

Causal framework

Tobias Kurth

Zagreb, 2019



BERLIN SCHOOL OF
PUBLIC HEALTH

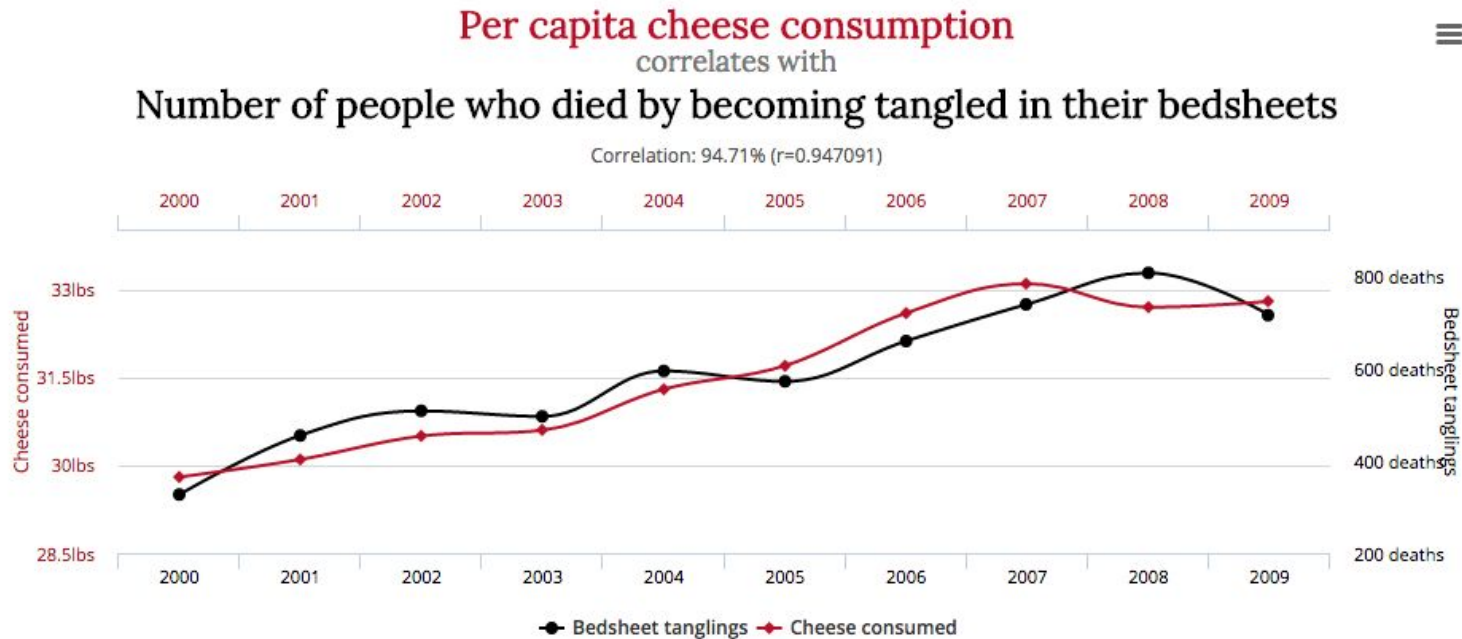
Content

- Association, correlation, causation
- How to define a “cause”
- Introduction to directed acyclic graphs (DAGs)
- Exploring concepts of colliders, mediation, and counterfactuals
- Effect measure modification, interaction

Association, correlation, causation

- We need to make a clear distinction between association, correlation and causation
- Correlation implies association, but **not** causation
- Conversely, causation implies association, but **not** correlation

Correlation? Association? Causation?



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

tylervigen.com

Causation

- Causation is a clear conception of cause and consequences
- It requires a clear direction of the effect and specific analytic strategies

Causal thinking in epidemiology

- Epidemiologic studies investigate groups of individuals (= population) who are or who are not exposed to causes of a disease
- We are interested in understanding the number of excess cases of disease that can be removed if we remove a particular cause in a specific population

What is a cause? What is a causal relationship?

- How to define?
-challenging!

Bradford Hill's criteria for a causal relationship

1. Strength of the association (effect size)
2. Consistency of findings (reproducibility)
3. Specificity (specific population, specific exposure)
4. Temporal sequence of association (1st cause, then effect)
5. Biological gradient (e.g. dose-response).
6. Biological plausibility (mechanism plausible?)
7. Coherence (does it fit with what we already know?)
8. Experiment (does empirical research agree?)
9. Analogy (do similar factors work the same way?)

Bradford Hill's criteria: problems & debate

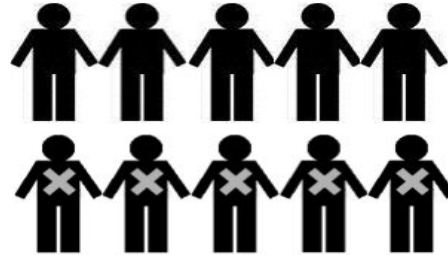
1. Strength → if bias = wrong regardless of strength; many small effects
2. Consistency of findings → replication can be difficult/impossible
3. Specificity → diseases caused by multiple factors and single factor can cause multiple diseases
4. Temporal sequence of association (1st cause then effect)
5. Biological gradient → presence alone can trigger effect
6. Biological plausibility → limited by current knowledge
7. Coherence → can't confirm everything in a laboratory/trial
8. Experiment → again, can't confirm everything in a lab/trial
9. Analogy → what to compare?

What is a cause? → a newer definition

- “A factor that contributes, **at least in part**, to the development (or prevention) of illness, **at least in some individuals**”
 - Rothman, Greenland, Schwartz, Susser, Keyes, Galea & others
- Caters to multifactorial diseases
 - Most diseases are complex, multiple causes (e.g. myocardial infarction)
 - Individual causes = “component causes”
 - E.g. family history, smoking, obesity, lack of preventive care

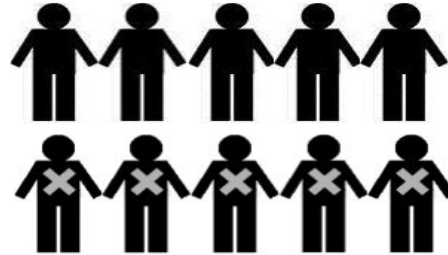
Excess cases

Ten people, all exposed

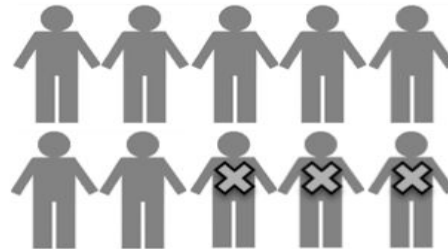


Excess cases, cont.

Ten people, all exposed

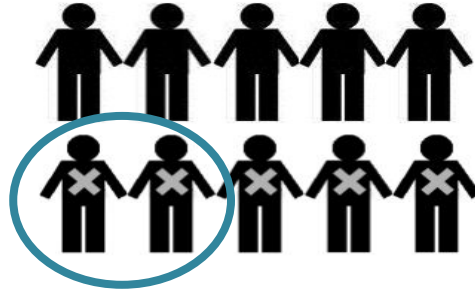


The same ten people, observed at the same time, without the exposure

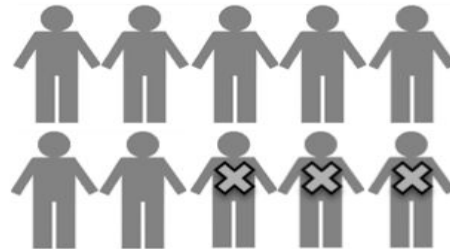


Excess cases, cont.

Ten people, all exposed



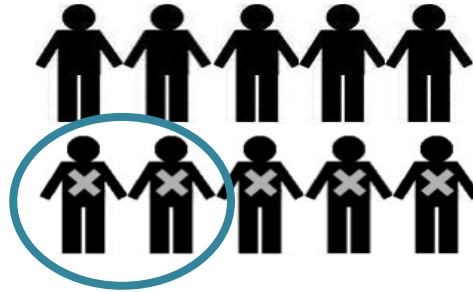
The same ten people, observed at the same time, without the exposure



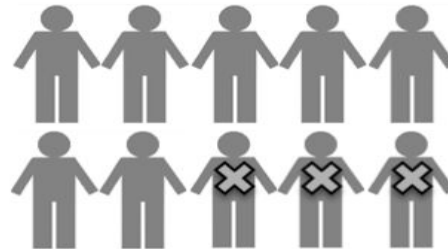
Comparing groups, cont.

Excess cases of disease due to causal effect of the exposure on the outcome

Ten people, all exposed



The same ten people, observed at the same time, without the exposure



Causal study design considerations

- It is impossible to observe the same individuals over the same time period both with and without the exposure
- Instead, we use comparison of exposed and unexposed groups, often observed in parallel over a similar time period
- Ideally we want the **unexposed** group in a population studied to **represent the experience of the exposed group had they not been exposed**

Counterfactual thinking: “thought experiment”

Person 1:



[exposed]



[not exposed]



[exposed & disease]



[not exposed & disease]

Person 2:



[exposed]



[not exposed]

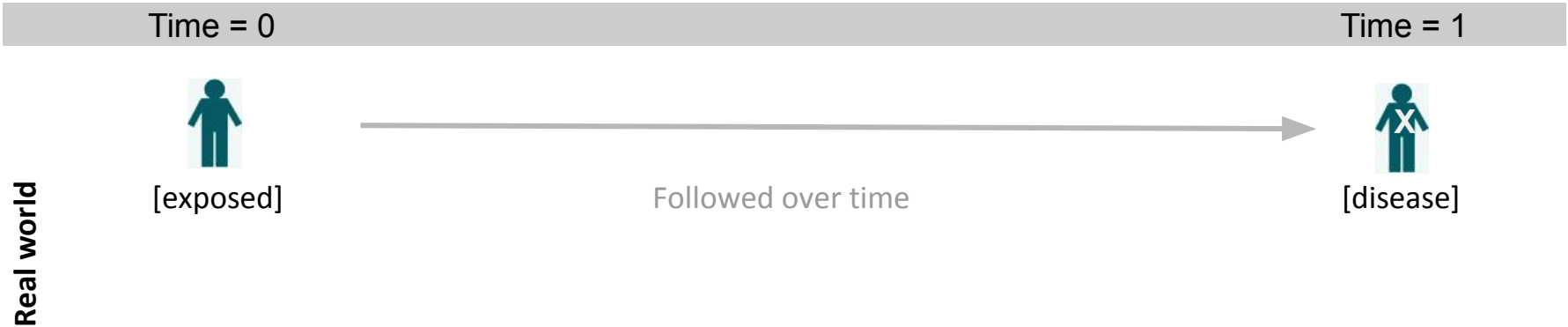


[exposed & disease]



[not exposed & disease]

Counterfactual thinking, cont.



Counterfactual thinking, cont.

Counterfactual



[not exposed]



Followed over time



[disease]

Time = 0

Time = 1

Real world



[exposed]



Followed over time



[disease]

Counterfactual thinking, cont.

Counterfactual

Not observable

Time = 0

Time = 1



[exposed]



Followed over time



[disease]

Real world



[not exposed]



Followed over time



[disease]

Counterfactual thinking, cont.

Counterfactual

Not observable

Time = 0

Time = 1



[exposed]



Is the contrast causal?



[disease]



[not exposed]

Followed over time



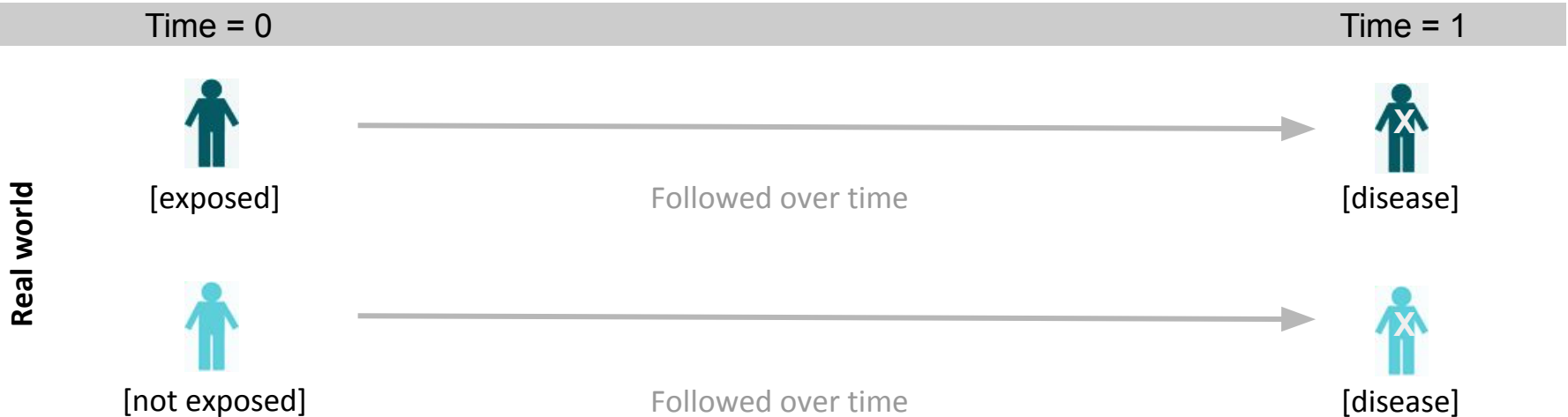
[disease]

Real world

Counterfactual thinking, cont.

Counterfactual

Counterfactual causal effect observable in real world if the difference between person 1 and person 2 is only the exposure



Introduction to Directed Acyclic Graphs (DAGs)

DAGs

- DAGs are a tool to **visualize** causal links between variables (imply causal structure)
- Help to setup the correct (causal) statistical analysis
 - Minimum set of variables needed to *correctly* control for covariates
- Setup of DAGs requires **subject-matter** knowledge
 - **No** purely statistical rules exist to guide setting up DAGs!

DAGs - Think of as 'chain reaction' in one direction



- Causal flow in only **one** direction (directed)
- Flow only possible once (no loops, circles)
- Only forward in time
 - A cause today cannot be affected by future events
- No jumps or skips between variables
 - Need to follow the path(s)

DAGs - Syntax

- A** = Exposure (sometimes “E” or “Z”)
- Y** = Outcome (sometimes “D” or “O”)
- L** = Confounder (sometimes “C”)
- C, W, Z** = Other covariates
- U** = Unmeasured variable (no information)

It does not matter what letters are used, just be sure to check what they mean!

DAGs - The basics



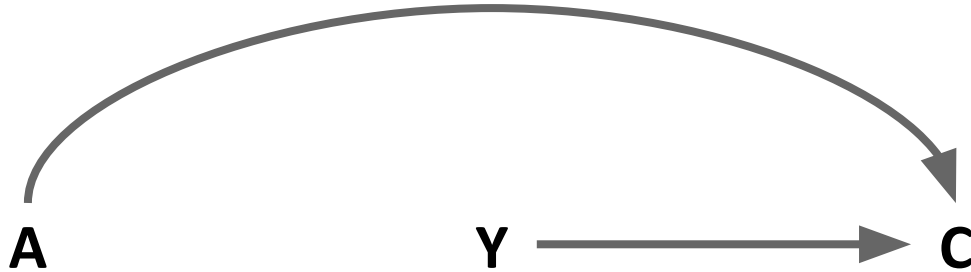
- **A** and **Y** are variables (“knots”)
- **A** is directly causing **Y**
- **A** is the *parent* of **Y**
- **Y** is the *child/descendent* of **A**
- On this DAG, there is only one *path* from **A** to **Y**
- The arrows indicate that the *causal flow* is only from **A** to **Y**
- Ideally, arrows should point from left to right to reflect temporal arrangement

DAGs - The basics, cont.



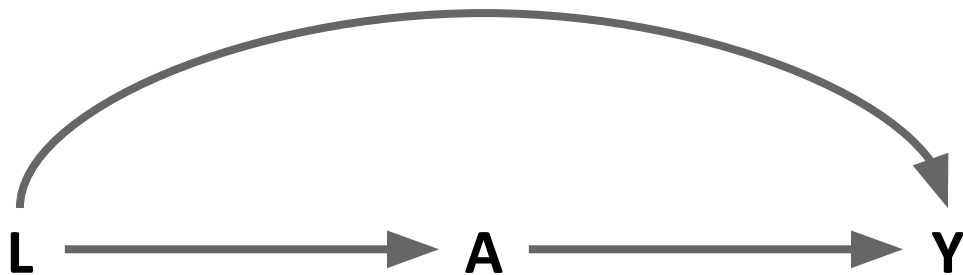
- **A** is a cause of **C** and **C** is a cause of **Y**
- **A** is causing **Y** only via **C**
- There is a causal path from **A** via **C** to **Y**

DAGs - The basics, cont.



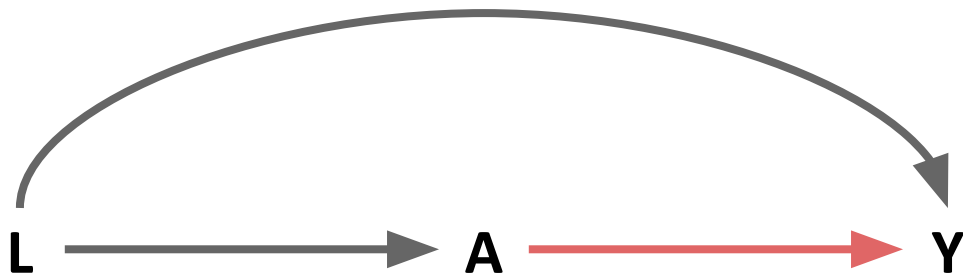
- **A** is a cause of **C** and **Y** is a cause of **C**
- **A** has no causal effect on **Y**
- There is **NO** causal path from **A** via **C** to **Y** (**stops at C**)

DAGs - The basics, cont.



- **L** is a common cause (*parent*) of **A** and **Y**
- There are two *paths* (effects) into **Y**:
 - Direct: from **L** to **Y**
 - Indirect: **L** via **A** to **Y**
- Total effect into **Y** = Sum of direct and indirect effects

Effect of A on Y

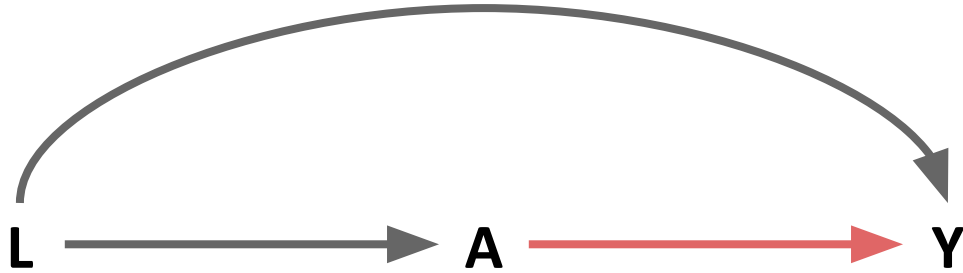


- Can we estimate the *direct* causal effect of **A** on **Y**?
- In the language of Judea Pearl, we say that the *association* between **A** and **Y** fails to identify the *causal effect* of **A** on **Y** because there is an open “*backdoor path*” path from **A** via **L** to **Y**.

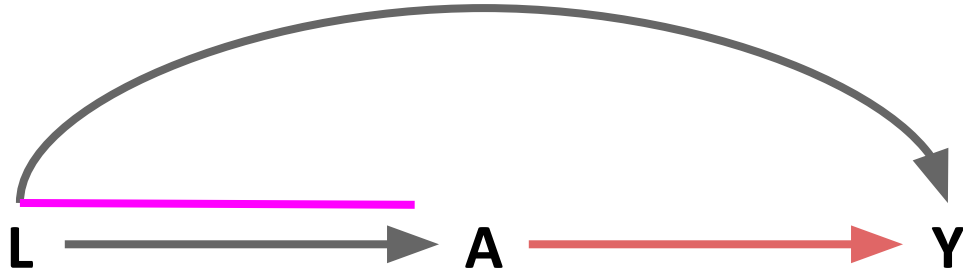
Backdoor criterion

- Go from the exposure of interest (**A**) in the other direction (i.e. not the causal direction; via the backdoor) and see whether there is an *open path* on which you can reach your outcome of interest (**Y**)
- A *backdoor path* is any pathway, without consideration of directionality, that connects **Y** and **A**
- If there is an *open path*: can you *block* the path?

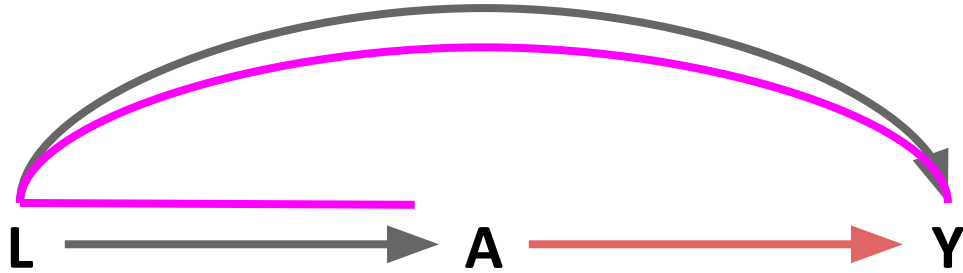
Can we estimate the causal effect of A on Y?



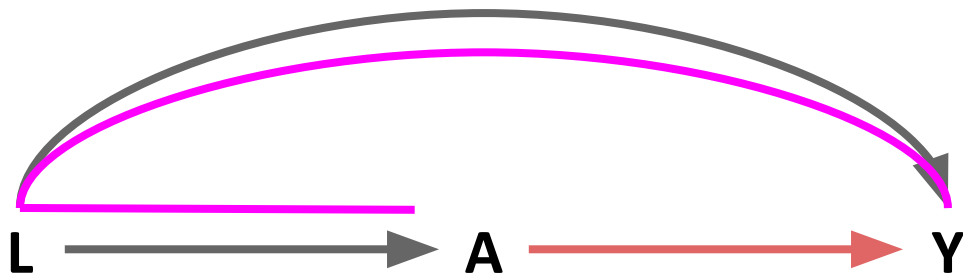
Can we estimate the causal effect of A on Y?



Can we estimate the causal effect of A on Y?

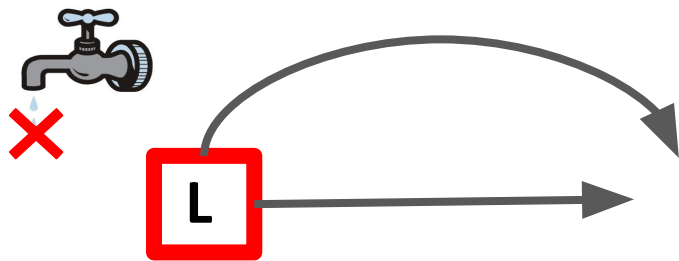


Can we block the *backdoor path*?



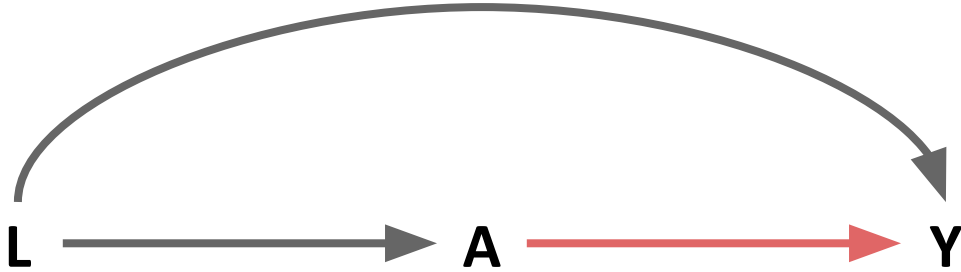
- No, the effect of **A** on **Y** is not *causal*, unless we can *block* the backdoor path!

Blocking a variable

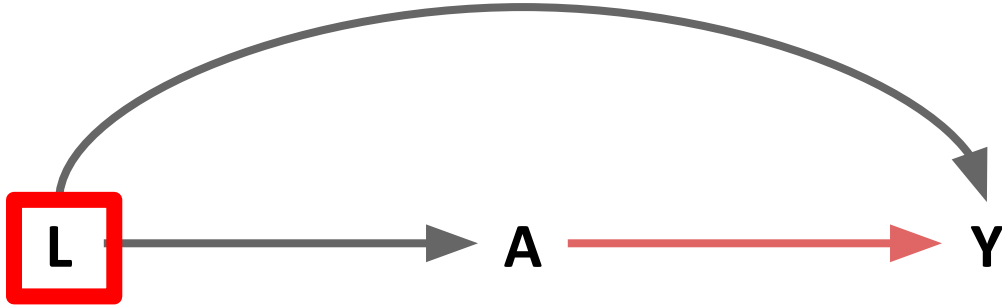


- Blocking a variable is indicated in a DAG by putting a box around that variable
- Blocking = conditioning, controlling, adjusting, etc.
- Unless a variable has two or more *direct* causes, blocking a variable will stop any flow through that variable

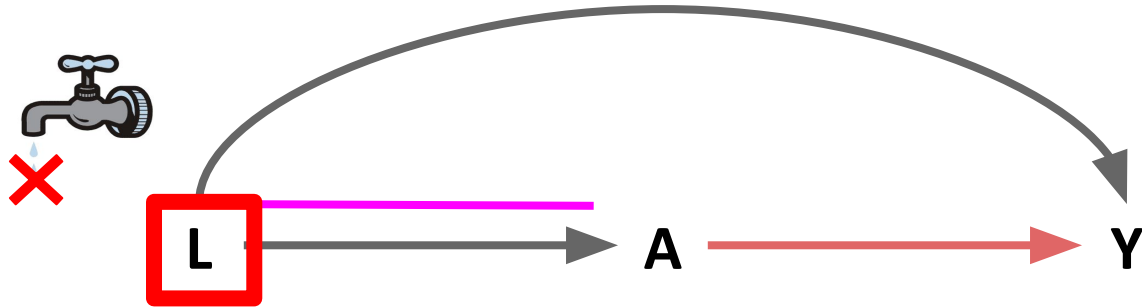
Blocking a variable removes the paths from DAG



Blocking a variable removes the paths from DAG



Can we block the *backdoor path* = YES



- Conditioning (controlling, adjusting) on **L** will block the *backdoor path* from **A**, via **L** to **Y**
- This removes the indirect effect of **A** via **L** on **Y**
- By blocking **L**, we can estimate the causal effect of **A** on **Y**

Blocking a variable removes the paths from DAG



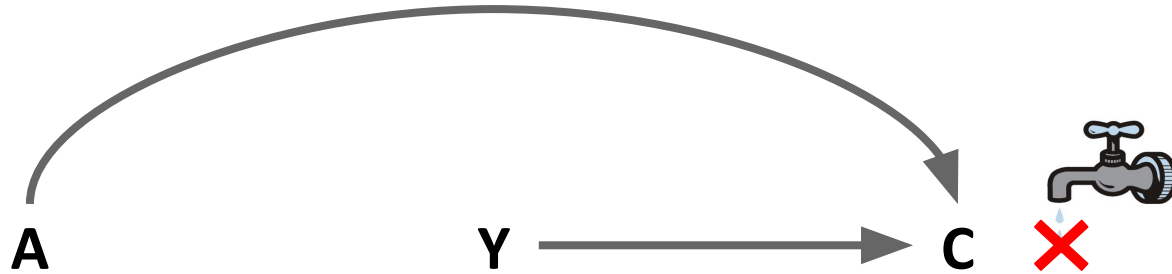
“d-Separation” of two variables

- d-separation between A and Y
 - Corresponds to counterfactual, no confounding

Can occur two ways:

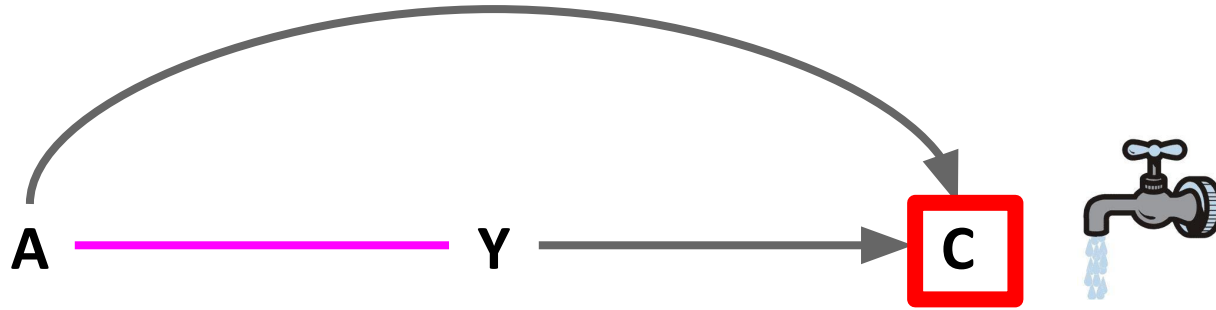
1. Occurs if there is no backdoor path at all
 - a. Usually only the case in randomized controlled trials
2. Occurs if all backdoor paths between A and Y are blocked
 - a. e.g. in a cohort study where we have collected information on all confounders that are necessary to close that path

Collider



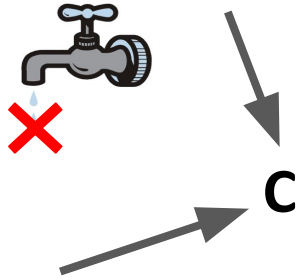
- A collider is a variable that is the consequence of at least two causes
 - Two arrow heads intersect at that variable
- Here, **C** is caused by **A** and **Y**
- The causal flow **stops** at a collider because one cannot 'exit' the collider (arrow heads both point into that variable)

Collider, cont.

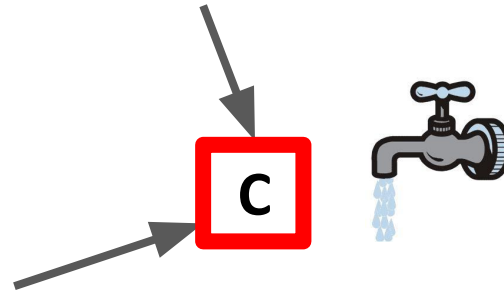


- Blocking (conditioning on, controlling for, adjusting for, etc.) a collider will **allow a non-causal flow** of association through the collider
- Here, by conditioning on **C**, we have created an **association** between **A** and **Y**

Collider flow, summary



Closed

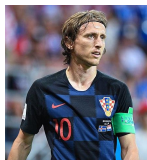


Open

But what does this mean?



Collider: World Cup example



Referee

Football player

**On field at
World Cup**



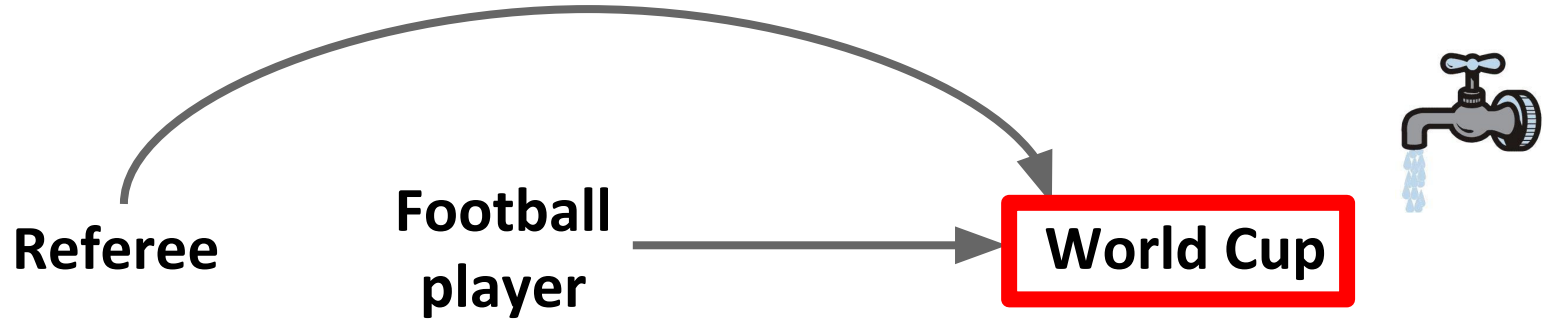
- Causal DAG (both excellent referees and excellent football players are on the field for a World Cup game)
- To be on the field, I either have to be an excellent player or referee
- Being a good referee does not make a person an excellent football player, nor vice versa (independent skills; no link or arrow between referee and player)

Collider: World Cup example



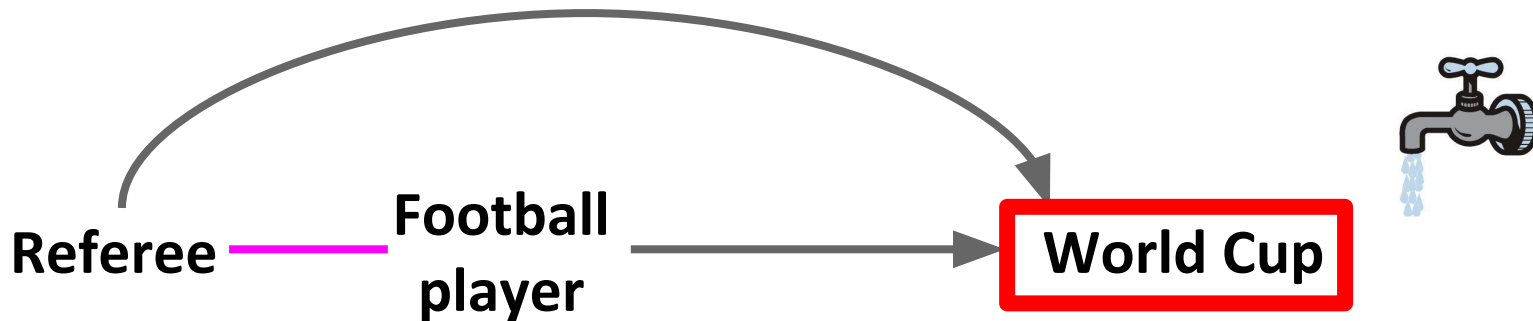
- Not conditioning on “World Cup” stops the flow at “World Cup”

Collider: World Cup example



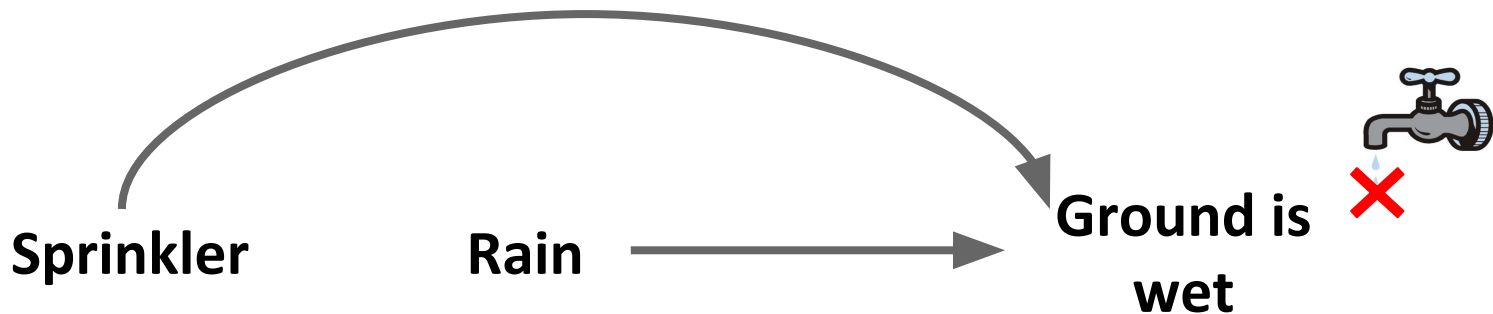
- Let's condition on World Cup

Collider: World Cup example



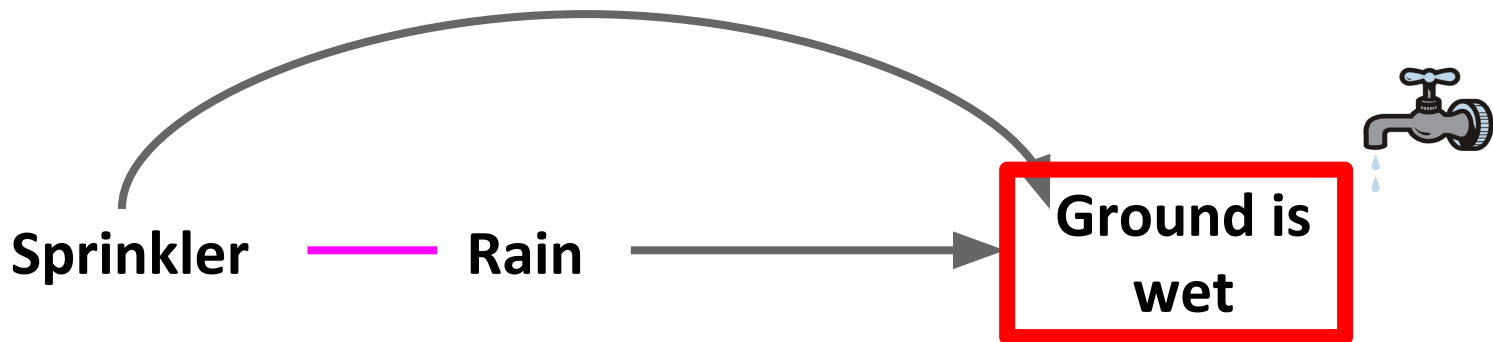
- Let's condition on World Cup
- This creates “spurious” association between referee and player
 - Statistical association that is **not** causal
- By conditioning on a collider, we have created something that is not there in reality!

Conditioning on a collider: example 2



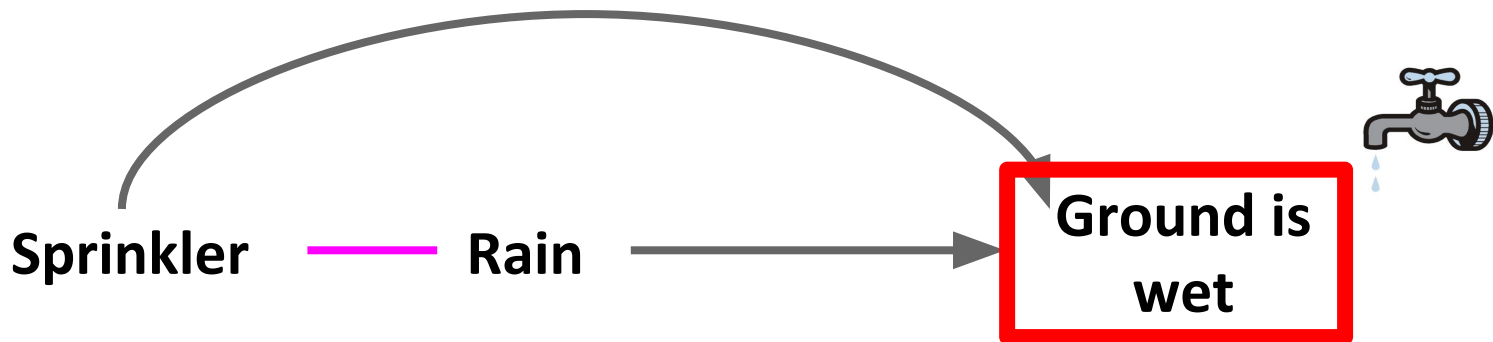
- Only 2 reasons ground can be wet:
 - It can rain
 - Sprinkler (on 1x per week schedule, unrelated to weather)

Conditioning on collider: opens the flow of info.



- We collect data on all 3 variables over time
- If we look at the days when ground is wet (e.g. strata wet = 1)...
 - If we know it rained, is it more or less likely the sprinkler was on?

Conditioning on collider: opens the flow of info.



- We collect data on all 3 variables over time
- If we look at the days when ground is dry (e.g. strata wet = 0)...
 - If we know it rained, is it more or less likely the sprinkler was on?
 - If we know sprinkler was on, is it more or less likely it rained?

From intuition to equations



[exposed]



[not exposed]



[exposed & disease]



[not exposed & disease]

For dichotomous exposure and disease/outcome

- Exposure = A (1: exposed, 0: unexposed)
- Disease/Outcome = Y (1: diseased, 0: not diseased)

For dichotomous (0,1) exposure and outcome



[exposed]

[$a = 1$]



[not exposed]

[$a = 0$]



[exposed & disease]

[$Y^{a=1} = 1$]

(read: observed
outcome Y under
exposure $a=1$)



[not exposed & disease]

[$Y^{a=0} = 1$]

(read: observed
outcome Y under
exposure $a=0$)


Formal definition of a *causal effect*

The exposure **A** has a causal effect on **Y** if:

$$\gamma^{a=1} \neq \gamma^{a=0}$$

Formal definition of a *causal effect*

The exposure **A** has a causal effect on **Y** if:

$$Y^{a=1} \neq Y^{a=0}$$


A causal diagram consisting of a horizontal arrow pointing from the letter **A** on the left to the letter **Y** on the right. The arrow is a solid black line with a triangular arrowhead pointing towards **Y**.

Formal definition of a *causal effect*

The exposure **A** has a causal effect on **Y** if:

$$\gamma^{a=1} \neq \gamma^{a=0}$$

A \longrightarrow **Y**

The exposure **A** has no causal effect on **Y** if:

$$\gamma^{a=1} = \gamma^{a=0}$$

Formal definition of a *causal effect*

The exposure **A** has a causal effect on **Y** if:

$$\begin{array}{ccc} & \gamma^{a=1} \neq \gamma^{a=0} & \\ \mathbf{A} & \longrightarrow & \mathbf{Y} \end{array}$$

The exposure **A** has no causal effect on **Y** if:

$$\begin{array}{ccc} & \gamma^{a=1} = \gamma^{a=0} & \\ \mathbf{A} & & \mathbf{Y} \end{array}$$

Average *causal effects* in populations

- We are usually after **average** causal effects in our population of interest
- The average is equal to the “expectation,” denoted with the letter **E**
- An average *causal effect* of the exposure on the outcome is present if:

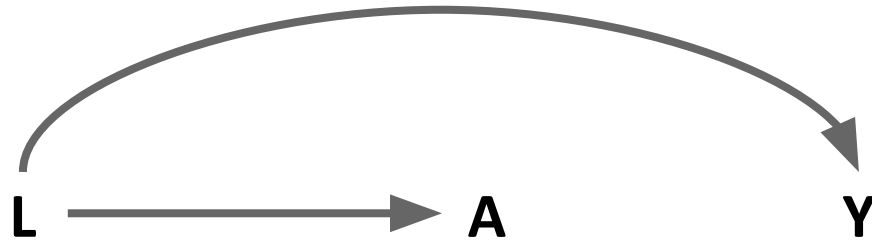
$$E[Y^{a=1}] \neq E[Y^{a=0}]$$

Absolute null hypothesis and confounding

- To be able to estimate causal effects, we need to be sure that in case of *no* causal effect, the exposure and the outcome are *not* associated
- In statistical terms, we say that **A** needs to be *independent*(\perp) of **Y**

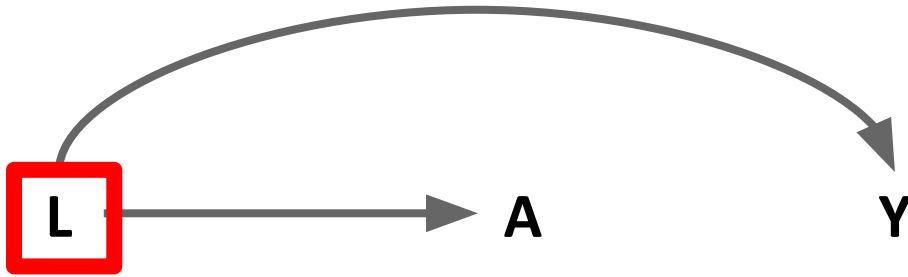
$$Y^{a=1} = Y^{a=0} \quad \text{or} \quad Y \perp A$$

Testing the absolute null hypothesis in observational studies



- Is **Y** independent of **A** (i.e., $Y \perp A$)?

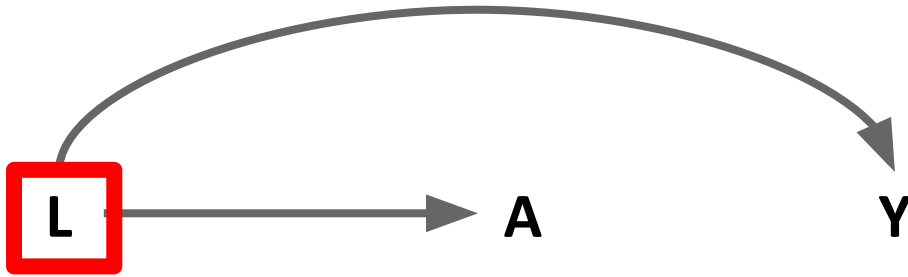
Testing the absolute null hypothesis in observational studies



- Is **Y** independent of **A** (i.e., $Y \perp A$)?

No, *unless* we block **L**

Testing the absolute null hypothesis in observational studies



- Is **Y** independent of **A** (i.e., $Y \perp A$)?

No, unless we block L

$$Y \perp A \mid L$$

Reads: **Y** is independent of **A** given **L**

Estimating causal effects

$$E[Y^{a=1} | L] - E[Y^{a=0} | L]$$

- Expected effect of exposure $a=1$ on outcome Y conditioned on L *minus* the expected effect of exposure $a=0$ on the outcome Y conditioned on L
- Causal risk difference

Confounder

What is a “confounder”?

Confounder, revisited

Usual definition:

- A confounder is associated with the exposure
- A confounder is associated with the outcome (independently of the exposure)
- A confounder is not in the path between exposure and outcome

Confounder, revisited

Usual definition:

- A confounder is associated with the exposure
- A confounder is associated with the outcome (independently of the exposure)
- A confounder is not in the path between exposure and outcome

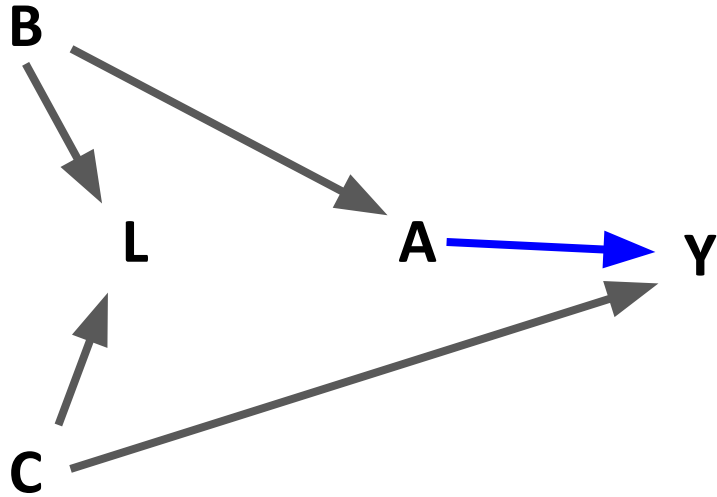
New definition:

- A confounder is a common cause (direct or indirect via another variable) of both the exposure and the outcome
 - Implies that a confounder cannot be a consequence of exposure

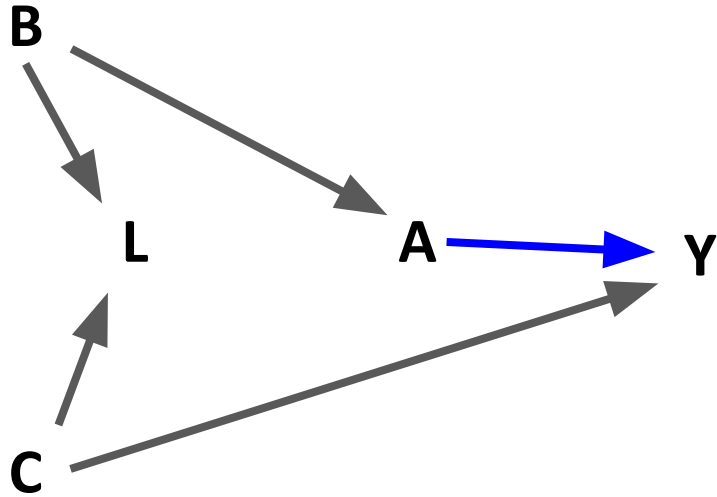
Confounding and confounder

- We must separate the concept of confounding from individual confounders!
- Confounding arises if the effect of exposure on an outcome is **not solely caused by the exposure**
 - We need to ensure the exposure is “d-separated” from the outcome
 - *No open backdoor paths* can exist!
- A *confounder* is a covariate for which we have information. We can **adjust** for this confounder to block *confounding*.

Is L a confounder of the effect of A on Y?

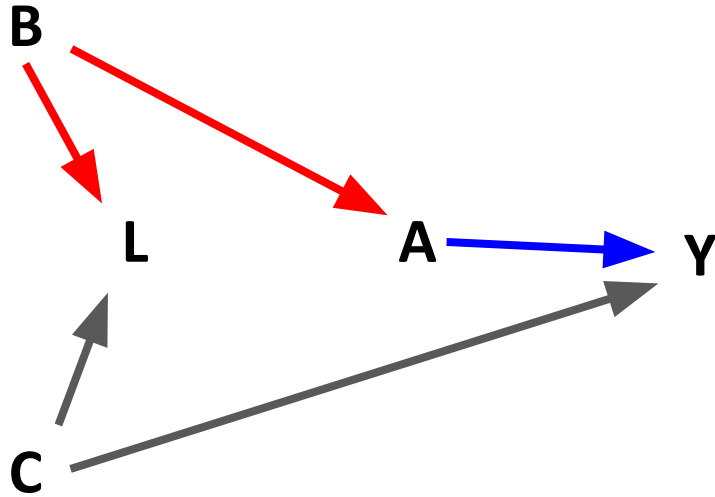


Confounder: old definition



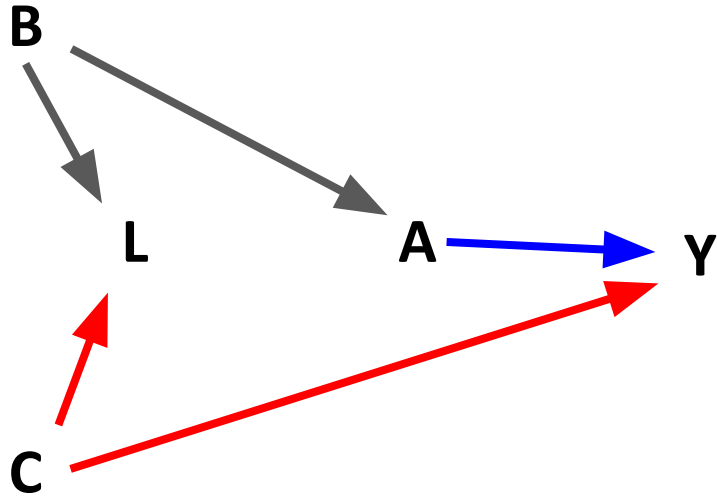
- Is L associated with A?
- Is L associated with Y?
- Is L in the path between A and Y?
- Consequence?

Confounder: old definition



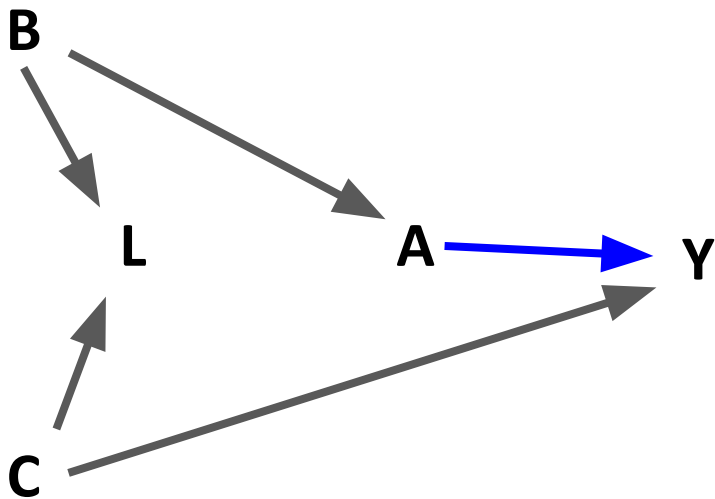
- Is L associated with A?
→ Yes, via the path L, B, A
- Is L associated with Y?
- Is L in the causal path between A and Y?
- Consequence?

Confounder: old definition



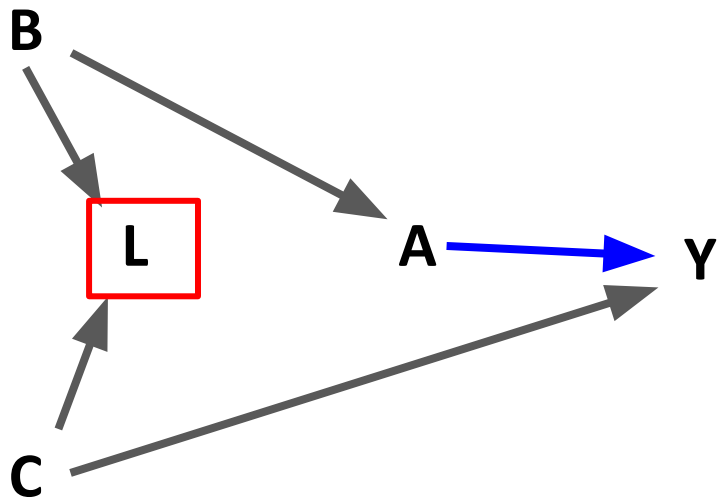
- Is L associated with A?
→ Yes, via the path L, B, A
- Is L associated with Y?
→ Yes, via the path L, C, Y
- Is L in the causal path between A and Y?
- Consequence?

Confounder: old definition



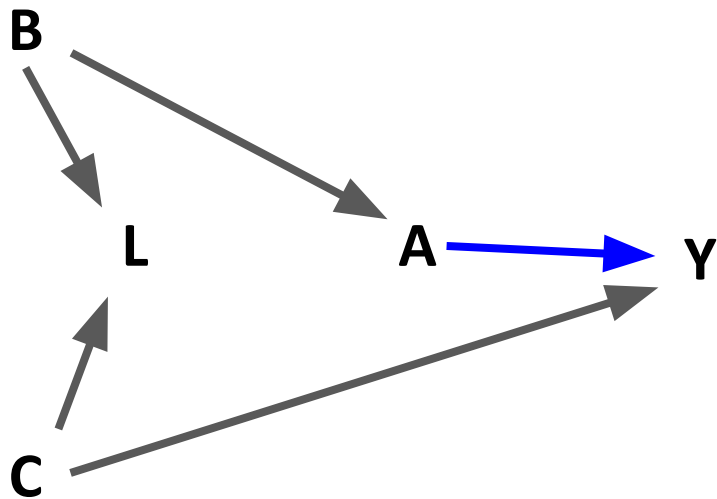
- Is L associated with A?
→ Yes, via the path L, B, A
- Is L associated with Y?
→ Yes, via the path L, C, Y
- Is L in the causal path between A and Y?
→ No
- Consequence?

Confounder: old definition



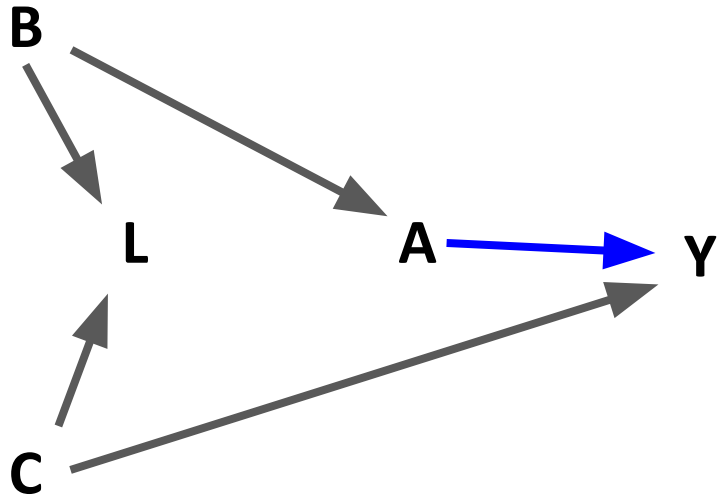
- Is L associated with A?
→ Yes, via the path L, B, A
- Is L associated with Y?
→ Yes, via the path L, C, Y
- Is L in the causal path between A and Y?
→ No
- Consequence?
→ L is a confounder, need to adjust

Confounder: new definition



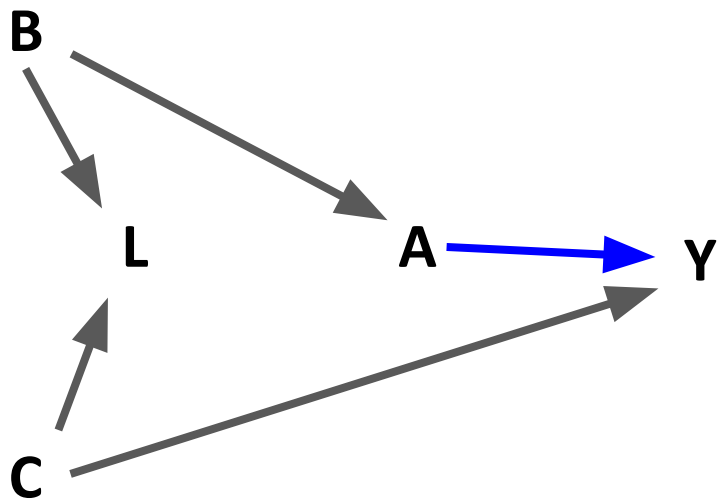
- Is L a common cause of A and Y?
- Is L a consequence of A?
- Is there an open *backdoor path* from A to Y?
- Consequence?

Confounder: new definition



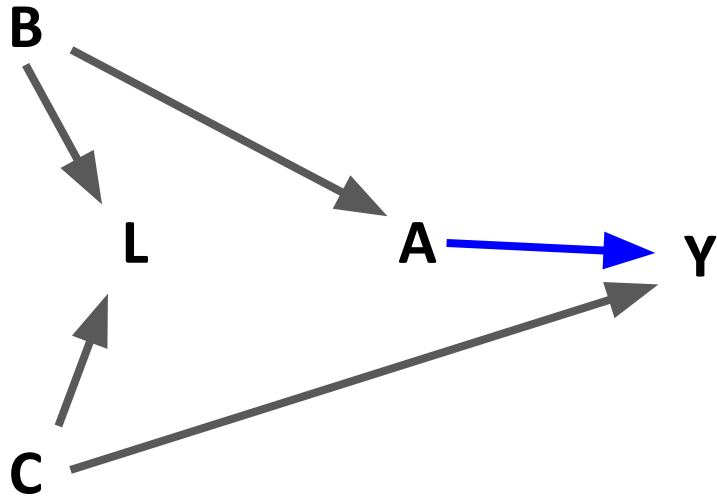
- Is L a common cause of A and Y?
→ No
- Is L a consequence of A?
- Is there an open *backdoor path* from A to Y?
- Consequence?

Confounder: new definition



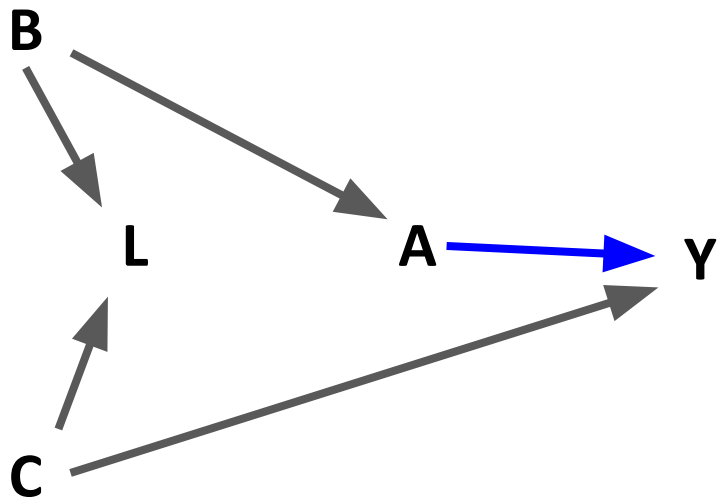
- Is L a common cause of A and Y?
→ No
- Is L a consequence of A?
→ No
- Is there an open *backdoor path* from A to Y?
- Consequence?

Confounder: new definition



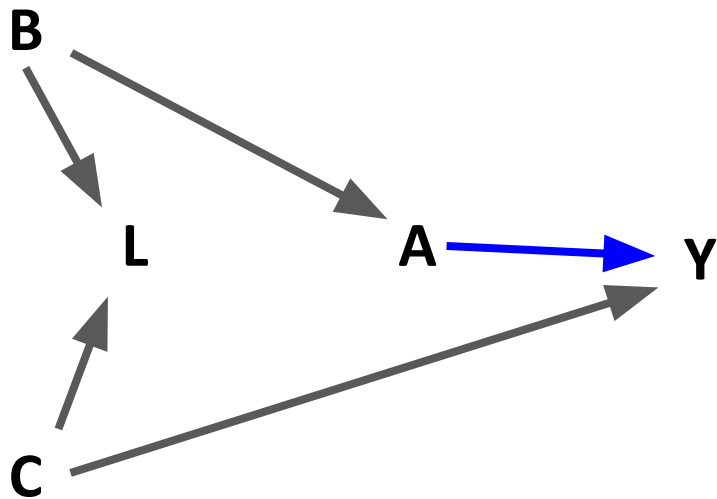
- Is L a common cause of A and Y?
→ No
- Is L a consequence of A?
→ No
- Is there an open *backdoor path* from A to Y?
→ No
- Consequence?

Confounder: new definition



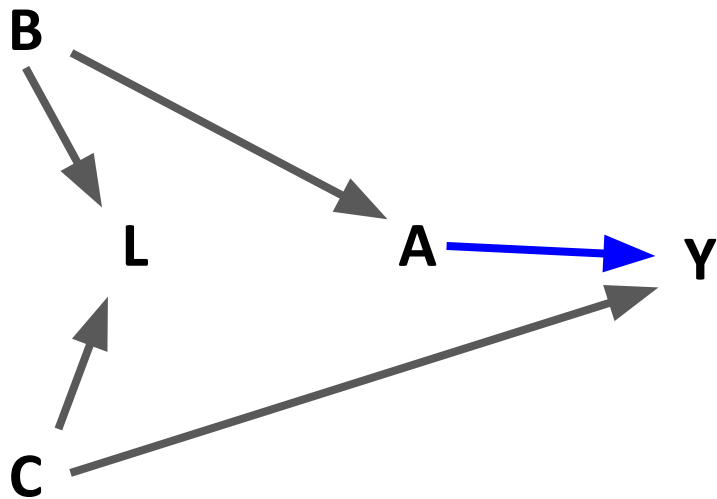
- Is L a common cause of A and Y?
→ No
- Is L a consequence of A?
→ No
- Is there an open *backdoor path* from A to Y?
→ No
- Consequence?
→ L is **not** a confounder, no need to adjust

Is there confounding?



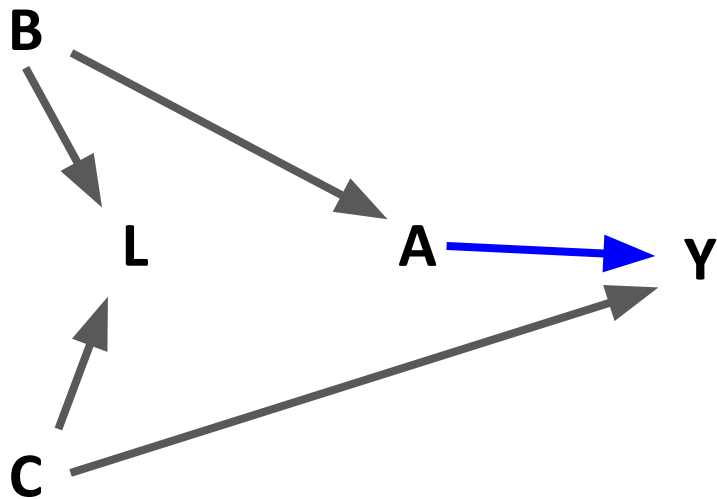
- Is there an open *backdoor path* from A to Y?
- Are A and Y “d-separated”?
- Consequence

Is there confounding?



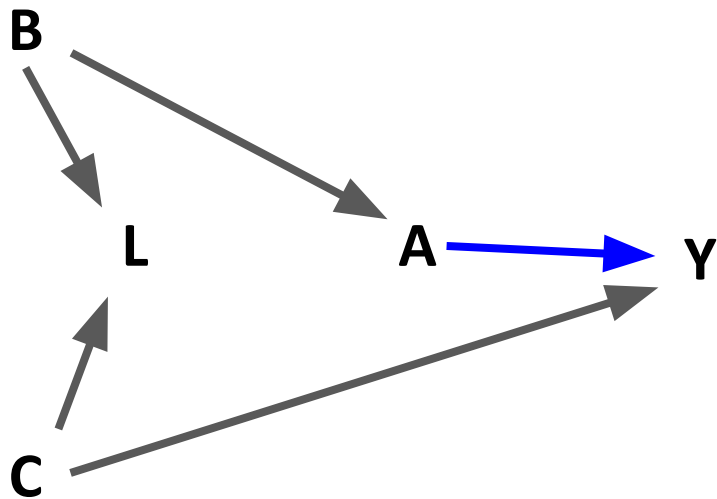
- Is there an open *backdoor path* from A to Y?
→ No
- Are A and Y “d-separated”?
- Consequence

Is there confounding?



- Is there an open *backdoor path* from A to Y?
→ No
- Are A and Y “d-separated”?
→ Yes
- Consequence

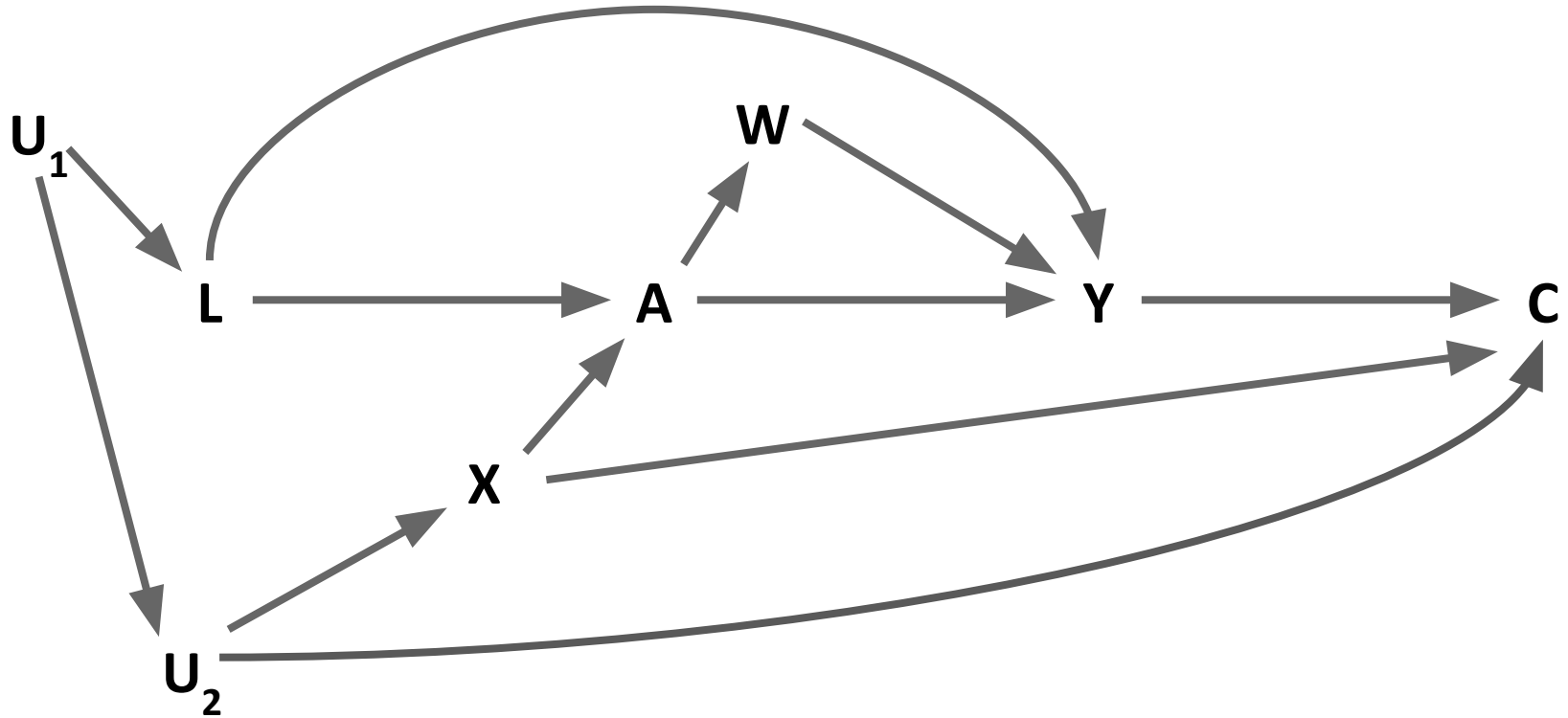
Is there confounding?



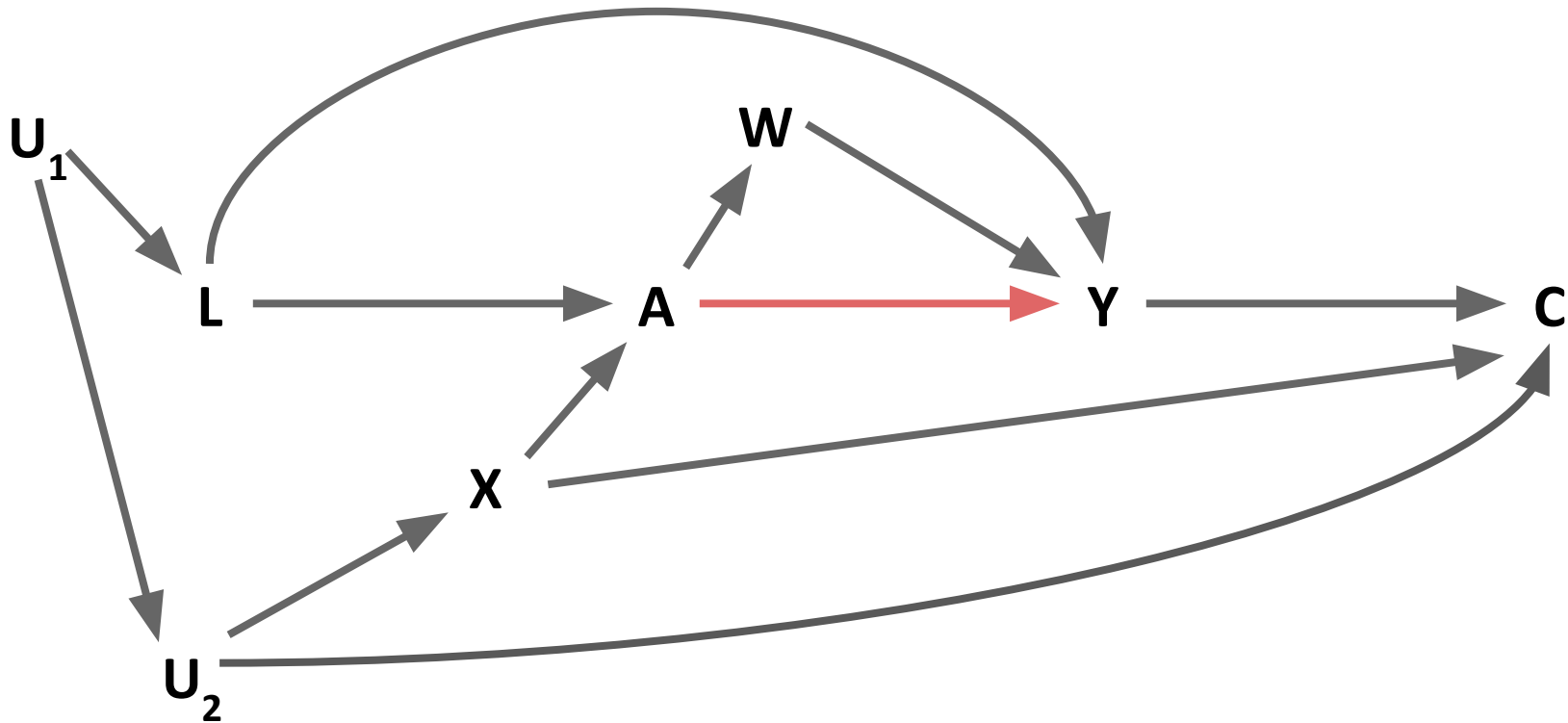
- Is there an open *backdoor path* from A to Y?
→ No
- Are A and Y “d-separated”?
→ Yes
- Consequence
→ There is no confounding
→ The **crude** association is causal

In this scenario, adjusting for L actually INTRODUCES a bias!

Reduction to only the essentials



Effect A on Y? For which variable must we control?

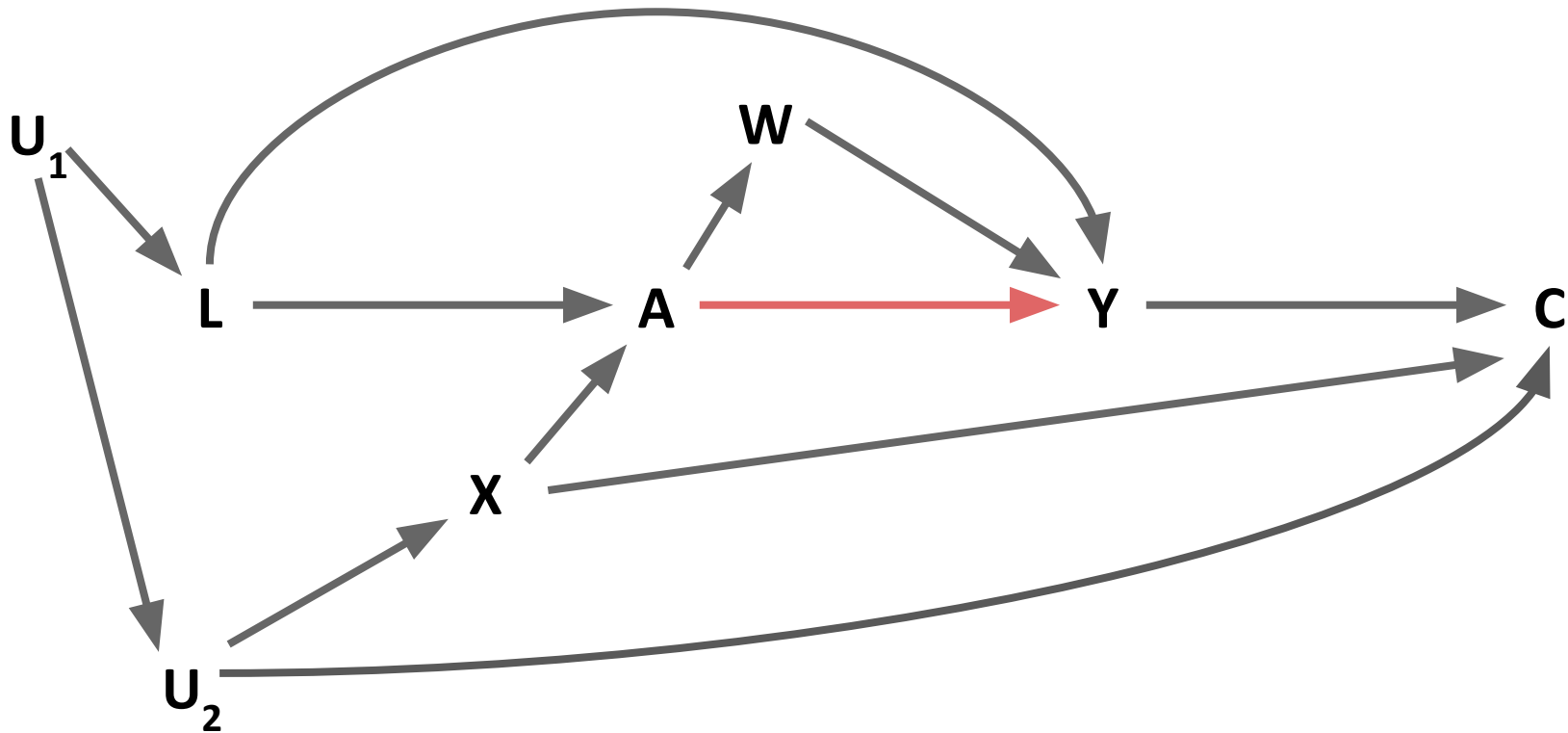


Adjust for or not?

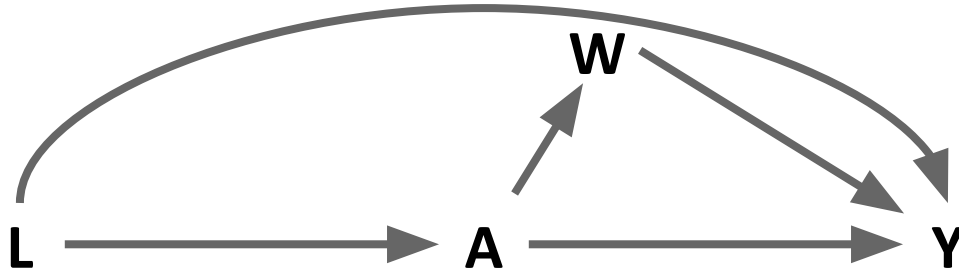
- U_1 & U_2** not possible since unmeasured
- L** Yes, since it's a confounder
- W** No, since it's an intermediate on path between A to Y
- X** No, since it's not a confounder
- C** No, since it's a consequence (child) of outcome & a collider

What does the reduced DAG look like?

Effect A on Y? Minimally sufficient set to control for?

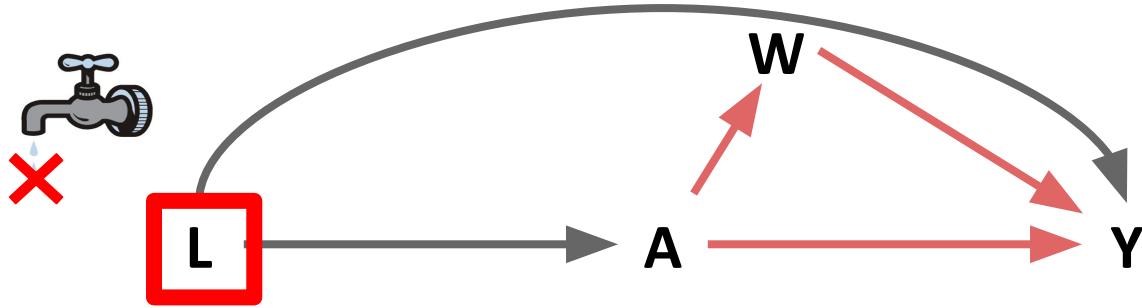


Reduced DAG for causal effect of A on Y



- We only have to block L to get to estimate the causal effect of A on Y

Estimating the causal effect of A to Y



Effect (*measure*) modification

- Is different from confounding
- Occurs when the magnitude of the effect of the primary exposure on an outcome differs depending on the *level* of a *third variable*
- Strongly depends where we **measure** it, hence the term effect *measure* modification
 - Most often effect modification is used without specification of where it has been measured

Semantic: effect modification vs. interaction

- Effect modification and interaction are often used interchangeably
- Misconception:
 - Interaction = statistical definition
 - Effect modification = based on biological ground
- Difference between statistical effect measure modification and interaction is conceptual
 - Do I know what to look for and test it statistically? or
 - Do I test any/all combinations in my dataset for possible interaction?

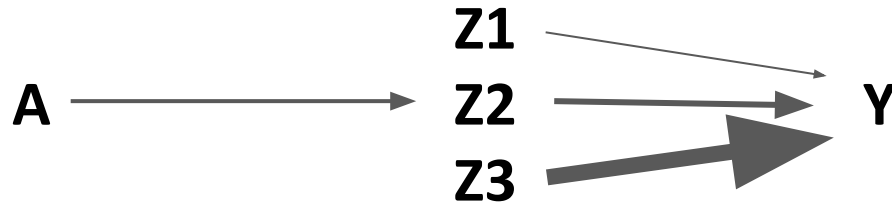
Difference between interaction and effect modification

- **Interaction** refers to an interaction of **two effects** (ie, treatments) on an outcome (ie, they interact with each other)
- As composed to **effect measure modification**, where the effect of the primary exposure *differs* in levels of a third variable
- Thus,
 - Effect modification can be present with no interaction
 - Interaction can be present with no effect modification
 - There are settings in which it is possible to assess effect modification but not interaction, or to assess interaction but not effect modification

When to look for effect (measure) modification

- If there is a clear biological mechanism by which the effect of the exposure on the outcome differs in level of a third variable
 - Hormones work differently in men and women
 - Smoking affects brain function differently in the young than in the old
- If we would like to understand (but there is little biological evidence) that exposure-outcome effect is magnified in subgroups
 - CAVE: the interpretation of such findings fall within the concept of hypothesis generation

Effect measure modification

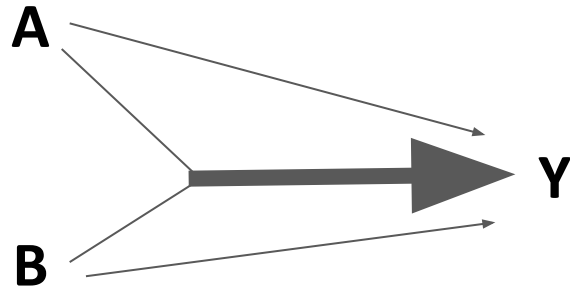


The effect of **A** on **Y** is modified by levels of **Z**

THIS IS NOT A DAG!!

Effect measure modification cannot simply be included in a DAG

Interaction

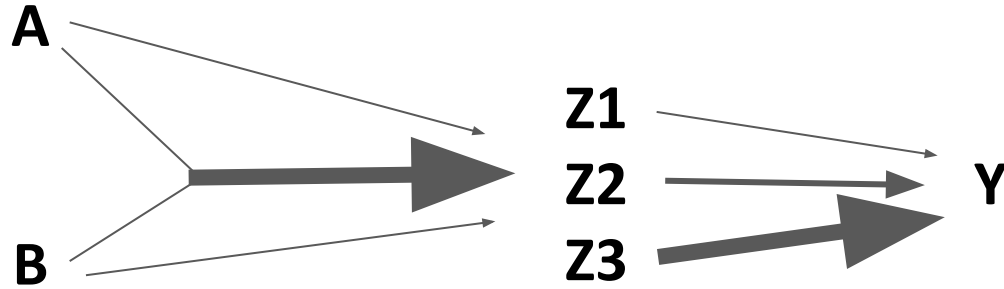


THIS IS NOT A DAG!!

Interactions cannot simply be included in a DAG

- The effect of **A** and **B** on **Y** interact
- Given A or B alone has less of an effect than given both together

Both, interaction and effect measure modification



THIS IS NOT A DAG!!

Effect measure modification and confounding

- Confounding is a harmful effect that we want to completely eliminate in our study when reporting (causal) effects
- Effect measure modification is describing important variation of the exposure - outcome effect in levels of a third variable
 - We should report this
- Complex issue: “testing” for effect measure modification vs. “confounder”
 - If a variable is modifying the exposure effect on the outcome, it cannot be part of confounding based on causal structures!

Confounding and effect measure modification

- The causal conception of confounding must happen before the exposure (open *backdoor path*...)
 - Temporality is crucial
- Effect measure modification can only happen **after** exposure
 - To evaluate whether confounding or effect measure modification is present **cannot** be decided solely based on inference from the data!
 - Cannot test for this

One Research Question, One DAG

- Problematic to use same confounding adjustment scheme for multiple RQs!

Thank You!



BERLIN SCHOOL OF
PUBLIC HEALTH

