Causal Paths

Christoph Hanck

Summer 2023

Causal paths

Paths

The path between two variables on a causal diagram is a description of the set of arrows and nodes you visit when "walking" from one variable to another.

Causal paths

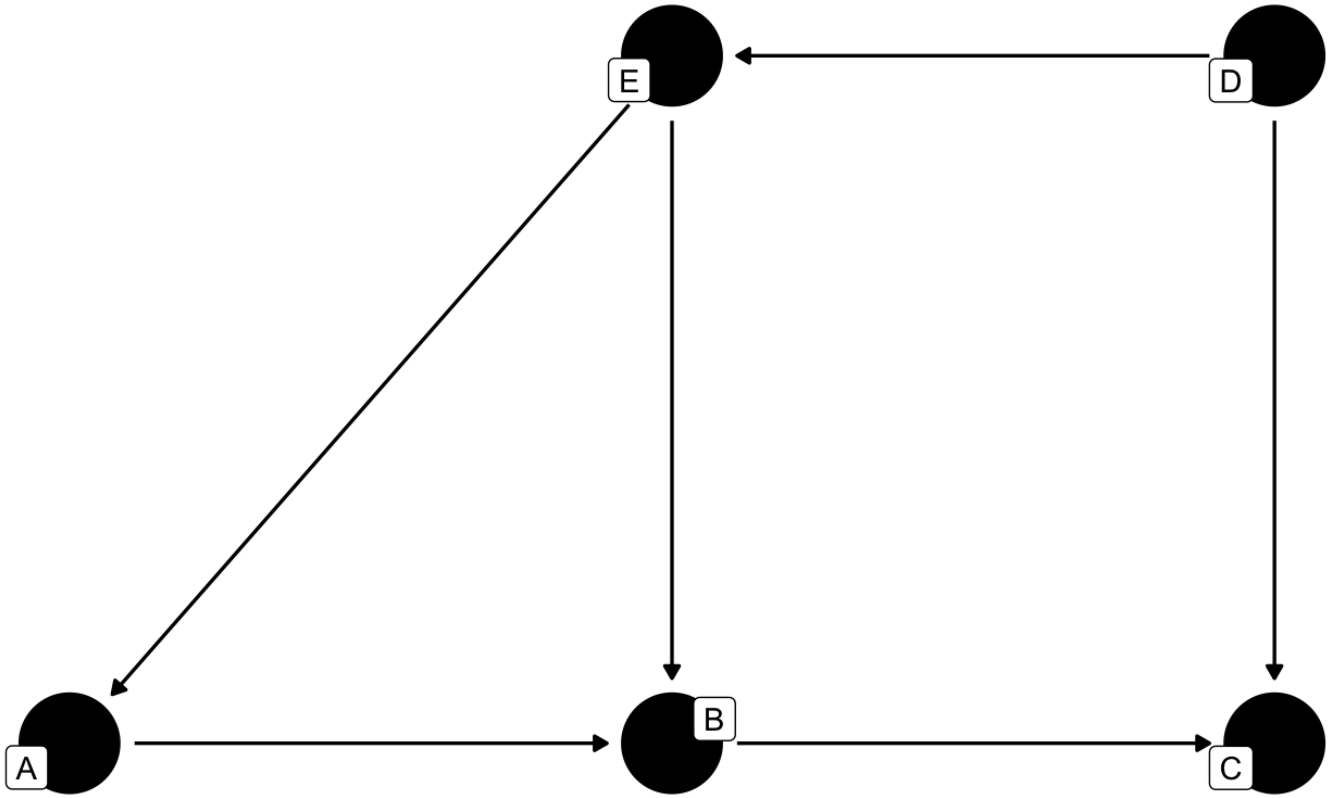


Figure 1: Causal paths

Causal paths

- B and C are related:

  B causes C directly—that is one path between B and C.

  D causes both E and C, and *E causes B*. We have the path

  $$B \leftarrow E \leftarrow D \rightarrow C.$$

  If our research question is about the effect of B on C, then the pathway D is responsible for is another reason we would see B and C being related other than $B \rightarrow C$.

- We thus have an alternate explanation for why B and C might be related, other than the explanation that answers our research question of whether (and how much) B causes C.
- The paths can tell us the road we want to walk on, and also the road we want to avoid.

Finding all paths

Why is it important?

- We want to be able to write out every single path that starts with the treatment variable and ends with the outcome variable.
- This is because, each path explains one way in which the treatment and outcome variables might be related.
- The alternative paths are alternate explanations for the causation.
- If we want to really show how much your treatment causes the outcome, we have to be able to find those alternate explanations to account for them in your research and identify just the explanation of interest.

Finding all paths

How can we find every path from treatment to outcome?

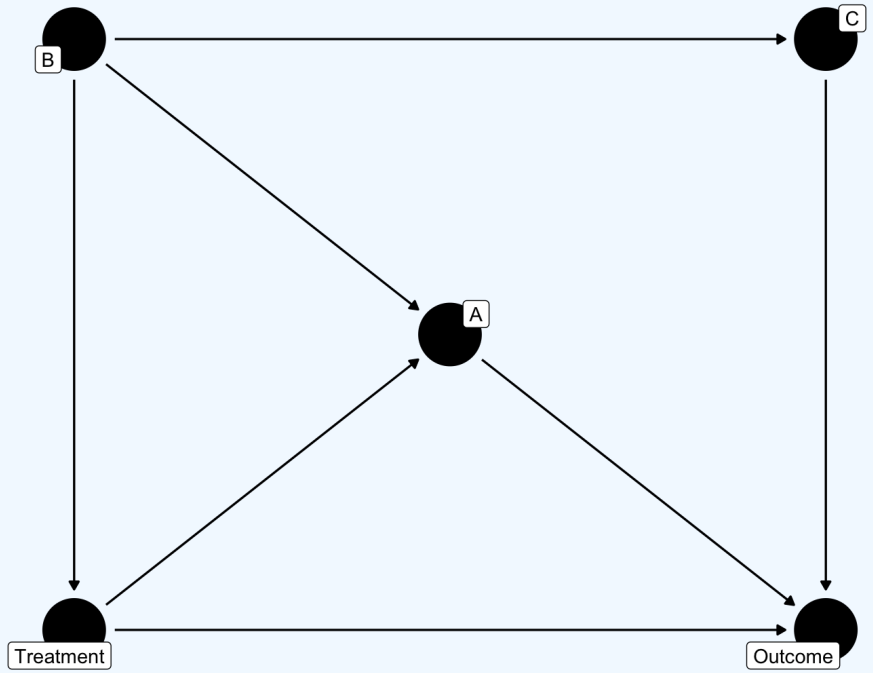
1. Begin at the treatment variable and follow an in- or outcoming arrow to find another variable
2. Then, follow one of the arrows coming in or out of that variable
3. Repeat step 2 until you either...

   come to a variable that has been already visited (*a loop that is not a path*)

   or find the outcome variable, *which is a path — write it down!*
4. Every time you either find a path or a loop, back up and try a different arrow in/out until tried them all. Then, back up again and try all those arrows
5. Stop when you have tried all the ways out of the treatment variable and all the eventual paths.

Finding all paths

Example: Can you find all the paths?



- Treatment → Outcome
- Treatment → A → Outcome
- Treatment → A ← B → C → Outcome
- Treatment ← B → A → Outcome
- Treatment ← B → C → Outcome

Finding all paths

Example: Wine-drinking and lifespan

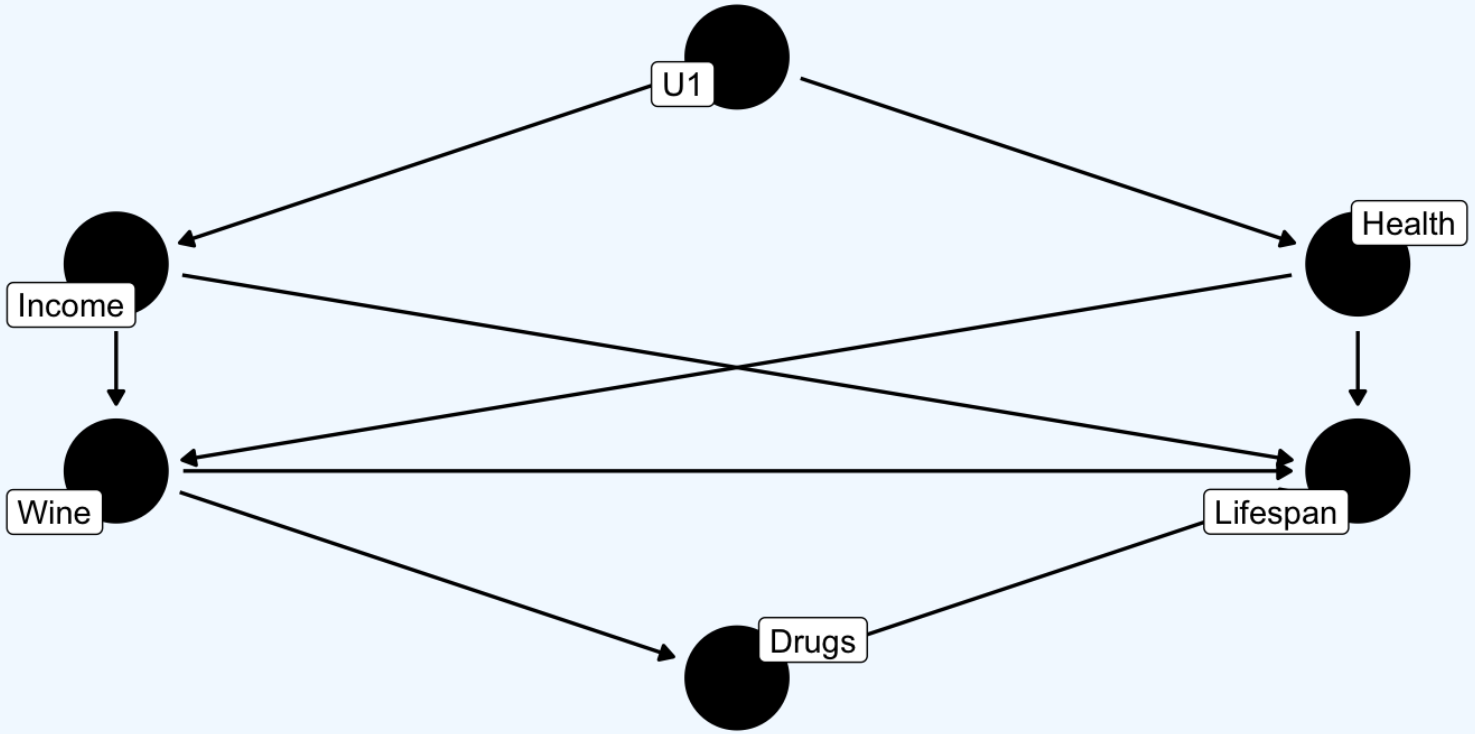


Figure 2: Causal paths

Finding all paths

Example: Wine-drinking and lifespan

1. *Wine* → *Lifespan*.
2. Back to *Wine*. *Drugs* only goes to *Lifespan*: *Wine* → *Drugs* → *Lifespan*.
3. From *Income* to *Lifespan*:

   Wine ← Income → Lifespan

   Wine ← Income ← U1 → Health → Lifespan
4. *Health* remains:

   Wine ← Health → Lifespan

   Wine ← Health ← U1 → Income → Lifespan

Good paths vs. bad paths & front doors vs. back doors

- **Good paths** are the ones that relate treatment and outcome variables and are thus relevant for answering our research question. **Bad paths** are alternate explanations.

  Every path in which all the arrows face away from treatment are good paths, the rest are Bad Paths
- Paths that face away from treatment are also known as **front door** paths. The rest would then be **back door** paths

  Paths with at least one arrow pointing towards treatment are back door paths!
- Usually, all the front door paths are good, and all the back door paths are bad.

Good paths vs. bad paths & front doors vs. back doors

Example: Wine-drinking and lifespan revisited

- There are two paths where all the arrows are going away from *Wine*:

  Wine → *Lifespan*

  Wine → *Drugs* → *Lifespan*

- These are all the ways in which a change in *Wine* would cause a change in *Lifespan* and thus are front door paths.
- If our research question is "does Wine cause Lifespan?" then these (front door) paths are the good paths, and the other (back door paths) are bad ones.

Open and closed paths

Definition: Open vs. clodsed paths

A path is *open* if all of the variables along that paths are allowed to vary. A path is *closed* if at least one of the variables along that path has no variation.

Open and closed paths

Example: Wine-drinking and lifespan revisited

Assumption: *Wine* → *Drugs* → *Lifespan*

- If we had data with wine drinkers and non-wine drinkers, drug-users and non-drug users, and people with shorter and longer lifespans, then all the variables along this path *have variation*.
- If we had data in which nobody uses drugs, we would have no variation in *Drugs*, and thus none of the relationship between *Wine* and *Lifespan* could possibly be driven by *Drugs*. This is a *closed* path.

Open and closed paths

Why we wish to close paths

- *Closing a path* means that we can remove all the variation due to a variable along that path:

  We eliminate a threat to the identification of the variation we are interested in!

- If we can control for at least one variable on each of our Bad Paths without controlling for anything on one of our Good Paths, we have identified the answer to our research question.

Open and closed paths

Example: Wine-drinking and lifespan revisited

- To find the effect of *Wine* on *Lifespan*, we want all the ways in which *Wine* can cause *Lifespan* to change.
- We can identify the answer to our research question by picking at least one variable along each Bad Path to control, without controlling for anything on a good Path.

  We have two good (front door) paths:

  Wine → *Lifespan* *Wine* → *Drugs* → *Lifespan*

  The rest are bad paths, i.e, control for *Health* and *Income*

Colliders

Definition: Collider

A variable is a **collider** on a particular path if, on that path, both arrows point at it. A path is *closed* by default when there is a collider.

Colliders

Colliders hinder closing of paths

- We have defined paths to be *open* as long as every variable along the path is allowed to vary. Removing the variation from a variable on the path (controlling/adjusting for it) *closes* the path.
- If we are looking for alternate explanations of why treatment and outcome might be related, the collider shuts down that alternate explanation.
    - If the path were *Treatment* $\leftarrow$ *C* $\rightarrow$ *Outcome* without a collider, then one reason why *Teatment* and *Outcome* vary together is because *C* causes them both. But with a collider, *Treatment* $\leftarrow$ *A* $\rightarrow$ *B* $\leftarrow$ C $\rightarrow$ Outcome , C can affect *Outcome*, and *C* can affect *B*, but because *B* doesn't affect *Treatment*, *C* can no longer induce a relationship between *Treatment* and *Outcome*. *B* saved us.
- Once we control for the collider, the two variables pointing to the collider become related: the alternate explanation returns.

Colliders

Dealing with colliders

- We may just not control for the colliders. However, two problems are associated with that:
  1. We need to figure out that a variable is a collider so we know not to control for it.
  2. One common way we control for a collider, say Z, is by selecting a sample: picking a sample with no variation in Z is one way of controlling for Z.

     So if we do a study, say, of college students, then we are inevitably controlling for college attendance. However, college attendance can be a collider!
- To identify the answer to our research question, what we need is to close all the bad paths while leaving all the good paths open.
- Once we accomplish that, any remaining relationship between treatment and control can only be going through the good paths, and all of the good paths we want are included.
- This is exactly what we want.

Using paths to test a causal diagram

Look for paths between any two variables to determine whether we have the right diagram in the first place:

1. Pick two variables on our diagram other than Treatment and Outcome. Let us call them A and B.
2. List all of the paths between A and B. Then, we do what we need to do to ensure they are all closed.
3. If A and B are still related to each other, that means there must be some other path we did not account for.

   Our diagram is *deficient*, and perhaps in an important way.

Using paths to test a causal diagram

Example: Wine-drinking and lifespan revisited

Consider *Income* and *Drugs*. We list all paths between *Wine* and *Income*

Drugs ← Wine ← Income

Drugs ← Wine ← Health ← U1 → Income

Drugs ← Wine ← Health → Lifespan ← Income

Drugs ← Wine → Lifespan ← Income

Drugs ← Wine → Lifespan ← Health ← U1 → Income

Drugs → Lifespan ← Income

Drugs → Lifespan ← Wine ← Income

Drugs → Lifespan ← Wine ← Health ← U1 → Income

Drugs → Lifespan ← Health → Wine ← Income

Drugs → Lifespan ← Health ← U1 → Income

Using paths to test a causal diagram

Example: Wine-drinking and lifespan revisited

- The list of open paths is much smaller, since *Lifespan* is a collider everywhere it turns up!

  Drugs ← *Wine* ← *Income*

  Drugs ← *Wine* ← *Health* ← *U1* → *Income*

  If we could control for *Wine* then both paths would close, too.

- Our diagram makes the unrealistic claim that *Income* and *Drugs* only relate to each other through *Wine*.

  We can if *Income* and *Drugs* are related after controlling for *Wine* using our data. If they are, our model is incomplete.

  This is an example of a **placebo test**.

Using paths to test a causal diagram

Definition: Placebo test

Tests like these, where we expect that a relationship should be zero because our diagram says there are no Open Paths, and we see whether it's actually zero, are called placebo tests.

Using paths to test a causal diagram

What if we *fail* a placebo test?

- Failing a placebo test proves that the model is incorrect and incomplete.
- If we find a small nonzero relationship that, according to the diagram, should not be there, that might be a minor case for concern.
- If there is an enormous and strong relationship that should not be there, we should really worry and revise the causal diagram to come up with an improved model.