A sampling collider

In the previous chapter, we reviewed four mechanisms that can induce, affect, or alter the crude association between two variables, A and B:
1) A causes B
2) B causes A
3) C causes both A and B
4) Conditioning on a collider opens a confounding path between A and B

Besides confounding, the fourth mechanism can cause another trouble, which is called selection bias. Similar to collider-related confounding, selection bias is caused by conditioning on a collider, but the circumstances are different: First, the collider here is not a "usual" variable such as physical activity; it is a special kind of a binary variable that indicates whether a person could have belonged to our sample or not. Second, the conditioning is often not intentional; we did not deliberately stratify on that collider with de-confounding in mind. Third, we end up computing a measure of association for only one stratum of the collider, not both. Fourth, we cannot always eliminate the bias by conditioning wisely on some other variable; sometimes the only remedy is to conduct another study on a different sample.

To understand the mechanism of selection bias, let's imagine again a world that contains only three variables: the exposure A, the effect of interest B, and a third, binary variable C whose values are "0" and "1". Suppose that A does not cause B and that both are a cause of C—two assumptions that are displayed in the top part of Figure 7–1. Evidently, C is a collider.

As shown in the bottom part of the figure, conditioning on C will induce an association between A and B in the strata of C and therefore the stratum-specific measures of association will not estimate the effect of A on B (here, no effect). They will be biased.
Figure 7–1. Causal diagrams showing a binary collider C on the path from A to B and associations between A and B following stratification on C.

But why would we want to condition on C when trying to estimate the effect of A on B? Well, we never want to condition on C as depicted in Figure 7–1, but we may end up doing so unintentionally or mistakenly when C is a *sampling variable*: a binary variable whose value is "1" for those who could have participated in our study and "0" for those who could not. In some cases, it is possible to argue that a sampling variable has played the role of a collider, and therefore, the study sample makes up a collider's stratum: it is contained in the circle of Figure 7–1 (C=1).

Figure 7–2 shows an example. Suppose that we study the possible effect of marital status on dementia status in a sample of nursing home residents, yet neither variable is a cause of the other. If both marital status and dementia status determine whether a person will be living in a nursing home, the binary variable "place of residence" (nursing home or other location) becomes a collider on the path that connects them. Now, having studied nursing home residents, our sample makes up one stratum of the collider, and in that stratum marital status and dementia status will be associated even though neither causes the other. Every measure of their association from a sample of nursing home residents is "confounded" with no hope for de-confounding. The only remedy is to conduct another study in a different sample.

![Causal diagram](image)

**Figure 7–2.** Causal diagrams showing that marital status and dementia status are associated in a sample of nursing home residents because the sampling variable, place of residence, is a collider.
The idea of selection bias seems to conflict with the indeterministic intuition of confounding. Even if our sample makes up a stratum of a sampling collider, why shouldn’t it be possible, in principle, to condition on confounding variables and eventually ensure the same background propensity to cause dementia in each causal assignment of marital status?

I don’t have a satisfactory answer but maybe the following thought would help. If it were possible to construct a measure of effect from nothing, rather than de-confound a measure of association, we could have argued that only confounding within the sample matters. But we don’t have that luxury. We start with a measure of association and have to find our way to a measure of effect with the help of a causal diagram and conditioning. When our measure of association comes from one stratum of a sampling collider, it already contains a component called selection bias, and sometimes, no action can remove this bias. As you will see in the next example, however, other times conditioning on a variable will help.

Figure 7–3 shows an example of selection bias in which the sampling variable is "hospitalization status" (hospitalized, not hospitalized). Under the assumptions of the top diagram, pneumonia causes coughing, ulcer causes abdominal pain, and both pneumonia and ulcer are a cause of hospitalization but not a cause of each other. If we study the possible effect of coughing on abdominal pain in a sample of hospitalized patients, we will, in effect, be conditioning on a sampling collider, induce an association between pneumonia and ulcer, and open a so-called confounding path between coughing and abdominal pain (dotted line.) Unlike the previous example, however, we do have a remedy: conditioning on either pneumonia (status) or ulcer (status) will block that path and allow us to estimate the effect of coughing on abdominal pain in a sample of hospitalized patients.
Figure 7–3. Causal diagrams showing why the crude association between coughing and abdominal pain in a sample of hospitalized patients is affected by selection bias. The sampling variable, hospitalization status, is a collider.

**Analytical selection bias**

Sometimes our study sample does not make up a stratum of a sampling collider, but we inadvertently create such a stratum when we restrict the analysis to part of the sample. The most common trap is, perhaps, the studying of an early causal link in a chain of cause and effect—after excluding those who had reached a later effect.
Suppose, for example, that we are interested in the effect of fibrinogen (a coagulation protein) on atherosclerosis, which is a known cause of clinical cardiovascular disease. According to our background knowledge, fibrinogen causes cardiovascular disease by mechanisms that do not pass through atherosclerosis because it has a pro-thrombotic effect (Figure 7–4, top part). Evidently, cardiovascular disease status plays the role of a collider on a path that connects fibrinogen and atherosclerosis:

fibrinogen $\rightarrow$ coagulation $\rightarrow$ cardiovascular disease status $\leftarrow$ atherosclerosis

![Causal diagram](image)

Figure 7–4. A causal diagram showing why the association between fibrinogen and atherosclerosis is affected by selection bias when the analysis is restricted to those who are free of cardiovascular disease. The variable cardiovascular disease status is a collider.

Unaware of the pitfall of conditioning on a collider, many researchers (including me) have studied the cross-sectional associations between atherosclerosis and risk factors for
cardiovascular disease after excluding people who had already developed the disease. From the standpoint of causal diagrams, it might have been an unfortunate methodological mistake. Restricting the analysis to one stratum of the variable "cardiovascular disease status" is a form of selection bias because we condition on a collider and thereby induce or alter the association between atherosclerosis and its putative cause (Figure 7–4, bottom part). A similar kind of bias is lurking in cross-sectional studies of risk factors for sub-clinical disease that have recruited only people who did not develop clinical disease.

The decision to condition on a late effect (by restriction) is often driven by fear of estimating a reversed causal pathway. For example, if clinical cardiovascular disease can somehow change the plasma concentration of fibrinogen, we are facing the following diagram:

atherosclerosis→cardiovascular disease status→fibrinogen

which implies a marginal association between atherosclerosis and fibrinogen due to reversed causality of no immediate interest. That association may be blocked by conditioning on cardiovascular disease status (for example, by restricting the analysis to disease-free people). Unfortunately, in a cross-sectional sample it is impossible to distinguish between the path above (reversed causality) and the path below (an intermediary collider) which was depicted in Figure 7–4:

atherosclerosis→cardiovascular disease status←fibrinogen

That means that we can't tell which evil is smaller: restricting the sample to disease-free people or analyzing the entire sample? In some examples, however, the mechanism of reverse causality is nonsensical and, therefore, restriction is unquestionably wrong. For instance, cardiovascular disease status cannot (yet) affect one's genotype or one's sex group. To sum up, we may cause more harm than good by naively assuming that a study of an early link (A→B) in a causal chain (A→B→C) would benefit from excluding people who developed a later effect (C) in that chain.

The idea of selection bias applies to several other situations: 1) selection of controls for a case-control study; 2) selection of prevalent cases rather than incidence cases; 3) and losses to follow up in a cohort study. In all three situations it is possible to explain how the sampling of controls or cases (in a case-control study) or a special type of losses to follow up (in a cohort study) amounted to conditioning on a collider. Examples will be provided in chapter 21.

The dark side of selection bias

Contrary to its theoretical importance, selection bias often has little practical importance because only rarely can we conjecture how it might have been generated. It is much simpler to add the name of another confounder to a causal diagram than to add the
name of one sampling collider. This fact, however, has not deterred many from invoking the argument of selection bias in the name of rigorous science.

When researchers are accused, for example, of having studied a "convenient sample", the accuser has selection bias in mind but he does not bother to explain how the convenience level of recruiting for the study has formed a sampling collider. Indeed, it is often hard to explain, perhaps because a convenient sample and an inconvenient sample do not make up two strata of a sampling collider. Another accusation of selection bias comes in the guise of "volunteerism bias", directed at researchers who have recruited a sample of volunteers. This criticism is deficient twice over: theoretically and morally. On the theoretical end, the critic has to explain how the decision to volunteer for a study, or to not volunteer, has formed a sampling collider—often a difficult task. On the moral end, the question is this: Are researchers expected to force people to participate in their research?

Unless the critic names a sampling collider and displays the colliding variables, the statement "Selection bias could have affected the results" is as useless as the statement "There might have been another confounder." Both are true and both are useless truisms. Which brings up a key principle of the scientific method. We do not accept into the realm of science any criticism that has immunized itself against the possibility of being shown false. If you want to cast doubt on an empirical result, you must be prepared to show an empirical method that could cast doubt on your doubt—a rule that separates untestable truisms from testable criticism. Of course, no empirical method can cast doubt on the truth of "There might have been another confounder" or on the truth of "Selection bias could have affected the results". It is impossible to show that no other confounding path exists and that we have not conditioned unintentionally on any sampling collider.

But the worst practical side of selection bias is, perhaps, associating a non-random sample with selection bias and a random sample with its absence, a misconception that can be traced to the fields of demography, sociology, and even public health. In these fields, random sampling from a well-defined population often serves to estimate the distribution of a variable of interest, such as poverty level, election vote, and blood pressure. What does estimation of a causal parameter have in common with estimating the distribution of a variable? Not much, if anything at all.

In an effort to rebut criticism of selection bias, many researchers feel compelled to show that participants in their studies do not differ much from non-participants on those variables that were somehow measured in both groups. To my mind, the rationale for this practice is as vague as the criticism it tries to rebut. We cannot infer the absence of a sampling collider from similarity of participants and non-participants (nor can we infer its presence from their dissimilarity.)

**A target population, a sample, and a causal diagram**

In this section, I will try to shed light on the link between the three ideas that make up the section's title: target population, sample, and causal diagram. The topic may deserve a chapter of its own, but I decided to include it under "Selection Bias" because it has something to do with samples and selection. Here, again, we will have to follow two trails of thought: deterministic and indeterministic.

As you may recall, deterministic reasoning required us to specify a target population whose causal parameter is estimated in the study, because causal parameters are not
universal entities (chapter 4). When the study is a randomized trial, the framework is clear: the study sample makes up the target population; the causal diagram shows two variables, $A \rightarrow B$; and a measure of their crude association in the sample estimates a causal parameter of interest. But what happens when the causal assignments of $A$ were not randomized and we have to condition on confounders? Which population is described in the diagram and whose causal parameter is estimated from the sample?

Not having randomized, the determinist is no longer obligated to equate the target population with the study sample, but he cannot escape the commitment to make up some arbitrary "target of interest" for his causal parameter. Then, he can draw a causal diagram to display assumptions about confounding paths, pretend that he drew his sample from that (finite) target population, and add arrows to display a sampling collider, if any. These steps will set the stage for de-confounding, assuming he is willing to ignore a few missing links between the four classes of deterministic causation (chapter 4) and a causal diagram.

To my fallible mind, his framework for science is a walk in a minefield. First, I have yet to meet a scientist who has specified a finite target population for her study. No one that I know of has ever stated that she was estimating a causal parameter for the 2,537,816 residents of some city or for any other finite population. Nor have I heard of a scientist who had drawn a sample from a sampling frame and then announced that his causal parameter belongs to that sampling frame alone. Either ignorance of method is entrenched in science or the deterministic insistence on a target population is epistemologically wrong. Moreover, I don’t even dare interrogating some of my colleagues about the target populations for their studies because they will start laughing. What is the target population for a cell culture or for a sample of mice, both of which are commonly used to study the causes of human disease and its cures? Should in vitro studies and animal models play no role in causal inquiry of human illness because no target population is in sight?

Second, it is not clear to me how the determinist reconciles empirical associations in the sample with theoretical relations in a target population—even when the sample is a randomized trial and the two concepts overlap. Recall, for example, that trial of streptokinase and placebo for ischemic stroke and assume for a moment that despite randomization the size of the stroke was associated with the randomized treatment: a larger proportion of streptokinase recipients than placebo recipients had suffered a large stroke. Now, since the treatment was assigned at random, no arrow or line should connect the variable called "stroke size" with the variable called "treatment group" in the target population (which is the sample at hand!) But reality says otherwise; so what should we do? Hold to the causal diagram and not condition on the stroke size, or revise the diagram to match the sample’s reality and condition on the stroke size? I know the indeterministic answer to this question—revise and condition because the background propensity is unbalanced—but I don’t know the deterministic answer because there is no one answer. Every deterministic writer will state that the crude association is unbiased because the study was randomized, but many will still argue for conditioning in the name of "efficiency" (a statistical idea whose connection to deterministic causation seems loose to me.) And if the answer remains uncertain for a randomized trial, what is the answer for a non-randomized study? Does a theoretical diagram for the target override the reality of the sample, or not?

Finally, although determinism ties the causal parameter to a target population, causal diagrams rarely display relations that are unique to one arbitrary target. When we draw an arrow that emanates from asthma and points to FEV, our mind entertains universal
causal reality—asthma affects the value of FEV, *in general*—not the unique reality of some target. To me, the hybrid of a target-specific causal parameter and a diagram that displays universal causal assumptions seems peculiar, if not incoherent.

Indeterminism does not have all of the answers, but it does have one fewer problem to solve because there is no target population and no target-specific causal parameter. The indeterministic causal parameter compares the causal propensity of two causal assignments, and the sample at hand—whether a randomized trial, a non-randomized study, or a cell culture—is a specimen of universal causal reality: sometimes a good specimen and sometimes a bad one. Since no scientist will ignore the quality or the properties of the specimen she is studying, we should not ignore the presence and absence of empirical associations in the sample at hand.

When the sample suggests that a variable creates a confounding path between A and B and the diagram says that it doesn’t, we should revise the diagram accordingly (adding, for example, a line to connect the size of the stroke with the treatment group.) On the other hand, if the diagram suggests a confounding variable and the sample does not affirm its expected associations with A and B, we should ask ourselves why and decide whether to condition based on our answer. If we don’t have an answer, as might well be the case, and decide to "try it both ways", then the data might not provide just one estimated effect of A on B.

Which brings up another key idea. In the process of de-confounding we derive inference not only from data but also from *assumptions*, some of which might be wrong. But we cannot replace an assumption about the need to condition (fearing it might be wrong), with a claim of uncertainty: we cannot derive inference from not knowing which assumption to choose. Causal inference, like all scientific inference, is *conditional on premises*, not on ignorance.

To sum up: A causal diagram is an important addition to our inventory of methods for causal inquiry and its rules for conditioning deserve to be taught side-by-side with statistical ideas. I might even dare equating the contribution of its developers with the contribution of the founders of some schools of statistics. It is not, however, a sacred algorithm that allows us to ignore the reality of a sample, nor an automated procedure that leads to the Truth. Any claim to have found a road to the Truth, even if miraculously true, is bound to be epistemologically false.

Why false?

Because such a claim implies that scientific knowledge could become certain.
Information bias

Back in chapter 2, we realized that the variable of interest might be beyond our reach. Even when we know precisely what we wish to measure, our measurement usually contains some error so we end up having to rely on a surrogate variable for the "real one". In particular, measured (classified) disease status often differs from true disease status because of misclassification error. As a result, we face the causal structure in Figure A.

\[
\text{E} \rightarrow \text{D} \rightarrow \text{D*} \quad \text{Figure A}
\]

where D is true disease status and D* is its surrogate—the measured version. The arrow from D to D* signifies that true disease status is, undoubtedly, a cause of classified disease status, though not the only cause.

Both the rationale behind a surrogate variable, such as D*, and the drawback of using it are derived from the principles of causal diagrams. Although the effect of interest is E \rightarrow D, we can estimate only the marginal association between E and D*. If D* were perfectly correlated with D, then we could have simply substituted D* for D in the diagram. But since this is rarely the case, if ever, the marginal association between E and D* inevitably reflects two pieces: a causal segment of interest (E \rightarrow D), and a causal segment of no interest (D \rightarrow D*).

Our unavoidable need to use D*, the measured outcome, instead of D, the true outcome, can lead to another category of bias called "information bias". If E affects D* by pathways that do not involve D (Figure B), the marginal association of E with D* reflects not only the path E \rightarrow D \rightarrow D*, but also the path E \rightarrow Z \rightarrow D*, which is often of no causal interest. We call this situation "information bias" because E provides information on D* (allows us to predict D*), apart from any effect it might have on D.

\[
\text{E} \rightarrow \text{D} \rightarrow \text{D*} \quad \text{Figure B}
\]

Suppose, for example:

- E = estrogen use
- D = endometrial cancer status (truth)
- D* = endometrial cancer diagnosis status (measured)
- Z = frequency of gynecological exams
It is easy to rationalize the existence of a path from estrogen use (E) to endometrial cancer diagnosis (D*) via the frequency of gynecological exams (Z): Gynecologists likely examine their patients frequently after prescribing estrogen, and frequent gynecological exams increase the chances of detecting (diagnosing) endometrial cancer.

Conditioning on Z will block the information path between E and D* and eliminate the bias—unless D happens to be a cause of Z as well (Figure C). In that case, Z becomes a collider on the path E→Z→D and conditioning on it will induce an association between E and D, thereby opening a confounding path between E and D* (E→D→D*). Vaginal bleeding status, for example, may play the role of Z in a study of estrogen use and endometrial cancer, because taking estrogen (E) and endometrial cancer (D) are both causes of vaginal bleeding.

As you might have realized, the last situation is hopeless: We cannot eliminate the bias. If we don’t condition on Z, the marginal association of E and D* contains a path of information bias (E→Z→D*). And if we do condition on Z, the conditional association of E and D* contains a confounding path (E→D→D*), apart from the causal path, E→D→D* (if it exists).

Figure C

 Depending on the nature of Z, information bias goes by various specific names such as "unmasking bias", "diagnosis suspicion bias", and "ascertainment bias". In all cases, however, the mechanism involves at least one causal path of "information" about D*.

Although we are usually interested in D, it is possible to come up with examples where D* itself is also of causal interest. For example, we may be interested in the overall effect of a widely prescribed drug on the detection of some disease, perhaps because detected disease affects medical care, subsequent medical expenses, and psychological wellbeing. In that case, the marginal association of E and D* provides all that we need; there is no bias.

Let’s turn next to the exposure variable, E, and its surrogate E*. Again, E* is the measured exposure, which may differ from E, the true exposure, in part because of measurement error. The rationale for using E* is a little different, however (Figure D).

Figure D
If E* were perfectly correlated with E, we could have simply substituted E* for E in the diagram, but this is rarely the case, of course. Nonetheless, notice that E is a common cause of E* and D, and that E* itself is not a cause of D. Therefore, the observed marginal association between E* and D is generated by the path E*←E→D, embedding two causal segments: E→D, which is of interest, and E→E*, which is not. Interestingly, the idea of a common cause—the classical property of a confounder—serves us well here because the common cause happened to be the exposure itself.

As always, the surrogate variable E* should be a strong correlate of the true exposure, E. If the error in E* is so large such that E* and E are not associated, then E* and D will not be associated, no matter how strong of an effect E may have on D. To reach the last conclusion, you don’t even need to follow the logic of a causal diagram: if E* has nothing in common with E, it cannot tell us anything about the effects of E.

A classical kind of exposure information bias occurs if D→E*—if disease status supplies information on the measured exposure (Figure E). What is usually labeled "recall bias" in case-control studies follows that causal structure: for various psychological reasons, cases may report their historical exposure differently from controls. Consequently, the marginal association between E* and D would reflect two paths:

E*←E→D (a path that contains the segment of interest: E→D)
D→E* (information path).

Of course, the second path is of no interest to us; it produces information bias.

To sum up, the principles of causal diagrams clearly explain how key categories of bias (confounding, selection, and information) interfere with our attempts to estimate effects by marginal or conditional associations. These principles teach us that three kinds of paths may contribute to a measure of association between the exposure (E) and the disease (D)—besides the causal paths of interest:

- Naturally occurring confounding paths (by common causes of E and D)
- Induced confounding paths due to conditioning on colliders (some of which are "sampling colliders")
- Causal paths of no interest to us due to our using surrogates for E or D (information bias)

No other methodological discovery has contributed so much to our understanding of sources of bias in causal inquiry.