

Exposure-induced mediator–outcome confounders in causal mediation: Implications and visualization

Etsuji Suzuki, MD, PhD¹, Tomohiro Shinozaki, PhD^{2, 3}, Eiji Yamamoto, PhD⁴

¹ Department of Epidemiology, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan

² Interfaculty Initiative in Information Studies, the University of Tokyo, Tokyo, Japan

³ Department of Biostatistics, School of Public Health, Graduate School of Medicine, the University of Tokyo, Tokyo, Japan

⁴ Okayama University of Science, Okayama, Japan

Corresponding author:

Etsuji Suzuki, Department of Epidemiology, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, 2-5-1 Shikata-cho, Kita-ku, Okayama 700-8558, Japan.

Email: etsuji-s@cc.okayama-u.ac.jp

16-digit ORCID:

Etsuji Suzuki: 0000-0002-1290-5793

1 Tomohiro Shinozaki: 0000-0003-3395-9691

2

3 **Acknowledgments:** None.

4 **Contributors:** ES drafted the manuscript and prepared the figure. TS and EY critically revised the
5 manuscript for important intellectual content. All authors approved the final version for publication.

6 **Funding:** ES received funding from the Japan Society for the Promotion of Science (KAKENHI Grant
7 Number JP23K09740). TS received funding from the Japan Society for the Promotion of Science
8 (KAKENHI Grant Number JP24K14864).

9 **Competing interests:** None declared.

10

1 In their thoughtful review on causal mediation methods, Pearce et al.¹ highlighted the challenges posed
 2 by exposure-induced mediator–outcome confounders for identifying natural direct and indirect effects.²
 3 Here, we briefly provide a broader perspective on this issue using causal directed acyclic graphs (DAGs).

4 We let A denote an exposure, Y an outcome, M a mediator, and L an exposure-induced
 5 mediator–outcome confounder (Figure 1a). In the counterfactual framework, we let $Y(a, m)$ denote
 6 the potential outcomes of Y if, possibly contrary to fact, there had been interventions to set A to a and
 7 to set M to m . Similarly, we let $M(a)$ and $L(a)$ denote the potential outcomes of M and L ,
 8 respectively, had A been set to a . We assume that positivity and consistency hold.

9 In nonparametric structural equation models with independent errors (NPSEM-IE), the natural
 10 direct and indirect effects (i.e., $E[Y(a, M(a'))] - E[Y(a', M(a'))]$ and $E[Y(a, M(a))] -$
 11 $E[Y(a, M(a'))]$, respectively) can be identified if the following four assumptions hold:³

12

$$13 \quad Y(a, m) \perp\!\!\!\perp A \ (\forall a, m), \quad (1)$$

$$14 \quad Y(a, m) \perp\!\!\!\perp M(a) | A = a \ (\forall a, m), \quad (2)$$

$$15 \quad M(a) \perp\!\!\!\perp A \ (\forall a), \quad (3)$$

$$16 \quad Y(a, m) \perp\!\!\!\perp M(a') \ (\forall a, a', m). \quad (4)$$

17

18 Unlike assumptions (1) to (3), assumption (4) is the so-called cross-world independence assumption

1 because it involves counterfactuals referring to two different “worlds” or scenarios. In Figure 1a,
 2 although assumptions (1) and (3) hold, assumptions (2) and (4) *generally* do not. This point can be
 3 visually illustrated by incorporating the potential outcomes and the error terms ε in the causal DAG,
 4 which is constructed under the NPSEM-IE (Figure 1b).⁴ For example, assumption (1) holds because there
 5 are no open paths between $Y(a, m)$ and A . By contrast, assumption (2) does not generally hold
 6 because, among those with $A = a$, there is an open path between $Y(a, m)$ and $M(a)$: $Y(a, m) \leftarrow$
 7 $L(a) \rightarrow M(a)$. This is based on the rule of *weak faithfulness* of causal DAGs, while assumption (2) never
 8 holds under the assumption of *faithfulness*. Similarly, assumption (4) does not generally hold because
 9 there is an open path between $Y(a, m)$ and $M(a')$: $Y(a, m) \leftarrow L(a) \leftarrow \varepsilon_L \rightarrow L(a') \rightarrow M(a')$. Note
 10 that the L -conditional cross-world independence assumption does not generally hold because,
 11 conditional on L , there are two open paths between $Y(a, m)$ and $M(a')$: $Y(a, m) \leftarrow L(a) \leftarrow \varepsilon_L \rightarrow$
 12 $L(a') \rightarrow M(a')$ and $Y(a, m) \leftarrow L(a) \rightarrow L \leftarrow L(a') \rightarrow M(a')$.

13 In conclusion, to clarify the statement in their Table 1,¹ if there is an exposure-induced
 14 mediator–outcome confounder L , the natural direct and indirect effects are not *generally* identified
 15 irrespective of whether data are available on L , except under strong assumptions, such as no interaction
 16 between the exposure and mediator at the individual level.⁵ For further details, see Suzuki et al.⁴

17

1 **REFERENCES**

- 2 1 Pearce A, Hope S, Green MJ, et al. Causal mediation approaches for understanding pathways to
3 inequalities and policy entry points: examples from early years health and development. *J Epidemiol*
4 *Community Health* 2025;jech-2025-224260.
- 5 2 Suzuki E. L or M_1 —Critical challenges in mediation analysis. *Epidemiology* 2025;**36**:686–9.
- 6 3 Pearl J. Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty in*
7 *Artificial Intelligenc*. San Francisco, CA: Morgan Kaufmann 2001:411-20.
- 8 4 Suzuki E, Shinozaki T, Yamamoto E. Natural effects and separable effects: Insights into mediation
9 analysis. *Current Epidemiology Reports* 2025;**12**:20.
- 10 5 Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. In:
11 Green PJ, Hjort NL, Richardson S, eds. *Highly Structured Stochastic Systems*. New York, NY: Oxford
12 University Press 2003:70-81.

13

14

1 **Figure legend**

2 **Fig. 1** (a) A causal directed acyclic graph (DAG) with exposure A , mediator M , and outcome Y when there
3 is an exposure-induced mediator–outcome confounder L . (b) A causal DAG incorporating the potential
4 outcomes as well as error terms. Note that an arrow exerts from each potential outcome to the
5 corresponding observed variable, such that each observed variable has its direct causal variable(s) and
6 the corresponding potential outcomes as parents. The observed variables and their potential outcomes
7 are endogenous variables, whereas the error terms are exogeneous variables in the system of structural
8 equations. For details, see Figure 3 and the related explanations in Suzuki et al.⁴

9