Statistica Sinica Preprint No: SS-2025-0165	
Title	Discussion on “Causal and Counterfactual Views of Missing Data Models”
Manuscript ID	SS-2025-0165
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0165
Complete List of Authors	Shu Yang and Jae Kwang Kim
Corresponding Authors	Jae-Kwang Kim
E-mails	jkim@iastate.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0165

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

Shu Yang and Jae Kwang Kim

North Carolina State University and Iowa State University

1. Introduction

This paper presents a new perspective on missing data by integrating causal and counterfactual frameworks, reinterpreting missing data as a causal inference problem. By drawing an analogy between missing data and unobserved counterfactuals, the paper advances a structured approach to understanding and identifying parameters under different missingness mechanisms, particularly in Missing Not At Random (MNAR) settings. Utilizing Directed Acyclic Graphs (DAGs) and their extensions to missing data DAGs (m-DAGs), this work provides a rigorous framework for characterizing dependencies and assumptions in missing data problems. A key contribution is the extension of the g-formula to accommodate counterfactual distributions in missing data models, offering new insights and methodological

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0165

advancements for addressing MNAR challenges.

Challenges in Rubin's Hierarchy of Missing Data Mechanisms. Rubin's classification (Rubin, 1976) - *Missing Completely at Random (MCAR)*, *Missing at Random (MAR)*, and *Missing Not at Random (MNAR)* - has long served as the foundational framework for modeling missingness. However, this hierarchy introduces both conceptual and practical limitations, especially under the MAR assumption. Specifically, MAR requires that the missingness indicator R depends only on observed outcomes Y^{obs} and auxiliary covariates X, but not on the missing components Y^{mis}:

$$\mathbb{P}(R \mid Y^{\text{obs}}, Y^{\text{mis}}, X) = \mathbb{P}(R \mid Y^{\text{obs}}, X).$$

Simulating data that strictly conform to MAR is challenging due to the implicit dependencies between Y^{obs} and R. Since the observed data are themselves a function of the missingness mechanism, defining a coherent MAR structure across different patterns is nontrivial. These interdependencies expose foundational limitations in Rubin's taxonomy.

Contributions of this paper. This work advances both the theory and practice of missing data analysis by addressing the limitations of traditional frameworks. The main contributions include:

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0165

- **Conceptual Innovation:** By reframing missingness as a causal inference problem, the paper bridges the gap between missing data methodologies and causal inference, drawing on established tools such as counterfactual reasoning and graphical modeling. This unified view improves interpretability and conceptual coherence.

- **Graphical Modeling via m-DAGs:** The introduction of m-DAGs provides a transparent way to encode assumptions and represent complex dependencies. This approach enhances model specification and facilitates clearer communication with stakeholders.

- **Identification Theory:** The framework extends identification theory by applying counterfactual reasoning in MNAR settings, enabling identification without restrictive parametric assumptions. Notably, it demonstrates identification of the propensity score $P(R \mid L^{(1)})$, highlighting the potential benefits of m-DAGs over conventional causal inference DAGs (CI-DAGs).

2. Discussions

While the m-DAG framework provides conceptual clarity and theoretical rigor, it also presents several challenges that warrant further discussion.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0165

This section discusses practical challenges and outlines areas for future research.

2.1 Practicality and Assumption Validation

The utility of m-DAGs in applied settings depends critically on the accurate specification of causal structures, which may not be readily feasible. Compared to conventional missing data assumptions, m-DAGs often impose stronger and more explicit conditional independence requirements. These assumptions are encoded graphically and analyzed using tools such as d-separation and do-calculus.

This mirrors the broader debate in causal inference between DAG-based and potential outcomes approaches. As noted by Imbens (2020), the DAG framework has not been widely adopted in economics, partly due to a lack of empirical applications demonstrating its advantages. Moreover, key identification strategies in econometrics - such as instrumental variables and monotonicity - can be difficult to represent within DAGs.

These concerns also apply to m-DAGs. Their adoption may be hindered by the difficulty of encoding complex real-world assumptions and the lack of established estimation and inference tools aligned with graphical models. The challenge intensifies in high-dimensional settings, where the number

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0165

of variables and potential missingness patterns grows rapidly. Validating assumptions in such contexts requires substantial domain knowledge and computational resources, raising questions about scalability and feasibility.

2.2 Sensitivity Analysis

Sensitivity analysis plays a crucial role in assessing the robustness of causal assumptions. Recent work by Ding et al. (2023) reviews methods for evaluating the no unmeasured confounding assumption in observational studies. Similarly, DAGs can help researchers identify potential sources of bias before data collection (e.g., Faries et al., 2025).

However, implementing sensitivity analysis within m-DAGs is nontrivial. A central question is how robust identification results are to plausible violations of the assumed graph. For instance, if an edge is mistakenly omitted or added - representing an incorrect independence assumption - what is the impact on inference? Furthermore, when unmeasured confounders affect both the missingness indicator and the outcome, how can one quantify the resulting bias? If extensive sensitivity analyses are required to account for these possibilities, does the approach become infeasible? Developing a structured methodology for sensitivity analysis within the m-DAG framework is crucial for ensuring its reliability and accountability in practical

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0165

applications.

2.3 Comparison with Alternative Identification Strategies

In the current framework, rank preservation has been used to strengthen identification results. However, an open question remains: are there alternative assumptions that could further enhance identification? Common assumptions in econometrics and statistics, such as monotonicity and convexity, may offer useful alternatives (e.g., the use of non-response shadow variables). Other strategies involve structural equation modeling (Miao et al., 2022), where identification is assessed based on the number of equations and parameters, or leveraging inverse problem-solving techniques (Yang et al., 2019). Each identification strategy comes with distinct strengths and limitations. A key challenge is determining how to balance these methods in practice and under what conditions m-DAGs offer the most advantages.

2.4 Integrating Causal Inference and m-DAGs

An intriguing insight from the paper is that m-DAGs may provide enhanced identification results compared to conventional causal inference DAGs (CI-DAGs). This raises an important question: how can causal inference methods be effectively integrated with m-DAGs to address missing data prob-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0165

lems, which are increasingly prevalent in applied research?

Prior work suggests that the timing of missingness and treatment assignment is critical in causal analysis. For example, Yang et al. (2019) employed a conditional independence assumption and completeness to establish identification. Other studies, such as Chu et al. (2025), introduced shadow variables to estimate individual treatment regimes with one-sided feedback, incorporating additional assumptions to identify the expected potential outcome under a given treatment regime. A key question is whether m-DAGs retain their advantages under such scenarios. Further research is needed to determine the extent to which m-DAGs can be effectively combined with existing causal inference methodologies to improve identification in complex missing data problems.

3. Conclusion

The paper "Causal and Counterfactual Views of Missing Data Models" offers a substantial contribution to the theory and practice of missing data analysis by reframing it through the lens of causal inference. By introducing missing data DAGs (m-DAGs) and leveraging counterfactual reasoning, the authors develop a principled framework for addressing challenges specific to Missing Not at Random (MNAR) mechanisms. This integration of causal

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0165

and missing data methodologies provides new insights into identification and offers a pathway to more interpretable and robust analyses.

Future research should prioritize practical applications, including the integration of m-DAGs with modern computational methods and machine learning techniques. Additionally, addressing challenges related to scalability and assumption validation remains critical for broader adoption. While this framework presents a promising theoretical foundation, further empirical studies are necessary to assess its practical advantages over traditional missing data methods.

Acknowledgements

We thank professor John Stufken for inviting us to write a discussion paper. Yang was partially supported by the National Science Foundation grant SES 2242776 and the National Institutes of Health grants 1R01ES031651 and 1R01HL169347. Kim was partially supported by a Cooperative Agreement between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University and by a grant from National Science Foundation (Award No: 2242820).

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0165

References

Chu, J., S. Yang, W. Lu, and P. Ghosh (2025). Efficient causal decision making with one-sided feedback. In *The Proceedings of the 30th International Conference on Learning Representations (ICLR)*.

Ding, P., Y. Fang, D. Faries, S. Gruber, H. Lee, J.-Y. Lee, P. Mishra-Kalyani, M. Shan, M. van der Laan, S. Yang, and X. Zhang (2023). Sensitivity analysis for unmeasured confounding in medical product development and evaluation using real world evidence. *arXiv preprint arXiv:2307.07442*.

Faries, D., C. Gao, X. Zhang, C. Hazlett, J. Stamey, S. Yang, P. Ding, M. Shan, K. Sheffield, and N. Dreyer (2025). Real effect or bias? best practices for evaluating the robustness of real-world evidence through quantitative sensitivity analysis for unmeasured confounding. *Pharmaceutical Statistics 24*, e2457.

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature 58*, 1129–1179.

Miao, W., W. Hu, E. L. Ogburn, and X. H. Zhou (2022). Identifying

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0165

effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association 118*, 1953–1967.

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

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