Causal inference is not a statistical problem

Lucy D'Agostino McGowan Malcolm Barrett

Abstract

This paper introduces a collection of four data sets, similar to Anscombe's Quartet, that aim to highlight the challenges involved when estimating causal effects. Each of the four data sets is generated based on a distinct causal mechanism: the first involves a collider, the second involves a confounder, the third involves a mediator, and the fourth involves the induction of M-Bias by an included factor. The paper includes a mathematical summary of each data set, as well as directed acyclic graphs that depict the relationships between the variables. Despite the fact that the statistical summaries and visualizations for each data set are identical, the true causal effect differs, and estimating it correctly requires knowledge of the data-generating mechanism. These example data sets can help practitioners gain a better understanding of the assumptions underlying causal inference methods and emphasize the importance of gathering more information beyond what can be obtained from statistical tools alone. The paper also includes R code for reproducing all figures and provides access to the data sets themselves through an R package named quartets.

Introduction

Anscombe's quartet is a set of four data sets with the same summary statistics (means, variances, correlations, and linear regression fits) but exhibit different distributions and relationships when plotted on a graph (Anscombe 1973). Often used to teach introductory statistics courses, Anscombe created the quartet to illustrate the importance of visualizing data before drawing conclusions based on statistical analyses alone. Here, we propose a different quartet, where statistical summaries do not provide insight into the 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 capture the relationship between the available factors correctly. This proposed quartet can help practitioners better understand the assumptions underlying causal inference methods, further driving home the point that we require more information than can be gleaned from statistical tools alone to estimate causal effects accurately.

The data generated to create the figures displayed here are available in an R package titled `quartets` (D'Agostino McGowan 2023).

Causal inference primer

In causal inference, we are often trying to estimate the effect of some exposure, X, on some outcome Y. One framework we use to think through this problem is the "potential outcomes" framework (Rubin 1974). Here, you can imagine that each individual has a set of potential

outcomes under each possible exposure value. For example, if there are two levels of exposure (exposed: 1 and unexposed: 0), we could have the potential outcome under exposure ($Y(1)$) and the potential outcome under no exposure ($Y(0)$) and look at the difference between these, $Y(1) - Y(0)$ to understand the impact on the exposure on the outcome, Y. Of course, at any moment in time, only one of these potential outcomes is observable, the potential outcome corresponding to the exposure the individual actually experienced. Under certain assumptions, we can borrow information from individuals who have received different exposures to compare the average difference between their observed outcomes. We make the assumption that one individual's exposure does not impact the outcome of any other individual. We assume that everyone has a non-zero probability of having each level of the exposure. And finally, we assume that the potential outcomes are independent of the exposure level given the observed covariate(s) *that are adjusted for*. The purpose of this paper is to focus on this Z. Given you have three variables, an exposure, X, an outcome, Y, and some measured factor, Z, how do you decide whether you should estimate the average treatment effect adjusting for Z?

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:

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

(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. Or, equivalently, the estimated average treatment effect (ATE) without adjusting for Z is 1. The correlation between X and the additional known factor Z is also 0.70.

We have simulated 100 data points from each of the four mechanisms; we display each in Figure 2. This set of figures demonstrates that despite the very different data-generating mechanisms, there is no 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. The unadjusted models are 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). 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). The four datasets are available for use in the `quartets` R package (D'Agostino McGowan 2023).

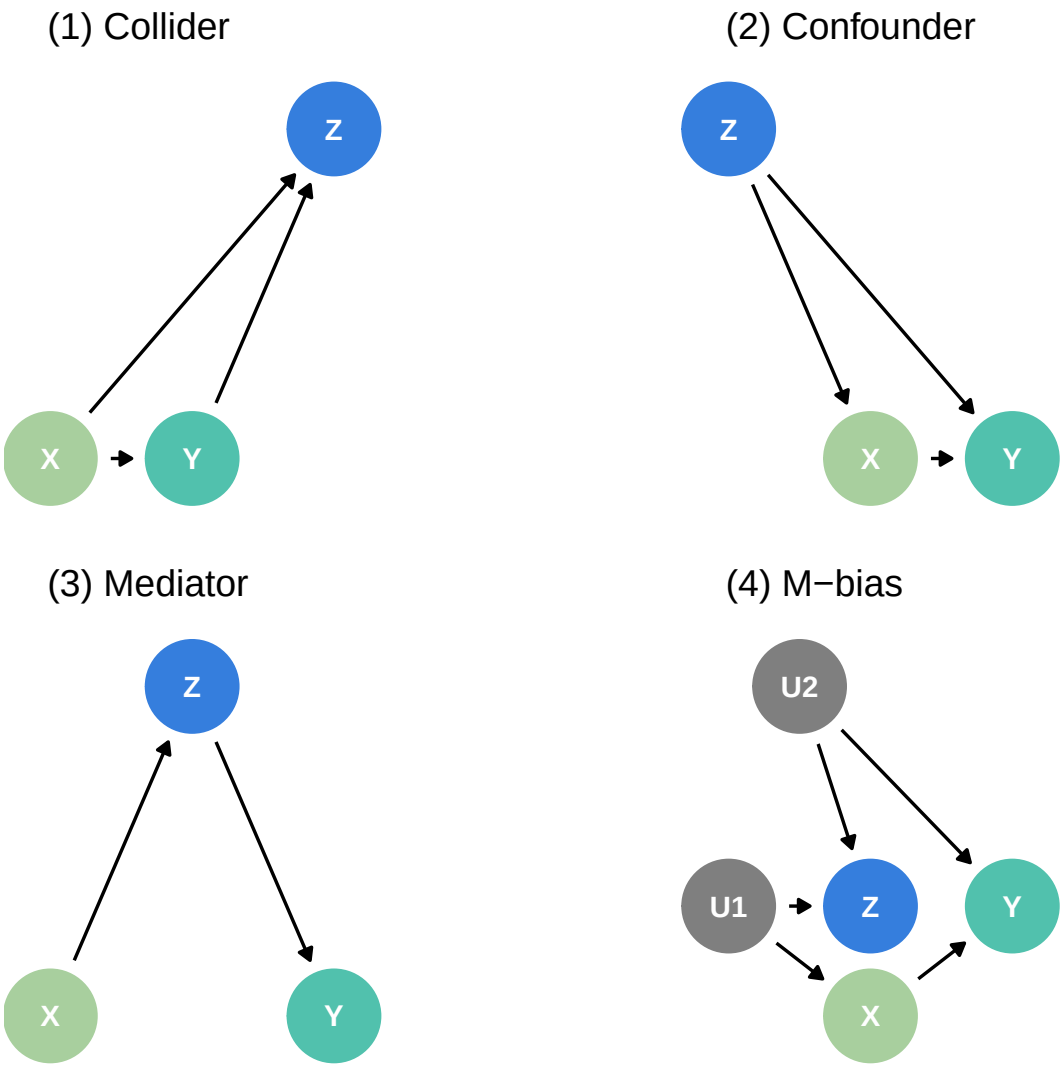


Figure 1: Directed Acyclic Graphs describing the four data generating mechanisms: (1) Collider (2) Confounder (3) Mediator (4) M-Bias.

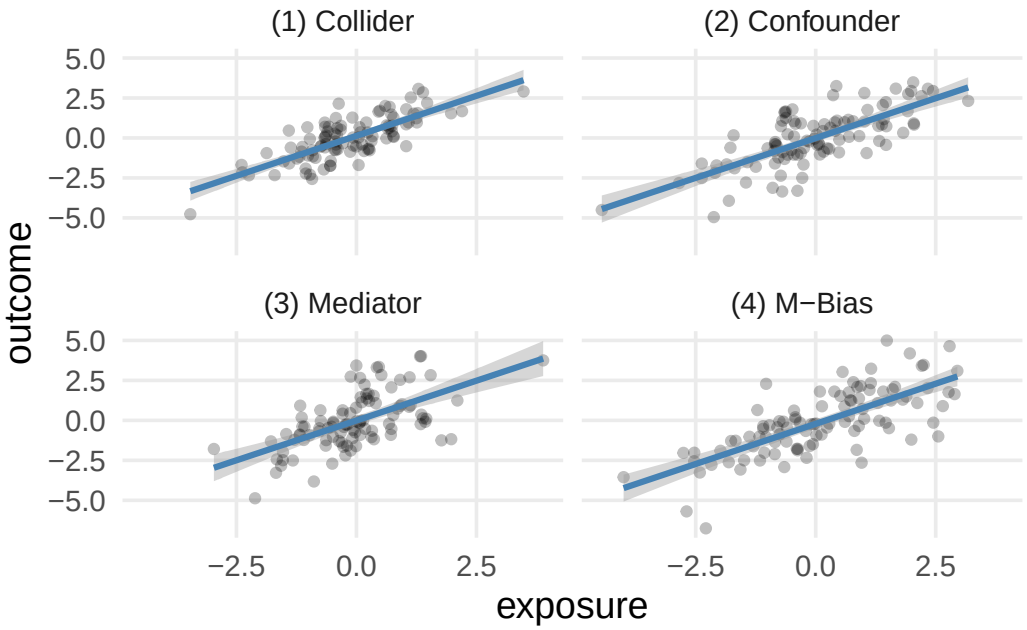


Figure 2: 100 points generated using the data generating mechanisms specified (1) Collider (2) Confounder (3) Mediator (4) M-Bias. The blue line displays a linear regression fit estimating the relationship between X and Y; in each case, the slope is 1.

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	ATE not adjusting for Z	ATE adjusting for Z	Correlation of X and Z
(1) Collider	1	0.55	0.7
(2) Confounder	1	0.50	0.7
(3) Mediator	1	0.00	0.7
(4) M-Bias	1	0.88	0.7

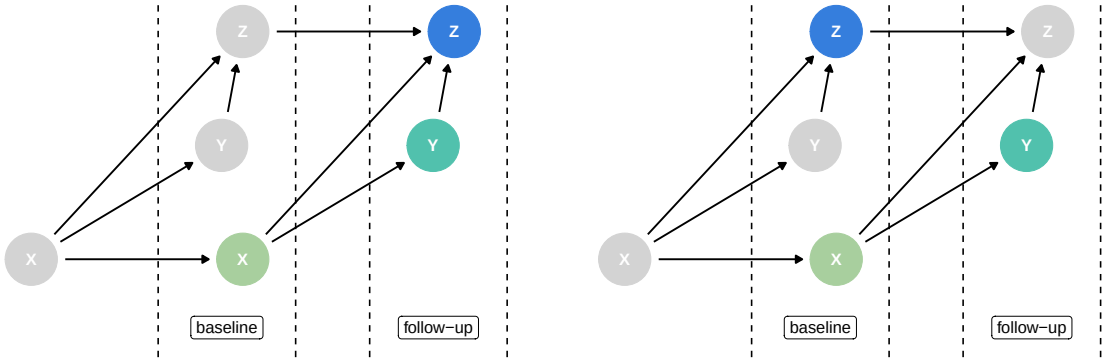
Table 1: Correct causal models and causal effects for each data-generating mechanism.

Data generating mechanism	Correct causal model	Correct causal effect
(1) Collider	Y ~ X	1
(2) Confounder	Y ~ X + Z	0.5
(3) Mediator	Direct effect: Y ~ X + Z	Direct effect: 0
	Total Effect: Y ~ X	Total effect: 1
(4) M-Bias	Y ~ X	1

The Solution

Here we have demonstrated that when presented with an exposure, outcome, and some measured factors, statistics alone, whether summary statistics or data visualizations are insufficient to determine the appropriate causal estimate. Analysts need additional information about the data generating mechanism to draw the correct conclusions. While knowledge of the data generating process is necessary to estimate the correct causal effect in each of the cases presented,

an analyst can take steps to make mistakes such as those shown here less likely. The first is discussing understood mechanisms with content matter experts before estimating causal effects. Drawing the proposed relationships via causal diagrams such as the directed acyclic graphs shown in Figure 1 before calculating any statistical quantities can help the analyst ensure they are only adjusting for factors that meet the "backdoor criterion," that is, adjusting for only factors that close all backdoor paths between the exposure and outcome of interest (Pearl 2000).



(a) Adjusting for Z as shown here would induce collider bias.

(b) Adjusting for this pre-exposure Z as shown here would **not** induce collider bias.

Figure 3: Time-ordered collider DAG where each factor is measured twice. X is the exposure, Y is the outcome, and Z is the measured factor. The highlighted Z node indicates which time point is being adjusted for when estimating the average treatment effect of the highlighted X on the highlighted Y

Absent subject matter expertise, the analyst can at least consider 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 establish a causal relationship plausibly. 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

Table 3: 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	ATE not adjusting for pre-exposure Z	ATE adjusting for pre-exposure Z	Correct causal effect
(1) Collider	1	1.00	1.0
(2) Confounder	1	0.50	0.5
(3) Mediator	1	1.00	1.0
(4) M-Bias	1	0.88	1.0

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 show that the factor Z falls between the exposure and the outcome in time. For example, if we drew the second panel of Figure 1 (the Collider) as a time ordered DAG, we would see something like Figure 3. If we carefully adjust only for factors that are measured pre-exposure, we would not induce the bias we see in Table 2 (Figure 3b). The Causal Quartets data sets are accompanied by a set of four data sets with time-varying measures for each of the factors, X, Y, and Z, generated under the same data generating mechanisms. Here, as long as a pre-exposure measure of Z is adjusted for, the correct causal effect is estimated in all scenarios except M-Bias (Table 3).

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. Still, even then, caution is advised (Groenwold, Palmer, and Tilling 2021). 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 before 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 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 sensitive to any deviations from this form (Ding and Miratrix 2015).

Discussion

The use of small sets of data to demonstrate a key concept in the spirit of Anscombe's Quartet has been applied to a wide variety of data analytic problems. Recent examples include an extension of the original idea proposed by Anscombe called the "Datasaurus Dozen" (Matejka and Fitzmaurice 2017), an exploration of varying interaction effects (Rohrer and Arslan 2021), a quartet of model types fit to the same data that yield the same performance metrics but fit very different underlying mechanisms (Biecek, Baniecki, and Krzyznski 2023), and a set of conceptual causal quartets that highlight the impact of treatment heterogeneity on the average treatment effect (Gelman, Hullman, and Kennedy 2023). While similar in name, the conceptual causal quartets are different from what we present here as they provide excellent insight into how variation in a treatment effect / treatment heterogeneity can impact an average treatment effect (by plotting the latent true causal effect). We believe both sets provide important and

complementary understanding for data analysis practitioners.

We have presented four example data sets demonstrating the importance of understanding the data-generating mechanism when attempting to answer causal questions. These data indicate that more than statistical summaries and visualizations are needed to provide insight into the underlying relationship between the variables. An understanding or assumption of the data-generating mechanism is required to capture causal relationships correctly. 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.

Biecek, Przemyslaw, Hubert Baniecki, and Mateusz Krzyznski. 2023. "Performance Is Not Enough: A Story of the Rashomon's Quartet." *arXiv Preprint arXiv:2302.13356*.

D'Agostino McGowan, Lucy. 2023. *Quartets: Datasets to Help Teach Statistics*.

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.

Gelman, Andrew, Jessica Hullman, and Lauren Kennedy. 2023. "Causal Quartets: Different

Ways to Attain the Same Average Treatment Effect." *arXiv Preprint arXiv:2302.12878*.

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.

Matejka, Justin, and George Fitzmaurice. 2017. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1290–94.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Rohrer, Julia M, and Ruben C Arslan. 2021. "Precise Answers to Vague Questions: Issues with Interactions." *Advances in Methods and Practices in Psychological Science* 4 (2): 25152459211007368.

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

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies." *Journal of Educational Psychology* 66 (5): 688.

———. 2008. "For Objective Causal Inference, Design Trumps Analysis."

———. 2009. "Should Observational Studies Be Designed to Allow Lack of Balance in Covari-

ate 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:

```
library(tidyverse)

# devtools::install_github("r-causal/quartets")

library(quartets)


## Figure 2


ggplot(causal_quartet, aes(x = exposure, y = outcome)) +
  geom_point(alpha = 0.25) +
  geom_smooth(
    method = "lm",
    formula = "y ~ x",
    linewidth = 1.1,
    color = "steelblue"
  ) +
```

```
  facet_wrap(~dataset)

## Table 2


causal_quartet |>
  nest_by(dataset) |>
  mutate(`ATE not adjusting for Z` = round(coef(lm(outcome ~ exposure, data = data))[2], 2
         `ATE adjusting for Z` = round(coef(lm(outcome ~ exposure + covariate, data = data
         `Correlation of X and Z` = round(cor(data$exposure, data$covariate), 2)) |>
  select(-data, dataset) |>
  knitr::kable("latex",
               booktabs = TRUE,
               escape = FALSE,
               col.names =
                 kableExtra::linebreak(c("Data generating mechanism",
                                         "ATE\nnot adjusting for Z",
                                         "ATE\nadjusting for Z",
                                         "Correlation of\nX and Z"), align = "c"))
```