FAST FACTS

Selection bias due to conditioning on a collider

Effect estimates may be biased when the study design or the data analysis is conditional on a collider—a variable that is caused by two other variables. Causal directed acyclic graphs are a helpful tool to identify colliders that may introduce selection bias in observational research.

Miguel A Hernán,[1] Susana Monge[2, 3]

[1] CAUSALab and Departments of Epidemiology and Biostatistics, Harvard T H Chan School of Public Health, Boston, MA, USA

[2] Department of Communicable Diseases, National Centre of Epidemiology, Institute of Health Carlos III, 28029 Madrid, Spain

[3] Consortium for Biomedical Research in Infectious Diseases (CIBERINFEC), Spain

Correspondence to: S Monge
smonge@isciii.es
https://orcid.org/0000-0003-1412-3012

Cite this as: *BMJ* 2023;381:p1135
http://dx.doi.org/10.1136/bmj.p1135
Published: 07 June 2023

Definition of a collider

In a causal graph, a collider is a variable that is affected by two or more variables on the graph.[1 2] For example, suppose a causal diagram has three variables: vaccine V (yes or no) at baseline, infection I (yes or no) during the subsequent six months, and individual, pre-baseline susceptibility S to infection (high, medium, low) (fig 1, top graph). There would be an arrow from V to I because the vaccine lowers the risk of infection, and an arrow from S to I because susceptibility increases the risk of infection. Therefore, the variable I is a collider because arrows go into it from two other variables: the arrows from V and S "collide" into I. Identifying colliders is important because conditioning on colliders is expected to lead to selection bias.[1 3]

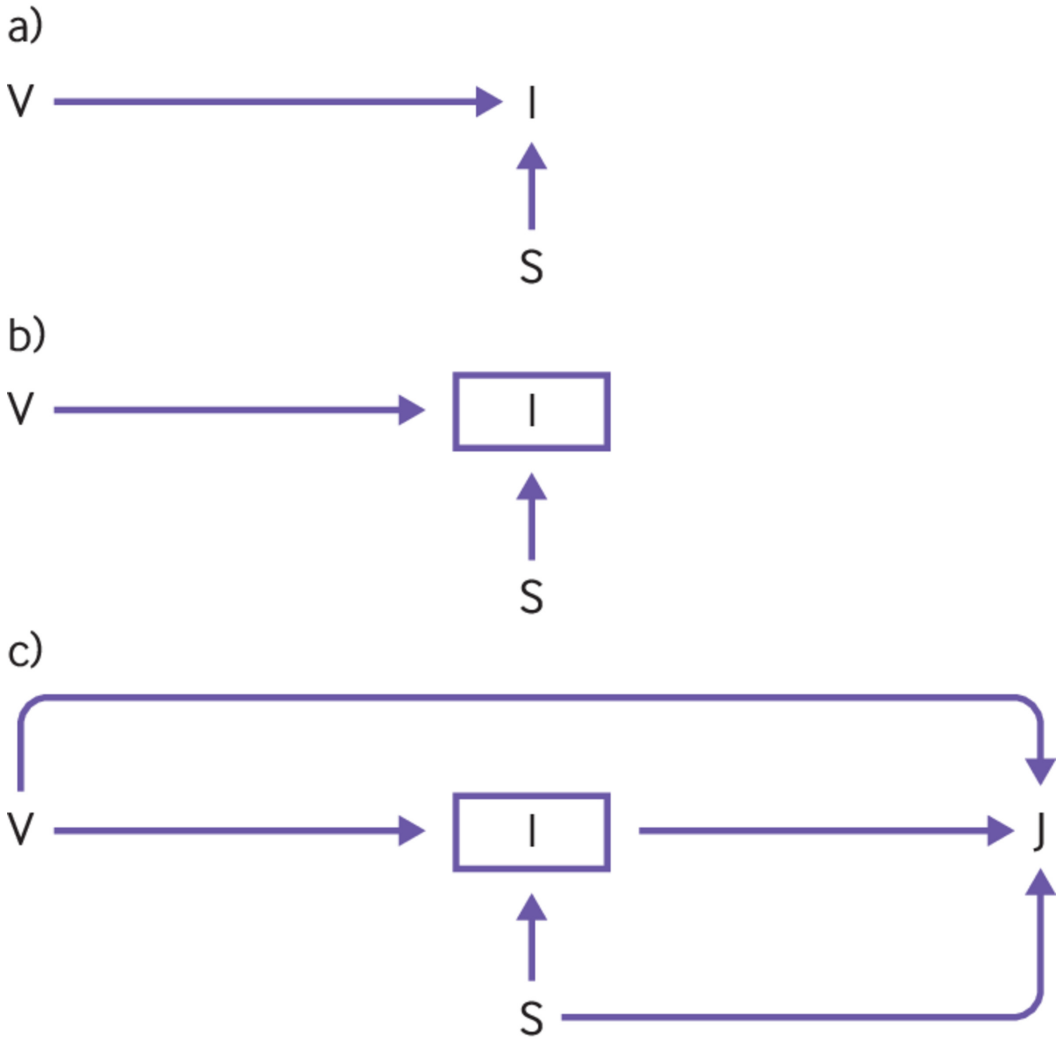

Fig 1 | Three types of causal graphs. I=infection; J=fourth variable; S=susceptibility to infection; V=vaccine

Colliders and selection bias

Suppose that a randomised trial is performed in which V is randomly assigned to individuals in the study population. In the data, no association is expected between V and S because the random assignment at baseline implies that the distribution of pre-baseline susceptibility is the same between the vaccinated group and unvaccinated group—that is, even though V and S are graphically connected through the collider I, V and S are not associated because two variables measured at baseline cannot become associated by a third variable, the collider, that occurs in the future (fig 1, top graph). In the jargon of graph theory, it is said that a path between two variables is blocked by a collider on the path.

Now consider what happens if only people who were infected in the first six months are selected for analysis—that is, the analysis is conditional on I=yes (fig 1, middle graph)—as was done in the analysis of recent studies of immune imprinting of covid-19 vaccines.[4 5] An individual who becomes infected despite receiving a vaccine dose is more likely to have high susceptibility to infection than an unvaccinated individual who becomes infected. Therefore, among infected people, a greater proportion of people with high susceptibility to infection would be expected among those who received the vaccine than among those who did not receive the vaccine—that is, V and S become associated when the analysis is conditional on I=yes even though V and S were not associated when the analysis included the entire population of the randomised trial. In the jargon of graph theory, it is said that a path between two variables is not blocked by a collider on the path when the collider is conditioned on.

When conditioning on the collider I, a V-S association arises due to selecting a common effect of V and S. This conditional association does not correspond to the effect of V on S (there is no effect of V on S because S predates V) or to the effect of S on V (there is no effect of S on V because V was randomly assigned); rather, the conditional V-S association in the stratum I=yes is a selection bias with no causal interpretation. The conditional association is also expected to arise when conditioning on I=no, although there are mathematical conditions under which conditioning on one (and only one) of the values of a collider may not induce an association between its causes.[6]

Colliders and direct effects

If a fourth variable, J, is added to the causal graph: infection between six months and 12 months after randomisation (yes or no) (fig 1, bottom graph), an arrow from I and S points into J because both previous infection and individual susceptibility affect the risk of subsequent infection. Suppose there is interest in the direct effect of V on J that is not mediated by I. Let it be assumed that researchers are unaware that the direct effect is null because the vaccine effect wanes until disappearing at six months, and thus no direct arrow goes from V to J. In an attempt to estimate the direct effect of the vaccine V on late infection J that is not mediated by early infection I, an analysis restricted to individuals who were infected early could naively be carried out. However, selecting those with I=yes leads to an association between V and S and, because S is associated with J, between V and J. An incorrect interpretation of that conditional association would be that V has a direct effect on J because, in reality, the association between V and J is the result of selection bias.

The big picture

Associations created by colliders are everywhere. For example, conditioning on a collider explains the following two statements[7]: "Among successful actors, being physically attractive is inversely related to being a good actor," and "Among American college students, being academically gifted is inversely related to being good at sport." The collider is acting success (yes or no) in the first example and admission to a US college (yes or no) in the second example.

In fact, selection biases such as the one described may arise in any research setting in which the study design or the data analysis is conditional on a collider. This form of selection bias largely explains, for example, the association reported between postmenopausal hormone treatment and coronary heart disease,[8] the birth weight paradox,[9] and the obesity paradox.[10] Also, many commonly used methods for adjustment of confounders, including regression, rely on estimating associations conditional on covariates. As a result, a causally blind selection of adjustment covariates may introduce selection bias if some of those covariates are colliders rather than confounders.[6]

Funding and competing interests available in the linked paper on bmj.com.

1 Pearl J. *Causality*. Cambridge University Press, 2009doi: 10.1017/CBO9780511803161.
2 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;10:-48. doi: 10.1097/00001648-199901000-00008 pmid: 9888278
3 Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15:-25. doi: 10.1097/01.ede.0000135174.63482.43 pmid: 15308962
4 Monge S, Pastor-Barriuso R, Hernán MA. The imprinting effect of covid-19 vaccines: an expected selection bias in observational studies. *BMJ* 2023;381:e074404.
5 Chemaitelly H, Ayoub HH, Tang P, etal. COVID-19 primary series and booster vaccination and immune imprinting.*medRxiv* 2022;2022.10.31.22281756doi: 10.1101/2022.10.31.22281756.
6 Hernán MA, Robins JM. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.
7 Sterne J. We should be cautious about associations of patient characteristics with COVID-19 outcomes that are identified in hospitalised patients. (Cited 2023 Feb 7.) https://www.hdruk.ac.uk/news/we-should-be-cautious-about-associations-of-patient-characteristics-with-covid-19-outcomes-that-are-identified-in-hospitalised-patients/
8 Hernán MA, Alonso A, Logan R, etal. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008;19:-79. doi: 10.1097/EDE.0b013e3181875e61 pmid: 18854702
9 Hernández-Díaz S, Schisterman EF, Hernán MA. The birth weight "paradox" uncovered?*Am J Epidemiol* 2006;164:-20. doi: 10.1093/aje/kwj275 pmid: 16931543
10 Lajous M, Bijon A, Fagherazzi G, etal. Body mass index, diabetes, and mortality in French women: explaining away a "paradox". *Epidemiology* 2014;25:-4. doi: 10.1097/EDE.0000000000000031 pmid: 24270963