Statistica Sinica Preprint No: SS-2025-0152	
Title	Discussion on "Causal and Counterfactual Views of Missing Data Models"
Manuscript ID	SS-2025-0152
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0152
Complete List of Authors	Shanshan Luo and Zhi Geng
Corresponding Authors	Zhi Geng
E-mails	zhigeng@pku.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

Discussion on "Causal and Counterfactual Views of Missing Data Models"

Shanshan Luo[1] and Zhi Geng[2,1,*]

[1] *School of Mathematics and Statistics, Beijing Technology and Business University*

[2] *School of Mathematical Sciences, Peking University*

Abstract: In response to the insightful work by Profs. Nabi, Bhattacharya, Shpitser, and Robins on formulating missing data problems within the potential outcomes framework, this discussion further investigates identifiability in missing not at random settings using missing data directed acyclic graphs (m-DAGs). Focusing on binary outcomes, we consider scenarios in which missingness indicators may be directly influenced by the outcome itself. We first analyze identifiability in the absence of causal dependencies among missingness indicators, and then examine how auxiliary variables can facilitate identification when

*Correspondence to: zhigeng@pku.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

such dependencies are present.

Key words and phrases: Causal inference, Missing data models, Missing not at random

1. Introduction

We sincerely congratulate Profs. Nabi, Bhattacharya, Shpitser, and Robins on their interesting contribution, which addresses challenges in missing data analysis through the framework of potential outcomes. In this paper, the authors explain how missing data problems can be framed as causal inference problems: the complete variables are viewed as counterfactual outcomes, the missingness indicators are treated as treatment variables, and the partially observed variables are interpreted as combinations of potential outcomes and treatments. The authors introduce missing data directed acyclic graphs (m-DAGs), review several missing data models from previous literature that can be represented using m-DAGs, and present identifiability results for various graph structures.

In this discussion, we explore more identifiable missing not at random (MNAR) models within the m-DAG framework, with a particular focus on causal graphs involving binary outcomes. We observe that in many of the figures presented in the main text, the complete counterfactual variables are

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

not allowed to have direct effects on their own missingness indicators. In our discussion, we consider alternative scenarios in which the missingness mechanism may be directly affected by the outcome variable itself, as in the self-censoring problem (Nabi et al., 2020; Li et al., 2023), and examine the associated identifiability issues. We first study identifiability without effects between missingness indicators, and then consider how auxiliary variables can facilitate identification with effects between missingness indicators. The supplementary material providing the complete proofs of the theoretical results is available online: `https://gitlab.com/pipishan95/mdag_SM`.

2. Identification in Self-Censoring m-DAGs

In this discussion, we focus on an MNAR mechanism in which the missingness of a variable depends on its unobserved outcome. This type of mechanism is common in social and biomedical studies. For example, individuals with higher income levels may be less likely to respond to questions about their income. In this case, whether the income is reported depends on its (partially unobserved) true value. Throughout this paper, let $L = (L_1, L_2)$ denote two partially observed binary outcomes, and let $R = (R_1, R_2)$ denote their corresponding missingness indicators. The associated potential outcomes are denoted by $L^{(1)} = (L_1^{(1)}, L_2^{(1)})$. Specifically, the link between

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

the potential outcomes and the observed variables is: for $k = 1, 2$, if $R_k = 1$, then $L_k = L_k^{(1)}$; if $R_k = 0$, L_k is unobserved, and $L_k = L_k^{(0)}$ (with $L_k^{(0)}$ trivially denoted as "?"). This is closely related to the consistency assumption in causal inference.

2.1 Identification without Effect between Missingness Indicators

In this subsection, we first consider MNAR mechanisms where the missingness mechanism of one variable does not directly affect another counterfactual outcome. This corresponds to certain conditional independences encoded in the mDAG. Let the symbol "$\perp\!\!\!\perp$" denote (conditional) independence between variables. The following assumption formally characterises these conditions in the setting with two outcome variables.

Assumption 1. (i) $R_1 \perp\!\!\!\perp L_2^{(1)} \mid (L_1^{(1)}, R_2)$ and (ii) $R_2 \perp\!\!\!\perp L_1^{(1)} \mid (L_2^{(1)}, R_1)$.

Assumption 1 characterizes a nonignorable missingness mechanism involving two binary outcomes. In particular, Assumption 1(i) implies that the missingness indicator R_1 is conditionally independent of the other outcome $L_2^{(1)}$, given its own outcome $L_1^{(1)}$ and the missingness indicator R_2. This allows R_1 to be directly affected by its corresponding counterfactual outcome $L_1^{(1)}$, and similarly for R_2 and $L_2^{(1)}$. Such a mechanism is commonly referred to as self-censoring (Nabi et al., 2020; Li et al., 2023), while this

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

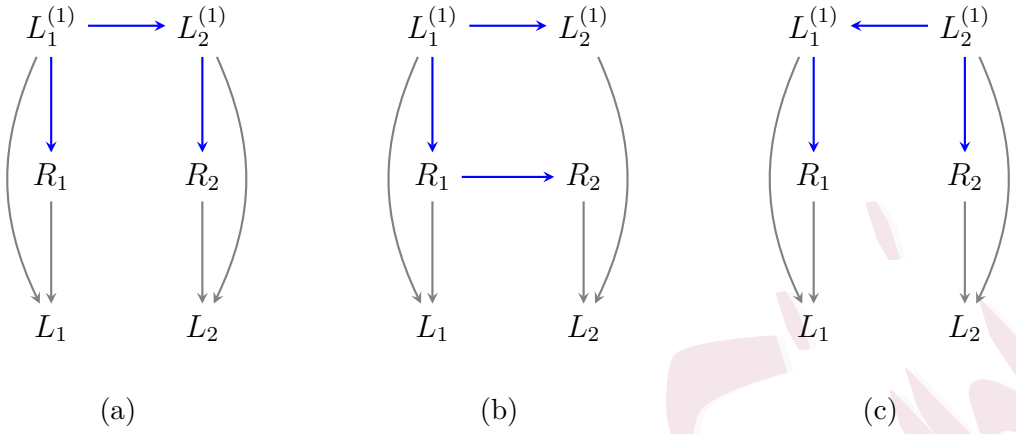

Figure 1: Three possible structures satisfy Assumption 1.

specific issue is not explicitly discussed in the main text. Figure 1 presents three possible MNAR structures under Assumption 1, each capturing some dependency patterns between $L_1^{(1)}$ and $L_2^{(1)}$, and how the missingness mechanisms R_1 and R_2 are influenced by other variables. A special case of Assumption 1, illustrated in Figures 1(a) and 1(c), is when the missingness of each outcome depends solely on its own value, that is, $R_1 \perp\!\!\!\perp (L_2^{(1)}, R_2) \mid L_1^{(1)}$ and $R_2 \perp\!\!\!\perp (L_1^{(1)}, R_1) \mid L_2^{(1)}$. To simplify the exposition, we impose positivity by assuming that $p(l_1, l_2 \mid r_1 = 1, r_2 = 1) > 0$ for all $l_1, l_2 \in \{0, 1\}$.

Lemma 1. *The full law* $p(r_1, r_2, l_1^{(1)}, l_2^{(1)})$ *is identifiable for the m-DAG under Assumption 1 if*

$$\frac{p(l_1 = 0, l_2 = 0 \mid r_1 = 1, r_2 = 1)}{p(l_1 = 0, l_2 = 1 \mid r_1 = 1, r_2 = 1)} \neq \frac{p(l_1 = 1, l_2 = 0 \mid r_1 = 1, r_2 = 1)}{p(l_1 = 1, l_2 = 1 \mid r_1 = 1, r_2 = 1)}. \quad (2.1)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

Lemma 1 shows that the full data law is identifiable if the two ratios in (2.1) are not equal, a testable condition from the observed data. For all causal structures depicted in Figure 1, the condition is equivalent to the following inequality:

$$\frac{p(l_1^{(1)}=0,l_2^{(1)}=0)}{p(l_1^{(1)}=0,l_2^{(1)}=1)} \neq \frac{p(l_1^{(1)}=1,l_2^{(1)}=0)}{p(l_1^{(1)}=1,l_2^{(1)}=1)},$$

which further implies that $L_1^{(1)}$ and $L_2^{(1)}$ must be associated in these graphs ($L_1^{(1)} \not\perp\!\!\!\perp L_2^{(1)}$). In the context of graphical discussions concerning binary nonresponse in longitudinal studies, Ma et al. (2003) points out that the full law becomes identifiable when a fully observable auxiliary variable does not directly affect the missingness mechanism. Lemma 1 further extends this idea by employing the incomplete variable L_i as the auxiliary variable for another missing outcome L_j (with $i \neq j$).

2.2 Identification with Effect between Missingness Indicators

In practice, the full law becomes unidentifiable when Assumption 1 is violated, which may occur in a class of graph structures illustrated in Figure 2. It can be observed that all graphs contain a collider structure, where each missingness mechanism may be affected by its corresponding outcome variable. A variable is referred to as a collider when it is causally influenced by two or more variables (Pearl, 2000). We have noticed that in the context

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

of m-DAGs, this variable is also referred to as "colluder" (Nabi et al., 2020), and we would like to avoid any ambiguity caused by this terminology. For instance, Figure 2(a) includes the structure $R_1 \to R_2 \leftarrow L_2^{(1)}$, while Figures 2(b) and 2(c) include the structure $R_2 \to R_1 \leftarrow L_1^{(1)}$. Specifically, when the structure $R_2 \to R_1 \leftarrow L_1^{(1)}$ is present, intervening on R_1 induces a correlation between R_2 and $L_1^{(1)}$, thereby violating the conditional independence assumption in Assumption 1(i). In this section, we will address the identifiability issues in Figure 2(a) by introducing auxiliary variables. To simplify the exposition, we always introduce the positivity assumption in the subsequent discussions where needed.

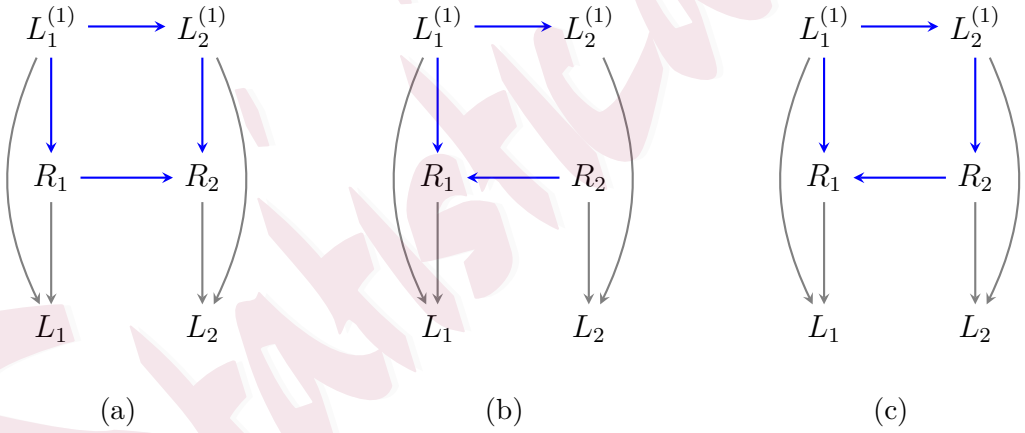

Figure 2: Three possible structures do not satisfy Assumption 1.

We first introduce the baseline self-censoring scenario in Figure 3(a), where a baseline outcome L_0, its missingness indicator R_0, and its poten-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

tial outcome $L_0^{(1)}$ are included. For example, in a survey involving sensitive questions, $L_0^{(1)}$ may represent the respondent's potential answer to whether they have ever engaged in a certain type of behaviour. The actual response is denoted by L_0, which may be observed or missing, and R_0 denotes the corresponding missingness indicator. These variables $(L_0, R_0, L_0^{(1)})$ are assumed to occur before $(L_1^{(1)}, L_2^{(1)}, R_1, R_2)$ and satisfy specific structural independence conditions. Although there is a collider structure among $(L_1^{(1)}, L_2^{(1)}, R_1, R_2)$, the subgraph consisting of $(L_1^{(1)}, L_0^{(1)}, R_1, R_0)$ additionally satisfies the conditional independencies $R_1 \perp\!\!\!\perp L_0^{(1)} \mid (L_1^{(1)}, R_0)$ and $R_0 \perp\!\!\!\perp L_1^{(1)} \mid (L_0^{(1)}, R_1)$. According to Lemma 1, the full law $p(r_0, r_1, l_0^{(1)}, l_1^{(1)})$ is identifiable under additional testable conditions analogous to those in (2.1), which contributes to identifying the full law $p(r_0, r_1, r_2, l_0^{(1)}, l_1^{(1)}, l_2^{(1)})$.

Theorem 1. *The full law $p(r_0, r_1, r_2, l_0^{(1)}, l_1^{(1)}, l_2^{(1)})$ is identifiable for the m-DAG in Figure 3(a) if*

$$\frac{p(l_0 = 0, l_1 = 0 \mid r_0 = 1, r_1 = 1)}{p(l_0 = 0, l_1 = 1 \mid r_0 = 1, r_1 = 1)} \neq \frac{p(l_0 = 1, l_1 = 0 \mid r_0 = 1, r_1 = 1)}{p(l_0 = 1, l_1 = 1 \mid r_0 = 1, r_1 = 1)}, \quad (2.2)$$

and

$$\frac{p(l_2 = 0 \mid r_1 = 1, r_2 = 1, l_1 = 0)}{p(l_2 = 0 \mid r_1 = 1, r_2 = 1, l_1 = 1)} \neq \frac{p(l_2 = 1 \mid r_1 = 1, r_2 = 1, l_1 = 0)}{p(l_2 = 1 \mid r_1 = 1, r_2 = 1, l_1 = 1)}. \quad (2.3)$$

Condition (2.2) in Theorem 1 is similar to condition (2.1), and allow us to identify the joint distribution $p(r_0, r_1, l_0^{(1)}, l_1^{(1)})$ via Lemma 1. Ad-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

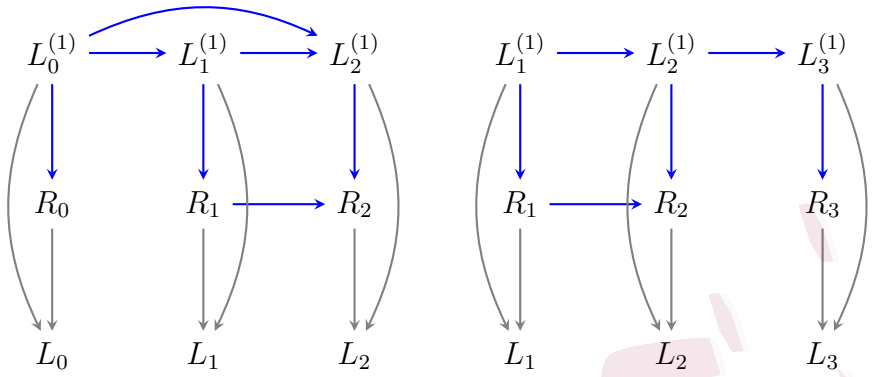

Figure 3: Some identifiable m-DAGs with baseline and follow-up measurements.

ditionally, to ensure identifiability of the full law, we further impose an inequality condition (2.3) between two ratios, based on the independence shown in Figure 3(a), which means that

$$\frac{p(l_1 = 0, l_2 = 0 \mid r_2 = 1)}{p(l_1 = 1, l_2 = 0 \mid r_2 = 1)} \neq \frac{p(l_1 = 0, l_2 = 1 \mid r_2 = 1)}{p(l_1 = 1, l_2 = 1 \mid r_2 = 1)}.$$

This implies that the counterfactual variables $L_1^{(1)}$ and $L_2^{(1)}$ must be dependent given $R_2 = 1$; that is, $L_1^{(1)} \not\perp\!\!\!\perp L_2^{(1)} \mid R_2 = 1$. It can be seen that the baseline measurement plays a key role in ensuring the identifiability of the joint distribution $p(l_1^{(1)}, r_1)$, which is then used to identify subgraphs with a collider structure among $(L_1^{(1)}, L_2^{(1)}, R_1, R_2)$. From this perspective, the potentially missing auxiliary variable $L_0^{(1)}$ and its missingness indicator R_0

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

provide additional information for identification.

We next show that, in addition to the baseline measurement L_0, the presence of an additional follow-up measurement L_3, along with its missingness indicator R_3 and associated potential outcome $L_3^{(1)}$, can also ensure the identifiability of the full data law. Let L_1 and L_2 represent the first-stage and second-stage health measurements, respectively, with corresponding missingness indicators R_1 and R_2. The follow-up measurement collected at a later time point is denoted by L_3, with its missingness indicator R_3. Such settings are common in practice, especially in longitudinal studies involving repeated measurements or long-term follow-up. Figure 3(b) illustrates this sequential scenario involving three outcomes. The following theorem shows that this structure allows for the identification of the full law.

Theorem 2. *The full law* $p(r_1, r_2, r_3, l_1^{(1)}, l_2^{(1)}, l_3^{(1)})$ *is identifiable for the m-DAG in Figure 3(b) if* (2.3) *holds and*

$$\frac{p(l_2 = 0, l_3 = 0 \mid r_2 = 1, r_3 = 1)}{p(l_2 = 0, l_3 = 1 \mid r_2 = 1, r_3 = 1)} \neq \frac{p(l_2 = 1, l_3 = 0 \mid r_2 = 1, r_3 = 1)}{p(l_2 = 1, l_3 = 1 \mid r_2 = 1, r_3 = 1)}. \quad (2.4)$$

The inequality condition between two ratios in (2.4) of Theorem 2 is similar to condition (2.1) in Lemma 1 and condition (2.2) in Theorem 1. In addition, Theorem 2 also requires that condition (2.3) holds. As in Theorem 1, the key idea is to leverage the subgraph over $(L_2^{(1)}, L_3^{(1)}, R_2, R_3)$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

to recover the joint distribution $p(r_2, r_3, l_2^{(1)}, l_3^{(1)})$. By exploiting the conditional dependence between $L_2^{(1)}$ and $L_3^{(1)}$, the distribution $p(r_2, l_2^{(1)})$ can further contribute to identifying the full data law $p(r_1, r_2, r_3, l_1^{(1)}, l_2^{(1)}, l_3^{(1)})$.

3. Discussion

We close by again congratulating the authors on their important contribution to m-DAGs. In this discussion, we consider some potential identification issues arising from m-DAGs with self-censoring scenario. Future work may enhance identifiability in more general settings (Li et al., 2023), such as continuous outcomes.

Acknowledgments

This discussion is supported by the National Natural Science Foundation of China (No. 12401378).

References

Li, Y., Miao, W., Shpitser, I., and Tchetgen Tchetgen, E. J. (2023). A self-censoring model for multivariate nonignorable nonmonotone missing data. *Biometrics*, 79(4):3203–3214.

Ma, W.-Q., Geng, Z., and Hu, Y.-H. (2003). Identification of graphical

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0152

models for nonignorable nonresponse of binary outcomes in longitudinal studies. *Journal of Multivariate Analysis*, 87(1):24–45.

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

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Shanshan Luo

School of Mathematics and Statistics, Beijing Technology and Business University

E-mail: shanshanluo@btbu.edu.cn

Zhi Geng

School of Mathematical Sciences, Peking University

School of Mathematics and Statistics, Beijing Technology and Business University

E-mail: zhigeng@pku.edu.cn