THE *DO*-OPERATOR AND THE BACK-DOOR CRITERION

To understand the back-door criterion, it helps first to have an intuitive sense of how information flows in a causal diagram. I like to think of the links as pipes that convey information from a starting point X to a finish Y. Keep in mind that the conveying of information goes in both directions, causal and noncausal, as we saw in Chapter 3.

In fact, the noncausal paths are precisely the source of confounding. Remember that I define confounding as anything that makes $P(Y \mid do(X))$ differ from $P(Y \mid X)$. The *do*-operator erases all the arrows that come into X, and in this way it prevents any information about X from flowing in the noncausal direction. Randomization has the same effect. So does statistical adjustment, if we pick the right variables to adjust.

In the last chapter, we looked at three rules that tell us how to stop the flow of information through any individual junction. I will repeat them for emphasis:

(a) In a chain junction, $A \rightarrow B \rightarrow C$, controlling for B prevents information about A from getting to C or vice versa.

(b) Likewise, in a fork or confounding junction, $A \leftarrow B \rightarrow C$, controlling for B prevents information about A from getting to C or vice versa.

(c) Finally, in a collider, $A \rightarrow B \leftarrow C$, exactly the opposite rules hold. The variables A and C start out independent, so that information about A tells you nothing about C. But if you control for B, then information starts flowing through the "pipe," due to the explain-away effect.

We must also keep in mind another fundamental rule:

(d) Controlling for descendants (or proxies) of a variable is like "partially" controlling for the variable itself. Controlling for a descendant of a mediator partly closes the pipe; controlling for a descendant of a collider partly opens the pipe.

Now, what if we have longer pipes with more junctions, like this:

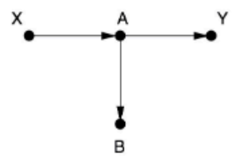
$$A \leftarrow B \leftarrow C \rightarrow D \leftarrow E \rightarrow F \rightarrow G \leftarrow H \rightarrow I \rightarrow J?$$

The answer is very simple: if a single junction is blocked, then J cannot "find out" about A through this path. So we have many options to block communication between A and J: control for B, control for C, don't control for D (because it's a collider), control for E, and so forth. Any one of these is sufficient. This is why the usual statistical procedure of controlling for everything that we can measure is so misguided. In fact, this particular path is blocked if we don't control for anything! The colliders at D and G block the path without any outside help. Controlling for D and G would open this path and enable J to listen to A.

Finally, to deconfound two variables X and Y, we need only block every noncausal path between them without blocking or perturbing any causal paths. More precisely, a *back-door path* is any path from X to Y that starts with an arrow pointing into X. X and Y will be deconfounded if we block every back-door path (because such paths allow spurious correlation between X and Y). If we do this by controlling for some set of variables Z, we also need to make sure that no member of Z is a descendant of X on a causal path; otherwise we might partly or completely close off that path.

That's all there is to it! With these rules, deconfounding becomes so simple and fun that you can treat it like a game. I urge you to try a few examples just to get the hang of it and see how easy it is. If you still find it hard, be assured that algorithms exist that can crack all such problems in a matter of nanoseconds. In each case, the goal of the game is to specify a set of variables that will deconfound X and Y. In other words, they should not be descended from X, and they should block all the back-door paths.

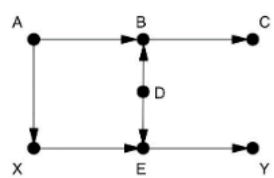
Game 1.



This one is easy! There are no arrows leading into X, therefore no back-door paths. We don't need to control for anything.

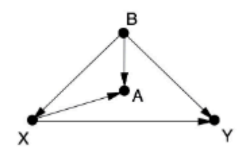
Nevertheless, some researchers would consider B a confounder. It is associated with X because of the chain $X \rightarrow A \rightarrow B$. It is associated with Y among individuals with $X = 0$ because there is an open path $B \leftarrow A \rightarrow Y$ that does not pass through X. And B is not on the causal path $X \rightarrow A \rightarrow Y$. It therefore passes the three-step "classical epidemiological definition" for confounding, but it does not pass the back-door criterion and will lead to disaster if controlled for.

Game 2.



In this example you should think of A, B, C, and D as "pretreatment" variables. (The treatment, as usual, is X.) Now there is one back-door path $X \leftarrow A \rightarrow B \leftarrow D \rightarrow E \rightarrow Y$. This path is already blocked by the collider at B, so we don't need to control for anything. Many statisticians would control for B or C, thinking there is no harm in doing so as long as they occur before the treatment. A leading statistician even recently wrote, "To avoid conditioning on some observed covariates... is nonscientific ad hockery." He is wrong; conditioning on B or C is a poor idea because it would open the noncausal path and therefore confound X and Y. Note that in this case we could reclose the path by controlling for A or D. This example shows that there may be different strategies for deconfounding. One researcher might take the easy way and not control for anything; a more traditional researcher might control for C and D. Both would be correct and should get the same result (provided that the model is correct, and we have a large enough sample).

Game 3.



In Games 1 and 2 you didn't have to do anything, but this time you do. There is one back-door path from X to Y, $X \leftarrow B \rightarrow Y$, which can only be blocked by controlling for B. If B is unobservable, then there is no way of estimating the effect of X on Y without running a randomized controlled experiment. Some (in fact, most) statisticians in this situation would control for A, as a proxy for the unobservable variable B, but this only partially eliminates the confounding bias and introduces a new collider bias.

GAME 4.

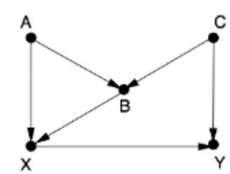
This one introduces a new kind of bias, called “M-bias” (named for the shape of the graph). Once again there is only one back-door path, and it is already blocked by a collider at *B*. So we don’t need to control for anything. Nevertheless, all statisticians before 1986 and many today would consider *B* a confounder. It is associated with *X* (via $X \leftarrow A \rightarrow B$) and associated with *Y* via a path that doesn’t go through *X* ($B \leftarrow C \rightarrow Y$). It does not lie on a causal path and is not a descendant of anything on a causal path, because there is no causal path from *X* to *Y*. Therefore *B* passes the traditional three-step test for a confounder.

M-bias puts a finger on what is wrong with the traditional approach. It is incorrect to call a variable, like *B*, a confounder merely because it is associated with both *X* and *Y*. To reiterate, *X* and *Y* are unconfounded if we do not control for *B*. *B* only becomes a confounder when you control for it!

When I started showing this diagram to statisticians in the 1990s, some of them laughed it off and said that such a diagram was extremely unlikely to occur in practice. I disagree! For example, seat-belt usage (*B*) has no causal effect on smoking (*X*) or lung disease (*Y*); it is merely an indicator of a person’s attitudes toward societal norms (*A*) as well as safety and health-related measures (*C*). Some of these attitudes may affect susceptibility to lung disease (*Y*). In practice, seatbelt usage was found to be correlated with both *X* and *Y*; indeed, in a study conducted in 2006 as part of a tobacco litigation, seat-belt usage was listed as one of the first variables to be controlled for. If you accept the above model, then controlling for *B* alone would be a mistake.

Note that it’s all right to control for *B* if you also control for *A* or *C*. Controlling for the collider *B* opens the “pipe,” but controlling for *A* or *C* closes it again. Unfortunately, in the seat-belt example, *A* and *C* are variables relating to people’s attitudes and not likely to be observable. If you can’t observe it, you can’t adjust for it.

GAME 5.



Game 5 is just Game 4 with a little extra wrinkle. Now a second back-door path $X \leftarrow B \leftarrow C \rightarrow Y$ needs to be closed. If we close this path by controlling for *B*, then we open up the *M*-shaped path $X \leftarrow A \rightarrow B \leftarrow C \rightarrow Y$. To close that path, we must control for *A* or *C* as well. However, notice that we could just control for *C* alone; that would close the path $X \leftarrow B \leftarrow C \rightarrow Y$ and not affect the other path.

Games 1 through 3 come from a 1993 paper by Clarice Weinberg, a deputy chief at the National Institutes of Health, called “Toward a Clearer Definition of Confounding.” It came out during the transitional period between 1986 and 1995, when Greenland and Robins's paper was available but causal diagrams were still not widely known. Weinberg therefore went through the considerable arithmetic exercise of verifying exchangeability in each of the cases shown. Although she used graphical displays to communicate the scenarios involved, she did not use the logic of diagrams to assist in distinguishing confounders from deconfounders. She is the only person I know of who managed this feat. Later, in 2012, she collaborated on an updated version that analyzes the same examples with causal diagrams and verifies that all her conclusions from 1993 were correct.

In both of Weinberg's papers, the medical application was to estimate the effect of smoking (*X*) on miscarriages, or "spontaneous abortions" (*Y*). In Game 1, *A* represents an underlying abnormality that is induced by smoking; this is not an observable variable because we don't know what the abnormality is. *B* represents a history of previous miscarriages. It is very, very tempting for an epidemiologist to take previous miscarriages into account and adjust for them when estimating the probability of future miscarriages. But that is the wrong thing to do here! By doing so we are partially inactivating the mechanism through which smoking acts, and we will thus underestimate the true effect of smoking.

Game 2 is a more complicated version where there are two different smoking variables: *X* represents whether the mother smokes now (at the beginning of the second pregnancy), while *A* represents whether she smoked during the first pregnancy. *B* and *E* are underlying abnormalities caused by smoking, which are unobservable, and *D* represents other physiological causes of those abnormalities. Note that this diagram allows for the fact that the mother could have changed her smoking behavior between pregnancies, but the other physiological causes would not change. Again, many epidemiologists would adjust for prior miscarriages (*C*), but this is a bad idea unless you also adjust for smoking behavior in the first pregnancy (*A*).

Games 4 and 5 come from a paper published in 2014 by Andrew Forbes, of Monash University, and Elizabeth Williamson, now at the London School of Hygiene and Tropical Medicine. They are interested in the effect of smoking on adult asthma. In Game 4, *X* represents an individual's smoking behavior, and *Y* represents whether the person has asthma as an adult. *B* represents childhood asthma, which is a collider because it is affected by both *A*, parental smoking, and *C*, an underlying (and unobservable) predisposition toward asthma. In Game 5 the variables have the same meanings, but they added two arrows for greater realism. (Game 4 was only meant to introduce the *M*-graph.)

In fact, the full model in their paper has a few more variables and looks like the diagram in Figure 4.7. Note that Game 5 is embedded in this model in the sense that the variables *A*, *B*, *C*, *X*, and *Y* have exactly the same relationships. So we can transfer our conclusions over and conclude that we have to control for *A* and *B* or for *C*; but *C* is an unobservable and therefore uncontrollable variable. In addition we have four new confounding variables: *D* = parental asthma, *E* = chronic bronchitis, *F* = sex, and *G* = socioeconomic status. The reader might enjoy figuring out that we must control for *E*, *F*, and *G*, but there is no need to control for *D*. So a sufficient set of variables for deconfounding is *A*, *B*, *E*, *F*, and *G*.