Statistica Sinica Preprint No: SS-2023-0382	
Title	Causal and Counterfactual Views of Missing Data Models
Manuscript ID	SS-2023-0382
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0382
Complete List of Authors	Razieh Nabi, Rohit Bhattacharya, Ilya Shpitser and James M. Robins
Corresponding Authors	Razieh Nabi
E-mails	razieh.nabi@emory.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

CAUSAL AND COUNTERFACTUAL VIEWS OF MISSING DATA MODELS

Razieh Nabi, Rohit Bhattacharya, Ilya Shpitser, James M. Robins

Emory University, Williams College, Johns Hopkins University, Harvard University

Abstract: It is often said that the fundamental problem of causal inference is a missing data problem—the comparison of responses to two hypothetical treatment assignments is made difficult because for every experimental unit only one potential response is observed. In this paper, we consider the implications of the converse view: that missing data problems are a form of causal inference. We make explicit how the missing data problem of recovering the complete data law from the observed law can be viewed as identification of a joint distribution over counterfactual variables corresponding to values had we (possibly contrary to fact) been able to observe them. Drawing analogies with causal inference, we show how identification assumptions in missing data can be encoded in terms of graphical models defined over counterfactual and observed variables. We review recent results in missing data identification from this viewpoint. In doing so, we note interesting similarities and differences between missing data and causal identification theories.

Key words and phrases: Causal inference, causal graphs, missing not at random

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

1. Introduction

Missing data is a common challenge in the analysis of survey, experimental, and observational data, both for the purpose of prediction and for drawing causal conclusions. Complete-case analysis is a popular and simple approach to handling missing data, but it is generally only justified when data entries are missing-completely-at-random (MCAR) (Rubin, 1976). When data entries are missing in a way that only depends on observed data values, the data are said to be missing-at-random (MAR) (Rubin, 1976). Under MAR assumptions, it is possible to identify target parameters of the underlying data distribution without the need for further parametric assumptions. Moreover, we can estimate parameters identified under MAR via likelihood-based methods such as expectation maximization (Dempster et al., 1977; Horton and Laird, 1999; Little and Rubin, 2002), multiple imputation (Rubin, 1987; Schafer, 1999), inverse probability weighting (Robins et al., 1994; Li et al., 2013), or semiparametric methods that exploit information about mechanisms that determine missingness and are closely related to methods for estimating causal parameters (Robins et al., 1995; Scharfstein et al., 1999; Robins and Rotnitzky, 2001; Tsiatis, 2006; Tchetgen, 2009).

However, it is often the case that missingness status depends on the underlying values that are themselves censored. This type of missingness is

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

known as missing-not-at-random (MNAR) (Rubin, 1976). Without any assumptions, parameters of interest in an MNAR model cannot be identified from the observed data distribution. A common approach to MNAR problems is to impose sufficient parametric or semiparametric restrictions on the underlying data distribution and missingness selection model, such that they yield identification (Wu and Carroll, 1988; Little and Rubin, 2002; Ma et al., 2003; Wang et al., 2014; Miao et al., 2016; Miao and Tchetgen Tchetgen, 2016; Sun et al., 2018). Other approaches to handling MNAR mechanisms include conducting sensitivity analysis and obtaining nonparametric bounds (Rotnitzky et al., 1998; Robins et al., 2000; Scharfstein and Irizarry, 2003; Vansteelandt et al., 2007; Mattei et al., 2014; Moreno-Betancur and Chavance, 2016; Scharfstein et al., 2021; Duarte et al., 2024).

Identification may be achieved by imposing a set of independence restrictions among variables in the full data distribution that are sufficient to express parameters of interest as functions of the observed data distribution. This approach is also taken in nonparametric identification theory developed in causal inference, where independence restrictions among variables in a full data distribution in a causal model are encoded via directed acyclic graphs (DAGs). This framework has led to sound and complete algorithms for identification of a wide set of causal parameters as functions of the ob-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

served data (Tian and Pearl, 2002; Shpitser and Pearl, 2006; Huang and Valtorta, 2006; Richardson et al., 2023; Bhattacharya et al., 2022). Completeness here means that failure of the algorithm on a particular parameter input implies that the parameter is, in fact, *provably* not identified given the set of restrictions encoded by the proposed model. These algorithms generalize many existing results regarding special cases, such as identification by covariate adjustment that relies on the stable unit treatment value assumption and conditional ignorability (Rubin, 1976), or the g-computation algorithm that relies on sequential ignorability (Robins, 1986).

Directed acyclic graphs have also been adapted to encode independence restrictions in full data distributions in missing data models. Using such representations, many complex scenarios have been described where it is possible to recover target parameters as functions of the observed data distribution (Glymour, 2006; Daniel et al., 2012; Martel García, 2013; Mohan et al., 2013; Thoemmes and Rose, 2014; Tian, 2015; Shpitser, 2016; Bhattacharya et al., 2019; Saadati and Tian, 2019; Nabi et al., 2020; Mohan and Pearl, 2021; Scharfstein et al., 2021; Nabi and Bhattacharya, 2023; Guo et al., 2023; Chen et al., 2023). In particular, this line of work has shed light on several classes of MNAR models that still permit identification of the target parameter without relying on any parametric assumptions on

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

the full data distribution. In addition to providing concise representations of statistical models by means of factorizations, graphs also illustrate the causal mechanisms responsible for missingness and provide a natural interpretation of such mechanisms in applied settings.

It has been noted by many authors that causal inference and missing data are analogous in terminology, theory of identification, and statistical inference. Causal inference has often been phrased as a missing data problem since responses to some treatment interventions are not observed (Rubin, 1974; Ding and Li, 2018), and missing data has been viewed as a form of causal inference where interventions on missingness indicators can be carried out (Robins, 1986; Shpitser et al., 2015; Bhattacharya et al., 2019). At the same time, not much discussion has been devoted to important differences between these frameworks. In this paper, we discuss identifiability of models with MNAR mechanisms and examine new developments in graphical missing data models. We show how identification theory may be understood by viewing missing data models counterfactually, by analogy with causal models, and discuss additional ingredients needed to augment causal identification theories to handle identification in missing data models.

The paper is organized as follows. In Section 2, we provide a brief

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

overview of classical missing data models and then redefine these models using causal and counterfactual terminology. In Section 3, we give an overview of statistical and causal models of DAGs. In Section 4, we formally define missing data DAG models. In Section 5, we discuss several unique techniques for nonparametric identification of complete data distributions in missing data DAGs. In Section 6, we conclude the paper by providing a discussion on whether ideas explored in missing data DAG models joined with rank preservation assumptions can be used to obtain novel identification results in the causal inference settings. We also consider missing data DAGs that account for the presence of unmeasured confounders in the supplementary materials.

2. Missing Data Models

2.1 Classical Missing Data Models

Inferring a parameter of interest in the presence of missing data often involves posing a statistical model that encodes a set of assumptions on the missingness mechanisms. Let $Z = (Z_1, \ldots, Z_K)^T$ be a vector of K random variables with finite support and probability distribution p_Z. Given a sample of the random vector Z, let $R = (R_1, \ldots, R_K)^T$ be the vector of binary missingness indicators with $R_k = 1$ if Z_k is observed and $R_k = 0$ if Z_k is

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

missing. Denote the conditional distribution of R given Z by $p_{R|Z}$ and the joint distribution of Z and R by $p_{(R,Z)}$, and assume p_Z, $p_{R|Z}$, and $p_{(R,Z)}$ are contained in statistical models $\mathcal{M}_Z$, $\mathcal{M}_{R|Z}$, and $\mathcal{M} = \mathcal{M}_Z \otimes \mathcal{M}_{R|Z}$, respectively, where $\otimes$ denotes the direct product of the two statistical models. $\mathcal{M}$ is referred to as a selection model in the missing data literature (Little, 2016). The observed data is often denoted by $O = (R, Z_{\text{obs}})$, where Z_{obs} is the subvector of Z corresponding to the subvector of R whose entries are 1, i.e., $Z_{\text{obs}} \coloneqq \{Z_k \in Z \text{ s.t. } R_k = 1\}$. If $\mathcal{M}$ imposes no restrictions on the observed data distribution p_O, then it is called a *nonparametric saturated* model (Robins, 1997).

2.2 A Counterfactual View of Classical Missing Data Models

To motivate causal and counterfactual views of missing data, we first provide a description of causal models. Causal models are often phrased in terms of counterfactual responses to interventions. A random variable of the form $Y^{(a)}$ is used to denote the response of an outcome Y when a treatment A is intervened on and set to the value a. The observed (factual) outcome Y is typically defined as coarsened versions of counterfactual outcomes via the *consistency* property. For example, for a binary treatment A with values 0 and 1, the observed outcome is obtained via the following

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

coarsening mechanism $Y := Y^{(a=1)} \times A + Y^{(a=0)} \times (1 - A)$. That is, the observed outcome Y gives us an imperfect view into the underlying counterfactuals $Y^{(a=1)}$ and $Y^{(a=0)}$: for individuals that received treatment $A = 1$, the observed outcome corresponds to the counterfactual response $Y^{(a=1)}$; for those that received $A = 0$, we gain information regarding the counterfactual response $Y^{(a=0)}$. In general, we see only one of the (potentially several for non-binary treatments) counterfactual responses for each individual. This complicates the task of computing causal parameters, which are phrased in terms of contrasts between different counterfactual responses. This forms the basis of the observation that the fundamental problem of causal inference is a missing data problem.

We can redefine the classical missing data terminology using the terminology of causal models described above. We may view each missingness indicator $R_k \in R$ as a treatment variable that can be intervened on. Each $Z_k \in Z$ can then be interpreted counterfactually—a random variable had we, possibly contrary to fact, intervened and set the corresponding missingness indicator R_k to 1. By analogy with causal models, from here onward, we refer to Z_k as $L_k^{(r_k=1)}$ to highlight the counterfactual nature of the variable. This notation explicitly encodes, in counterfactual language, the assumption implicit in classical missing data models, that the value of Z_k

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

remains the same regardless of whether any other $Z_j \in Z$ is observed or missing (or equivalently R_j is 1 or 0). We collect all these counterfactual variables into a vector $L^{(r=1)} := (L_1^{(r_1=1)}, \ldots, L_K^{(r_K=1)})^T$, and simplify the notation for $L_k^{(r_k=1)}, L^{(r=1)}$ via $L_k^{(1)}, L^{(1)}$.

The link between Z and Z_{obs} can be viewed as the link between the counterfactual variables $L^{(1)}$ that are of substantive interest, treatment variables R, and factual variables L (a.k.a. proxies) that we observe. Specifically, for any $L_k^{(1)} \in L^{(1)}$, its corresponding proxy $L_k \in L$ is *deterministically* defined as a function of $L_k^{(1)}$ and R_k as follows: $L_k = L_k^{(1)}$ if $R_k = 1$ and $L_k =$ "?" if $R_k = 0$. This link is closely related to the *consistency* property in causal inference, described above.

The state space of any $L_k \in L$ is equal to the state space of the corresponding $L_k^{(1)}$ in $L^{(1)}$ joined with the special value "?". Hence, we denote generic realizations of $L_k \in L$ and $L_k^{(1)} \in L^{(1)}$ as l_k and $l_k^{(1)}$, respectively. We denote realizations of $L^{(1)}$ and L by $l^{(1)}$ and l, respectively. By consistency, it is always true that $p(L_k^{(1)} \mid R_k)|_{R_k=1} = p(L_k \mid R_k)|_{R_k=1}$, however $p(l_k^{(1)} \mid r_k)|_{r_k=1} = p(l_k \mid r_k)|_{r_k=1}$ is true if and only if $l_k^{(1)} = l_k$. Here, we use the notation $p(.)|_{R_k=1}$ to mean that R_k in the probability expression $p(.)$ is evaluated at value 1. Our convention is that in any equation or probability expression where $l_k^{(1)}$ and l_k appear and $r_k = 1$, $l_k^{(1)} = l_k$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

Thus $p(l_k^{(1)} \mid r_k)|_{r_k=1} = p(l_k \mid r_k)|_{r_k=1}$ becomes always true. In this way, we can evaluate probability expressions at realizations rather than at the corresponding random variables while still imposing consistency.

We redefine Z_{obs} via $L = L^{(R)} = L^{(r)}|_{R=r}$. With this new notation, the observed data then changes from $O = (R, Z_{\text{obs}})$ to $O = (R, L)$. Note, however, that the state space of L is formed by augmenting the state space of Z with the special value "?".

We end this subsection by noting an important difference between causal and missing data models. A causal model posits the existence of two counterfactuals $Y^{(a=0)}$, $Y^{(a=1)}$ for any binary treatment A and observed variable Y. In contrast, in a missing data model, there exists only one counterfactual $L_k^{(r_k=1)}$ for any binary missingness indicator R_k and potentially censored variable L_k. As we will show later, this crucial difference allows additional identification theory to be developed specifically for missing data problems that has no analogue in causal inference identification theory.

2.3 Identification in Missing Data Models

A common goal in missing data problems is to determine whether the joint distribution of the complete data $L^{(1)} \coloneqq Z$, that is $p(l^{(1)}) \coloneqq p(z) \in \mathcal{M}_{L^{(1)}} \coloneqq \mathcal{M}_Z$, is identified from the observed data $O = (R, L) \coloneqq (R, Z_{\text{obs}})$,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

in a model $\mathcal{M} = \mathcal{M}_{R|L^{(1)}} \otimes \mathcal{M}_{L^{(1)}}$ defined over the joint distribution of $\left(R, L^{(1)}\right) \coloneqq (R, Z)$. When discussing identification in missing data problems below, we will refer to $p(l^{(1)})$ as the *target law*, $p(r \mid l^{(1)})$ as the *missingness mechanism*, and $p(l^{(1)}, r)$ as the *full law*. These distributions may also be extended with a set of auxiliary variables W that are always observed and/or variables U that are never observed.

Unless explicitly stated otherwise, we assume $\mathcal{M}_{R|L^{(1)}}$ encompasses positive distributions, meaning that for every $p(R = r \mid l^{(1)}) \in \mathcal{M}_{R|L^{(1)}}$, $p(R = r \mid l^{(1)}) > 0$ with probability 1 for all $r \in \{0, 1\}^K$. This assumption may be modified to allow pattern restrictions such as monotonicity, where $p(R = r \mid l^{(1)}) > 0$ for any pattern of values of R where, under some ordering of missingness indicators $R_1, \ldots, R_K$, if $R_k = 0$ then $R_{k+1} = 0$ with probability 1 for every $k \in \{1, \ldots, K-1\}$.

The model $\mathcal{M} = \mathcal{M}_{R|L^{(1)}} \otimes \mathcal{M}_{L^{(1)}}$ is said to be nonparametric just identified if: (i) $\mathcal{M}_{L^{(1)}} = \mathcal{M}^{\text{np}}_{L^{(1)}}$, where $\mathcal{M}^{\text{np}}_{L^{(1)}}$ is the set of all distributions of $L^{(1)}$, (ii) the distribution of the observed data L is unrestricted, and (iii) $p\left(l^{(1)}\right)$ is identified from the observed data distribution $p(l)$ (Robins, 1997).

Variables in L, by definition, contain all information in variables in R. However, in what follows, we will employ observed data distributions $p(l, r)$ which will allow ideas from causal inference to be used for identfiability,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

with R being viewed as treatment variables.

A *necessary and sufficient* condition for identification of the target law $p(l^{(1)})$ is that for all $p(r \mid l^{(1)}) \in \mathcal{M}_{R|L^{(1)}}$, $p(R = 1 \mid l^{(1)}) > 0$ with probability 1 and $p(R = 1 \mid l^{(1)})$ is identified from the observed distribution $p(l, r)$. This follows from the fact that $p\left(R = 1, l^{(1)}\right) = p\left(R = 1, l\right)$ by consistency and an application of chain rule:

$$p(l^{(1)}) = p(l, R = 1)/p(R = 1 \mid l^{(1)}). \tag{2.1}$$

A *necessary and sufficient* condition for identification of the full law $p(l^{(1)}, r)$ is that for all $p(r \mid l^{(1)}) \in \mathcal{M}_{R|L^{(1)}}$, $p(R = 1 \mid l^{(1)}) > 0$ with probability 1 and $p(R = r \mid l^{(1)})$ is identified from the observed distribution $p(l, r)$, for any missingness pattern $r \in \{0, 1\}^K$. This also follows from the chain rule of probability,

$$p(l^{(1)}, R = r) = \left\{p(l, R = 1)/p(R = 1 \mid l^{(1)})\right\} \times p(R = r \mid l^{(1)}). \tag{2.2}$$

In this paper, we study a general procedure for analyzing MNAR models by imposing a set of restrictions on the full data distribution (the target distribution and its missingness mechanism) that are sufficient to yield identification of the parameter of interest. In many models, the restrictions may be represented by a factorization of the full data law with respect to a DAG. Our objective is to identify the target law $p(l^{(1)})$ from factual data

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

on variables (R, L) where the missing data model is represented via a DAG with potentially hidden (never observed) variables. As we will see, restricting attention to such missing data models allows ideas inspired by causal identification theory to be brought to bear. To this end, we first elaborate on the use of DAGs in causal inference.

3. Directed Acyclic Graphs in Causal Inference

Progress in reasoning about counterfactual quantities is achieved by imposing restrictions on a causal model, which consists of sets of joint distributions defined over the factual and counterfactual random variables. Consider the task of identifying the average causal effect, defined as $\mathbb{E}(Y^{(a=1)} - Y^{(a=0)})$. Given a set of baseline covariates X we may restrict ourselves to distributions satisfying an independence assumption that $A \perp\!\!\!\perp Y^{(a)} \mid X$ for every value a, and positivity of the distribution of A conditioned on X, denoted by $p_{A|X}$. Under the assumptions of this causal model, the average causal effect is identified via the adjustment formula: $\mathbb{E}\{\mathbb{E}(Y|A = 1, X) - \mathbb{E}(Y|A = 0, X)\}$.

Assumptions in a causal model can often be encoded in a more intuitive fashion via DAGs. We formally introduce the statistical and causal models of a DAG below.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

3.1 Statistical DAG Models

Let $V = (V_1, \ldots, V_K)^T$ be a vector of K random variables with finite support and probability distribution p_V. We will abbreviate the joint probability $p_V(V = v)$ as simply $p(v)$. Restrictions on the distribution p_V can be encoded via a DAG as follows. Define a DAG $\mathcal{G}(V)$ consisting of a set of nodes V associated with each random variable $V_i \in V$, and a set of directed edges that form connections between these variables with the restriction that these edges do not form a directed cycle. We will sometimes abbreviate $\mathcal{G}(V)$ as simply $\mathcal{G}$ if the set of vertices V is assumed or obvious. For a given DAG $\mathcal{G}$, the statistical model $\mathcal{M}^{\mathcal{G}}$ is the set of distributions that factorize as $p(v) = \prod_{v_i \in v} p\left(v_i \mid \text{pa}_{\mathcal{G}}(v_i)\right)$, where $\text{pa}_{\mathcal{G}}(v_i)$ is the set of values of variables corresponding to the parents of V_i, $\text{pa}_{\mathcal{G}}(V_i)$, i.e., the set of vertices in $\mathcal{G}$ with directed edges into V_i. Distributions in $\mathcal{M}^{\mathcal{G}}$ are said to be Markov relative to $\mathcal{G}$.

We use the following notation for standard genealogical relations in DAGs. We denote the *children* of a vertex V_i in $\mathcal{G}$—the set of all vertices that have V_i as a parent—as $\text{ch}_{\mathcal{G}}(V_i)$. We denote the *descendants* of V_i—the set of all vertices V_j such that there exists a directed path from V_i to V_j—as $\text{de}_{\mathcal{G}}(V_i)$. By convention, $\text{de}_{\mathcal{G}}(V_i)$ is defined to include V_i itself.

When $\mathcal{G}$ is *complete*—all vertices are pairwise connected via a directed

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

edge—$\mathcal{M}^{\mathcal{G}}$ imposes no restrictions on p_V. When $\mathcal{G}$ is not complete, it is informative to compare the DAG factorization of p_V to the chain rule factorization to understand how missing edges entail restrictions on the observed distribution. Consider any valid topological ordering $\prec_{\mathcal{G}}$ of the variables—an ordering satisfying the property that whenever $V_i \prec_{\mathcal{G}} V_j$, V_i is not a descendant of V_j in $\mathcal{G}$. Then for every variable $V_i \in V$, define $\text{past}_{\prec_{\mathcal{G}}}(V_i)$ to be the set of vertices earlier than V_i under $\prec_{\mathcal{G}}$ (we suppress the explicit mention of $\mathcal{G}$ and write instead $\text{past}_{\prec}(V_i)$ to avoid notational clutter). Under any variable ordering $\prec_{\mathcal{G}}$, we have the following equality between the chain factorization and DAG factorization of $p(v)$: $p(v) = \prod_{v_i \in v} p(v_i \mid \text{past}_{\prec}(v_i)) = \prod_{v_i \in v} p(v_i \mid \text{pa}_{\mathcal{G}}(v_i))$.

Whenever $\text{pa}_{\mathcal{G}}(V_i) \subset \text{past}_{\prec}(V_i)$ (corresponding to missing edges in $\mathcal{G}$) the above equality implies that $V_i \perp\!\!\!\perp \text{past}_{\prec}(V_i) \setminus \text{pa}_{\mathcal{G}}(V_i) \mid \text{pa}_{\mathcal{G}}(V_i)$. That is, given a DAG $\mathcal{G}$, restrictions in p_V are characterized by the following ordered local Markov property: each variable V_i is independent of its non-parental past given its parents. All restrictions entailed by a DAG $\mathcal{G}$ are easily read via the d-separation criterion (Pearl, 2009).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

3.2 Causal DAG Models

In addition to a statistical DAG model $\mathcal{M}^{\mathcal{G}}$ representing restrictions on the factual variables V, it is possible to define a causal DAG model associated with $\mathcal{G}$. Causal DAG models can be generalized from statistical DAG models by equipping them with a special subset $A^{\dagger} \subseteq V$ referred to as treatment or action variables, with non-action variables $V \setminus A^{\dagger}$ denoted as Y.

We will index variables $X \subseteq V$ in the model via subscript indices that are consistent with a total ordering $\prec$ on V which must be topological with respect to $\mathcal{G}$, in the following sense: if X_i and X_j are both elements of $X \subseteq V$, and $i < j$, then $X_i \prec X_j$.

Given any variable X_i in $X \subseteq V$, define $\overline{X}_i$ to be the set consisting of X_i and all elements in X earlier than X_i in the ordering $\prec$. Similarly, define $\underline{X}_i$ to be the set consisting of X_i and all elements in X later than X_i in the ordering $\prec$. For any variable V_i and a set X, define $X_{j>i}$ to be the $\prec$-smallest element of X larger than V_i, and $X_{j<i}$ to be the $\prec$-largest element of X smaller than V_i. Define X_{-i} to be all elements in X other than X_i. Finally, for every $V_i \in V$, define $\text{past}_{\prec}(V_i)$ to be all variables in V earlier than V_i in the ordering. We will extend this notation in the natural way to values as well, e.g. $\overline{x}_i$ are values of $\overline{X}_i$.

Although a number of different causal DAG models have been proposed

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

(Robins, 1986; Pearl, 2009), they all satisfy the following assumptions, for a given fixed topological ordering $\prec$ of $\mathcal{G}$: (Note that any topological ordering yields the same model with different variable indexing.)

(i) **Counterfactual existence.** For $V_i \in V$, there exists a set of counterfactual variables $V_i^{(a^\dagger)}$ representing V_i's behavior when $A^\dagger$ is set to $a^\dagger$ by external intervention.

(ii) **No backwards causation.** For every $V_i^{(a^\dagger)}$, we have

$$V_i^{(a^\dagger)} \equiv V_i^{(\overline{a}^\dagger_{j<i},\, \underline{a}^\dagger_{j>i})} = V_i^{(\overline{a}^\dagger_{j<i},\, \underline{a}^\dagger_{j>i}{}')} \equiv V_i^{(\overline{a}^\dagger_{j<i})},$$

where $\overline{a}^\dagger_{j<i}$ are distinct values of variables in $\overline{A}^\dagger_{j<i} = \{A_j \in A^\dagger : A_j \prec V_i\}$, and $\underline{a}^\dagger_{j>i}{}'$ and $\underline{a}^\dagger_{j>i}$ are distinct values of variables in $\underline{A}^\dagger_{j>i} = \{A_j \in A^\dagger : V_i \prec A_j\}$. For $V_i = A_i \in A^\dagger$, this definition implies that A_i does not depend on its own counterfactually set values a_i. This follows since the above identity yields $A_i^{(a^\dagger)} = A_i^{(\overline{a}^\dagger_{j<i})}$, and $\overline{A}^\dagger_{j<i}$ does not include A_i.

(iii) **Recursive substitution.** Given any subset $A \subset A^\dagger$ and values a of A, we define $V_i^{(a)}$ as $V_i^{\left(a, \{A_j^{(a)} : A_j \in A^\dagger \setminus A\}\right)}$. This means $V_i^{(a)}$ is the response of V_i when A is set to a and A_j in $A^\dagger \setminus A$ is set to its value under this intervention. This recursive definition ensures each $V_i^{(a)}$ is definable in terms of existing counterfactuals, guaranteed by Property

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

(i) and acyclicity of $\mathcal{G}$.

(iv) **Consistency.** Every counterfactual variable $V_i^{(a)}$ is linked to the factual variable V_i by the consistency property, which states that $V_i = \sum_a \mathbb{I}(A = a) \times V_i^{(a)} = \sum_{\overline{a}_j} \mathbb{I}(\overline{A}_j = \overline{a}_j) \times V_i^{(\overline{a}_j)}$, or equivalently, $V_i = V_i^{(a)}$ if $A = a$. Thus, in any context where $A = a$, we assume $V_i = V_i^{(a)}$, allowing us to assert $p(v_i^{(a)} \mid a) = p(v_i \mid a)$ as always true.

(v) **Positivity.** For every $A_i^\dagger \in A^\dagger$, $p(a_i^\dagger \mid \text{past}_\prec(A_i^\dagger)) > 0$ for all values of variables in $\text{past}_\prec(A_i^\dagger)$.

(vi) **Sequential ignorability.** For every $A_i \in A \subseteq A^\dagger$, we assume $\underline{V}_{j>i}^{(a)} \perp\!\!\!\perp A_i \mid \text{past}_\prec(A_i)$, for values of $\text{past}_\prec(A_i)$ that are consistent with the treatment assignment a.

(vii) **Markov property.** The joint $p(v)$ is Markov relative to $\mathcal{G}$.

Given any $A \subseteq A^\dagger$, assumptions (i) and (iii) allows us to meaningfully discuss counterfactuals $L^{(a)}$, where $L = V \setminus A$. Furthermore, under the above causal model, the distribution $p(l^{(a)})$—the joint distribution over counterfactuals obtained by setting a subset A of the treatment variables to a set of values a—can be expressed as a function of the joint distribution $p(v)$ as follows. Assumptions (iii) and (iv), and our notational convention together imply $p(l^{(a)}) \times p(a \mid l^{(a)}) = p(l^{(a)}, a) = p(l, a)$, so

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

$$p(l^{(a)}) = \frac{p(l,a)}{p(a \mid l^{(a)})} = \frac{p(v)}{p(a \mid l^{(a)})}. \qquad \textit{(counterfactual g-formula)} \tag{3.3}$$

We refer to (3.3) as the counterfactual g-formula, since counterfactual variables appear in the denominator. Note that this expression is exactly the same as the expression for target law identification from (2.1) with $a = r$ and $r = 1$.

It follows that $p(l^{(a)})$ is identified from the distribution of the factuals V if and only if $p(a \mid l^{(a)})$ is identified. The additional assumptions (iv), (v), (vi), and (vii) imply $p(a \mid l^{(a)})$ is indeed identified as follows:

$$\begin{aligned} p(a \mid l^{(a)}) &= \prod_{a_k \in a} p(a_k | \overline{a}_{k-1}, \{l_i^{(a)} : l_i \prec a_k\}, \{l_i^{(a)} : a_k \prec l_i\}) \\ &= \prod_{a_k \in a} p(a_k | \text{past}_{\prec}(a_k), \{l_i^{(a)} : a_k \prec l_i\}) \\ &= \prod_{a_k \in a} p(a_k | \text{past}_{\prec}(a_k)) \\ &= \prod_{a_k \in a} p(a_k | \, \text{pa}_{\mathcal{G}}(a_k)). \end{aligned}$$

The first equality follows from the chain rule of probability, where for each conditional $p(a_k \mid .)$ we split $l^{(a)}$ into the sets $\{l_i^{(a)} : l_i \prec a_k\}$ and $\{l_i^{(a)} : a_k \prec l_i\}$. To see that the second equality follows, note that no backwards causation (ii) implies $\{l_i^{(a)} : l_i \prec a_k\} = \{l_i^{(\overline{a}_{k-1})} : l_i \prec a_k\}$. Conditioning on $\overline{a}_{k-1}$ then implies that $\{l_i^{(\overline{a}_{k-1})} : l_i \prec a_k\} = \{l_i : l_i \prec a_k\}$. Finally, if

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

we partition V into L and A, $\text{past}_{\prec}(a_k) = (\{l_i : l_i \prec a_k\}, \overline{a}_{k-1})$. The third equality follows from (vi). The fourth from (vii). Finally, the positivity assumption (v) implies the RHS of the last equality is a unique function of the distribution of V. We conclude that, by term cancellation, $p(l^{(a)}) = p(l,a)/p(a \mid l^{(a)})$ is identified by the g-formula (Robins, 1986) as follows:

$$\frac{\prod_{v_k \in l \cup a} p(v_k \mid \text{past}_{\prec}(v_k))}{\prod_{a_k \in a} p\big(a_k \mid \text{past}_{\prec}(a_k)\big)} = \frac{\prod_{v_k \in l \cup a} p(v_k \mid \text{pa}_{\mathcal{G}}(v_k))}{\prod_{a_k \in a} p\big(a_k \mid \text{pa}_{\mathcal{G}}(a_k)\big)} = \prod_{l_k \in l} p\left(l_k \mid \text{pa}_{\mathcal{G}}(l_k)\right), \tag{3.4}$$

where all values in the above expression are consistent with a, l, a functional we will denote by $p_{\mathcal{G}}(l \parallel a)$, following the notation in (Lauritzen, 1996). Note that by our notational convention, it follows that $p_{\mathcal{G}}(l^{(a)} \parallel a) = p_{\mathcal{G}}(l \parallel a)$.

The g-formula $p_{\mathcal{G}}(l \parallel a)$ in (3.4) can be viewed as a truncated factorization in the sense that the terms $p(a_k \mid \text{pa}_{\mathcal{G}}(a_k))$ that occur in the Markov factorization of the distribution $p(l,a)$ are no longer present in the g-formula factorization representing the intervention distribution. As we will see in Section 4, the g-formula is also closely connected to identification in missing data problems.

The truncated factorization $p_{\mathcal{G}}(l \parallel a)$ is still a distribution as it provides a mapping from values a of A to normalized densities over variables in L. We call objects of this type *kernels*. A kernel acts in most respects like a

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

conditional distribution. In particular, given a kernel $p(v \parallel w)$ and a subset $Z \subseteq V$, conditioning and marginalization are defined in the usual way as

$$p(z \parallel w) \coloneqq \sum_{v \setminus z} p(v \parallel w) \quad \text{and} \quad p(v \setminus z \mid z \parallel w) \coloneqq \frac{p(v \parallel w)}{p(z \parallel w)}. \tag{3.5}$$

Property (vi) above is the critical identifying assumption for obtaining the g-formula: in words, for each A_i there exists a factual past: $\text{past}_{\prec}(A_i) = \{V_j \in V : V_j \prec A_i\} \subseteq V$ such that treatment A_i is as if randomly assigned conditional on this past—and hence independent of future counterfactuals. In the language of epidemiologists, conditional on $\text{past}_{\prec}(A_i)$ the causal effect of A_i on $\underline{V}_{j>i} = \{V_j \in V : A_i \prec V_j\}$ is unconfounded.

In most observational studies, some of the variables that need to be included in $\text{past}_{\prec}(A_i)$ (and thus in V) to make A_i unconfounded may be unknown to the investigators and/or known but not measured for financial or logistical reasons. As a consequence $p(l^{(a)})$ will obviously not be identified from the factual distribution $p(o, a)$ of the observed variables (O, A), where $O \subset L$. Moreover, even the counterfactual distribution over observed outcomes,

$$\begin{aligned} p(o^{(a)}) &= \sum_{l^{(a)} \setminus o^{(a)}} p(l^{(a)}) = \sum_{l \setminus o} p_{\mathcal{G}}(l \parallel a) \equiv p_{\mathcal{G}}(o \parallel a) \\ &= \sum_{l \setminus o} \frac{\prod_{v_k \in l \cup a} p(v_k \mid \text{past}_{\prec}(v_k))}{\prod_{a_k \in a} p\big(a_k \mid \text{past}_{\prec}(a_k)\big)} = \sum_{l \setminus o} \frac{\prod_{v_k \in l \cup a} p(v_k \mid \text{pa}_{\mathcal{G}}(v_k))}{\prod_{a_k \in a} p\big(a_k \mid \text{pa}_{\mathcal{G}}(a_k)\big)}, \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

may not be identified from $p\,(o, a)$. This is because, in general, property (vi) (sequential ignorability) does *not* imply $\underline{O}^{(a)}_{j>i} \perp\!\!\!\perp A_i \mid \text{past-obv}_{\prec}(A_i)$ for all i, where $\underline{O}_{j>i} = \{O_j \in O : A_i \prec O_j\}$, and past-obv$_{\prec}(A_i)$ are the elements in past$_{\prec}(A_i)$ that are observed.

As an example illustrating Assumptions (i)-(vii), consider the DAG in Fig. 1(a). The statistical model $\mathcal{M}^{\mathcal{G}}$ of this DAG is the set of distributions $p(v)$ that factorize as: $p(u_1)p(u_2|u_1)p(r_1|u_2)p(r_2|l_1, r_1)p(l_1|r_1, u_1)p(l_2|r_2, u_2)$. Suppose we choose the treatment set $A^{\dagger} = \{R_1, R_2\}$ and the outcome set $Y \coloneqq \{U_1, U_2, L_1, L_2\}$. Assumption (i) entails, for every pair of values r_1, r_2, the existence of counterfactuals: $L_1^{(r_1,r_2)}$, $L_2^{(r_1,r_2)}$, $R_2^{(r_1,r_2)}$, $R_1^{(r_1,r_2)}$, $U_1^{(r_1,r_2)}$, $U_2^{(r_1,r_2)}$. Assumption (ii) entails the following identities: $R_2^{(r_1,r_2)} = R_2^{(r_1)}$, $R_1^{(r_1,r_2)} = R_1$, $U_1^{(r_1,r_2)} = U_1$, $U_2^{(r_1,r_2)} = U_2$. Note, in particular, that counterfactuals corresponding to R_1 and R_2 are not influenced by their own counterfactually set values. Assumption (iii) allows us to define all possible counterfactuals allowed by the model in terms of counterfactuals assumed to exist via (i). For example, we may define $L_2^{(r_1)}$ as $L_2^{(R_2^{(r_1)},r_1)}$. Assumption (iv) implies the following identities: $L_1 = \sum_{r_1} \mathbb{I}(R_1 = r_1)L_1^{(r_1)}$, $L_2 = \sum_{r_1,r_2} \mathbb{I}(R_1 = r_1, R_2 = r_2)L_2^{(r_1,r_2)}$. For binary R_1, R_2, these may be rewritten as $L_1 = L_1^{(1)}R_1 + L_1^{(0)}(1 - R_1)$, and $L_2 = L_2^{(1,1)}R_1R_2 + L_2^{(1,0)}R_1(1 - R_2) + L_2^{(0,1)}(1 - R_1)R_2 + L_2^{(0,0)}(1 - R_1)(1 - R_2)$. Assumption (v) implies that

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

$p(r_1|u_2)$ and $p(r_2|l_1, r_1)$ are positive for all values of u_2 and l_1, r_1, respectively. Assumption (vi) implies the following independence assumptions: $L_1^{(r_1)}, R_2^{(r_1)}, L_2^{(r_1)} \perp\!\!\!\perp R_1 \mid U_1, U_2$ and $L_2^{(r_1,r_2)} \perp\!\!\!\perp R_2 \mid R_1 = r_1, L_1, U_2, U_1$. Assumption (vii) implies the factorization that was described above. Under the above described causal DAG, the counterfactual distribution $p(u_1^{(r_1,r_2)}, u_2^{(r_1,r_2)}, l_1^{(r_1,r_2)}, l_2^{(r_1,r_2)})$ is identified via the g-formula as:

$$p_{\mathcal{G}}(y \parallel r_1, r_2) = p(u_1, u_2, r_1, l_1, r_2, l_2)/\{p(r_2 \mid r_1, l_1) \times p(r_1 \mid u_2)\}. \quad (3.6)$$

Similar to how the DAG factorization is the factorized representation of a factual probability distribution Markov relative to a given causal DAG, the g-formula $p_{\mathcal{G}}(y \parallel a)$, is the factorized representation of a counterfactual probability distribution Markov relative to a truncated version of the causal DAG, a.k.a. conditional causal DAG, in which the edges pointing into the treatment variables A that are intervened upon are removed. The conditional causal DAG corresponding to (3.6) is shown in Fig. 1(b).

Suppose U_1, U_2 are unobserved, and let $O = (L_1, L_2)$. The distribution of interest

$$p\left(o^{(r_1,r_2)}\right) \equiv p(l_1^{(r_1,r_2)}, l_2^{(r_1,r_2)}) \coloneqq p_{\mathcal{G}}(l_1, l_2 \parallel r_1, r_2) = \sum_{u_1,u_2} p_{\mathcal{G}}(l_1, l_2, u_1, u_2 \parallel r_1, r_2)$$

is not a function of the observed data distribution $\sum_{u_1,u_2} p(u_1, u_2, r_1, r_2, l_1, l_2)$ (Shpitser and Pearl, 2006). Intuitively, this is because the g-formula $p(l_1, l_2 \parallel r_1, r_2)$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

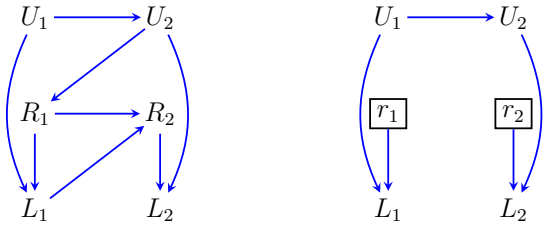


Figure 1: (a) A DAG where U_1 and U_2 may be unmeasured; (b) A conditional DAG illustrating interventions on R_1 and R_2.

(moving forward we drop the subscript $\mathcal{G}$, when it is clear from the context what graph we are referring to) contains the term $p(r_1 \mid u_2)$ in the denominator, and there is no information on U_2 in the observed data.

We now describe how DAGs are used to encode independence restrictions in a missing data model.

4. Missing Data DAG Models

Unlike a causal DAG, a missing data DAG (or m-DAG for short) includes counterfactual variables directly on the graph. Specifically, an m-DAG $\mathcal{G}_m(V)$ consists of a set of vertices V associated with variables $L^{(1)}, R, L$. There may exist additional vertices corresponding to variables that are always observed, and variables that are never observed. See Supplemental Section S2 for examples. In addition to acyclicity, an m-DAG restricts the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

presence of certain edges in the following ways:

(a) $\text{pa}_{\mathcal{G}_m}(L_k) = \{L_k^{(1)}, R_k\}$. That is, each proxy variable $L_k \in L$ has only two parents in $\mathcal{G}_m$ to represent the deterministic function that defines L_k in terms of $L_k^{(1)}$ and R_k. These edges are drawn in gray in all m-DAGs in this manuscript in order to distinguish deterministic relations from probabilistic ones.

(b) $L^{(1)} \cap \{\text{de}_{\mathcal{G}_m}(R) \cup \text{de}_{\mathcal{G}_m}(L)\} = \emptyset$. That is, variables in R and L cannot have directed paths to variables in $L^{(1)}$. In the presence of always observed variables W or always missing variables U, we also assume $(W \cup U) \cap \{\text{de}_{\mathcal{G}_m}(R) \cup \text{de}_{\mathcal{G}_m}(L)\} = \emptyset$.

A special case of missing data DAGs where $\text{ch}_{\mathcal{G}_m}(L_i) = \emptyset$ for every L_i was considered by (Mohan et al., 2013).

Restriction (a) is imposed by definition. Since every L_k is a deterministic function of $L_k^{(1)}$ and R_k, only those two variables can serve as causes of L_k, and thus as parents of L_k in the graph. Restriction (b) is imposed to ensure missingness indicators R, which are the missing data analogues of treatment variables in causal inference, cannot influence counterfactual variables. This restriction is a consequence of the missing data version of consistency, which implies observed variables are caused by their corre-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

sponding counterfactuals, and not vice versa. For a detailed discussion on the implication of relaxing this assumption see Srinivasan et al. (2023).

The above restrictions imply that $\text{ch}_{\mathcal{G}_m}(L_k) \subseteq \{R_i \in R \mid i \neq k\}$ and $\text{ch}_{\mathcal{G}_m}(R_k) \subseteq \{L_k\} \cup \{R_i \in R \mid i \neq k\}$. That is, the values of the observed R_k, L_k can only influence the decisions R_i as to which of the other variables $L_i^{(1)}$ will be observed.

The statistical model of an m-DAG $\mathcal{G}_m$, denoted by $\mathcal{M}^{\mathcal{G}_m}$, consists of a set of joint distributions that factorize with respect to $\mathcal{G}_m$. In the simple case where the variables in U, W are not present, the joint distributions $p(l, r, l^{(1)})$ in $\mathcal{M}^{\mathcal{G}_m}$ factorize as follows

$$\prod_{v_i \in l \cup r \cup l^{(1)}} p(v_i \mid \text{pa}_{\mathcal{G}_m}(v_i)) = \prod_{l_k \in l} p(l_k \mid l_k^{(1)}, r_k) \times \prod_{v_i \in r \cup l^{(1)}} p(v_i \mid \text{pa}_{\mathcal{G}_m}(v_i)). \quad (4.7)$$

The terms $p(l_k \mid l_k^{(1)}, r_k)$ are deterministically defined: $p(L_k = l_k \mid L_k^{(1)} = l_k^{(1)}, R_k = 1) = 1$ and $p(L_k = \text{“?”} \mid L_k^{(1)} = l_k^{(1)}, R_k = 0) = p(L_k = \text{“?”} \mid R_k = 0) = 1$ for any $L_k \in L$.

Generally, the joint distribution $p(l, r, l^{(1)})$ is assumed to satisfy *positivity*. This assumption is stated as follows: $p(R_k = r_k \mid \text{pa}_{\mathcal{G}_m}(r_k)) > 0$ with probability 1 for all $R_k \in R$. However, some popular missing data models, such as models assuming *monotone missingness* assume only a subset of possible missingness patterns have support. Some such models may be representible as DAGs, though they satisfy a weaker positivity assumption

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

that only applies to a subset of possible patterns. In particular, the monotone MAR model we discuss later is one example of a restricted pattern missing data model representible as a DAG.

We can view an m-DAG as a special case of a causal DAG (with hidden variables) described in the previous section where Y is taken to be L, and $A^{\dagger}$ is taken to be R, with a set of additional restrictions. Specifically, every variable L_k that is potentially missing corresponds to *one* non-trivial counterfactual $L_k^{(1)}$, with the other counterfactual $L_k^{(0)}$ trivially defined as "?". In addition, every treatment variable R_k in R can only affect exactly one outcome variable, namely L_k.

Given a missing data DAG model, our objective is to determine whether the target law $p(l^{(1)})$, or a fixed function of the target law, can be identified as a function of the observed data law $p\,(r, l)$, and if so, find the identifying functional. To aid subsequent developments we will reformulate the identification problem in missing data models using the language common in causal inference. In this view, given an underlying variable $L_k^{(1)}$, the corresponding proxy variable L_k is viewed as an observed version of "the outcome," and the corresponding missing indicator R_k as the observed version of "the treatment."

Since m-DAGs are a special case of causal DAGs, the following result

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

immediately follows.

Proposition 1. *Under the missingness model of an m-DAG* $\mathcal{G}_m$

$$p(l^{(1)}) = \prod_{v_k \in l^{(1)}} p(v_k \mid \text{pa}_{\mathcal{G}_m}(v_k)) = \left. \frac{p(l,r)}{\prod_{r_k \in r} p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))} \right|_{r=1}. \tag{4.8}$$

The expression $f(.)|_{r=1}$ is taken to mean that $f(.)$ is evaluated at $r = 1$. The second equality in (4.8) is the missing data DAG equivalent of the counterfactual g-formula in (3.3) for causal inference problems.

In missing data DAG models, any counterfactual variable $L_k^{(1)}$ is allowed to have elements of R as children. This means that the g-formula in (4.8) does not necessarily lead to identification of $p(l^{(1)})$ in terms of the observed data distribution $p(l, r)$. This is because $\text{pa}_{\mathcal{G}_m}(r_k)$ in $p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))$ may involve values in $l^{(1)}$ which are not always observed and cannot be immediately dropped via independence assumptions. This is analogous to why unmeasured variables lead to non-identification in causal inference problems, as discussed earlier. In the next section, we illustrate via a number of examples how identification may nevertheless be accomplished in some missing data DAG models representing MNAR mechanisms.

4.1 Hierarchy of missing data DAG models

Similar to Rubin's hierarchy of missingness mechanisms, it is possible to set up a hierarchy for missing data DAG models that define the complexity

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

of identification techniques required. The missing data DAG model for a graph $\mathcal{G}_m$ with vertices $\{L^{(1)}, R, L\}$ is said to be

- Missing Completely At Random (MCAR) if $\prod_{r_k \in r} p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))$ is not a function of variables in $L^{(1)}$ and L. Graphically speaking there are no edges that point to variables in R.
- Missing At Random (MAR) if $\prod_{r_k \in r} p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))$ is not a function of variables in $L^{(1)}$. Graphically speaking there are no edges from variables in $L^{(1)}$ to variables in R.
- Missing Not At Random (MNAR) otherwise.

Note that a graphical MAR model is a submodel of the Rubin's MAR model, with the added restriction that the full law factorizes with respect to an m-DAG. Rubin's definition of a MAR mechanism requires that, for any given missingness pattern, the missingness is independent of the missing values given the observed values. These restrictions pertain to missingness patterns $R = r$, with up to 2^K distinct patterns for models with K missing variables. This missingness mechanism cannot be represented graphically, and several authors have noted the difficulty in interpreting Rubin's MAR model in practice (Gill et al., 1997; Robins and Gill, 1997; Schafer and Graham, 2002; McKnight et al., 2007; Graham, 2012; Tian, 2015). The

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

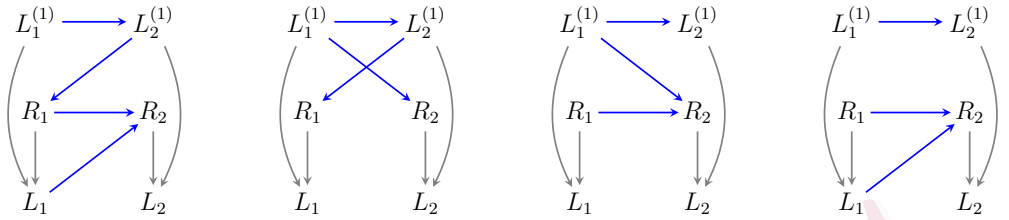
graphical hierarchy described above provides a more intuitive framework for understanding missingness mechanisms. In particular, any missing data model associated with an m-DAG inherently describes a data generating process where variables are generated sequentially according to a total order consistent with the m-DAG.

Further, while there exist MNAR models whose restrictions cannot be represented graphically, for instance the complete-case missing value model (Little, 1993), or the discrete choice model (Tchetgen et al., 2018), the restrictions posed in several popular MNAR models, such as the permutation model (Robins, 1997), the block-sequential MAR model (Zhou et al., 2010), and those in Martel García (2013); Mohan et al. (2013); Shpitser (2016); Saadati and Tian (2019); Bhattacharya et al. (2019) and Nabi et al. (2020) correspond to DAGs. Models described in (Shpitser, 2016; Sadinle and Reiter, 2017; Malinsky et al., 2022) correspond to graphical models that generalize DAG models, and are instead associated with chain graphs (Lauritzen, 1996).

4.2 Examples of missing data DAG models

We now present examples from prior literature of missing data models in which imposed restrictions can be encoded via m-DAGs.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382



(a) Permutation (b) Block-parallel (c) Block-sequential (d) MAR example

Figure 2: Examples of missing data DAG models.

Permutation model. Given an ordering (permutation) indexed by $k \in \{1, \ldots, K\}$ on variables in $L^{(1)}$, Robins (1997) defined a *permutation missingness* model with the restrictions that for any $R_k \in R$, R_k is independent of the current and past counterfactuals given the observed past and future counterfactual variables. We denote this model by $\mathcal{M}^{\text{per}}$. Formally, $\mathcal{M}^{\text{per}}$ is defined by the following conditional independence restrictions: $R_k \perp\!\!\!\perp \overline{L}_k^{(1)} \mid \overline{R}_{k-1}, \overline{L}_{k-1}, \underline{L}_{k+1}^{(1)}$ for all $k \in \{1, \ldots, K\}$. The graphical representation of $\mathcal{M}^{\text{per}}$ for two time points is shown in Fig. 2(a). The local Markov property for this m-DAG model yields the following set of independence restrictions: $R_1 \perp\!\!\!\perp L_1^{(1)} \mid L_2^{(1)}$ and $R_2 \perp\!\!\!\perp L_2^{(1)}, L_1^{(1)} \mid L_1, R_1$.

Block-parallel MNAR model. The block-parallel MNAR model, denoted by $\mathcal{M}^{\text{b-par}}$, was introduced by Mohan et al. (2013). It is defined by the following conditional independence restrictions: $R_k \perp\!\!\!\perp \{L_k^{(1)}, R_{-k}\} \mid L_{-k}^{(1)}$, for all $k \in \{1, \ldots, K\}$, where $V_{-k} \coloneqq V \setminus V_k$. The rationale for the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

name "block-parallel" will become clear in Section 5.2, where we discuss identification in this missing data model. The graphical representation of $\mathcal{M}^{\text{b-par}}$ for two time points is shown in Fig. 2(b). The local Markov property for this m-DAG model implies: $R_1 \perp\!\!\!\perp L_1^{(1)}, R_2 \mid L_2^{(1)}$ and $R_2 \perp\!\!\!\perp L_2^{(1)}, R_1 \mid L_1^{(1)}$.

Block-sequential MNAR model. The block-sequential MNAR model, denoted by $\mathcal{M}^{\text{b-seq}}$, was introduced by Zhou et al. (2010), under the name "block-conditional MAR." We use our alternative name to make it clear that this model is not MAR but MNAR. It is defined by the following conditional independence restrictions: $R_{k+1} \perp\!\!\!\perp \underline{L}_{k+1}^{(1)} \mid \overline{R}_k, \overline{L}_k^{(1)}$, $\forall k \in \{1, \ldots, K\}$. The graphical representation of $\mathcal{M}^{\text{b-seq}}$ for two time points is shown in Fig. 2(c). The local Markov property for this m-DAG model implies: $R_1 \perp\!\!\!\perp L_1^{(1)}, L_2^{(1)}$ and $R_2 \perp\!\!\!\perp L_2^{(1)} \mid R_1, L_1^{(1)}$.

An example of a MAR model. A MAR model can be defined via the following conditional independence restrictions (Rubin, 1976; Seaman et al., 2013): $\mathbb{I}(R = r) \perp\!\!\!\perp \left\{L_i^{(1)} \in L^{(1)} : r_i = 0, \forall r_i \in r\right\} \mid \left\{L_j^{(1)} \in L^{(1)} : r_j = 1, \forall r_j \in r\right\}$ for all $r \in \{0,1\}^K$, where $\mathbb{I}(.)$ is the indicator function. This MAR model cannot be represented by a DAG $\mathcal{G}_m$ with vertices $V = \{L^{(1)}, R, L\}$. This follows because the 2^K variables $\mathbb{I}(R = r)$ are not vertices on the graph. However, we can have more intuitive submodels of the MAR

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

model that can be represented graphically. The missing data DAG for two time points in Fig. 2(d) is one example. The local Markov property for this m-DAG model implies: $R_1 \perp\!\!\!\perp L_1^{(1)}, L_2^{(1)}$ and $R_2 \perp\!\!\!\perp L_2^{(1)} \mid R_1, L_1$.

An interesting observation is that under the monotonicity assumption, the MAR model in Fig. 2(d) and the block-sequential MNAR model in Fig. 2(c) are identical. The monotonicity assumption in the block-sequential MNAR model imposes restrictions on the univariate conditionals of each R_k given their parents on the graph: evaluating any of the parental missingness indicators at zero deterministically defines the conditional probability distribution of $p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))$. This is because if $i \prec_{\mathcal{G}_m} k$, and $R_i = 0$ then it must be the case that $R_k = 0$, otherwise the monotonicity assumption is violated. The only non-deterministic evaluation of the univariate conditionals occurs when $R_i = 1, \forall R_i \in \text{pa}_{\mathcal{G}_m}(R_k)$. Thus, by consistency, we can replace the pair $(R_i = 1, L_i^{(1)}) \in \text{pa}_{\mathcal{G}_m}(R_k)$ with $(R_i = 1, L_i)$. The equivalent graphical operation is replacing the set of edges $\{L_i^{(1)} \rightarrow R_k \leftarrow R_i\}$ with the set $\{L_i \rightarrow R_k \leftarrow R_i\}$. This renders the DAG in Fig. 2(c) and the one in Fig. 2(d) equivalent. The monotonicity assumption is a restriction on patterns, and cannot be directly represented in an m-DAG, unless it is augmented with additional edge markings denoting deterministic relationships among R variables that define monotonicity.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

We now compare the above missingness models on the basis of how they differ in telling a story about the underlying missingness mechanisms.

Fig. 2(a) is a two-variable version of the permutation model described in (Robins, 1997). Suppose a reform school offered HIV testing, with 35% of those tested being HIV positive among the 30% who accepted the offer, i.e., $p(L_1 = 1 \mid R_1 = 1) = 0.35$. Due to concerns about non-random non-response, data on HIV risks and fears were abstracted as variable L_2 from a stratified sample of counseling files, with sampling fractions of 20%, in the stratum $L_1 =$ "?", 50% in the stratum $L_1 = 1$, and 60% in the stratum $L_1 = 0$. By design, $p(R_2 = 1 \mid L_1^{(1)}, L_2^{(1)}, R_1)$ is only a function of L_1 (and thus also of R_1). Under the additional assumption that being tested (R_1) is independent of HIV status ($L_1^{(1)}$) given $L_2^{(1)}$, the data follow a non-monotone non-ignorable permutation model process where $(L_1^{(1)}, L_2^{(1)})$ are true HIV status, and the abstracted risks and fears of HIV infection, respectively.

Fig. 2(b) is a two variable version of the model first described in Mohan and Pearl (2014). In this model, the missingness of each of the two variables is not determined by the underlying value of the variable itself, but by the underlying value of the other variable. An example of such censoring processes occur in prisoner's dilemma situations, where L_1 represent recorded

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

parts of a criminal case against defendant 1 obtained from testimony, while $L_1^{(1)}$ represent the true facts about defendant 1. Many jurisdictions (such as the United States) forbid prosecutors from coercing self-incrimination, however co-conspirators are often induced to incriminate each other under plea deals. In such cases, observability of L_1, represented by R_1 may be causally influenced by the severity of the crime $L_2^{(1)}$ of a co-conspirator 2 of defendant 1, and similarly the observability status R_2 of L_2 would be influenced by $L_1^{(1)}$. Laws against self-incrimination prevent $L_1^{(1)}$ from influencing R_1, and $L_2^{(1)}$ from influencing R_2.

Fig. 2(d) is the classic monotone MAR model, which has been used to represent dropout in longitudinal observational studies. For instance, consider a study of the effect of highly active antiretroviral therapy on the CD4 outcome in HIV patients (Cain et al., 2016). In such a study, a set of covariates $L_i^{(1)}$ are measured at every followup period i for every patient that remains in the study. However, once a patient withdraws from the study starting at a particular time period j, for instance due to withholding informed consent, no measurements of $L_k^{(1)}$ for $k \geq j$ take place. Whether a patient withdraws from the study depends on covariates for the patient which have been measured so far. Fig. 2(d) represents a very simple such scenario for two time points, where the dropout indicator R_2 depends on

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

observed covariates L_1 measured at the previous time point. In addition, R_2 depends on the dropout indicator R_1 to enforce the monotonicity restriction, which states that $p(R_2 = 0 \mid R_1 = 0, L_1 = \text{“?”}) = 1$.

Fig. 2(c) represents a similar type of longitudinal study where missingness is instead *intermittent*. A set of covariates $L_i^{(1)}$, including determinants of health and socioeconomic factors, are surveyed at every followup period i for every study participant, provided that they show up for their appointment. However, not every participant shows up to every appointment, often for reasons caused by the covariates that were not measured, such as economic precarity. In the simple two timepoint version of such a study shown in Fig. 2(c), the censoring indicator R_2 depends on the indicator R_1 at the previous time point, as well as covariates $L_1^{(1)}$ at that time point, which may be observed for some participants, and not observed for others.

Note that the permutation model in Fig. 2(a) is untestable (i.e., though it implies restrictions on the full data law $p(l^{(1)}, r, l)$, it does not impose any restrictions on the observed data law $p(r, l)$) (Robins, 1997). However, the MAR model in Fig. 2(d) has a testable implication on the observed data distribution which can be used for falsification; see Nabi and Bhattacharya (2023). The block-parallel model in Fig. 2(b) also has a testable implication on the observed data distribution and hence can be falsified; see Malinsky

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

et al. (2022) and Nabi and Bhattacharya (2023).

In the next section, we discuss how identification strategies developed in causal inference may be applied to missing data problems, and how additional structures, unique to missing data, yields new identification strategies which are never encountered in causal inference problems.

5. Identification in Missing Data DAG Models

In missing data models associated with m-DAGs presented so far, the distribution $p(l^{(1)})$ can instead be viewed as $p(l \parallel r = 1)$—the distribution where all missingness indicators in R are intervened on and set to 1. Note that we have the following identities: $p(l^{(1)}) = p(l \parallel r = 1) = p(l^{(1)} \parallel r = 1)$.

Recall from Proposition (1) that identification of the target law $p(l^{(1)}) \coloneqq p(l \parallel r = 1)$ is equivalent to identification of the missingness mechanism evaluated at $R = 1$, i.e., the probability of observing the complete-case missing data pattern: $\prod_{R_k \in R} p(R_k = 1 \mid \text{pa}_{\mathcal{G}_m}(r_k))|_{R=1}$. If each conditional factor $p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))$ (evaluated at $R = 1$) is identified in this product, then the complete-case missing data pattern, and consequently the target law would be identified. We refer to the conditional factor $p(R_k = 1 \mid \text{pa}_{\mathcal{G}_m}(r_k))$ as the *propensity score* for R_k. Via a series of examples, we explore different identification strategies for identifying the distribution of the missingness

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

mechanism $p(r \mid \text{pa}_{\mathcal{G}_m}(r))$ evaluated at the complete-case pattern values: $R = 1$.

A Single Variable Interventional Reformulation of the Counterfactual G-formula. Given a joint distribution $p(l^{(1)}, r, l)$ that factorizes according to an m-DAG $\mathcal{G}_m$, the joint distribution after an intervention on $R_k \in R$ is equal to the truncated factorization where the joint distribution is divided by the identified propensity score of R_k. Letting r_k be a in (3.4), and $l^{(1)}_{-k} \cup r_{-k} \cup l$ be l in (3.4), we can write this distribution as

$$p(l^{(1)}_{-k}, r_{-k}, l \parallel r_k = 1) = \frac{p(l^{(1)}_{-k}, r, l)}{p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))}\bigg|_{R_k=1}. \tag{5.9}$$

When we intervene on R_k and set it to 1, it becomes redundant to include both $L^{(1)}_k$ and L_k in the joint distribution as they represent the same random variable when $R_k = 1$; hence we drop $L^{(1)}_k$ from both sides of the equation.

The propensity score of R_k is identified if we can replace each $L^{(1)}_j \in \text{pa}_{\mathcal{G}_m}(R_k)$ with $\{L_j, R_j = 1\}$. Such replacements are sometimes justified due to conditional independence restrictions in the full data distribution. For instance in Fig. 2(b), the propensity of R_1 is identified because we have $p(r_1 \mid l^{(1)}_2) = p(r_1 \mid l^{(1)}_2, R_2 = 1) = p(r_1 \mid l_2, R_2 = 1)$, since $R_1 \perp\!\!\!\perp R_2 \mid L^{(1)}_2$. When missingness indicators are connected, we may lose some of these convenient independence constraints. For instance in Fig. 2(a), R_1 and R_2 are no longer conditionally independent; thus identification of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

$p(r_1 \mid l_2^{(1)})$ is not as straightforward as it is in Fig. 2(b). However, due to an *invariance* property of the propensity scores, we can sometimes succeed in identification by exploring interventional distributions where a subset of observed variables are intervened on and consequently certain edges are removed. We formalize this property in the following lemma, which is analogous to the invariance property in causal inference.

Lemma 1 (Invariance property)**.** *Given the propensity score for $R_k \in R$, the conditioning set $\mathrm{pa}_{\mathcal{G}_m}(R_k)$ captures the direct causes of R_k and hence remains invariant under any set of interventions that disrupts other parts of the joint factorization. Given $R^* \subseteq R \setminus R_k$, we have*

$$p\left(r_k \mid \mathrm{pa}_{\mathcal{G}_m}(r_k)\right) = p\left(r_k \mid \mathrm{pa}_{\mathcal{G}_m}(r_k) \parallel r^* = 1\right). \tag{5.10}$$

Using this property, we now explore various strategies for target law identification in MNAR models. In Section 5.1, we use the permutation model as an example to illustrate how missingness mechanisms can sometimes be identified via a *sequence* of interventions on missingness indicators. This is a generalization of causal inference techniques used in longitudinal studies to sequentially identify the effect of multiple treatments, where missingness indicator interventions are identified by treating counterfactuals as confounders. An intervention in missing data, unlike interventions in causal inference, can sometimes induce *selection bias* due to conditioning

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

on a subset of missingness indicators that have not yet been intervened on. Introduction of selection during intervention operations may make identification by means of sequential applications of the truncated factorization impossible. In Section 5.2, we illustrate the selection issue of sequential interventions on the block-parallel model, and show how *parallel* (simultaneous) interventions can obtain identification even if selection is present when using sequential approaches. In some missing data models, such as the block-sequential model, we can identify the missingness probability of the complete cases pattern via either sequential or parallel application of interventions. In Sections 5.3 through 5.5, we go over examples where a combination of sequential and parallel interventions are needed to identify each propensity score. In Section 5.6, we unify ideas explored in this section to yield a general identification procedure for arbitrary missing data DAGs by exploring all possible *partial orders* of interventions defined on the observed variables on the graph. We also discuss some graphical structures in missing data DAG models that impede nonparametric identification of the model.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

5.1 Sequential Interventions

Consider the permutation model $\mathcal{M}^{\text{per}}$ with two time points, redrawn in Fig. 3(a). Our objective is to identify $p(l_1^{(1)}, l_2^{(1)}) := p(l_1, l_2 \parallel r_1 = 1, r_2 = 1)$ as a function of observed data. This can be done in two steps: we first intervene on R_2, then we intervene on R_1.

Step 1. Intervene on R_2 to get $p(. \parallel r_2 = 1)$. The propensity score of R_2, $p(r_2 \mid r_1, l_1)$, is a function of observed data and so the corresponding post intervention distribution is immediately identified. Intervening on R_2 and setting it to 1 yields the following kernel:

$$p(l_1^{(1)}, l_2, r_1, l_1 \parallel r_2 = 1) = \frac{p(l_1^{(1)}, l_1, l_2, r_1, r_2 = 1)}{p(r_2 = 1 \mid r_1, l_1)}. \tag{5.11}$$

By consistency in missing data, $p(l_2 \parallel r_2 = 1)$, $p(l_2^{(1)} \parallel r_2 = 1)$, and $p(l_2^{(1)}, l_2 \parallel r_2 = 1)$ all represent the same object. This kernel factorizes with respect to the (conditional) m-DAG shown in Fig. 3(b), obtained from Fig. 3(a) by removing edges $R_1 \to R_2$ and $L_1 \to R_2$, and denoting the vertex R_2 as a square showing $r_2 = 1$, indicating the intervention setting R_2 to 1.

In general, the graphical analogue of the intervention on R_k entails removing all edges with arrowheads into R_k in the corresponding missing data graph. We denote vertices corresponding to variables that have been intervened on with rectangles. Further, the pair $(L_k^{(1)}, L_k)$ is treated as a

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

single variable on the graph after intervening and setting $R_k = 1$ due to the deterministic relation of L_k with $L_k^{(1)}$ and R_k. We keep the proxy variable on the graph but in gray with dashed edges.

Step 2. Intervene on R_1 after intervening on R_2 to get $p(. \parallel r_1 = 1, r_2 = 1)$. In the second step, we want to intervene on R_1 in the kernel $p(. \parallel r_2 = 1)$ which is Markov relative to the conditional DAG in Fig. 3(b). In order to perform this intervention, we need to show that $p(r_1 \mid l_2^{(1)} \parallel r_2 = 1)$ is a function of $p(r, l)$. Using consistency, we have $p(r_1 \mid l_2^{(1)} \parallel r_2 = 1) = p(r_1 \mid l_2 \parallel r_2 = 1)$, and using kernel probability rules (provided in 3.5), we have $p(r_1 \mid l_2 \parallel r_2 = 1) = p(l_2, r_1 \parallel r_2 = 1)/\sum_{r_1} p(l_2, r_1 \parallel r_2 = 1)$. The numerator here is simply a marginal of the kernel $p(. \parallel r_2 = 1)$ in (5.11) and is identified as follows:

$$p(l_2, r_1 \parallel r_2 = 1) = \sum_{l_1, l_1^{(1)}} p(l_1^{(1)}, l_2, r_1, l_1 \parallel r_2 = 1) = \sum_{l_1, l_1^{(1)}} \frac{p(l_1^{(1)}, l_1, l_2, r_1, r_2 = 1)}{p(r_2 = 1 \mid r_1, l_1)}$$
$$= \sum_{l_1} \frac{p(l_1, l_2, r_1, r_2 = 1)}{p(r_2 = 1 \mid r_1, l_1)} = \sum_{l_1} p(r_1, l_1) \times p(l_2 \mid r_1, l_1, r_2 = 1).$$

Consequently, we are able to identify the propensity score of R_1 in the second step and proceed with the intervention on R_1, using the kernel in (5.11). This yields a new kernel that factorizes with respect to the conditional DAG in Fig. 3(c) and corresponds to the target law $p(l_1^{(1)}, l_2^{(1)})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

Putting all the pieces together, the target law is identified as:

$$
\begin{aligned}
p(l_1, l_2 \parallel r = 1) &= \frac{p(l_1^{(1)}, l_2, r_1 = 1, l_1 \parallel r_2 = 1)}{p(r_1 = 1 \mid l_2 \parallel r_2 = 1)} \qquad (5.12)\\
&= \frac{p(l_1, l_2, r_1 = 1, r_2 = 1)}{p(r_2 = 1 \mid r_1 = 1, l_1) \times p(r_1 = 1 \mid l_2 \parallel r_2 = 1)} \\
&= \frac{p(l_1, l_2, r_1 = 1, r_2 = 1)}{p(r_2 = 1 \mid r_1 = 1, l_1) \times \frac{\sum_{l_1} p(l_2 \mid r_2 = 1, r_1 = 1, l_1) \times p(r_1 = 1, l_1)}{\sum_{l_1, r_1} p(l_2 \mid r_2 = 1, r_1, l_1) \times p(r_1, l_1)}}.
\end{aligned}
$$

The identified forms of the two propensity scores in this example are quite different. The propensity score of R_2, $p(r_2 = 1 \mid r_1, l_1)$, corresponds to the conditional factor in the full factorization of the joint. The propensity score of R_1, $p(r_1 = 1 \mid l_2 \parallel r_2 = 1)$, is a complex function of $p(r, l)$ and corresponds to a conditional distribution in a hypothetical world where $L_2^{(1)}$ is rendered observed and equal to L_2 via an intervention that sets R_2 to 1.

The fact that R_1 had a counterfactual parent $L_2^{(1)}$ in Fig. 3(a), but $R_1 \not\perp\!\!\!\perp R_2 \mid \mathrm{pa}_{\mathcal{G}_m}(R_1)$ ensured that the propensity score $p(r_1 \mid \mathrm{pa}_{\mathcal{G}_m}(r_1))$ required to intervene on R_1 is not immediately identifiable. This induced a strict *total ordering* by which R_1 and R_2 had to be intervened on. A total order is a special case of a *strict partial order* which is defined as an irreflexible, anti-symmetric, and transitive binary relation. In our simple example above, to get identification for the target law we needed to follow the order $\{I_{r_2} < I_{r_1}\}$, where I_{r_k} denotes an intervention on R_k. This sequential procedure generalizes to an arbitrary number of time points in

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

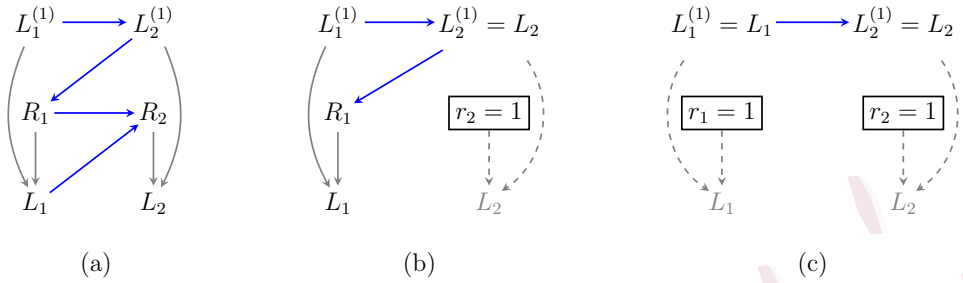


Figure 3: An illustration of the operation of a **sequential** identification algorithm. (a) Permutation model; (b) Intervention on R_2; (c) Intervention on R_1 after R_2.

the permutation model, where investigators take repeated measurements of some outcome over K time points. The total order for identifying the target law $p(l_1^{(1)}, \ldots, l_K^{(1)})$ is then given by a reverse topological ordering on the missingness indicators $\{I_{r_K} < \cdots < I_{r_1}\}$, i.e., we begin by intervening on the missingness indicator R_K corresponding to the final time point and proceed backwards (Robins, 1997).

Recall that in the causal model discussed earlier corresponding to Fig. 1, identification was not possible due to the presence of U_1 and U_2 instead of the counterfactuals $L_1^{(1)}$ and $L_2^{(2)}$. Clearly the key for obtaining identification was the additional information provided by counterfactuals $L^{(1)}$, which are sometimes observed, rather than U variables, which are never observed.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

In analogy with missing data, one may consider placing counterfactual versions of observed variables on the graph in causal inference (instead of, or in addition to, the hidden variables), with the hope that a pair of counterfactuals $L^{(1)}, L^{(0)}$ corresponding to L (for a binary treatment A) may lead to novel identification formulas. This is because the counterfactual pair $L^{(1)}, L^{(0)}$ is partially observed, due to the consistency property, unlike a purely hidden variable U. Unfortunately, unlike in the case of missing data, including counterfactuals on the graph does not lead to new identification formulas. In Section 6, we explain why the recent novel identification strategies in missing data models do not succeed for causal inference models.

5.2 Parallel Interventions

The sequential identification strategy in $\mathcal{M}^{\text{per}}$ coupled with consistency (i.e., $L_k^{(1)} = L_k$ if $R_k = 1$) resembles identification strategies employed in causal inference problems where the intervention distribution associated with a multivariate treatment is obtained sequentially by intervening on treatment variables one at a time. We now discuss an example that shows sequential strategies as in causal inference are insufficient and that there exists a class of missing data models where missingness indicators must be

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

intervened on "in parallel" to identify the target law.

The simplest example of a model in this class is the block-parallel MNAR model $\mathcal{M}^{\text{b-par}}$ with two time points, redrawn in Fig. 4(a). Identification of $p(l_1^{(1)}, l_2^{(1)}) := p(l_1, l_2 \parallel r_1 = 1, r_2 = 1)$ can be obtained in a single step as follows:

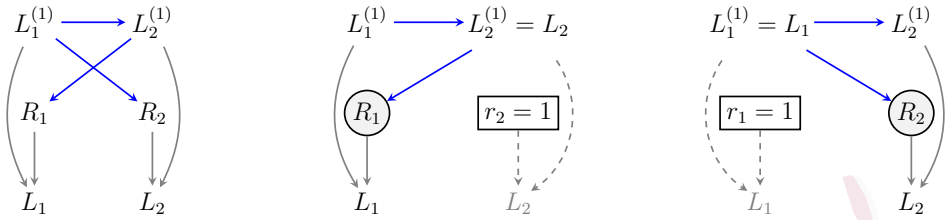
$$p(l_1^{(1)}, l_2^{(1)}) = \frac{p(l_1, l_2, r_1, r_2)}{p(r_1 \mid l_2^{(1)}) \times p(r_2 \mid l_1^{(1)})}\bigg|_{r=1} = \frac{p(l_1, l_2, r_1, r_2)}{p(r_1 \,|\, l_2, r_2) \times p(r_2 \,|\, l_1, r_1)}\bigg|_{r=1}. \tag{5.13}$$

The first equality holds by (4.8) and the second equality holds due to the independence restrictions implied by $\mathcal{M}^{\text{b-par}}$ which are: $R_1 \perp\!\!\!\perp R_2 \mid L_2^{(1)}$ and $R_1 \perp\!\!\!\perp R_2 \mid L_1^{(1)}$ and consistency. Given these independencies, both propensity scores in the denominator are easily identified as functions of the observed data.

We now illustrate why the sequential approach to identification fails in this example and that the simultaneous evaluation of the two propensity scores is necessary to yield target law identification. Assume we proceed with our two-step sequential procedure by first intervening on $R_2 = 1$. This results in the following kernel Markov relative to the conditional DAG in Fig. 4(b):

$$p(l_1^{(1)}, l_2, r_1, l_1 \parallel r_2 = 1) = \frac{p(l_1^{(1)}, l_1, l_2, r_1, r_2)}{p(r_2 \mid l_1^{(1)})}\bigg|_{r_2=1} = \frac{p(l_1^{(1)}, l_1, l_2, r_1, r_2 = 1)}{p(r_2 = 1 \mid l_1^{(1)}, r_1)}. \tag{5.14}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382



(a) (b) (c)

Figure 4: An illustration of the operation of a **parallel** identification algorithm. (a) Block-parallel; (b) Intervention on R_2 and selection on R_1; (c) Intervention on R_1 and selection on R_2.

The intervention on R_1 in the second step of the sequential procedure requires dividing the above kernel $p(. \parallel r_2 = 1)$ by the propensity score of R_1, $p(r_1 \mid l_2^{(1)}) = p(r_1 \mid l_2 \parallel r_2 = 1)$. Using Bayes rule, this propensity can be obtained from $p(r_1, l_2 \parallel r_2 = 1)$. However, we can only identify $p(r_1 = 1, l_2 \parallel r_2 = 1)$ from $p(. \parallel r_2 = 1)$. This is because $l_1^{(1)}$ appears in both the numerator and denominator, and when $R_1 = 0$, we cannot observe $l_1^{(1)}$. Hence, the kernel in (5.14) is identified only when evaluated at $R_1 = 1$. The fact that this kernel is not available at all levels of R_1 prevents us from sequentially obtaining $p(r_1 \mid \text{pa}_{\mathcal{G}_m}(r_1)) = p(r_1 \mid l_2 \parallel r_2 = 1) = p(l_2, r_1 \parallel r_2 = 1)/p(l_2 \parallel r_2 = 1)$ (where the first equality follows from invariance and the second from rules of kernel probability) due to our inability to sum out R_1 from $p(l_2, r_1 \parallel r_2 = 1)$ to obtain $p(l_2 \parallel r_2 = 1)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

The above is an example of what we term *selection* bias. In this case, division by $p(r_2 \mid l_1^{(1)})$ in (5.14) induces *selection* on R_1, a variable that is not yet intervened on, in the kernel $p(. \parallel r_2 = 1)$. We draw a gray circle around the vertices on the graph that are selected on (or conditioned on) to distinguish it from interventions. Attempting a sequential identification procedure starting with R_1 would similarly induce selection bias on R_2 in the kernel $p(. \parallel r_1 = 1)$, as shown in Fig. 4(c). Since it is not possible to obtain identification by performing the intervention operation in a sequence, no total ordering on missingness indicators can be imposed here. Instead, interventions on R_1 and R_2 are *incomparable*, and thus form a partial rather than a total ordering, which is simply denoted via $\{I_{r_1}, I_{r_2}\}$. Specifically, this notation denotes a partial order relationship on a set of two elements corresponding to interventions on indicators R_1 and R_2, where these elements are incompatible according to the order.

In some missingness models, we might be able to identify the target law in multiple ways. For instance, in the block-sequential MNAR model, the target law can be identified via either a sequence of interventions on missingness indicators in R, or via a one-step parallel intervention on all variables in R. These two strategies result in two different representations for the same identifying functional. These representations may, however,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

suggest different estimation strategies.

5.3 Sequential and Parallel Interventions

The previous two examples considered how identification of the target law in missing data problems entails evaluating the g-formula either sequentially or in parallel. These examples are special cases of a general identification algorithm for graphical missing data models, introduced in Shpitser et al. (2015). However, this algorithm is not complete in that it fails to account for the following example, and examples like it, where sequential and parallel applications of the g-formula must be combined in order to identify all of the propensity scores, and hence the target law. This strategy was introduced in Bhattacharya et al. (2019).

Consider the missing data model shown in Fig. 5(a) on three variables. The target law $p(l_1^{(1)}, l_2^{(1)}, l_3^{(1)}) := p(l_1, l_2, l_3 \parallel r_1 = 1, r_2 = 1, r_3 = 1)$ is equivalent to the following:

$$\left. \frac{p(l_1, l_2, l_3, r_1, r_2, r_3)}{p(r_1 \mid l_2^{(1)}, l_3^{(1)}) \times p(r_2 \mid l_3^{(1)}, r_1) \times p(r_3 \mid l_2^{(1)}, r_1)} \right|_{r=1}. \tag{5.15}$$

Neither of the approaches discussed in the previous two subsections would yield an identified missingness mechanism for the example of Fig. 5(a). First, the sequential application discussed in (5.1) fails since the selection induced on R_2 after intervening on R_3 impedes our next move in doing an

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

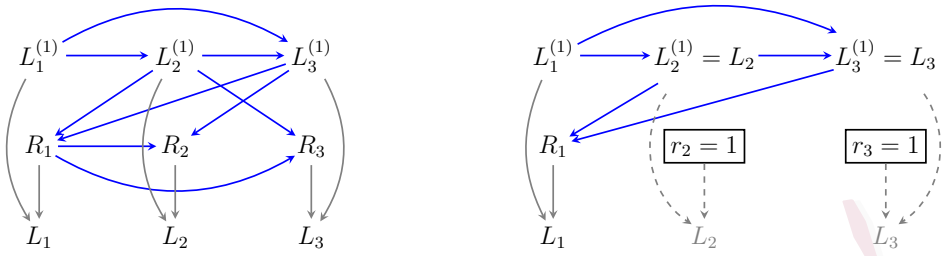
intervention on R_2. This is because obtaining $p(r_2 \mid \text{pa}_{\mathcal{G}_m}(r_2))$ from the kernel $p(. \parallel r_3 = 1)$ requires summing over R_2 (by Bayes rule), and that is not possible as R_2 is conditioned/forced to be one in the kernel. A similar problem persists if we intervene on R_2 first (R_3 is conditioned to be one in the kernel $p(. \parallel r_2 = 1)$). A sequential application starting with R_1 is not possible since we cannot immediately obtain $p(r_1 \mid \text{pa}_{\mathcal{G}_m}(r_1))$. This is because R_1 has counterfactual parents $L_2^{(1)}$ and $L_3^{(1)}$, but is not conditionally independent of the corresponding missingness indicators R_2 and R_3 given its parents. The parallel application discussed in (5.2) fails for the same reason, i.e., the propensity score for R_1 is not immediately identified.

We now show identification of this target distribution is still possible, but only by evaluating the g-formula in two sequential steps, with the first step consisting of a parallel evaluation of propensity scores for R_2 and R_3, and the second step consisting of the evaluation of propensity score for R_1 using the kernel distribution obtained from the first step.

Step 1. *Intervene on R_2 and R_3 in parallel to get $p(. \parallel r_2 = 1, r_3 = 1)$.* The factorization of join distribution in Fig. 5(a) implies $R_2 \perp\!\!\!\perp R_3 \mid L_3^{(1)}, R_1$ and $R_3 \perp\!\!\!\perp R_2 \mid L_2^{(1)}, R_1$. These independencies directly yield identified propensity scores for R_2 and R_3 as follows:

$$p(r_2 \mid \text{pa}_{\mathcal{G}_m}(r_2))|_{r=1} = p(r_2 \mid l_3^{(1)}, r_1)|_{r=1} = p(r_2 = 1 \mid l_3, r_1 = r_3 = 1), \quad (5.16)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382



(a) (b)

Figure 5: (a) An example m-DAG corresponding to a model where interventions must be applied both *sequentially and in parallel* to yield identification; (b) Graph derived from (a) representing the intermediate step of the identification algorithm where R_2 and R_3 are simultaneously set to 1.

$$p(r_3 \mid \mathrm{pa}_{\mathcal{G}_m}(r_3))|_{r=1} = p(r_3 \mid l_2^{(1)}, r_1)|_{r=1} = p(r_3 = 1 \mid l_2, r_1 = r_2 = 1). \quad (5.17)$$

This immediately implies that R_2 and R_3 can be intervened on in parallel, similar to the block-parallel example in the previous subsection. This results in the following kernel that factorizes with respect to the DAG in Fig. 5(b),

$$p(l_1^{(1)}, l_2, l_3, r_1, l_1 \parallel r_2 = r_3 = 1) = \frac{p(l_1^{(1)}, l_1, l_2, l_3, r_1, r_2 = 1, r_3 = 1)}{p(r_2 = 1 | l_3, r_1, r_3 = 1) \times p(r_3 = 1 | l_2, r_1, r_2 = 1)}.$$

The same strategy cannot be used for expressing $p(r_1 \mid \mathrm{pa}_{\mathcal{G}}(r_1)) = p(r_1 \mid l_2^{(1)}, l_3^{(1)})$ as a function of the factual distribution since $R_1 \not\perp\!\!\!\perp \{R_2, R_3\} \mid L_2^{(1)}, L_3^{(1)}$ under this model. However, a second, more involved, step leads to identifying the propensity score of R_1 and thus the target law $p(l_1^{(1)}, l_2^{(1)}, l_3^{(1)})$.

Step 2. *Intervene on R_1 after intervening on R_2, R_3 to get $p(. \parallel r_1 =$*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

$1, r_2 = 1, r_3 = 1)$. In the second step, we want to intervene on R_1 using the above kernel $p(. \parallel r_2 = 1, r_3 = 1)$ Markov relative to the DAG in Fig. 5(b). In order to perform this intervention, we need to show that the propensity of R_1, $p(r_1 \mid l_2^{(1)}, l_2^{(1)}) = p(r_r \mid l_2, l_3 \parallel r_2 = 1, r_3 = 1)$ is a function of $p(r, l)$. We have $p(r_1 \mid l_2, l_3 \parallel r_2 = r_3 = 1) = p(l_2, l_3, r_1 \| r_2 = r_3 = 1) / \sum_{r_1} p(l_2, l_3, r_1 \| r_2 = r_3 = 1)$. The numerator here is identified by simply marginalizing out $l_1^{(1)}$ from the kernel $p(. \| r_2 = r_3 = 1)$. Consequently, the propensity score of R_3 is rendered identified and we can identify the target law in (5.15) via:

$$p(l_1, l_2, l_3 \parallel r = 1) = \left. \frac{p(l_1^{(1)}, l_2, l_3, r_1, l_1 \parallel r_2 = 1, r_3 = 1)}{p(r_1 \mid l_2, l_3 \parallel r_2 = 1, r_3 = 1)} \right|_{r_1 = 1}.$$

In conclusion, the parallel interventions on R_2 and R_3 do not induce selection bias on R_1. Therefore, we are able to proceed with the sequential application of the g-formula and obtain $p(r_1 \mid \text{pa}_{\mathcal{G}_m}(r_1))$ from the kernel $p(. \parallel r_2 = 1, r_3 = 1)$. In other words, identification of the target law is only possible if R_1 was intervened on only after R_2 and R_3 were simultaneously intervened on. This identification procedure also induces a partial ordering for the interventions on the missingness indicators. The partial order of interventions in the above example can be written as $\{\{I_{r_2}, I_{r_3}\} < I_{r_1}\}$; that is interventions on R_2 and R_3 are incompatible, but both must occur prior to an intervention on R_1.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

5.4 Identifying Propensity Scores with Different Partial Orders

In the examples discussed so far, all missingness indicators were intervened on according to a *partial order* defined on the set $\{R_k \in R\}$. A procedure for target law identification may then aim at exploring the space of all possible partial orders. This effectively entails trying out all possible combinations of parallel and sequential application of the g-formula. Bhattacharya et al. (2019) showed that such a procedure remains incomplete, meaning that it fails to recognize a class of missing data models where the target law is indeed identified. They took an alternative view of the target law identification problem, where each propensity score $p(r_k \mid \text{pa}_{\mathcal{G}}(r_k))$ may be identified separately, using a potentially distinct partial order of intervention operations. This entails considering subproblems where selection bias, hidden variables, or both, are introduced, even if these complications are absent in the original problem.

Consider the missing data DAG in Fig. 6(a). According to Proposition 1, the target law $p(l_1^{(1)}, l_2^{(1)}, l_3^{(1)}) \coloneqq p(l_1, l_2, l_3 \parallel r_1 = 1, r_2 = 1, r_3 = 1)$ is equivalent to the following:

$$p(l_1, l_2, l_3 \parallel r = 1) = \frac{p(l_1, l_2, l_3, r_1, r_2, r_3)}{p(r_1|l_2^{(1)}) \times p(r_2|l_1^{(1)}, l_3^{(1)}, r_1) \times p(r_3|l_2^{(1)}, r_1)}\bigg|_{r=1}. \quad (5.18)$$

We can easily use the encoded independence restrictions, along with the consistency facts in missing data, to identify the propensity scores of R_2

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

and R_3 evaluated at $R = 1$. These restrictions are: $R_2 \perp\!\!\!\perp R_3 \mid R_1, L_1^{(1)}, L_3^{(1)}$ and $R_3 \perp\!\!\!\perp R_2 \mid R_1, L_2^{(1)}$. We have,

$$p(r_2 \mid \text{pa}_{\mathcal{G}_m}(r_2))|_{r=1} = p(r_2 = 1 \mid l_1, l_3, r_1 = 1, r_3 = 1), \tag{5.19}$$

$$p(r_3 \mid \text{pa}_{\mathcal{G}_m}(r_3))|_{r=1} = p(r_3 = 1 \mid l_2, r_1 = 1, r_2 = 1). \tag{5.20}$$

The propensity score of R_1 however, is not immediately identified since $R_1 \not\perp\!\!\!\perp R_2 \mid L_2^{(1)}$. A natural question to ask is whether this propensity score is identified from a kernel distribution where R_2 and R_3 are simultaneously intervened on. If so, then we can follow a similar argument as in the previous example with the two-step sequential procedure. Performing parallel interventions on R_2 and R_3 yield the following kernel distribution,

$$p(l_1^{(1)}, l_2, l_3, r_1 \parallel r_2 = r_3 = 1) = \frac{p(l_1^{(1)}, l_1, l_2, l_3, r_1, r_2 = r_3 = 1)}{p(r_2 = 1 | l_1^{(1)}, l_3, r_1, r_3 = 1) \times p(r_3 = 1 | l_2, r_1, r_2 = 1)}.$$

Following the second step of the sequential procedure, we need to first identify the propensity score of R_1 via the above kernel, i.e., identifying $p(r_1 \mid l_2 \parallel r_2 = 1, r_3 = 1)$ which by Bayes rule can be obtained from $p(r_1, l_2 \parallel r_2 = 1, r_3 = 1)$. Unfortunately, the above kernel $p(. \parallel r_2 = 1, r_3 = 1)$ exhibits selection on R_1. Consequently, the two-stage sequential strategy in the previous example fails to identify the target law in this current example. Nevertheless, the target law is still identifiable using a more involved argument outlined below. Since the propensity scores of R_2

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

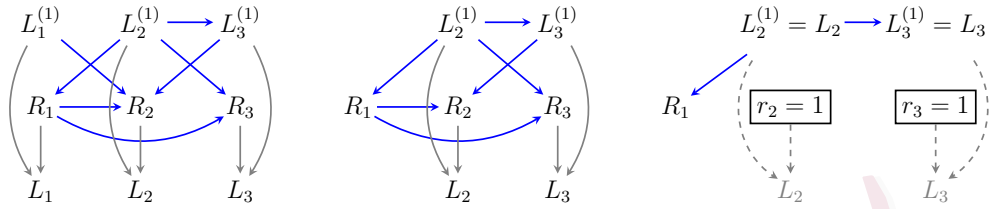


Figure 6: (a) An example m-DAG corresponding to a model where intervention on a single missingness indicator entails following a partial order of interventions; (b), (c) Selection bias on R_1 is avoidable by following a partial order intervention on a graph induced by projecting out $L_1^{(1)}, L_1$.

and R_3 are identified via (5.19) and (5.20), all we need to do to identify the target law is to find a way to identify the propensity score of R_1, $p(r_1 \mid l_2^{(1)})$.

Consider the marginal distribution $\sum_{l_1, l_1^{(1)}} p(l^{(1)}, r, l)$, where the full law factorizes according to the m-DAG in Fig. 6(a). The non-deterministic portion of this marginal distribution factorizes as: $p(l_2^{(1)}, l_3^{(1)}, r_1, r_2, r_3) = p(l_2^{(1)}) \times p(l_3^{(1)} | l_2^{(1)}) \times p(r_1 | l_2^{(1)}) \times p(r_2 | r_1, l_3^{(1)}) \times p(r_3 | l_2^{(1)}, r_1)$. The above factorization is Markov relative to the m-DAG in Fig. 6(b), where $L_1, L_1^{(1)}$ are projected out (or in other words, treated as hidden variables) from Fig. 6(a). Aside from absence of $p(l_1^{(1)})$ and $p(l_1 \mid r_1, l_1^{(1)})$ factors, the difference between the above factorization and the one from original m-DAG in Fig. 6(a) is the difference in propensity score of R_2, i.e., $p(r_2 \mid r_1, l_1^{(1)}, l_3^{(1)})$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

vs. $p(r_2 \mid r_1, l_3^{(1)})$. We now illustrate that this change in the propensity score of R_2, which we call a *pseudo-propensity score* for R_2, can overcome the selection bias issue on R_1 that resulted from an intervention on R_2.

The pseudo-propensity score of R_2, denoted by $\widetilde{p}(r_2 \mid r_1, l_3^{(1)})$, is easily identified via $\widetilde{p}(r_2 \mid r_1, l_3, r_3 = 1)$ since $R_2 \perp\!\!\!\perp R_3 \mid R_1, L_3^{(1)}$ in Fig. 6(b). Now to identify the propensity score of R_1, we proceed as follows. Using the pseudo-propensity score of R_2 and the propensity score of R_3, we obtain the following kernel distribution:

$$\widetilde{p}(l_2, l_3, r_1 \parallel r_2 = 1, r_3 = 1) = \frac{p(l_2, l_3, r_1, r_2 = 1, r_3 = 1)}{\widetilde{p}(r_2 = 1 \mid r_1, l_3, r_3 = 1) \times p(r_3 = 1 \mid l_2, r_1, r_2 = 1)}.$$

The above kernel $\widetilde{p}(. \parallel r_2 = 1, r_3 = 1)$, which factorizes according to the m-DAG in Fig. 4(c), is a direct function of observed data law without inducing any selection bias on R_1. Consequently, we can identify the propensity score of R_1, $p(r_1 \mid l_2 \parallel r_2 = 1, r_3 = 1)$ using the kernel $\widetilde{p}(. \parallel r_2 = 1, r_3 = 1)$. Using Bayes rule, we have $p(r_1 \mid l_2 \parallel r_2 = 1, r_3 = 1) = \sum_{l_3} p(r_1, l_2, l_3 \parallel r_2 = 1, r_3 = 1) / \sum_{l_3, r_1} p(r_1, l_2, l_3 \parallel r_2 = 1, r_3 = 1)$. This concludes that the target law in (5.18) is identified, since we showed each term in the denominator is identified.

The idea of focusing on identifying each propensity score separately leads to new identification strategies. Specifically, in the above example, the propensity score of R_1 was identified only after treating $L_1^{(1)}$ as a hid-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

den variable and marginalizing it out from the original distribution. Bhattacharya et al. (2019) developed a general procedure based on these observations which has significantly narrowed the identifiability gap in graphical models of missing data. Their procedure generalizes the notion of finding a partial order of interventions on the set $\{R_k \in R\}$ to finding partial orders of interventions for each individual $R_k \in R$ by exploring subproblems where a set of variables are treated as hidden or unmeasured.

In the above example, the partial orders for R_2 and R_3 are trivial—the corresponding propensity scores are immediately identified from the observed distribution. We can summarize the identification procedure for obtaining the propensity score of R_1 via the following partial order executed in a graph where $\{L_1^{(1)}, L_1\}$ are treated as hidden variables, which we will denote as $\mathcal{G}_m(V \setminus \{L_1^{(1)}, L_1\})$: the partial order of interventions for R_1 can be summarized via $\left\{\{I_{r_2}, I_{r_3}\} < I_{r_1}\right\}$ in $\mathcal{G}_m(V \setminus \{L_1^{(1)}, L_1\})$. That is, intervention on R_1 must occur after interventions on R_2 and R_3 in a graph where $L_1^{(1)}$ and L_1 are marginalized out, and as mentioned earlier interventions on R_2 and R_3 are incompatible.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

5.5 Intervention on Variables Outside of R

A feature common to all previous examples is that all propensity scores were obtained via partial orders of interventions defined only on the missingness indicators. This however, is also not always sufficient. In general the propensity score of R_k might be identified only by intervening on variables outside of R, including variables in $L^{(1)}$ that become observed after intervening or conditioning on relevant elements in R. As an example, consider the m-DAG in Fig 7(a) where L_3 is fully observed. This graph was considered in Bhattacharya et al. (2019) as a generalization of a model described in Shpitser et al. (2015). According to Proposition 1, we can write down the target law $p(l_1^{(1)}, l_2^{(1)}, l_4^{(1)}, l_3)$ as follows:

$$p(l_1, l_2, l_3, l_4 \parallel r = 1) = \left. \frac{p(l_1, l_2, l_3, l_4, r_1, r_2, r_4)}{p(r_1 | l_2^{(1)}, l_4^{(1)}, r_2) \times p(r_2 | l_1^{(1)}, l_4^{(1)}) \times p(r_4 | l_3)} \right|_{r=1}. \quad (5.21)$$

The propensity score of R_4, $p(r_4 \mid l_3)$, is a direct function of observed data. The propensity score of R_1, $p(r_1 \mid r_2, l_2^{(1)}, l_4^{(1)})$ evaluated at $R = 1$ is also a function of observed data via $p(r_1 = 1 \mid r_2 = 1, l_2, l_4, r_4 = 1)$ since $R_1 \perp\!\!\!\perp R_4 \mid R_2, L_2^{(1)}, L_4^{(1)}$. Thus, identification of the target law relies on whether the propensity score of R_2, $p(r_2 \mid l_1^{(1)}, l_4^{(1)})$ is identified or not, which is not immediately clear since $R_2 \not\perp\!\!\!\perp R_1 \mid L_1^{(1)}, L_4^{(1)}$. We now outline a procedure to identify this propensity score.

We first intervene on R_4, i.e., dropping the term $p(r_4 \mid l_3)$ from the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

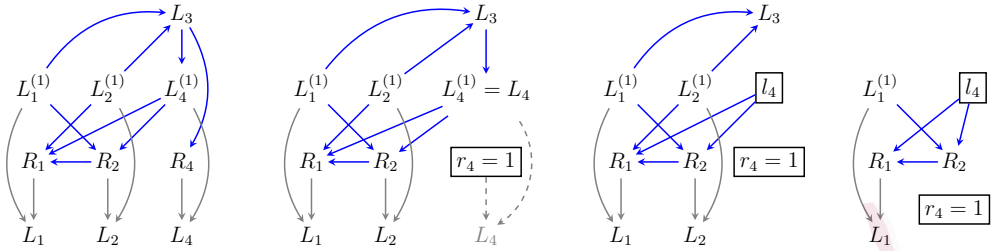


Figure 7: (a) An example m-DAG corresponding to a model where variables besides Rs are required to be intervened on in order to identify the propensity scores; (b) A conditional m-DAG where R_4 is intervened on; (c) A conditional m-DAG where L_4 is intervened on after an intervention on R_4; (d) A conditional m-DAG where $L_2^{(1)}$ and L_3 are marginalized out from the kernel $p(. \parallel r_4 = 1, l_4)$.

original factorization, yielding the kernel below which is Markov relative to the graph in Fig. 7(b),

$$p(l_1^{(1)}, l_2^{(1)}, l_3, l_4, r_1, r_2, l_1, l_2 \parallel r_4 = 1) = \frac{p(l_1^{(1)}, l_2^{(1)}, l_3, l_4, l_1, l_2, r_1, r_2, r_4 = 1)}{p(r_4 = 1 \mid l_3)}.$$

After intervening on R_4, $L_4^{(1)}$ is fully observed. We then intervene on L_4, i.e., dropping the term $p(l_4 \mid l_3 \parallel r_4 = 1)$ from the above kernel and obtain the following kernel which is Markov relative to the graph in Fig. 7(c),

$$p(l_1^{(1)}, l_2^{(1)}, l_3, r_1, r_2, l_1, l_2 \parallel r_4 = 1, l_4) = \frac{p(l_1^{(1)}, l_2^{(1)}, l_3, l_4, l_1, l_2, r_1, r_2, r_4 = 1)}{p(r_4 = 1 \mid l_3) \times p(l_4 \mid l_3, r_4 = 1)}.$$

Note that in this example, marginalizing out L_3 is equivalent to intervening on L_3. We can thus safely marginalize out $L_2^{(1)}, L_2$, and L_3 from the above

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

expression without changing the propensity score of R_2, yielding a new kernel that is Markov relative to the graph in Fig. 7(d),

$$p(l_1^{(1)}, r_1, r_2, l_1 \parallel r_4 = 1, l_4) = \sum_{l_3} \frac{p(l_1^{(1)}, l_3, l_4, r_1, r_2, r_4 = 1)}{p(r_4 = 1 \mid l_3) \times p(l_4 \mid l_3, r_4 = 1)}.$$

The propensity score of R_1 in the above kernel is different than the one in original factorization of the m-DAG in Fig. 7(a). We refer to this as the pseudo-propensity score of R_1 and denote it by $\widetilde{p}(r_1 \mid r_2 \parallel r_4, l_4)$. We now intervene on R_1 by dropping this term from the above kernel, yielding

$$p(l_1, r_2 \parallel r_1 = 1, r_4 = 1, l_4) = \frac{p(l_1, r_1 = 1, r_2 \parallel r_4 = 1, l_4)}{\widetilde{p}(r_1 = 1 \mid r_2 \parallel r_4 = 1, l_4)}.$$

The desired propensity score of R_2 (which remains invariant despite all previous operations as its direct causes are still present in Fig. 7(d)) is then identified in the above kernel as $p(l_1, r_2 \parallel r_1 = 1, r_4 = 1, l_4) / \sum_{r_2} p(l_1, r_2 \parallel r_1 = 1, r_4 = 1, l_4)$. Since all the propensity scores are identified, then the target law in (5.21) is identified as well.

In the above example, the partial orders for R_1 and R_4 are trivial—the corresponding propensity scores are immediately identified from the observed distribution. We can summarize the identification procedure for obtaining the propensity score of R_2 via the following partial order executed in a graph where R_4 then L_4 are intervened on and $\{L_2^{(1)}, L_2, L_3\}$ are treated as hidden variables, which we will denote as $\mathcal{G}_m(V \setminus \{L_2^{(1)}, L_2, L_3\}, \{r_4, l_4\})$:

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

the partial order of interventions for R_2 can be summarizes via $\{I_{r_4} < I_{l_4} < I_{r_1} < I_{r_2}\}$ in $\mathcal{G}_m(V \setminus \{L_2^{(1)}, L_2, L_3\}, \{r_4, l_4\})$. That is, intervention on R_2 must occur after interventions on R_4, L_4 and R_1 in a graph where $L_2^{(1)}, L_2$, and L_3 are marginalized out, and as mentioned earlier interventions on R_1 and R_4 are incompatible.

5.6 A Unifying Identification Procedure

Bhattacharya et al. (2019) proposed a procedure for target law identification that combines all the ideas discussed above. It proceeds as follows. For each missingness indicator $R_k \in R$, it proceeds to identify its propensity score $p(r_k | \operatorname{pa}_{\mathcal{G}_m}(r_k))$ evaluated at $R = 1$. It does so by checking if R_k is conditionally independent (given its parents) of the corresponding missingness indicators of its counterfactual parents. If this is the case, the propensity score is identified by a simple conditional independence argument (d-separation). Otherwise, the procedure checks if this condition holds in any intervention distribution where a subset of missingness indicators are intervened on, in either the original model or marginals of the model where the direct causes of R_k are still preserved. If the procedure succeeds in identifying the propensity score for each missingness indicator in this manner, then the target law is declared as being identified.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

Necessary conditions for target law identification have been discussed in the literature. A well-known result states that if an underlying variable causes its own missingness status, known as a self-censoring mechanism or nonignorable mechanism, then the target law is provably not identified. This means we can construct two missing data models that differ in target law distributions but both map to the same observed data distribution. The graphical structure simply corresponds to existence of an edge of the form $L_k^{(1)} \to R_k$ in the m-DAG. Nabi and Bhattacharya (2023); Guo et al. (2023) have also shown that the so-called "criss-cross" structure prevents target law identification. These structures involve a pair of variables $L_i^{(1)}, L_j^{(1)}$ that are directly connected as $L_i^{(1)} \to L_j^{(1)}$ or $L_i^{(1)} \leftarrow L_j^{(1)}$ and these edges exist simultaneously: $L_i^{(1)} \to R_j \leftarrow R_i \leftarrow L_j^{(1)}$. On the other hand, Nabi et al. (2020) have provided sufficient conditions under which a target law is identified. They show under the absence of self-censoring edges and so-called "colluder" structures, the target law is identified. A colluder is a special type of collider where there exists $R_i, R_j \in R$ such that $L_i^{(1)} \to R_j \leftarrow R_i$. Finding necessary and sufficient conditions (a sound and complete algorithm) for target law identification remains an open problem.

In many applied problems, some variables are not just missing but completely unobserved. We can use missing data DAG models with hidden

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

variables to encode the presence of unmeasured confounders and generalize the aforementioned identification strategies to such settings. We discuss extensions to identification in the presence of both missing data and hidden variables in Appendix S2.

6. Discussion

We have shown how unique features of missing data models represented by DAGs allow identification in seemingly counterintuitive situations, by taking advantage of Markov restrictions linking missingness indicators and underlying counterfactual variables. For instance, the target law is identified in the bivariate permutation model (Fig. 3(a)), but not in the analogous hidden variable causal model (Fig. 1(a)). Similarly, the target law is identified in the bivariate block-parallel model, shown in Fig. 9(a), by (5.13), but not in the analogous causal model shown in Fig. 9(b).

We described how graphical missing data models are a special case of hidden variable graphical causal models, where hidden variables are replaced by counterfactual variables, which are sometimes observed. It is this partial observability which allowed identification to be derived. Just as in missing data models, observed variables in causal models are derived from counterfactual variables by means of consistency, and thus may be viewed as

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

causes (parents) of observed variables in a particular type of causal graphs. This leads to a natural question: would expressing causal models via graphs which make the relationship between observed and counterfactual variables explicit by placing counterfactuals as vertices on the graph yield new types of identification results? We now show the answer is 'no' without further very strong assumptions.

Consider an elaboration of the graph in Fig. 1(a) shown in Fig. 8(a), where each U_i is replaced by two counterfactuals $L_i^{(1)}, L_i^{(0)}$ for $i = 1, 2$. These counterfactuals are associated, represented by the addition of a common parent ϵ_i. These counterfactuals are shown as parents of L_i, to indicate that L_i is determined from these counterfactuals by the value of R_i via consistency: $L_i = L_i^{(1)} R_i + L_i^{(0)}(1 - R_i)$. Furthermore, to preserve the structure of confounding in Fig. 1(a), $L_2^{(1)}, L_2^{(0)}$ are children of $L_1^{(1)}, L_1^{(0)}$. In fact, the graphs in Fig. 8(a) and Fig. 1(a) represent the same causal model.

Note that in this causal model, the intervention $r_2 = 1$ is identified from the observed data, since R_2 only has observed parents, namely L_1 and R_1. The graph where this intervention is performed is shown via the conditional causal DAG in Fig. 8(b), with the black edge representing the deterministic identity relationship between $L_2^{(1)}$ and L_2.

Just as in missing data, intervention $r_2 = 1$ renders the counterfactual

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

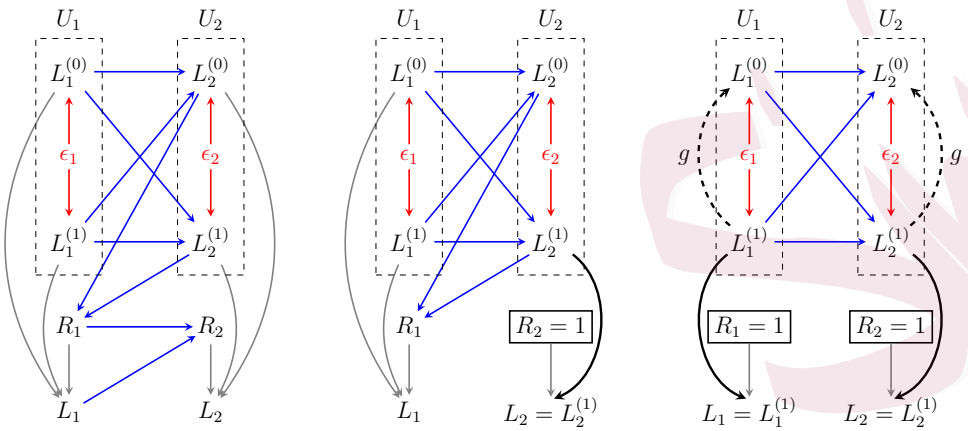


Figure 8: (a) A causal inference version of the bivariate permutation model with two counterfactual versions of L_1 and L_2 on the graph; (b) A world where an intervention $r_2 = 1$ is performed, yielding a situation where L_2 and $L_2^{(1)}$ coincide, i.e., $L_2 = L_2^{(1)}$; (c) A causal model where we impose a *rank preservation* relationship where a known bijective function $g(.)$ exists, such that $L_2^{(0)} = g(L_2^{(1)})$. In this model, the joint distribution $p(L_1^{(r_1)}, L_2^{(r_2)})$ is identified by sequentially intervening on r_2 then r_1.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

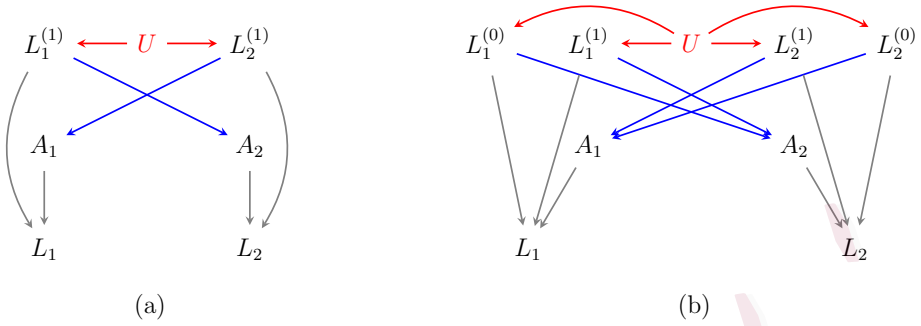
$L_2^{(1)}$ and the observed proxy variable L_2 to be the same, meaning that $L_2^{(1)}$ becomes observable. Since both $L_2^{(1)}$ and $L_2^{(0)}$ serve as parents of R_1, the fact that $L_2^{(1)}$ becomes observed after the intervention $r_2 = 1$ does not suffice to render the subsequent intervention on R_1 identified.

In particular, since two counterfactual versions of L_2, namely $L_2^{(1)}$ and $L_2^{(0)}$, exist, observing the former does not suffice to eliminate all confounding. This is illustrated by a non-causal path in Fig. 8(b) from R_1 to L_1 via $L_2^{(0)}$ and either $L_1^{(1)}$ or $L_1^{(0)}$.

However, if $L_2^{(1)}$ and $L_2^{(0)}$ are continuous random variables and we make the strong additional assumption of *rank preservation*, where $L_2^{(0)}$ is equal to $g(L_2^{(1)})$ for some function $g(.)$ bijective on the statespace of $L_2^{(1)}$ (and $L_2^{(0)}$), then identification is recovered since $p(R_1 \mid L_2^{(1)}, L_1^{(0)}) = p(R_1 \mid L_2^{(1)}, g(L_2^{(1)})) = p(R_1 \mid L_2^{(1)})$ becomes a functional of the observed data law once an intervention $r_2 = 1$ is performed. This is illustrated in Fig. 8(c), where the bijective relationship between $L_2^{(1)}$ and $L_2^{(0)}$ is represented by a dashed edge indexed by g. Note that once $L_2^{(1)}$ is observed, so is $L_2^{(0)}$, which implies that conditioning on one counterfactual implicitly conditions on the other, due to their deterministic relationship.

We discuss additional causal models in which rank preservation yields identification in the Appendix S4.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382



(a) (b)

Figure 9: (a) The bivariate block parallel missing data model. (b) The causal model analogue of the model in (a) where identification of causal effects is not possible without further assumptions, but is possible with rank preservation.

As a second example for illustrating why identification strategies in missing data models do not easily translate to causal models, consider the causal analogue of the bivariate block-parallel model shown in Fig. 9(b), with two binary treatments A_1, A_2 instead of missingness indicators, and two observed outcomes L_1, L_2, which implies two counterfactual versions of each observed outcome: $L_1^{(1)}, L_1^{(0)}$, and $L_2^{(1)}, L_2^{(0)}$. The model is defined by the following restrictions:

$$A_2 \perp\!\!\!\perp L_2^{(1)}, L_2^{(0)}, A_1 \mid L_1^{(1)}, L_1^{(0)} \quad \text{and} \quad A_1 \perp\!\!\!\perp L_1^{(1)}, L_1^{(0)}, A_2 \mid L_2^{(1)}, L_2^{(0)}. \quad (6.22)$$

These restrictions do not yield nonparametric identification since the propen-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

sity scores in the model depend on both versions of the counterfactuals:

$$p(a_2 \mid l_2^{(1)}, l_2^{(0)}, a_1, l_1^{(1)}, l_1^{(0)}) = p(a_2 \mid l_1^{(1)}, l_1^{(0)}),$$

$$p(a_1 \mid l_2^{(1)}, l_2^{(0)}, a_1, l_1^{(1)}, l_1^{(0)}) = p(a_1 \mid l_2^{(1)}, l_2^{(0)}).$$

However, we can recover an argument for identification of $p(l_1^{(i)}, l_2^{(j)})$ for any $(i, j) \in \{0, 1\}^2$ via rank preservation, which states that for $k \in \{0, 1\}$ there exist bijections $g_k(.)$ such that $L_1^{(1-i)} = g_1(L_1^{(i)})$ and $L_2^{(1-j)} = g_2(L_2^{(j)})$. Identification of $p(l_1^{(i)}, l_2^{(j)})$ then proceeds as follows:

$$\begin{aligned}
p(l_1^{(i)}, l_2^{(j)}) &= \frac{p(l_1^{(i)}, l_2^{(j)}, A_1 = i, A_2 = j)}{p(A_1 = i, A_2 = j \mid l_1^{(i)}, l_2^{(j)})} \\
&= \frac{p(l_1^{(i)}, l_2^{(j)}, A_1 = i, A_2 = j)}{\sum_{l_1^{(1-i)}, l_2^{(1-j)}} p(A_1 = i, A_2 = j \mid l_1^{(i)}, l_2^{(j)}, l_1^{(1-i)}, l_2^{(1-j)}) \times p(l_1^{(1-i)}, l_2^{(1-j)} \mid l_1^{(i)}, l_2^{(j)})} \\
&= \frac{p(l_1^{(i)}, l_2^{(j)}, A_1 = i, A_2 = j)}{\sum_{l_1^{(1-i)}, l_2^{(1-j)}} p(A_1 = i \mid l_2^{(j)}, l_2^{(1-j)}, A_2 = j) \times p(A_2 = j \mid l_1^{(i)}, l_1^{(1-i)}, A_1 = i) \times p(l_1^{(1-i)}, l_2^{(1-j)} \mid l_1^{(i)}, l_2^{(j)})} \\
&= \frac{p(l_1^{(i)}, l_2^{(j)}, A_1 = i, A_2 = j)}{\sum_{l_1^{(1-i)}, l_2^{(1-j)}} p(A_1 = i | l_2^{(j)}, l_2^{(1-j)}, A_2 = j) \times p(A_2 = j | l_1^{(i)}, l_1^{(1-i)}, A_1 = i) \times \mathbb{I}(l_1^{(1-i)} = g_1(l_1^{(i)}), l_2^{(1-j)} = g_2(l_2^{(j)}))} \\
&= \frac{p(l_1^{(i)}, l_2^{(j)}, A_1 = i, A_2 = j)}{p(A_1 = i \mid l_2^{(j)}, A_2 = j) \times p(A_2 = j \mid l_1^{(i)}, A_1 = i)} \\
&= \frac{p(l_1, l_2, A_1 = i, A_2 = j)}{p(A_1 = i \mid l_2, A_2 = j) \times p(A_2 = j \mid l_1, A_1 = i)}.
\end{aligned}$$

Here, the first and second equalities follow by the rules of probability, the third by (6.22), the fourth and fifth by the rank preservation assumption, and the last by consistency. This derivation is structurally very similar to the derivation for the bivariate block-parallel model, except for the last two steps which explicitly rely on rank preservation. In Appendix S4, we show that this structural similarity is quite general, and a similar identification strategy can be defined for a K variable causal analogue of the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

block-parallel model endowed with rank preservation.

While much of the discussion in this paper has focused on how causal identification techniques can be applied or extended to missing data settings, the above example demonstrates missing data techniques can only be applied to causal settings given much stronger untestable assumptions, such as rank preservation, than those plausibly assumed in causal models.

Supplementary Materials

The supplementary materials contain discussions on identification of missing data DAG models using the odds ratio parameterization extension of identification techniques to m-DAG models with unmeasured confounders, results on identification of the full law, proofs, and some additional results on identification in causal models using rank preservation.

Acknowledgements

Rohit Bhattacharya acknowledges support from the NSF CRII 2348287. Ilya Shpitser acknowledges support from the ONR N000142412701, NIH R01HS027819, and NSF CAREER 1942239.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

References

Bhattacharya, R., R. Nabi, and I. Shpitser (2022). Semiparametric inference for causal effects in graphical models with hidden variables. *Journal of Machine Learning Research 23*(295), 1–76.

Bhattacharya, R., R. Nabi, I. Shpitser, and J. Robins (2019). Identification in missing data models represented by directed acyclic graphs. In *Proceedings of the Thirty Fifth Conference on Uncertainty in Artificial Intelligence (UAI-35th)*. AUAI Press.

Cain, L. E., M. S. Saag, M. Petersen, M. T. May, S. M. Ingle, R. Logan, J. M. Robins, S. Abgrall, B. E. Shepherd, S. G. Deeks, et al. (2016). Using observational data to emulate a randomized trial of dynamic treatment-switching strategies: An application to antiretroviral therapy. *International Journal of Epidemiology 45*(6), 2038–2049.

Chen, J. M., D. Malinsky, and R. Bhattacharya (2023). Causal inference with outcome-dependent missingness and self-censoring. In *Proceedings of the 39th Uncertainty in Artificial Intelligence*, pp. 358–368. PMLR.

Daniel, R. M., M. G. Kenward, S. N. Cousens, and B. L. De Stavola (2012). Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research 21*(3), 243–256.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

REFERENCES

Ding, P. and F. Li (2018). Causal inference: A missing data perspective. *Statistical Science 33*, 214–237.

Duarte, G., N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser (2024). An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association 119*(547), 1778–1793.

Gill, R. D., M. J. Van Der Laan, and J. M. Robins (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, pp. 255–294. Springer.

Glymour, M. M. (2006). Using causal diagrams to understand common problems in social epidemiology. In J. M. Oakes and J. S. Kaufman (Eds.), *Methods in Social Epidemiology*, pp. 393–428. Jossey-Bass/Wiley.

Graham, J. W. (2012). *Missing Data: Analysis and Design*. Springer Science & Business Media.

Guo, A., J. Zhao, and R. Nabi (2023). Sufficient identification conditions and semiparametric estimation under missing not at random mechanisms. In *Proceedings of the 39th Uncertainty in Artificial Intelligence*, pp. 777–787. PMLR.

Horton, N. J. and N. M. Laird (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research 8*(1), 37–50.

Huang, Y. and M. Valtorta (2006). Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 217–224.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

REFERENCES

Lauritzen, S. L. (1996). *Graphical Models*. Oxford, U.K.: Clarendon.

Li, L., C. Shen, X. Li, and J. M. Robins (2013). On weighting approaches for missing data. *Statistical Methods in Medical Research 22*(1), 14–30.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association 88*(421), 125–134.

Little, R. J. (2016). *Selection Model (Missing Data)*, pp. 1–5. John Wiley & Sons, Ltd.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley.

Ma, W.-Q., Z. Geng, and Y.-H. Hu (2003). Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies. *Journal of Multivariate Analysis 87*(1), 24–45.

Malinsky, D., I. Shpitser, and E. J. Tchetgen Tchetgen (2022). Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association 117*(539), 1415–1423.

Martel García, F. (2013). Definition and diagnosis of problematic attrition in randomized controlled experiments. *Available at SSRN 2302735, https://ssrn.com/abstract=2302735*.

Mattei, A., F. Mealli, and B. Pacini (2014). Identification of causal effects in the presence of nonignorable missing outcome values. *Biometrics 70*(2), 278–288.

McKnight, P. E., K. M. McKnight, S. Sidani, and A. J. Figueredo (2007). *Missing Data: A*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

Gentle Introduction. Guilford Press.

Miao, W., P. Ding, and Z. Geng (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association 111*(516), 1673–1683.

Miao, W. and E. J. Tchetgen Tchetgen (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika 103*(2), 475–482.

Mohan, K. and J. Pearl (2014). Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in Neural Information Processing Systems 27*, pp. 1520–1528. Curran Associates, Inc.

Mohan, K. and J. Pearl (2021). Graphical models for processing missing data. *Journal of the American Statistical Association 116*(534), 1023–1037.

Mohan, K., J. Pearl, and J. Tian (2013). Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems 26*, pp. 1277–1285. Curran Associates, Inc.

Moreno-Betancur, M. and M. Chavance (2016). Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: Methodology and application in a clinical trial with drop-outs. *Statistical Methods in Medical Research 25*(4), 1471–1489.

Nabi, R. and R. Bhattacharya (2023). On testability and goodness of fit tests in missing data models. In *Uncertainty in Artificial Intelligence*, pp. 1467–1477. PMLR.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

REFERENCES

Nabi, R., R. Bhattacharya, and I. Shpitser (2020). Full law identification in graphical models of missing data: Completeness results. In *Proceedings of the Twenty Seventh International Conference on Machine Learning*, pp. 7153–7163. PMLR.

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

Richardson, T. S., R. J. Evans, J. M. Robins, and I. Shpitser (2023). Nested Markov properties for acyclic directed mixed graphs. *The Annals of Statistics 51*(1), 334–361.

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

Robins, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine 16*, 21–37.

Robins, J. M. and R. D. Gill (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine 16*(1), 39–56.

Robins, J. M. and A. Rotnitzky (2001). Comment on inference for semiparametric models: Some questions and an answer, by PJ Bickel and J. Kwon. *Statistica Sinica 11*, 920–936.

Robins, J. M., A. Rotnitzky, and D. O. Scharfstein (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 1–94. Springer.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

REFERENCES

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

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*(429), 106–121.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association 93*(444), 1321–1339.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology 66*, 688–701.

Rubin, D. B. (1976). Causal inference and missing data (with discussion). *Biometrika 63*, 581–592.

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

Saadati, M. and J. Tian (2019). Adjustment criteria for recovering causal effects from missing data. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Sadinle, M. and J. P. Reiter (2017). Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika 104*(1), 207–220.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

REFERENCES

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research 8*(1), 3–15.

Schafer, J. L. and J. W. Graham (2002). Missing data: our view of the state of the art. *Psychological methods 7*(2), 147–177.

Scharfstein, D. O. and R. A. Irizarry (2003). Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data. *Biometrics 59*(3), 601–613.

Scharfstein, D. O., J. J. Lee, A. McDermott, A. Campbell, E. Nunes, A. G. Matthews, and I. Shpitser (2021). Markov-restricted analysis of randomized trials with non-monotone missing binary outcomes: Sensitivity analysis and identification results. *arXiv preprint arXiv:2105.08868*.

Scharfstein, D. O., R. Nabi, E. H. Kennedy, M.-Y. Huang, M. Bonvini, and M. Smid (2021). Semiparametric sensitivity analysis: Unmeasured confounding in observational studies. *arXiv preprint arXiv:2104.08300*.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semi-parametric nonresponse models. *Journal of the American Statistical Association 94*, 1096–1146.

Seaman, S., J. Galati, D. Jackson, and J. Carlin (2013). What is meant by "missing at random"? *Statistical Science 28*(2), 257–268.

Shpitser, I. (2016). Consistent estimation of functions of data missing non-monotonically and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

not at random. *Advances in Neural Information Processing Systems 29*, 3144–3152.

Shpitser, I., K. Mohan, and J. Pearl (2015). Missing data as a causal and probabilistic problem. In *Proceedings of the Thirty First Conference on Uncertainty in Artificial Intelligence (UAI-15)*, pp. 802–811. AUAI Press.

Shpitser, I. and J. Pearl (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, pp. 1219–1226. AAAI Press.

Srinivasan, R., R. Bhattacharya, R. Nabi, E. L. Ogburn, and I. Shpitser (2023). Graphical models of entangled missingness. *arXiv preprint arXiv:2304.01953*.

Sun, B., L. Liu, W. Miao, K. Wirth, J. Robins, and E. J. T. Tchetgen (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica 28*(4), 1965.

Tchetgen, E. J. T. (2009). Commentary: A simple implementation of doubly robust estimation in logistic regression with covariates missing at random. *Epidemiology 20*(3), 391–394.

Tchetgen, E. J. T., L. Wang, and B. Sun (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica 28*(4), 2069–2088.

Thoemmes, F. and N. Rose (2014). A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research 49*(5), 443–459.

Tian, J. (2015). Missing at random in graphical models. In *Proceedings of the Artificial Intel-*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

ligence and Statistics, pp. 977–985. PMLR.

Tian, J. and J. Pearl (2002). On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02)*, Volume 18, pp. 519–527. AUAI Press, Corvallis, Oregon.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data* (1st edition ed.). Springer-Verlag New York.

Vansteelandt, S., A. Rotnitzky, and J. Robins (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika 94*(4), 841–860.

Wang, S., J. Shao, and J. K. Kim (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica 24*(3), 1097–1116.

Wu, M. C. and R. J. Carroll (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics 44*(1), 175–188.

Zhou, Y., R. J. A. Little, and K. J. D. (2010). Block-conditional missing at random models for missing data. *Statistical Science 25*(4), 517–532.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0382

REFERENCES

Razieh Nabi, Department of Biostatistics and Bioinformatics, Emory University

E-mail: (razieh.nabi@emory.edu)

Rohit Bhattacharya, Department of Computer Science, Williams College

E-mail: (rb17@williams.edu)

Ilya Shpitser, Department of Computer Science, Johns Hopkins University

E-mail: (ilyas@cs.jhu.edu)

James Robins, Department of Epidemiology, Harvard T. H. Chan School of Public Health

E-mail: (robins@hsph.harvard.edu)