Causal inference is not a statistical problem

Lucy D'Agostino McGowan

Introduction

Anscombe's quartet is a set of four datasets that have the same statistical properties in terms of summary statistics (means, variances, correlations, and linear regression fits), but exhibit different distributions and relationships when plotted on a graph (Anscombe 1973). The quartet, often used to teach introductory statistics courses, was created to illustrate the importance of visualizing data before drawing conclusions based on statistical analyses alone. Here, we propose a different quartet, where again statistical summaries do not provide insight into the true underlying mechanism, but even visualizations do not solve the issue. In these examples an understanding or assumption of the data generating mechanism is required to correctly capture the relationship between the available factors. This proposed quartet is meant to help practitioners better understand the assumptions underlying causal inference methods, further driving home the point that in order to accurately estimate causal effects we require more information than can be gleaned from statistical tools alone.

The data generated to create the figures displayed here are included in an R package titled `quartet`.

Methods

We propose the following four data generation mechanisms, summarized by the equations below as well as the directed acyclic graphs displayed in Figure 1. Here, X is presumed to be some continuous exposure of interest, Y a continuous outcome, and Z a known, measured factor. The M-Bias equation includes two additional unmeasured factors, U_1 and U_2.

(1) Collider:

$$\begin{aligned} &X \sim N(0,1) \\ &Y = X + \varepsilon_y,\ \varepsilon_y \sim N(0,1) \\ &Z = 0.45X + 0.77Y + \varepsilon_z,\ \varepsilon_z \sim N(0,1) \end{aligned} \tag{1}$$

(2) Confounder:

$$\begin{aligned}
&Z \sim N(0,1)\\
&X = Z + \varepsilon_x,\ \varepsilon_x \sim N(0,1)\\
&Y = 0.5Z + \varepsilon_y,\ \varepsilon_y \sim N(0,1)
\end{aligned} \tag{2}$$

(3) Mediator:

$$\begin{aligned}
&X \sim N(0,1)\\
&Z = X + \varepsilon_z,\ \varepsilon_z \sim N(0,1)\\
&Y = Z + \varepsilon_y,\ \varepsilon_y \sim N(0,1)
\end{aligned} \tag{3}$$

(4) M-Bias:

$$\begin{aligned}
&U_1 \sim N(0,1)\\
&U_2 \sim N(0,1)\\
&Z = 8U_1 + U_2 + \varepsilon_z,\ \varepsilon_z \sim N(0,1)\\
&X = U_1 + \varepsilon_x,\ \varepsilon_x \sim N(0,1)\\
&Y = X + U_2 + \varepsilon_y,\ \varepsilon_y \sim N(0,1)
\end{aligned} \tag{4}$$

In each of these scenarios, a linear model fit to estimate the relationship between X and Y with no further adjustment will result in a $\hat{\beta}$ coefficient of 1. Additionally, the correlation between X and the additional known factor Z is 0.70.

We have simulated 100 data points from each of the four mechanisms, each is displayed in Figure 2. This set of figures demonstrates that despite the very different data generating mechanisms, there is not a clear way to determine the "appropriate" way to model the effect of the exposure X and the outcome Y without additional information. For example, the unadjusted models are displayed in Figure 2, showing a relationship between X and Y of 1. This is the correct causal model for data generating mechanisms (1) and (4), however it overstates the effect of X for data generating mechanism (2), and describes the total effect of X on Y for data generating mechanism (3), but not the direct effect (Table 1). Indeed, even examining the correlation between X and the known factor Z does not help us determine whether adjusting for Z is appropriate, as it is 0.7 in all cases (Table 2).

Table 1: Correct causal models and causal effects for each data generating mechanism.

Data generating mechanism	Correct causal model	Correct causal effect
(1) Collider	Y ~ X	1
(2) Confounder	Y ~ X + Z	0.5

Data generating mechanism	Correct causal model	Correct causal effect
(3) Mediator	Direct effect: Y ~ X + Z Total Effect: Y ~ X	Direct effect: 0 Total effect: 1
(4) M-Bias	Y ~ X	1

Table 2: Coefficients for the exposure under each data generating mechanism depending on the model fit as well as the correlation between X and Z.

Data generating mechanism	Y ~ X	Y ~ X + Z	Correlation of X and Z
(1) Collider	1	0.41	0.7
(2) Confounder	1	0.50	0.7
(3) Mediator	1	0.00	0.7
(4) M-Bias	1	0.88	0.7

Discussion

Here we have demonstrated that when presented with an exposure, outcome, and some measured factors, statistics alone, whether summary statistics or data visualizations, are not sufficient to determine the appropriate causal estimate. Additional information about the data generating mechanism is needed in order to draw the correct conclusions. While knowledge of the data generating process is necessary in order to estimate the correct causal effect in each of the cases presented, there are steps an analyst can take in order to make mistakes such as those presented here less likely. The first is considering the time ordering of the available factors. Fundamental principles of causal inference dictate that the exposure of interest must precede the outcome of interest to plausibly establish a causal relationship. In addition, to account for potential confounding, any covariates adjusted for in the analysis must precede the exposure in time. Including this additional timing information would omit the potential for two of the three misspecified models above (Equation 1 the "collider" and Equation 3 the "mediator") as the former would demonstrate that the factor Z falls after both the exposure and outcome and the latter would demonstrate that the factor Z falls between the exposure and the outcome in time. Indeed, adjusting for only pre-exposure factors is widely recommended. The only exception is when a known confounder is only measured after the exposure in a particular data analysis, in which case some experts recommend adjusting for it, but even then, caution is advised. (Groenwold, Palmer, and Tilling 2021) In fact many causal inference methodologists would recommend conditioning on *all* measured pre-exposure factors. (Rosenbaum 2002; Rubin 2009, 2008; Rubin and Thomas 1996) Including timing information alone (and thus adjusting for all pre-exposure factors) does not preclude one from mistakenly fitting the adjusted model under the fourth data generating mechanism (M-bias), as Z can fall temporally prior to X and Y and still induce bias. It has been argued, however, that this strict M-bias (e.g. where U_1 and U_2 in Equation 4 have no relationship with each other and

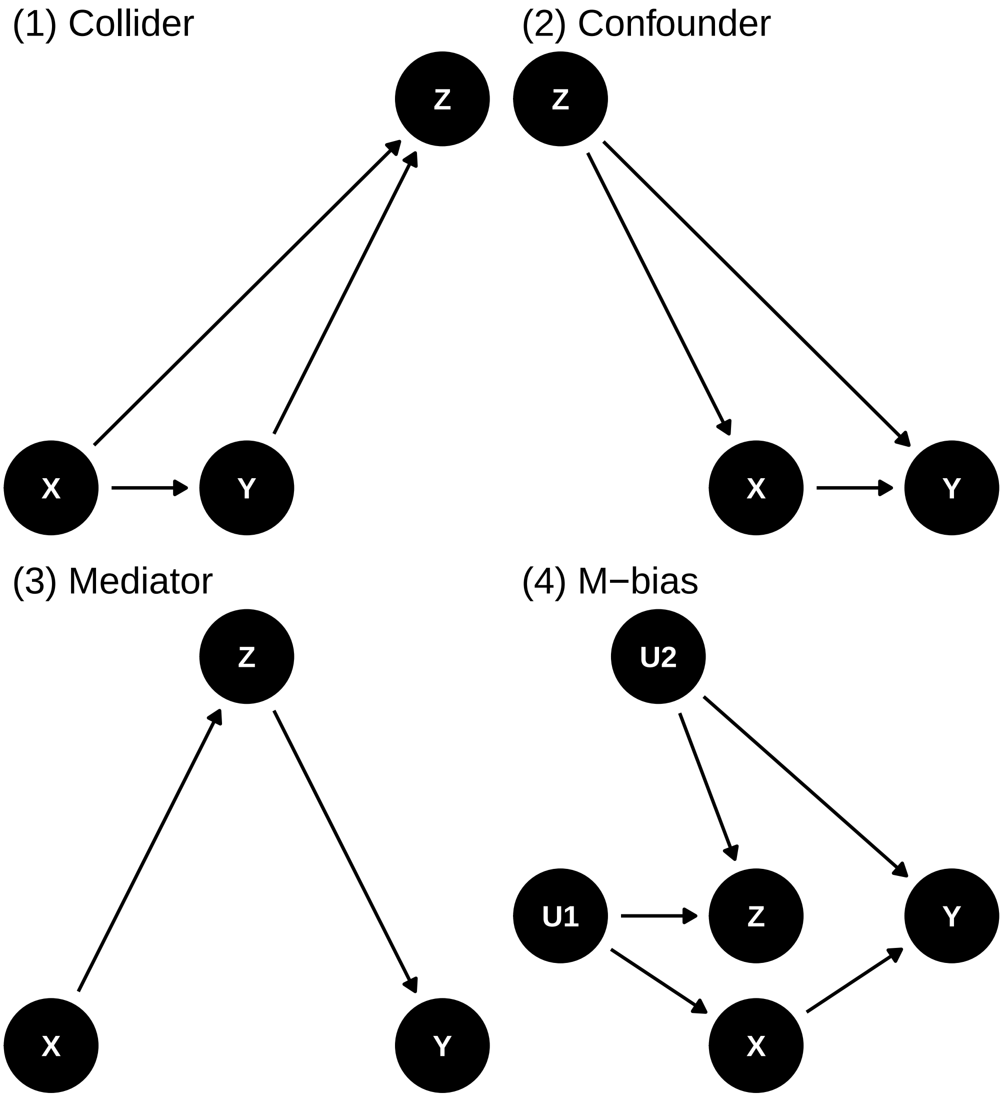

Figure 1: Directed Acyclic Graphs describing the four data generating mechanisms: (1) Collider (2) Confounder (3) Mediator (4) M-Bias.

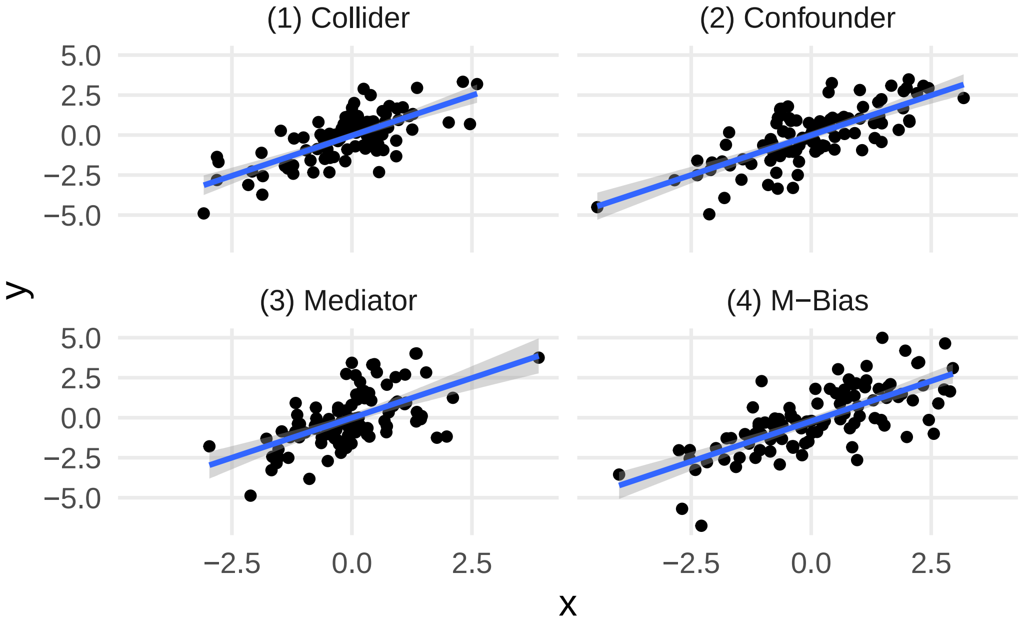

Figure 2: 100 points generated using the data generating mechanisms specified (1) Collider (2) Confounder (3) Mediator (4) M-Bias The blue line displays a linear regression fit estimating the relationship between X and Y, in each case the slope is 1.

Z has no relationship with X or Y other than via U_1 and U_2) is very rare in most practical settings. (Liu et al. 2012; Rubin 2009; Gelman 2011) Indeed, even theoretical results have demonstrated that bias induced by this data generating mechanism is very sensitive to any deviations from this form. (Ding and Miratrix 2015)

We have presented four example datasets demonstrating the importance of understanding the data-generating mechanism when attempting to answer causal questions. These data demonstrate that statistical summaries and visualizations alone will not provide insight into the true underlying relationship between the variables, and that an understanding or assumption of the data-generating mechanism is required to correctly capture causal relationships. These examples underscore the limitations of relying solely on statistical tools in data analyses, and highlight the crucial role of domain-specific knowledge. Moreover, they emphasize the importance of considering the timing of factors when deciding what to adjust for.

References

Anscombe, Francis J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21.

Ding, Peng, and Luke W Miratrix. 2015. "To Adjust or Not to Adjust? Sensitivity Analysis of m-Bias and Butterfly-Bias." *Journal of Causal Inference* 3 (1): 41–57.

Gelman, Andrew. 2011. "Causality and Statistical Learning." University of Chicago Press Chicago, IL.

Groenwold, Rolf HH, Tom M Palmer, and Kate Tilling. 2021. "To Adjust or Not to Adjust? When a 'Confounder' Is Only Measured After Exposure." *Epidemiology (Cambridge, Mass.)* 32 (2): 194.

Liu, Wei, M Alan Brookhart, Sebastian Schneeweiss, Xiaojuan Mi, and Soko Setoguchi. 2012. "Implications of m Bias in Epidemiologic Studies: A Simulation Study." *American Journal of Epidemiology* 176 (10): 938–48.

Rosenbaum, PR. 2002. "Constructing Matched Sets and Strata. Observational Studies." New York, Springer-Verlag.

Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis."

———. 2009. "Should Observational Studies Be Designed to Allow Lack of Balance in Covariate Distributions Across Treatment Groups?" *Statistics in Medicine* 28 (9): 1420–23.

Rubin, Donald B, and Neal Thomas. 1996. "Matching Using Estimated Propensity Scores: Relating Theory to Practice." *Biometrics*, 249–64.

Appendix

R code to generate the tables and figures:

```r
library(tidyverse)
# devtools::install_github("LucyMcGowan/quartet")
library(quartet)

## Figure 2

ggplot(causal_quartet, aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", formula = "y ~ x") +
  facet_wrap(~dataset)

## Table 2

causal_quartet %>%
  nest_by(dataset) %>%
  mutate(`Y ~ X` = round(coef(lm(y ~ x, data = data))[2], 2),
         `Y ~ X + Z` = round(coef(lm(y ~ x + z, data = data))[2], 2),
         `Correlation of X and Z` = round(cor(data$x, data$z), 2)) %>%
  select(-data, `Data generating mechanism` = dataset) %>%
  knitr::kable()
```