American Journal of Epidemiology, 2025, 00, 1–3
https://doi.org/10.1093/aje/kwae337
Advance access publication date September 4, 2024
Commentary

A note on "Simple graphical rules to assess selection bias in general-population and selected-sample treatment effects" by M. B. Mathur and I. Shpitser

Elena Stanghellini*[,1], Marco Doretti[2], Taiki Tezuka[3]

[1]Department of Economics, University of Perugia, 06100 Perugia, Italy
[2]Department of Statistics, Computer Science, and Applications, University of Florence, 50134 Florence, Italy
[3]Department of Mathematics, Physics, Electrical Engineering and Computer Science, Yokohama National University, Yokohama 240-8501, Japan

*Corresponding author: Elena Stanghellini, Department of Economics, University of Perugia (elena.stanghellini@unipg.it)

Abstract

This short note is a commentary on a 2024 article by Mathur and Shpitser in the *Journal*, with the aim to enlarge the class of graphs for which the conditional average treatment effect is nonparametrically identified, by allowing the outcome to be on the pathway between the treatment and the selection indicator. A first straightforward generalization is possible when (1) the outcome Y is binary, and (2) the population prevalence of Y is known a priori or can be made the object of a sensitivity analysis. Furthermore, identification of the effect is possible also for Y having any nature, provided that a selection bias breaking node V exists and the population prevalence of V is known.

Key words: selection bias; causal estimands.

Let A be the treatment and Y be the outcome of interest. Let R be the selection indicator that takes a value of 1 if the unit is selected into the analysis. The interest is the conditional average treatment effect $\delta_C = E[Y(a_1)|C] - E[Y(a_0)|C]$, where a_i, $i \in \{0, 1\}$, are two treatment levels of interest. Sufficient conditions for the identification of δ_C are provided in the article by Mathur and Shpitser.[1] It immediately can be seen that whenever Y is on the pathway between A and R, the sufficient conditions are not met. However, if additional information is available, identification of δ_C is possible. A notable example comes from observational studies on the determinants of post-acute or long COVID (PLC) syndrome, in which patients who are experiencing the syndrome are more likely to participate than others.[2] In this situation, R is influenced by the outcome, which, in turn, is influenced by the treatment. Later in this article, we show that, provided other less stringent conditions are met, δ_C can be identified.

A simple example

Figure 1 presents two simple directed acyclic graphs (DAGs) such that the one in Figure 1a meets the sufficient conditions in the Mathur and Shpitser article[1] for identification of δ_C, $C = \varnothing$, whereas the DAG in Figure 1b does not. However, if Y is binary, then from standard probability results:

$$\frac{P(Y=1|A=a_i)}{P(Y=0|A=a_i)} = \frac{P(Y=1|A=a_i,R=1)}{P(Y=0|A=a_i,R=1)} \frac{P(R=1|Y=0,A=a_i)}{P(R=1|Y=1,A=a_i)}. \quad (1)$$

It then follows from the case-control literature[3] that, if $R \coprod A \mid Y$, as in Figure 1b, when additional knowledge of the prevalence of Y in the population, $\pi^Y = P(Y=1)$, is available or can be made the object of sensitivity analysis, then $P(Y=1|A=a_i)$ is identified via inversion of eqn (1), as follows:

$$\frac{P(R=1|Y=0,A=a_i)}{P(R=1|Y=1,A=a_i)} = \frac{P(R=1|Y=0)}{P(R=1|Y=1)} = \frac{\pi^Y}{1-\pi^Y} \frac{P(Y=0|R=1)}{P(Y=1|R=1)},$$

where the last term is available from the observable information. Therefore δ_C is identified: $C = \varnothing$. Notice that because Y is binary, other causal estimands can be of interest. With reference to the odds ratio of A against Y [OR(A, Y)], one sees that, in both DAGs, the OR(A, Y) can be recovered from the observable distribution.[4,5] Theorem 1 in an article by Bareinboim and Pearl[6] provides a complete graphical condition for OR(A, Y | C) to be identified, with

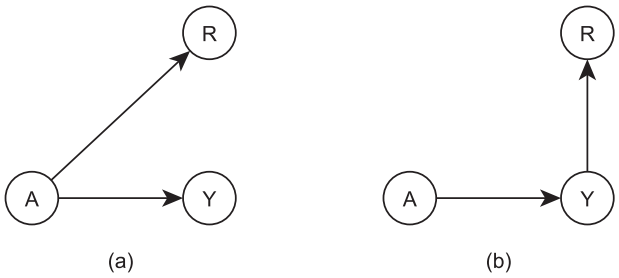

Figure 1. Two data-generating processes with selection of units into the analysis. A is the treatment; Y is the binary outcome; R is the indicator of selection. The average treatment effect is (a) identified without additional knowledge or (b) identified if the population prevalence of Y is known.

Received: April 4, 2024. **Accepted:** August 14, 2024

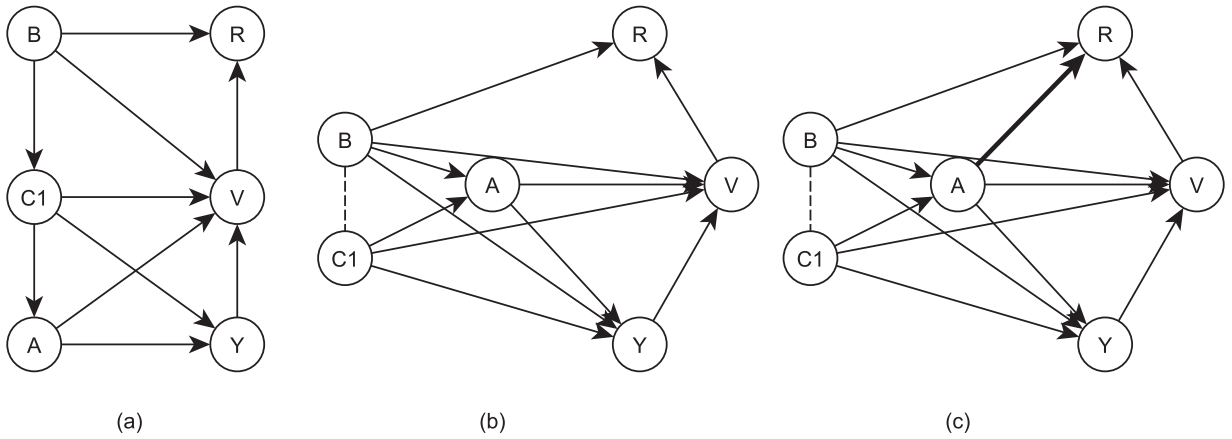

Figure 2. Three data-generating processes with selection of units into the analysis. A is the treatment; Y is the outcome; R is the indicator of selection; V is a selection bias breaking node; and B, C_1 are background covariates. Let C be the confounding-sufficient set of covariates. For V binary and π_b^V known: (a) a directed acyclic graph (DAG) such that $C = C_1$ and δ_C is identifiable; (b) a DAG such that $C = B \cup C_1$, with the marginal distribution of B and C_1 left unspecified, and such that δ_C is identifiable; (c) the same DAG as in (b) with the forbidden arrow marked thick (see "More Complex Structures" in the text).

C being an opportunely defined set of covariates. Extension to Y ordinal is straightforward.

More complex structures

Following the work of Doretti et al.,[7] we introduce the notion of selection bias breaking node. Let C be a set of covariates influencing Y and B be a set of discrete/categorical covariates directly influencing the selection node R. Let $C' = C \backslash B$. V is a selection bias breaking node with respect to B if: (1) V is binary (2) $R \coprod \{A, C'\} \mid \{V, B\}$ and (3) $Y \coprod R \mid \{V, B, A, C'\}$. If the population prevalence $\pi_b^V = P(V = 1 | B = b)$ is known, then the following algorithm, adapted from Doretti et al.,[7] shows that nonparametric identification of $E(Y|A = a_i, B = b, C' = c')$ is possible for Y having any nature. Let $k_b = \frac{1-\pi_b^V}{\pi_b^V} \frac{P(V=1|R=1,B=b)}{P(V=0|R=1,B=b)}$. For $i \in \{0, 1\}$:

1) Take the odds transforms of the observable probabilities $P(V = 1|a_i, b, c', R = 1)$;
2) Use knowledge of k_b to identify the odds transforms of $P(V = 1|a_i, b, c')$ via

$$\frac{P(V = 1|a_i, b, c')}{P(V = 0|a_i, b, c')} = \frac{P(V = 1|a_i, b, c', R = 1)}{P(V = 0|a_i, b, c', R = 1)} \times \frac{1}{k_b};$$

3) Revert to the probability scale to identify $P(V = 1|a_i, b, c')$;
4) Identify $E(Y|a_i, b, c', v) = E(Y|a_i, b, c', v, R = 1)$;
5) Identify

$$E(Y|a_i, b, c') = \sum_v E(Y|a_i, b, c', v) P(V = v|a_i, b, c').$$

If C is the confounding-sufficient set of covariates such that $Y(a) \coprod A \mid C$ and $D = B \backslash W$, with $W \coprod Y \mid \{A, D, C'\}$, then $E(Y|a_i, b, c') = E(Y|a_i, d, c')$, and then the causal estimand δ_S, $S = C' \cup D$, can be identified. Notice that W is the empty set if and only if $B \subseteq C$ and then $S = C$ (as $D = B$). Furthermore, $W = B$ if and only if D is the empty set and then $S = C\,(= C')$. Notice that the set B has to include all confounders that are directly influencing R, not necessarily all parents of R.

In Figure 2a, a possible DAG leading to identification of δ_C is shown, with $C = C_1$, and Figure 2b presents a DAG leading to identification of δ_C, with $C = B \cup C_1$, such that the marginal distribution of B and C_1 is left unspecified, because the two covariates may be causally ordered or on equal footing. Finally, in Figure 2c, the same DAG, with the forbidden arrow marked thick. If R is directly influenced by A, then A should be added to the set of variables forming the strata for which the prevalence V is available, an instance that is rarely met. Notice that the arrow $C_1 \to R$ may be present. In that case, knowledge of the prevalence $\pi_{b'}^V$ is required to identify $\delta_{B'}$, $B' = C_1 \cup B$.

Furthermore, if Y is binary, the arrow $Y \to R$ is permitted. This is because, conditionally on B and C_1 and marginally with respect to V, the corresponding DAG is as in Figure 1b; therefore, $\delta_{\{B,C_1\}}$ is identified provided the prevalence of Y is known in the strata formed by B and C_1. Notice that this instance is not covered by theorem 4 in the article by Bareinboim and Pearl,[6] in which different sources of additional information are postulated. Furthermore, conditioning on a covariate that is influenced by A is not admitted in the article by Mathur and Shpitser,[1] because additional external knowledge is not considered in their work. Nonetheless, as the following example on PLC data shows, there may be instances where the prevalence of V either is known or can be made the object of a sensitivity analysis.

Determinants of PLC symptoms

Post-acute or long COVID syndrome is a direct sequelae of SARS-CoV-2 infection that can highly compromise quality of life. It includes symptoms that persist from the acute COVID-19 phase or its treatment, symptoms that have resulted in a new health limitation, new symptoms that have occurred after the end of the acute phase but are understood to be a consequence of COVID-19 disease, and worsening of a preexisting underlying condition. See the report by Parotto et al.[8] for the definitions.

Estimating the prevalence, as well as understanding the determinants, of PLC, is crucially important because SARS-CoV-2 continues to circulate. To this end, several studies have been designed. Typically, participants are recruited by simply advertising the study, and people enter on a voluntary basis. This creates a possible bias, because people who are experiencing lasting symptoms after COVID-19 may be more motivated than others to participate. Moreover, patients who are recovering from the syndrome are more likely not to participate at follow-up visits, giving rise to informative dropout.

Let Y be the duration of a symptom of interest in days, taking the value 0 if there are no symptoms. The aim is to assess whether

one treatment strategy $A = a_1$, adopted during the infection, is more prone to have lasting effects on that particular symptom than another treatment strategy, $A = a_0$. Let V be an indicator variable taking value 1 if any PLC symptom is present and R the indicator taking the value 1 if a COVID-19 survivor participates in the study. We focus on assumptions for identification of $E[Y(a_i)]$ in the population of survivors, possibly after conditioning on a set of covariates.

Assume we have a sufficiently rich set of covariates C such that in the population of survivors (1) $Y(a) \coprod A \mid C$. We assume further that (2) the only bias is induced by selection into the study, and no measurement error, truncation, and censoring are present; and (3) a set of background covariates B exists, $B \subseteq C$, such that V is a selection bias breaking node with respect to B. As already mentioned, many studies provide information on the prevalence of PLC symptoms, both in the general population or in strata formed by B. Therefore $E[Y(a_i)|C]$ can be estimated from the observable distribution. Due to the selection bias issues, the prevalence can be overestimated.[2] The procedure we have outlined allows us to address, via sensitivity analysis, the impact on the estimands of π_b^V.

Discussion

We have enlarged the instances of DAGs such that $E[Y(a)|C]$, with C a possible empty set, can be nonparametrically identified from the observable distribution. When causal estimands are the object of inference, careful examination of the assumptions is strongly recommended. With reference to the PLC example, the existence is postulated of a set of individuals who would have survived under both treatment strategies. Because death itself is an outcome of interest, other assumptions may be considered.[9,10] Notice that some problems are better addressed via a parametric modeling strategy, which accommodates the nature of the response variable, such as a truncated normal or zero-inflated Poisson distribution, as well as of the treatment, such as a discrete or a continuous one. Extension to an ordinal selection bias breaking node V is also possible.

Funding

Support by "Fondazione Perugia" within the project "Danni permanenti dell'infezione da SARS-CoV-2" is gratefully acknowledged.

Conflict of interest

The authors declare no conflicts of interest.

References

1. Mathur MB, Shpitser I. Simple graphical rules for assessing selection bias in general-population and selected-sample treatment effects. *Am J Epidemiol*. 2024. https://doi.org/10.1093/aje/kwae145
2. Høeg TB, Ladhani S, Prasad V. How methodological pitfalls have created widespread misunderstanding about long COVID. *BMJ Evid Based Med*. 2024;29(3):142-146. https://doi.org/10.1136/bmjebm-2023-112338
3. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979;66(3):403-411. https://doi.org/10.1093/biomet/66.3.403
4. Cornfield J. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst*. 1951;11(6):1269-1275. https://doi.org/10.1093/jnci/11.6.1269
5. Whittemore A. Collapsibility of multidimensional contingency tables. *J R Stat Soc Series B Stat Methodol*. 1978;40(3):328-340. https://doi.org/10.1111/j.2517-6161.1978.tb01046.x
6. Bareinboim E, Pearl J. Controlling selection bias in causal inference. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, PMLR*. MLRPress; 2012;22:100-108.
7. Doretti M, Genbäck M, Stanghellini E. Mediation analysis with case–control sampling: identification and estimation in the presence of a binary mediator. *Biom J*. 2024;66(1):e2300089. https://doi.org/10.1002/bimj.202300089
8. Parotto M, Gyöngyösi M, Howe K, et al. Post-acute sequelae of COVID-19: understanding and addressing the burden of multisystem manifestations. *Lancet Respir Med*. 2023;11(8):739-754. https://doi.org/10.1016/S2213-2600(23)00239-4
9. Tcheghen Tcheghen E. Identification and estimation of survivor average causal effects. *Stat Med*. 2014;33(21):3601-3628. https://doi.org/10.1002/sim.6181
10. Tong G, Li F, Chen X, et al. Bayesian approach for estimating the survivor average causal effect when outcomes are truncated by death in cluster-randomized trials, 2023. *Am J Epidemiol*. 2023;192(6):1006-1015. https://doi.org/10.1093/aje/kwad038

American Journal of Epidemiology, 2025, **00**, 1–3
https://doi.org/10.1093/aje/kwae337
Commentary