Y_0} = 1\right]$ across all units. The DR.F estimator is then given by:

$$
\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\tau}(W_i) + \frac{R_{Y_0,i}}{\widehat{\mathbb{P}}(R_{Y_0} = 1 \mid W_i)}\left(\widehat{\tau}(W_i) - \Delta_a \widehat{Q}_1(W_i, Y_{0,i}, a)\right)\right].
$$

An estimate of its variance, to construct confidence intervals, can be obtained either through bootstrap methods or via an estimate of the asymptotic variance given by:

$$
\frac{1}{n^2}\sum_{i=1}^{n}\left[\widehat{\tau}(W_i) + \frac{R_{Y_0,i}}{\widehat{\mathbb{P}}(R_{Y_0} = 1 \mid W_i)}\left(\widehat{\tau}(W_i) - \Delta_a \widehat{Q}_1(W_i, Y_{0,i}, a)\right) - \hat{\phi}\right]^2.
$$

Both the DR.N and DR.F estimators exhibit desirable asymptotic properties under several technical conditions. These include: *(i)* a fully nonparametric or saturated model for the data-generating process, *(ii)* smoothness of model paths, *(iii)* positivity of the relevant propensity scores, *(iv)* boundedness of the outcome mean, Gâteaux derivative, and their variances, *(v)* the Donsker class condition (i.e., bounded estimator complexity), and *(vi)* sufficiently fast convergence of nuisance estimates and their interaction terms. For further details on one-step corrections, we refer the reader to Hines et al. (2022).

The data and code utilized for the simulation study are accessible in a personal GitHub repository at `https://github.com/johandh2o`.

B APPLICATION CASE DETAILS

We assess the impact of pharmacological treatment with stimulant medication upon the numeracy test scores at grade 8 obtained by Norwegian children diagnosed with ADHD. By integrating information from national registries, we compile data on the medication history and national test scores of all children diagnosed with ADHD born between 2000 and 2007 in Norway, who would go to take the national test up to 2021. We exclude those with severe comorbid disorders (totaling $8\,450$ individuals). Variables at the student, family, and school levels are linked from the Norwegian Prescription Database (NorPD), the Norwegian Patient Registry (NPR), the Database for Control and Payment of Health Reimbursement (KUHR), Statistics Norway (SSB), and the Medical Birth Registry of Norway (MBRN). We leverage data on students' and parents' diagnoses and their consultations with medical services during the pre-exposure period. To operationalize relevant variables, we employ the following grouping:

- **Pre-exposure covariates** W: sex at birth, birth year/month cohorts, birth parity number, raw score at grade 5 national test for numeracy, mother's education level, mother's age at birth, student's and parents' diagnoses and medical consultations for related comorbid disorders, school identification (fixed effect), prior dispensations of ADHD stimulant medication for at least 90 days, and duration of prior treatment.
- **Exposure** A: having received dispensations of ADHD stimulant medication for at least 75% of the prescribed treatment period between the start of grade 6 and the national test in grade 8.
- **Outcomes** Y: raw scores at grade 8 national test for numeracy.

The missingness rate on test scores at grade 5 is about 10%. This rate is not expected to induce a disproportionate amount of selection bias but, from a public health standpoint, it could still lead to shifts in conclusions and policy design that may not be fully aligned with the target population. All estimators were fitted with sample-splitting and super-learning schemes, with a battery of four base algorithms: generalized linear models (GLM), penalized GLMs, random forest (RF), and boosted decision trees (BDT).

Due to the sensitive nature of the research topic, data from the application case cannot be obtained from the authors, but have to be requested from the relevant Norwegian authorities.