

Causality and DAGs

Core principles of epidemiology - Lecture 1

Public Health Data Science

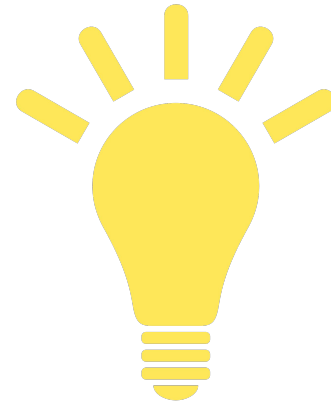
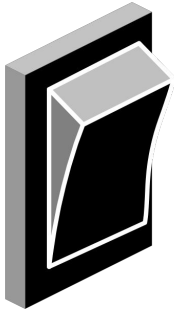
Arthur Hughes

9/9/2024

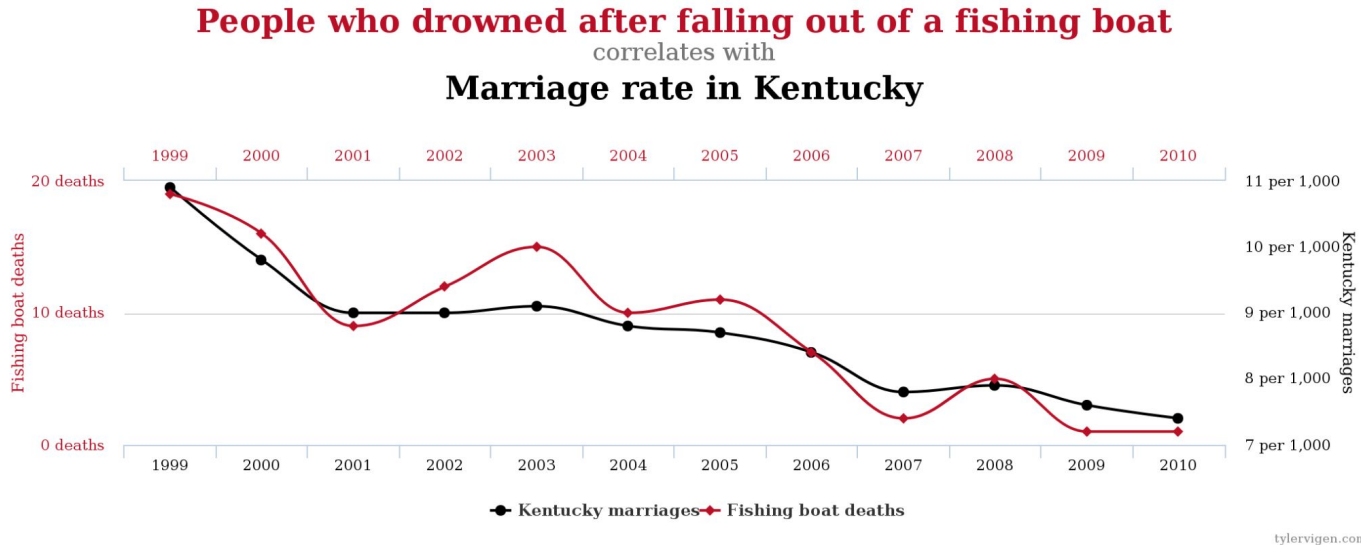
Causality

Etiological studies -> **causes of health outcomes**

The knowledge of anything, since all things have causes, is not acquired or complete unless it is known by its causes. Therefore in medicine we ought to know the causes of sickness and health. - Avicenna, 1012



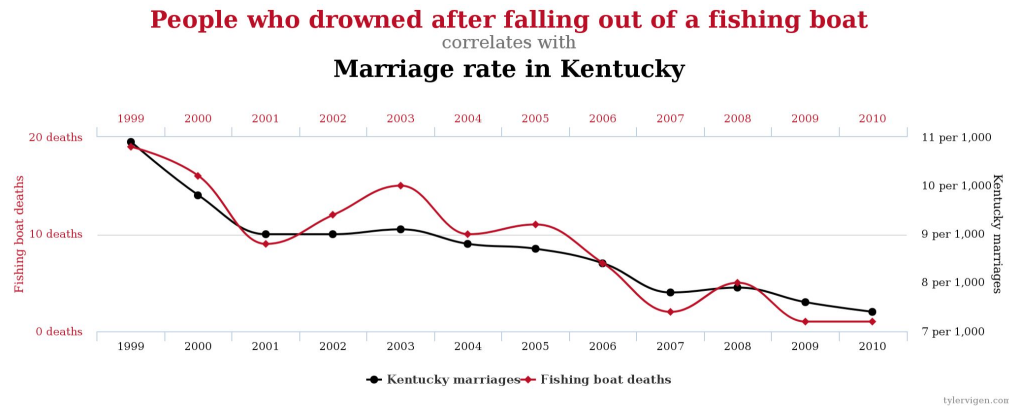
Causation vs association



Why could an exposure and outcome be associated?

Causation vs association

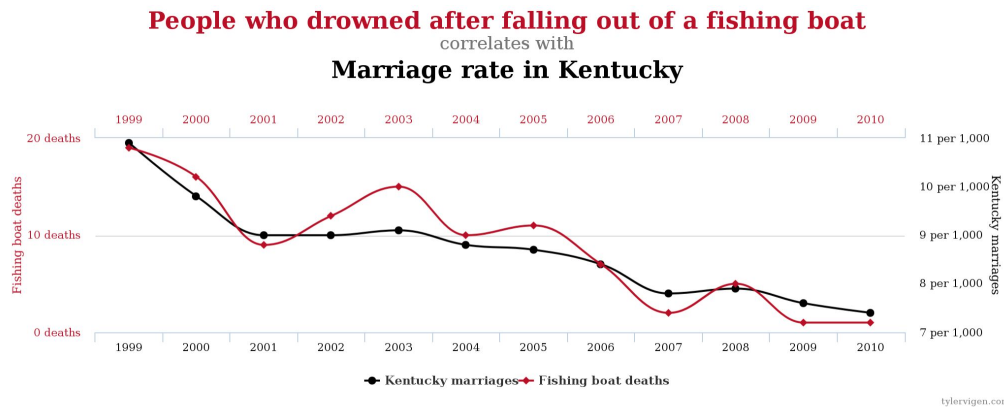
Why could an exposure and outcome be associated?



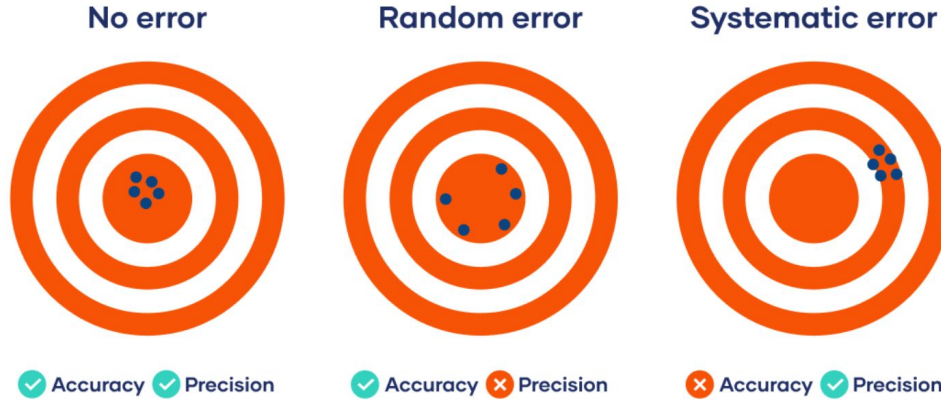
Causation vs association

Why could an exposure and outcome be associated?

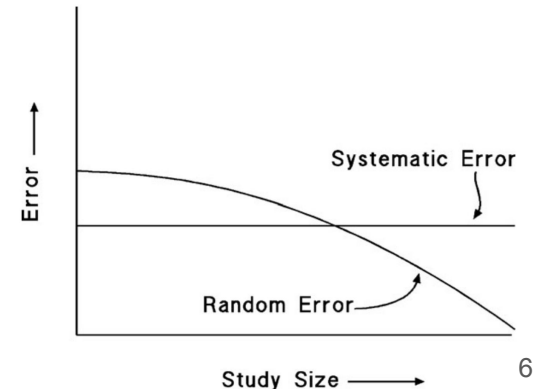
- Exposure causes outcome
- Outcome causes exposure
- Spurious association
 - Random error
 - Systematic Error
 - Confounding
 - Selection bias
 - Information bias



Random vs Systematic Error



- As sample size increases, random error $\rightarrow 0$
 - e.g. estimating a person's height
- Systematic error cannot be eliminated by increasing sample size
 - e.g. systematically faulty instrumentation
 - **Goal : reduce through valid study design and/or analysis choices!**



Selection Bias

- **Self-selection bias** : individuals with a certain characteristic more likely to volunteer
 - Health-conscious people
 - Those with family history of an illness
- **Healthy-worker effect** : Workers of a particular occupation compared to non-workers to identify risk of occupational hazard
 - Workers are typically healthier than non-workers
- **Informative loss-to-follow-up** : participants with a certain characteristic more likely to drop out of the study
 - Depressed people less likely to attend doctors' visits

Information bias

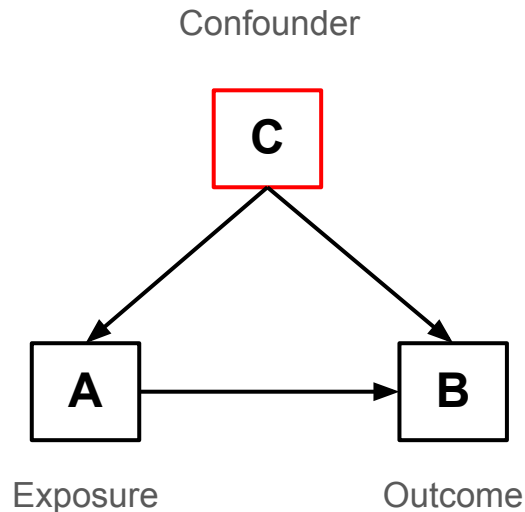
- **Misclassification bias** : Participants put in wrong categories
- **Recall bias** : participants misremembering past events
- **Differential** if error depends on exposure or outcome
 - Maternal recall bias : mothers who miscarry more likely to remember past exposures than mothers with healthy babies
 - Exposed individuals more likely to be diagnosed
- **Non-differential** if error does not depend on exposure or outcome
 - Effect of exposure “diluted”

Confounder bias

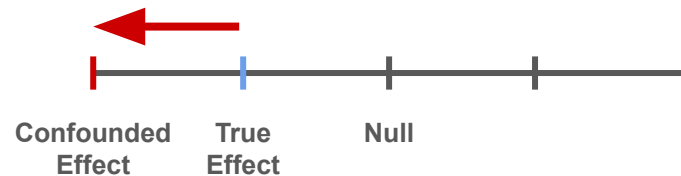
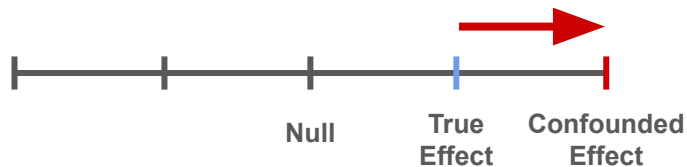
- “Common cause of exposure and outcome”
- Distorts the apparent exposure-outcome relationship

Three conditions for a confounder

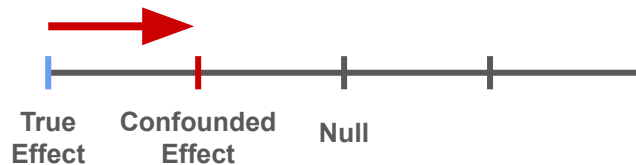
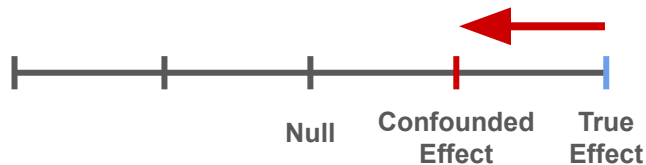
1. Associated with exposure
2. Associated with outcome
3. Not be an effect of the exposure or outcome



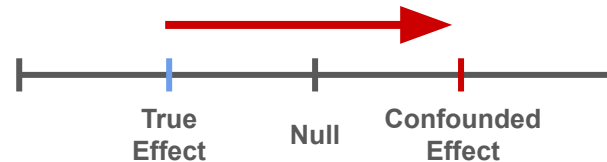
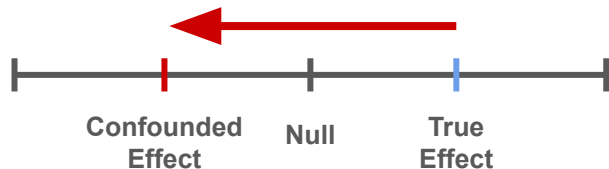
Positive confounding



Negative confounding



Qualitative confounding



Dealing with confounding through **study design**

Randomisation of exposure

- Ensures same **expected** distribution of characteristics between exposure groups
- Renders confounding error random instead of systematic

Restriction

- Study restricted to one value of a confounding variable (e.g. participants same age)

Matching

- Exposed and control “matched” on certain characteristics
- Can be done by matching distributions or on individual level
- Induces bias in case-control studies

Dealing with confounding **analytically**

Regression

- Include confounders as covariates in regression model

Stratification

- Analyse data within subgroups of confounding variable (e.g. per-age-group analysis)

Inverse probability of treatment weighting (IPTW)

- Create pseudo-population where confounding variables are balanced between groups

G-computation

- Predict response for each individual under different values of confounders and contrast

Bradford-Hill criteria for causality

1. **Strength** - are exposure and outcome strongly associated?
2. **Consistency** - have other studies found the same?
3. **Specificity** - is the exposure specific to the cause?
4. **Temporality** - did the exposure occur before the cause?
5. **Biologic gradient** - is there a “dose-response”?
6. **Plausibility** - is there biological plausibility?
7. **Coherence** - coherent with existing knowledge?
8. **Experimental evidence** - change in exposure = change in effect in experimental setting?
9. **Analogy** - do other similar relationships exist?



Austin Bradford-Hill
(1897-1991)

What are the problems with the BH criteria?

Criterion	Criticism	Example
Strength		
Consistency		
Specificity		
Temporality		
Biological gradient		
Plausibility		
Coherence		
Experimental evidence		
Analogy		

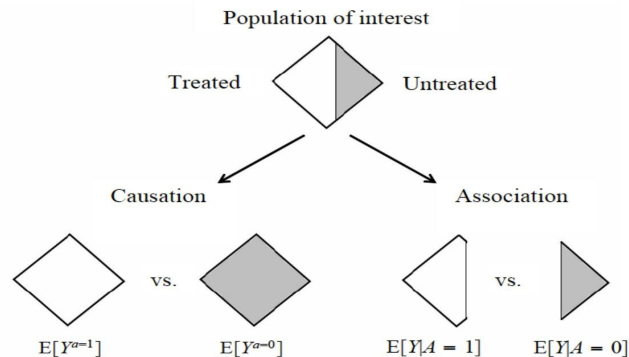
What are the problems with the BH criteria?

Criterion	Criticism	Example
Strength	Weak associations can be causal, strong associations can be non-causal (confounding)	Any confounded relationship (ice cream sales and shark attacks)
Consistency	Many studies may suffer from the same biases, differences in study design makes results differ	
Specificity	An exposure may be non-specific but still causal	Smoking is a risk factor for many diseases
Temporality	Hard to establish temporal sequence	Many cancers have a long 'latent' undiagnosed period
Biological gradient	Threshold effects	Alcohol consumption may not be a risk factor for liver disease until some threshold is passed
Plausibility	Subjective, mechanisms are often complex and defy current knowledge	Air pollution and cognitive decline
Coherence	Ill defined, existing knowledge may be faulty	
Experimental evidence	Not all exposures are able to be studied in experimental settings	Unethical to randomise harmful exposures like smoking
Analogy	Causal mechanisms may be too complex or specific to find similarities	Asbestos and lung cancer - does not apply to all carcinogens.

Modern approaches to causality

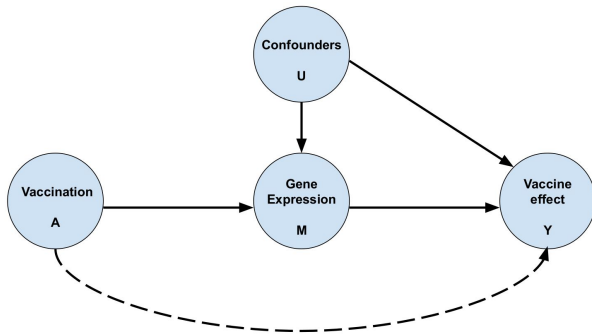
Counterfactual approach

- **Concept** : *What if we could observe individuals under both exposure conditions?*
- Counterfactual quantities can be estimated from data under certain assumptions

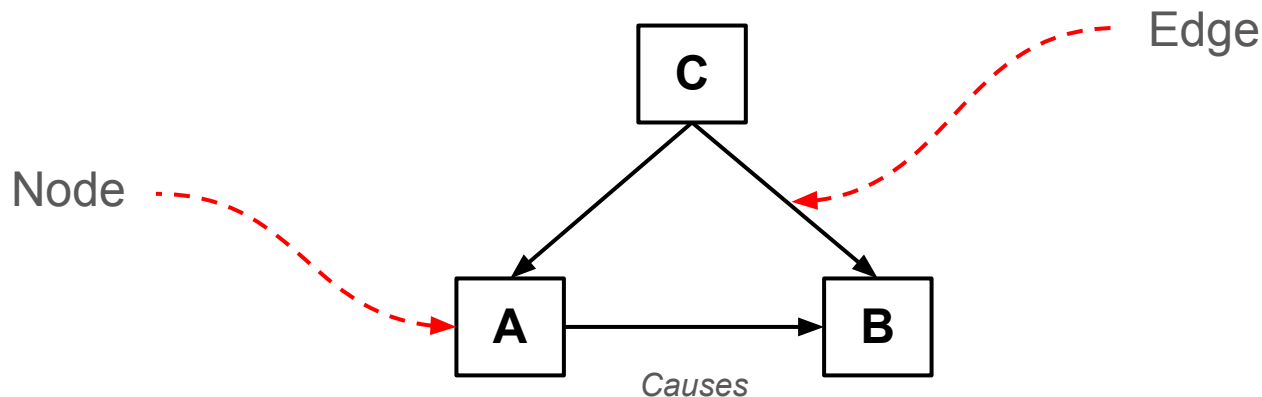


Directed Acyclic Graphs

- Graphic tool to depict **hypothesised causal relationships** in a system
- Can be used to decide **which variables to control** for to avoid biases



Anatomy of a DAG

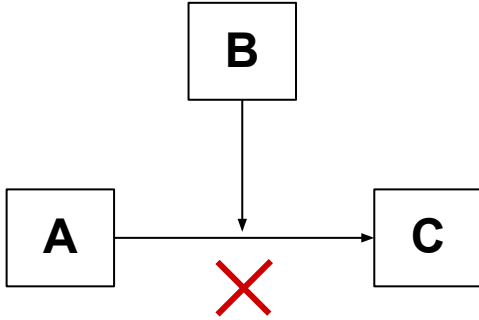


DAGs **do not** tell us the...

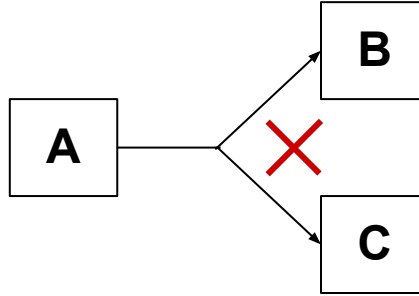
- Form of relationships (non-parametric)
- Direction of the association
- Strength of the association

Edge Rules

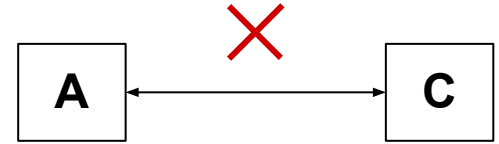
Edges between two variables only



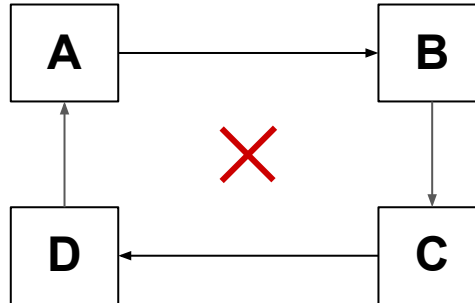
No splitting or joining



One direction only

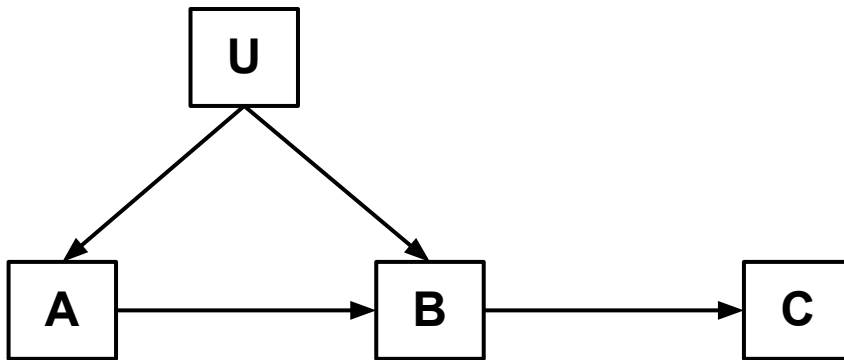


No closed loops



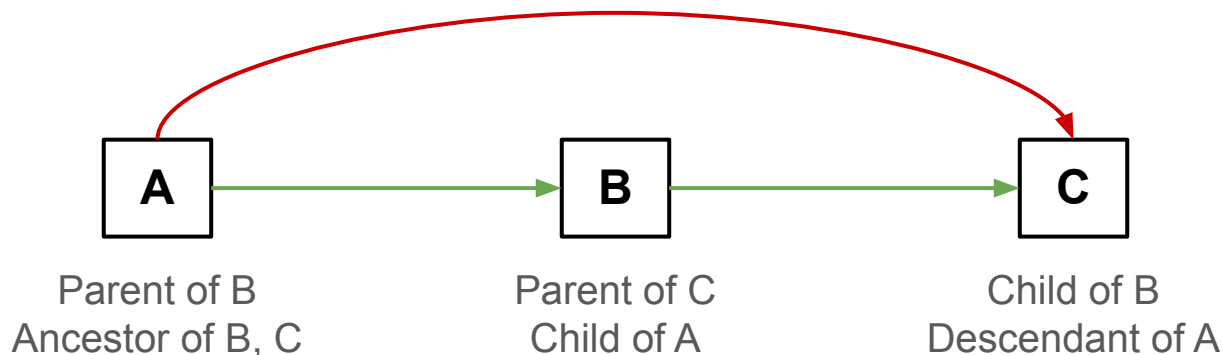
General DAG rules

- **Common causes must be included** - even if unmeasured (denoted U)
- Absence of an edge implies no relationship
- **Faithfulness** : connected variables are dependent (no cancellation of effects)

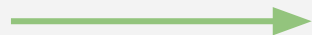


Terminology

- **Path** : Sequence of connected nodes, regardless of direction
- **Directed causal path** : path following the direction of edges
- **Direct effect** : directed causal path with one single edge
- **Indirect effect** : directed causal path with more than one edge
- **Parent (Child)** : direct cause (effect) of a particular variable
- **Ancestor (Descendant)** : direct or indirect cause (effect) of a variable



Direct effect of A on C



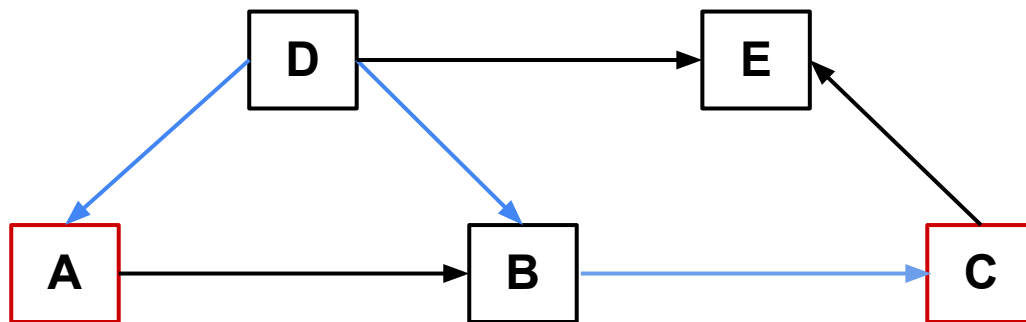
Indirect effect of A on C

Terminology

- **Confounder** : Ancestor of two variables
- **Mediator** : Variable on the directed causal path between two variables
- **Collider** : Child of two other variables
- **Backdoor path** : Path starting with arrow going in to some variable

We are interested in the relationship between **A** and **C**.

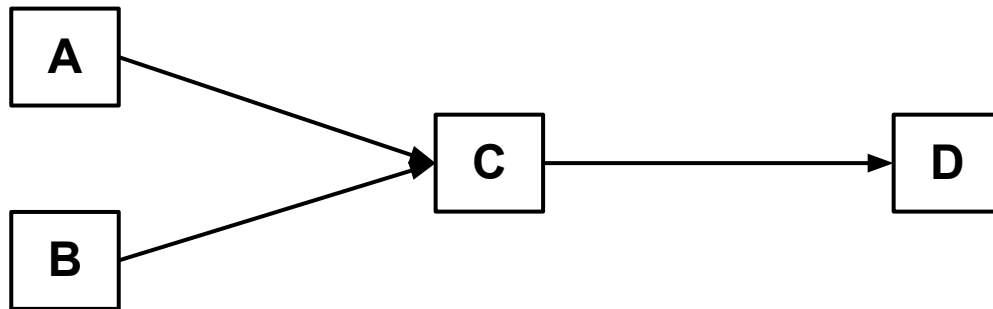
1. What do the variables **D**, **E**, and **B** represent?
2. What is the blue path an example of?
3. Does **A** have a direct effect on **C**?



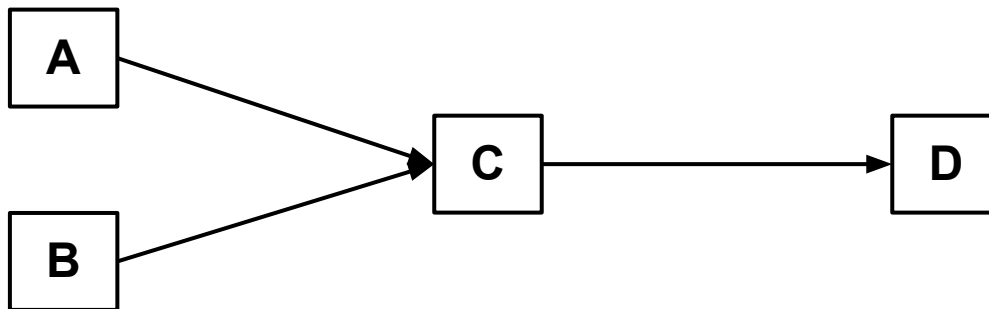
Causal vs Marginal Effects

Marginal association : association regardless of values of other variables

1. What causal assumptions are encoded in this DAG?
2. Which marginal associations are encoded?

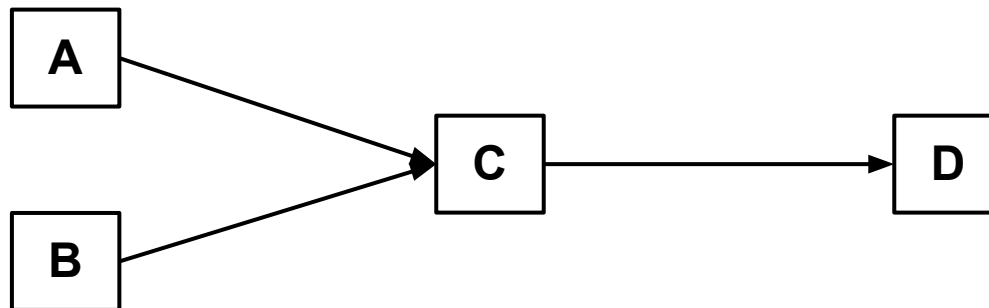


What information is encoded in the DAG?



Causal assumptions	Marginal associations
<ul style="list-style-type: none">• A is a direct cause of C• A is an indirect cause of D	<ul style="list-style-type: none">• A and C

What information is encoded in the DAG?



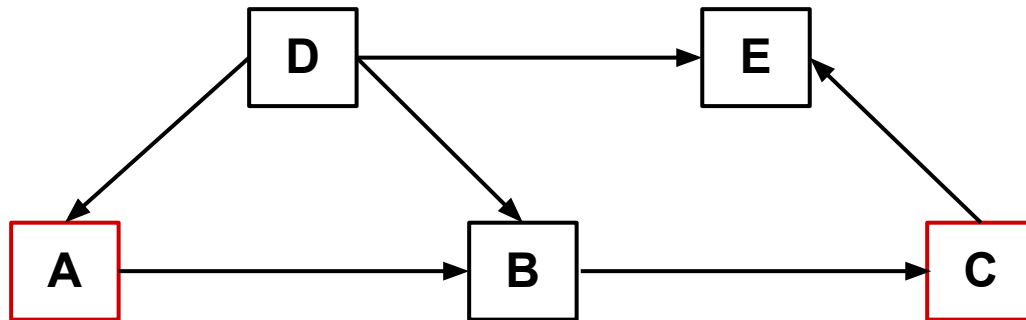
Causal assumptions	Marginal associations
<ul style="list-style-type: none">• A is an direct cause of C• B is an direct cause of C• C is a direct cause of D• A is an indirect cause of D• B is an indirect cause of D• <i>No common causes have been omitted</i>	<ul style="list-style-type: none">• A and C• A and D• B and C• B and D• C and D

Blocked and unblocked paths

A path is **blocked** if there is a collider on the path

A path is **unblocked** if there is **NOT** a collider on the path

- I.e. path unblocked if a line can be traced without two arrows colliding head-to-head



Which paths between A and C are unblocked?

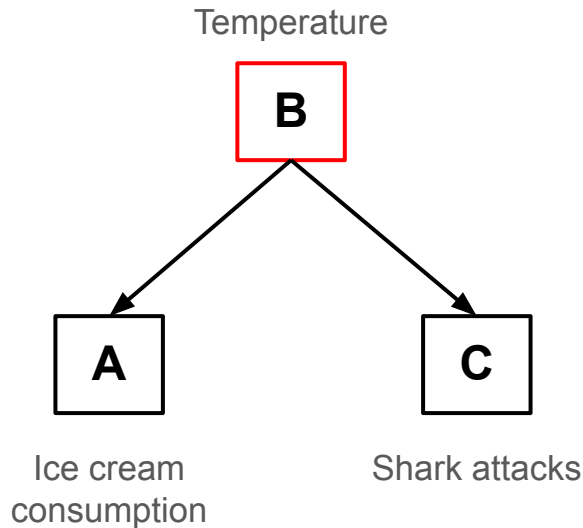
When does confounding occur?

An **unblocked backdoor path confounds** the relationship between A and C

$$A \leftarrow B \rightarrow C$$

- ❑ is a backdoor path
- ❑ Does not contain a collider

The relationship between ice cream consumption and shark attacks is **confounded**



How can we control for confounding?

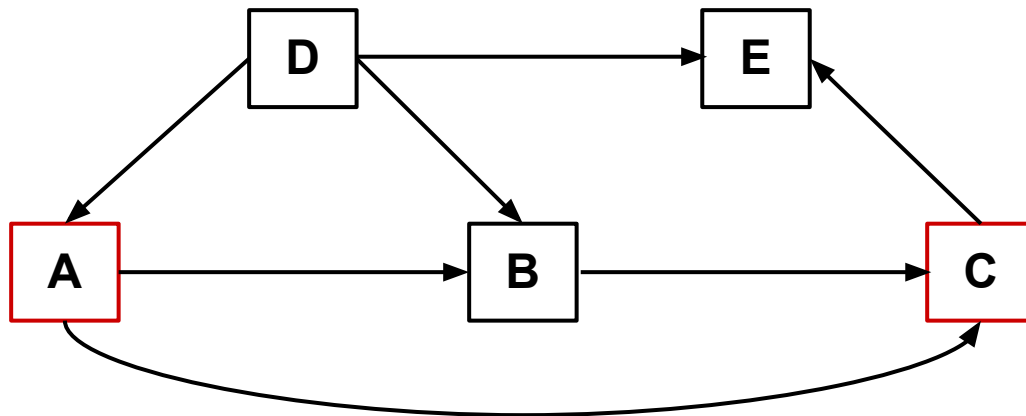
Reminder : all backdoor paths must be blocked

1. Identify backdoor paths
 - These start with arrows pointing **into** the exposure
2. Remove blocked paths
 - A path is blocked if it has a collider
3. **Block paths by conditioning on confounders** (or mediators on the path)

An example

How to control for confounders of the relationship between A and C?

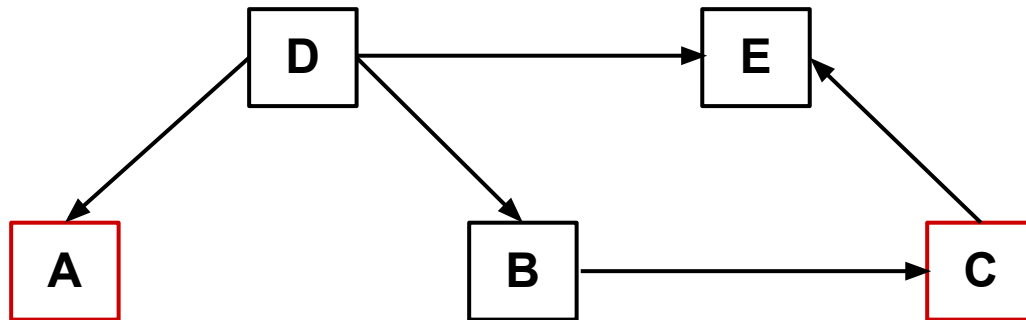
1. Identify backdoor paths



An example

How to control for confounders of the relationship between A and C?

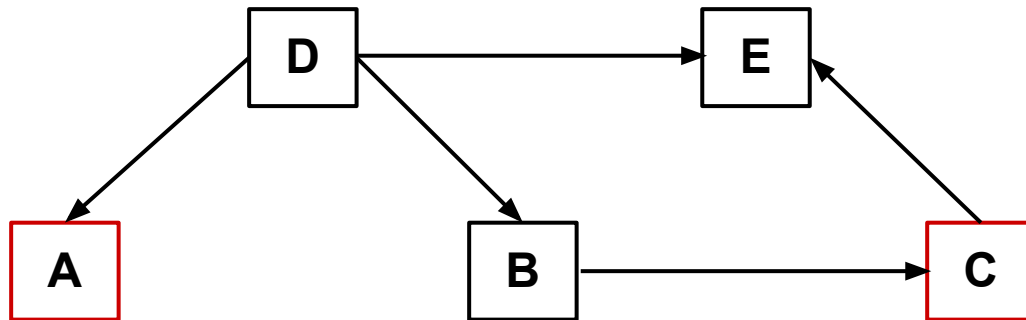
1. Identify backdoor paths



An example

How to control for confounders of the relationship between A and C?

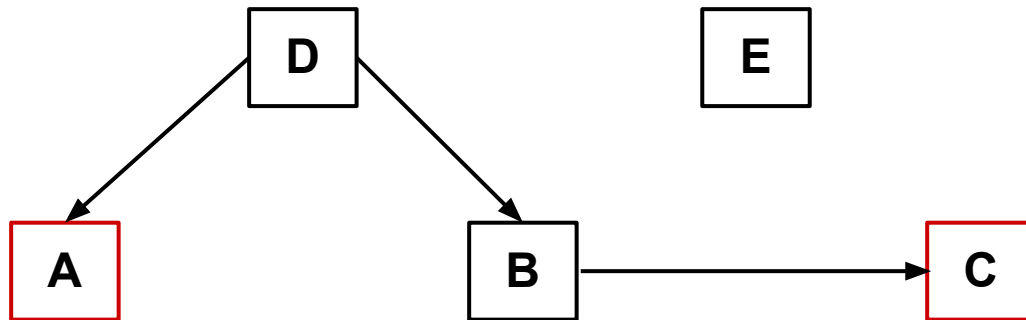
1. Identify backdoor paths
2. Identify unblocked paths



An example

How to control for confounders of the relationship between A and C?

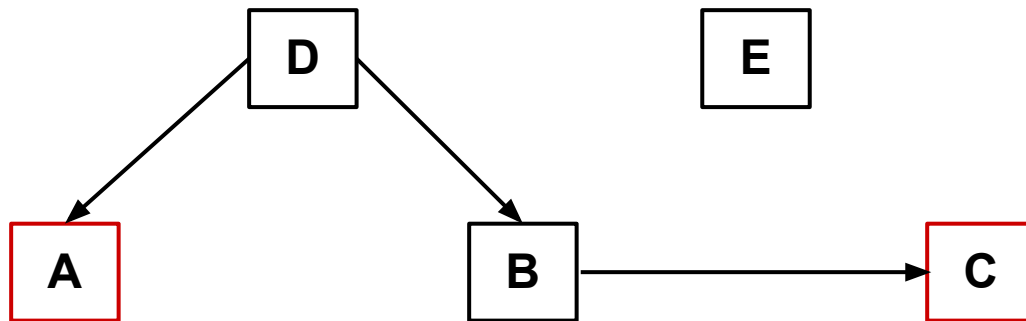
1. Identify backdoor paths
2. Identify unblocked paths



An example

How to control for confounders of the relationship between A and C?

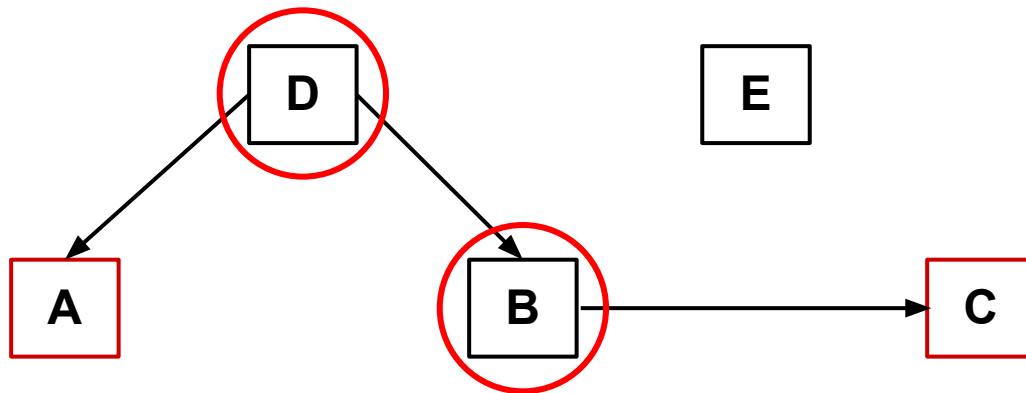
1. Identify backdoor paths
2. Identify unblocked paths
3. Choose set of variables to block paths



An example

How to control for confounders of the relationship between A and C?

1. Identify backdoor paths
2. Identify unblocked paths
3. Choose set of variables to block paths



We need to control for **D** or **B** to avoid confounding

Conditional independence

X and Y are independent conditional on Z (written $X \perp Y \mid Z$) if **X** does not provide information on **Y** given we know the value of **Z**.

Example :

Height and vocabulary are dependent as smaller people tend to be children. However, if two people are both 25, their heights do not give us information about their vocabularies. That is, **height is conditionally independent of vocabulary given age**.

$$\text{Height} \perp \text{Vocabulary} \mid \text{Age}$$

Independence rules in DAGs

d-separation : criterion for deciding if two variables are independent conditional on some set of other variables.

1. Two variables are (unconditionally) **d-connected** if there is an unblocked path between them.
2. Two variables are **d-separated by Z** if Z blocks all unblocked paths between them.
3. If Z contains a collider or descendant of a collider, paths which trace Z are no longer blocked.

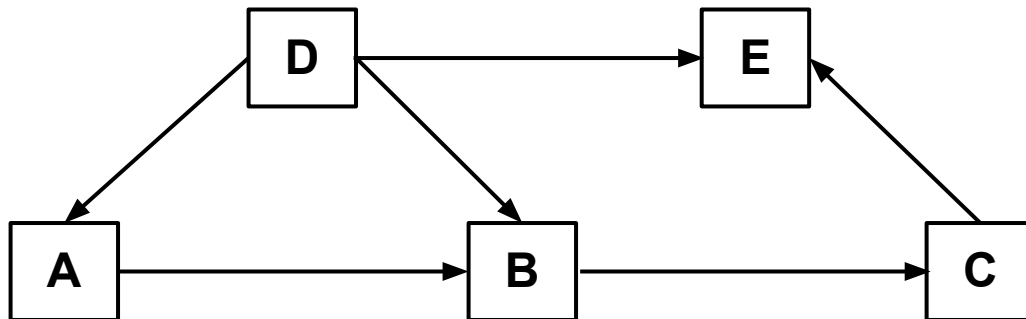
Translation

X and Y are unconditionally dependent if they are **d-connected**.

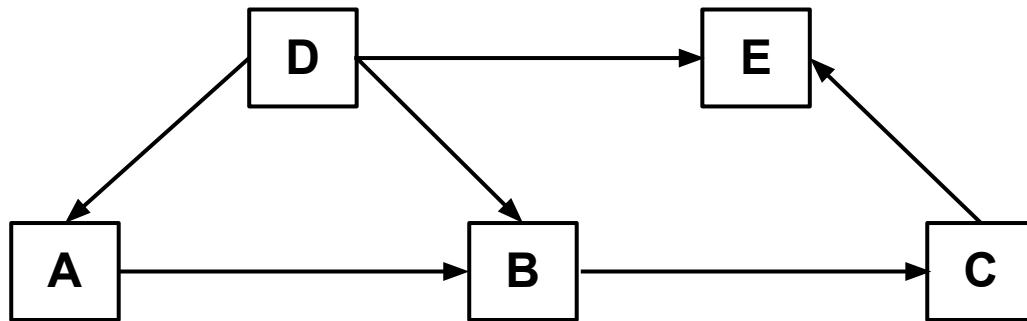
X and Y are conditionally independent given Z if they are **d-separated by Z** .

An example - rule 1

1. Two variables are (unconditionally) **d-connected** if there is an unblocked path between them.



Which variables are d-connected in this DAG?
What can we interpret from this?

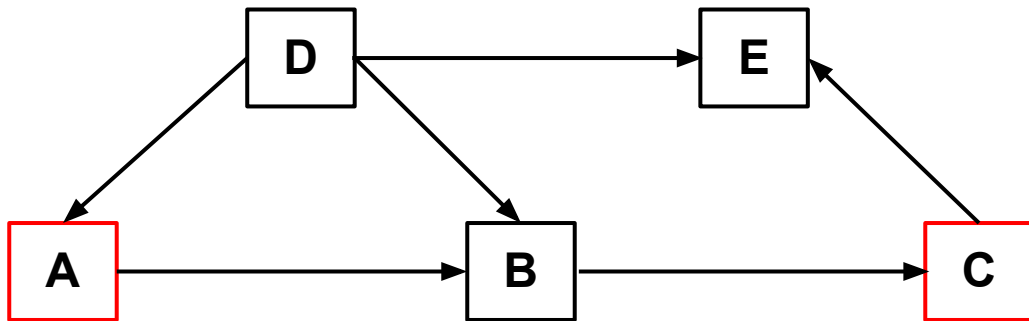


1. Which variables are d-connected in this DAG?
2. What can we interpret from this?

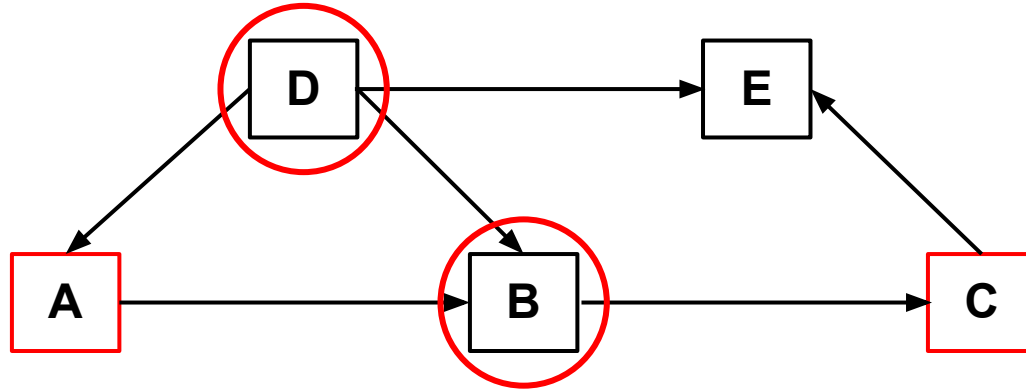
Answer : All pairs of variables are d-connected in this DAG as unblocked paths can be found between each one. This means **all of the variables are unconditionally dependent** (i.e. knowing any one gives information about all the others).

An example - rule 2

2. Two variables are **d-separated by Z** if **Z** blocks all unblocked paths between them.



1. Are **A** and **C** **d-separated** by some set **Z**?
2. Is this the only option?
3. What can we interpret from this?

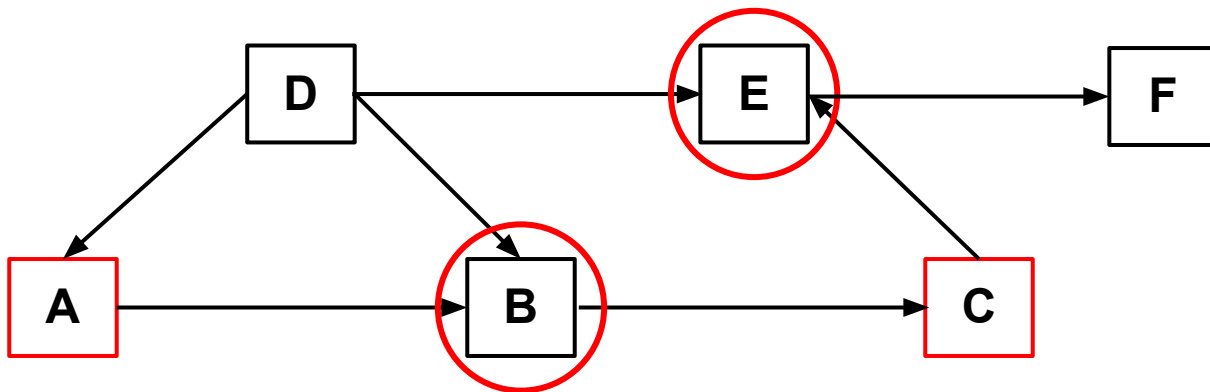


1. Are **A** and **C** **d-separated** by some set **Z**?
2. Is this the only option?
3. What can we interpret from this?

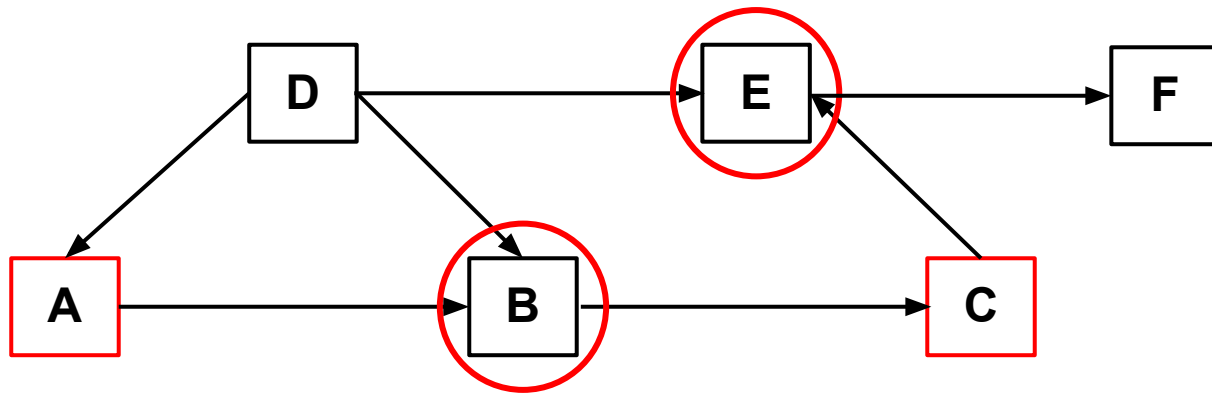
Answer : **A** and **C** are **d-separated** by **$Z = \{D, B\}$** , or **$Z = \{B\}$** . This means that **A** and **C** are **conditionally independent given B** (or, D and B). That is, if we know the value of B, knowing A does not provide information on the value of C.

An example - rule 3

3. If \mathbf{Z} contains a collider or descendant of a collider, paths which trace \mathbf{Z} are no longer blocked.



1. If we condition on $\mathbf{Z} = \{\mathbf{B}, \mathbf{E}\}$, are \mathbf{A} and \mathbf{C} d-separated or d-connected?
2. What if we condition on $\mathbf{Z} = \{\mathbf{B}, \mathbf{F}\}$?

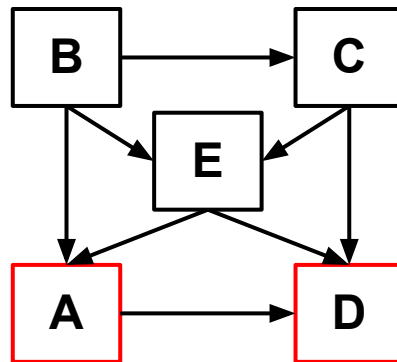
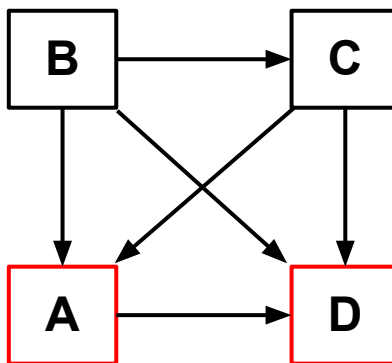
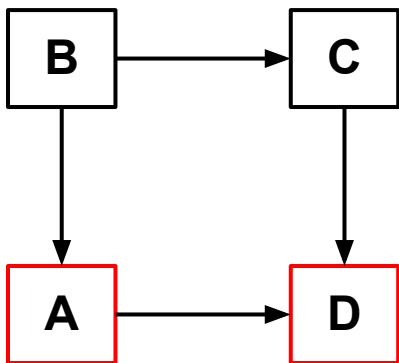


1. If we condition on $\mathbf{Z} = \{\mathbf{B}, \mathbf{E}\}$, are \mathbf{A} and \mathbf{C} d-separated or d-connected?
2. What if we condition on $\mathbf{Z} = \{\mathbf{B}, \mathbf{F}\}$?

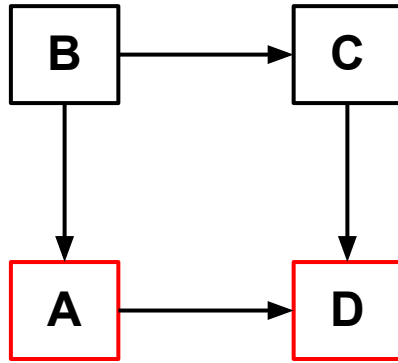
Answer : \mathbf{E} is a collider on the path $\mathbf{A} \rightarrow \mathbf{D} \rightarrow \mathbf{E} \leftarrow \mathbf{C}$. Therefore, conditioning on \mathbf{E} unblocks this path and \mathbf{A} and \mathbf{C} are d-connected. \mathbf{F} is a descendant of \mathbf{E} , so conditioning on \mathbf{F} also unblocks the path $\mathbf{A} \rightarrow \mathbf{D} \rightarrow \mathbf{E} \leftarrow \mathbf{C}$, so \mathbf{A} and \mathbf{C} are d-connected.

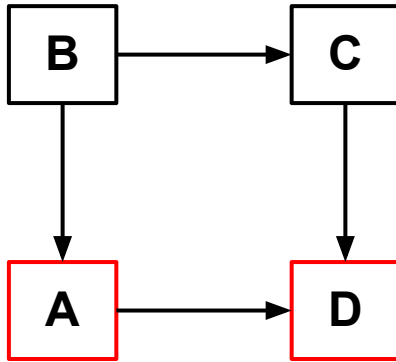
Sufficient adjusting sets

- A set **Z** is **sufficient** if after adjusting for **Z**, no open backdoor paths remain
- **Z** is **minimally sufficient** if no subset of **Z** is sufficient
 - i.e. **Z** is the **smallest possible adjusting set** which blocks unblocked backdoor paths
 - There may exist multiple minimally sufficient adjusting sets

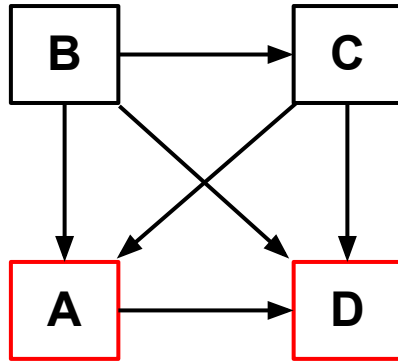


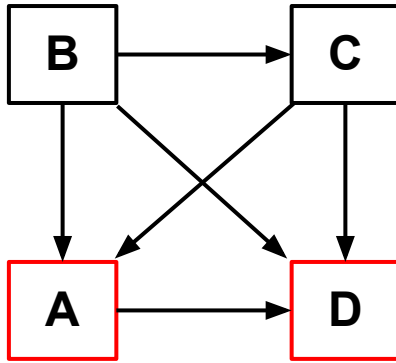
What are the minimally sufficient adjusting set(s)
for the relationship between **A** and **D**?



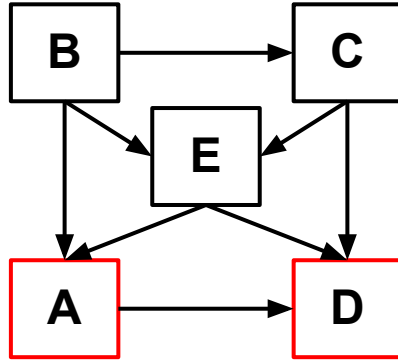


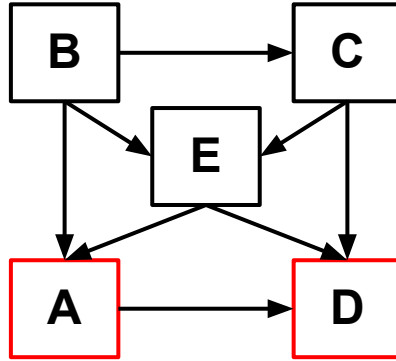
$\{B, C\}$ is a sufficient adjusting set, and $\{B\}$ and $\{C\}$ alone are minimal adjusting sets





$\{B, C\}$ is a minimally sufficient adjusting set.

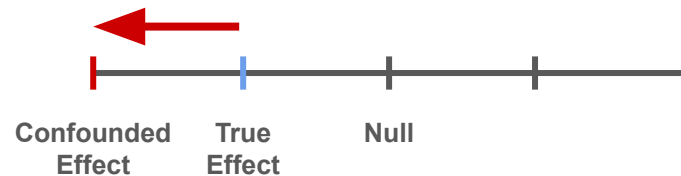
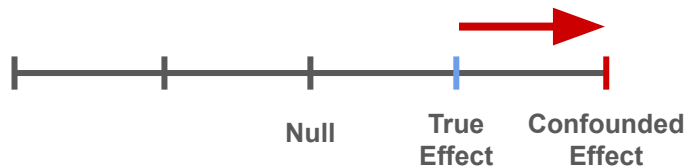




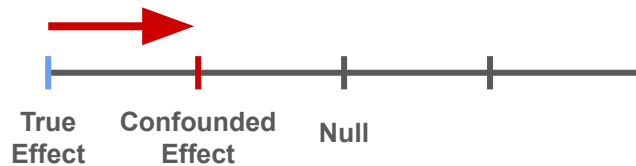
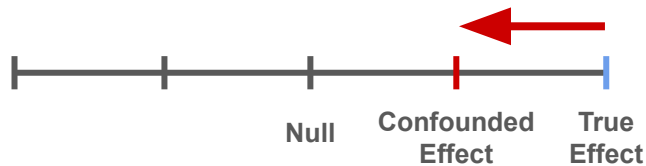
$\{E, B, C\}$ is a sufficient adjusting set.

$\{E, B\}$ and $\{E, C\}$ are minimally sufficient adjusting sets.

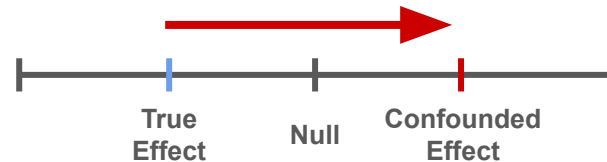
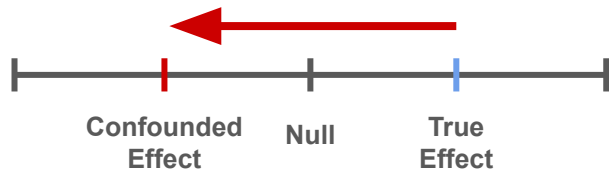
Positive confounding



Negative confounding



Qualitative confounding



Direction of confounding bias

- The direction of the confounding bias is the product of the component effect signs

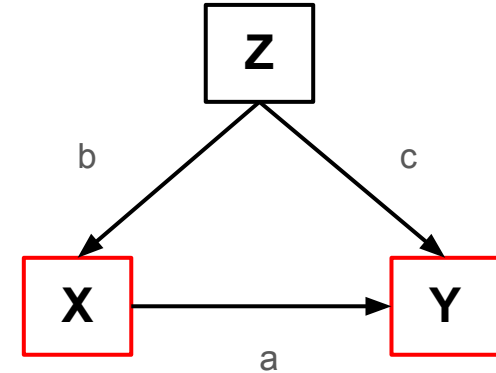
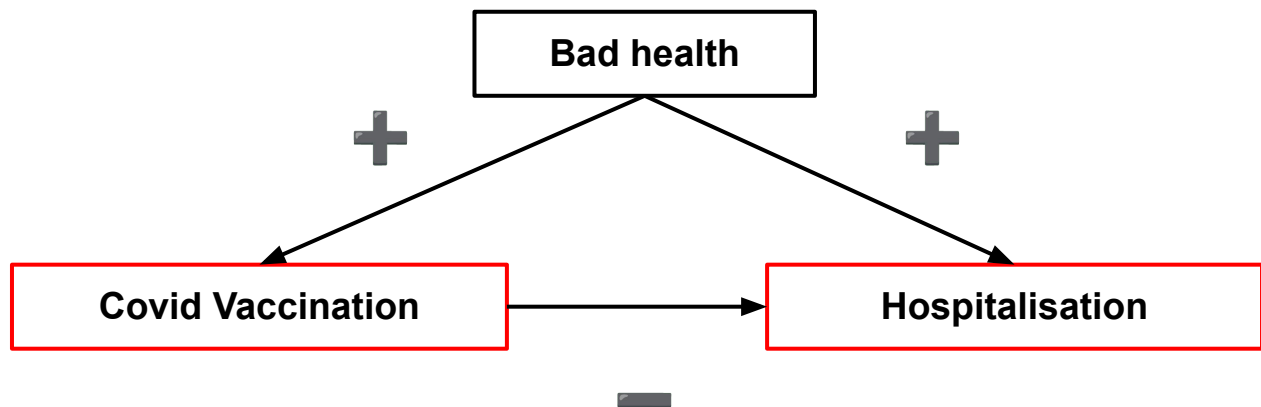


TABLE 1. Anticipating the direction of the confounder based on the direction of the associations between exposure, outcome, and covariate

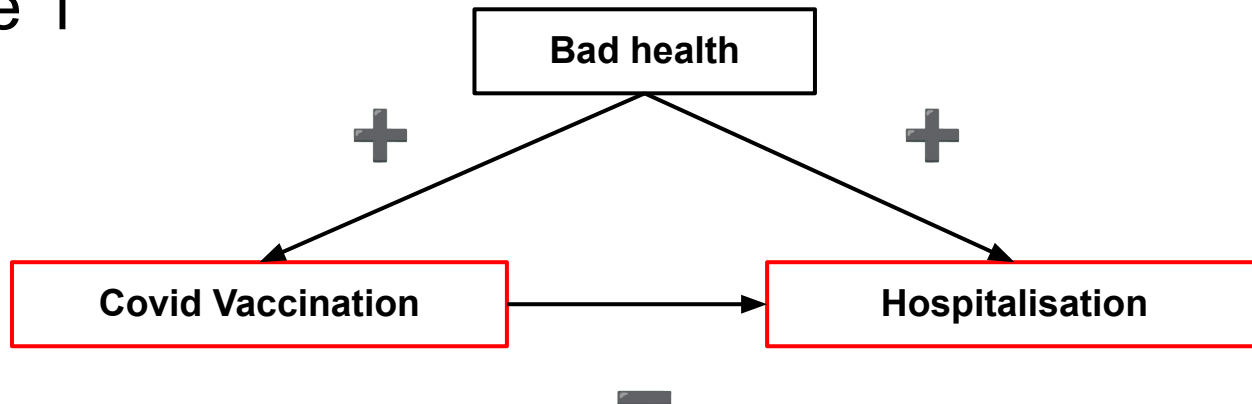
Case	Direction of association*			Sign of confounding bias (b*c)	Sign of triple product (a*b*c)	Direction of confounder†
	a	b	c			
1.	+	+	+	+	+	Positive
2.	+	+	-	-	-	Negative
3.	+	-	+	-	-	Negative
4.	+	-	-	+	+	Positive
5.	-	+	+	+	-	Negative
6.	-	+	-	-	+	Positive
7.	-	-	+	-	+	Positive
8.	-	-	-	+	-	Negative

Example 1



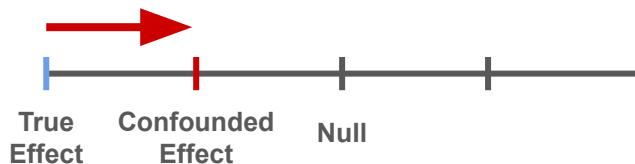
1. If unadjusted for, **what is the direction of the confounding bias** of bad health on the covid vaccination-hospitalisation relationship?
2. How does the direction of the confounding bias **affect our interpretation** of the effect of covid vaccination on hospitalisation?

Example 1

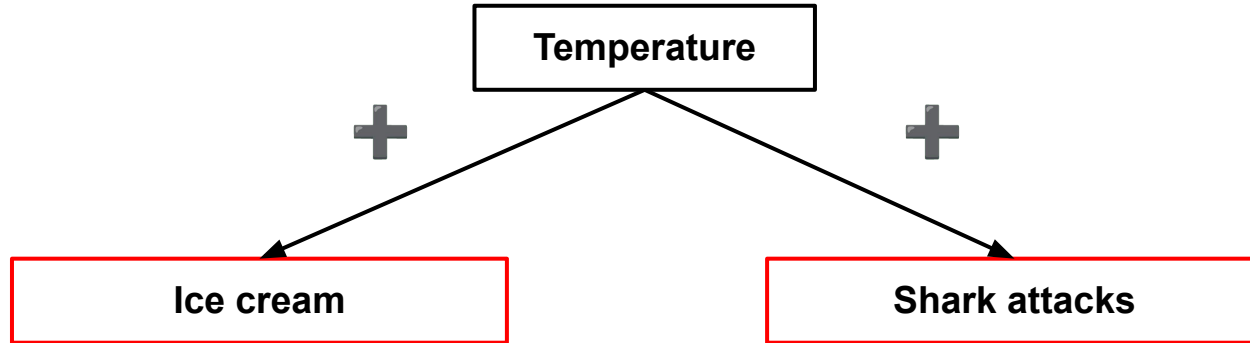


$$(+ \times + \times -) = -$$

Therefore, if unadjusted for, bad health negatively confounds the relationship between Covid vaccination and hospitalisations. This dilutes the apparent protective effect of vaccination and may even make vaccination seem harmful.

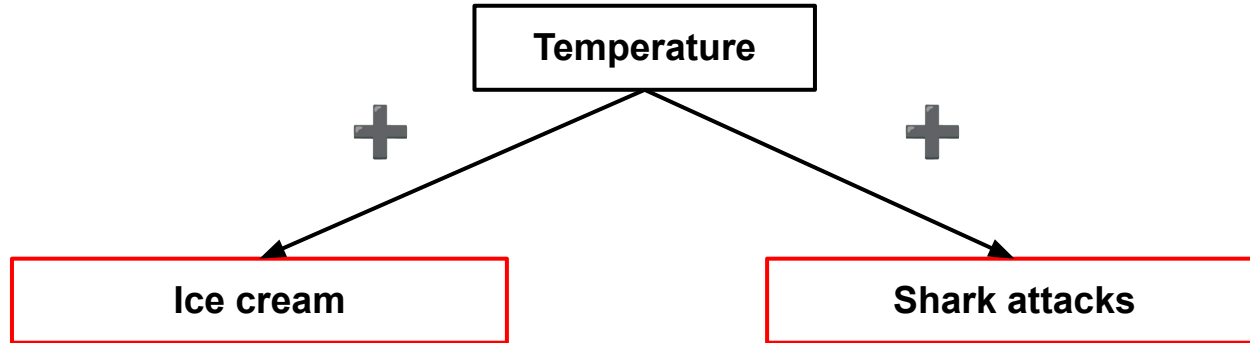


Example 2

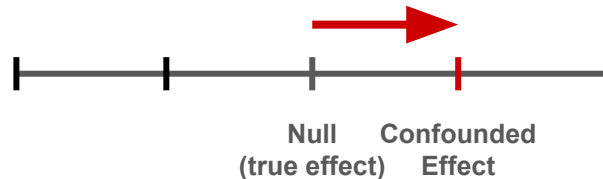


1. If unadjusted for, **what is the direction of the confounding bias** between shark attacks and hospitalisation? Assume no true association between ice cream and shark attacks.

Example 2

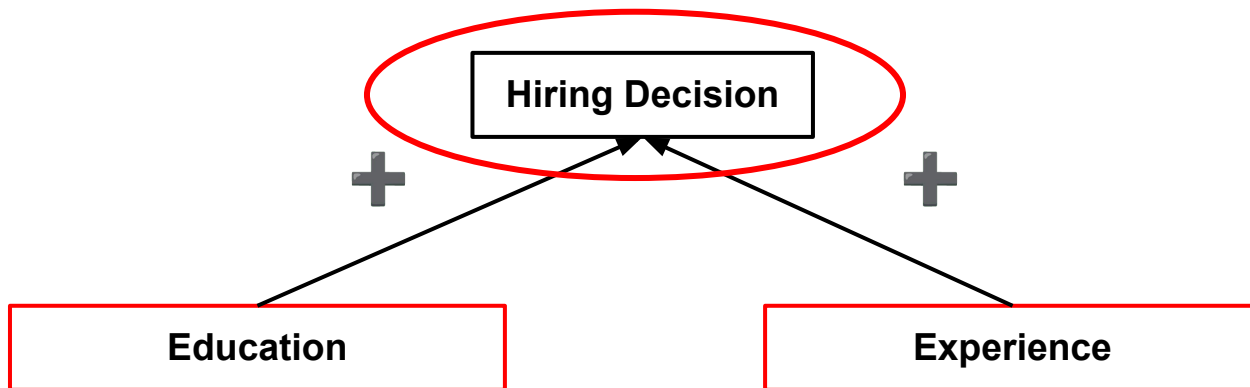


1. If unadjusted for, **what is the direction of the confounding bias** between shark attacks and hospitalisation?



Why should we avoid conditioning on colliders? Intuition

- We are evaluating candidates for a job position based on education and level of experience.
- Suppose that education and experience are independent.
- What happens if we condition on the hiring decision?

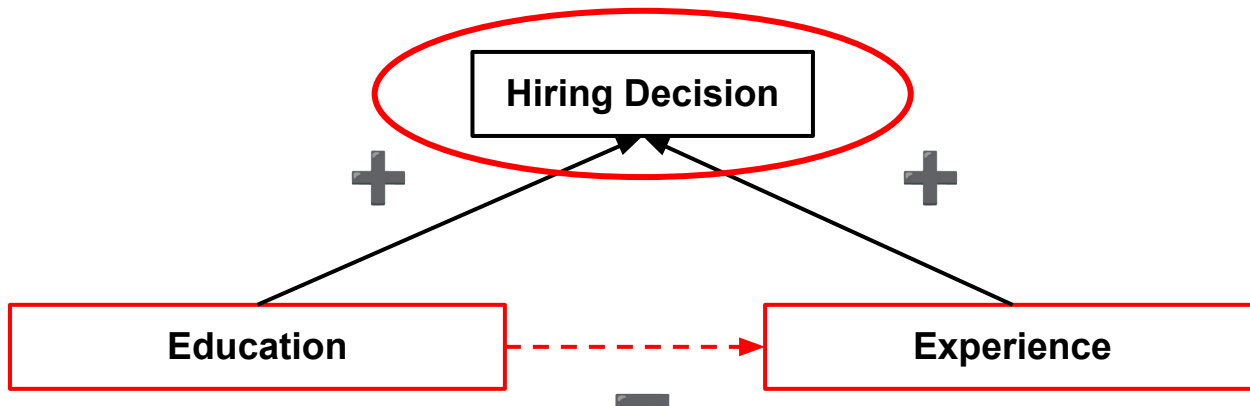


What happens if we condition on the hiring decision?

Amongst people that are hired,

- Those that are **less educated** probably have **more experience** to compensate
- Those that are **less experienced** probably have **more education** to compensate

Therefore, **knowing a hired person's education informs us about their experience and vice versa**. That is, a spurious association between education and experience has been induced by conditioning on the hiring decision.



A real-life example of collider bias



Menu Weekly edition Search

Subscribe Log in

Science & technology | Covid-19

Smokers seem less likely than non-smokers to fall ill with covid-19

That may point towards a way of treating it

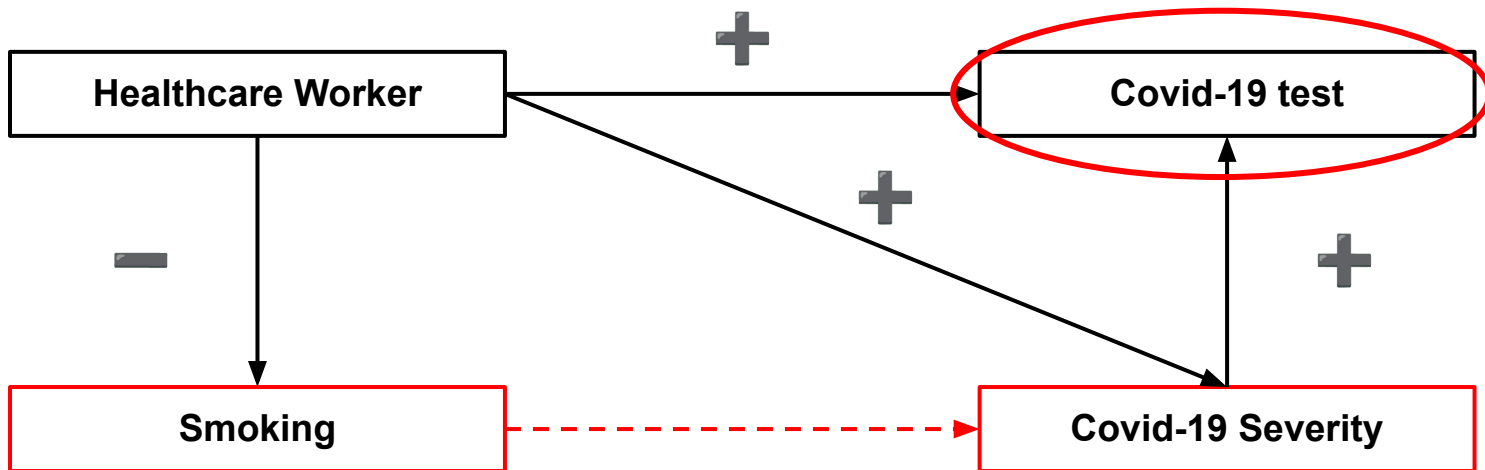


Getty Images

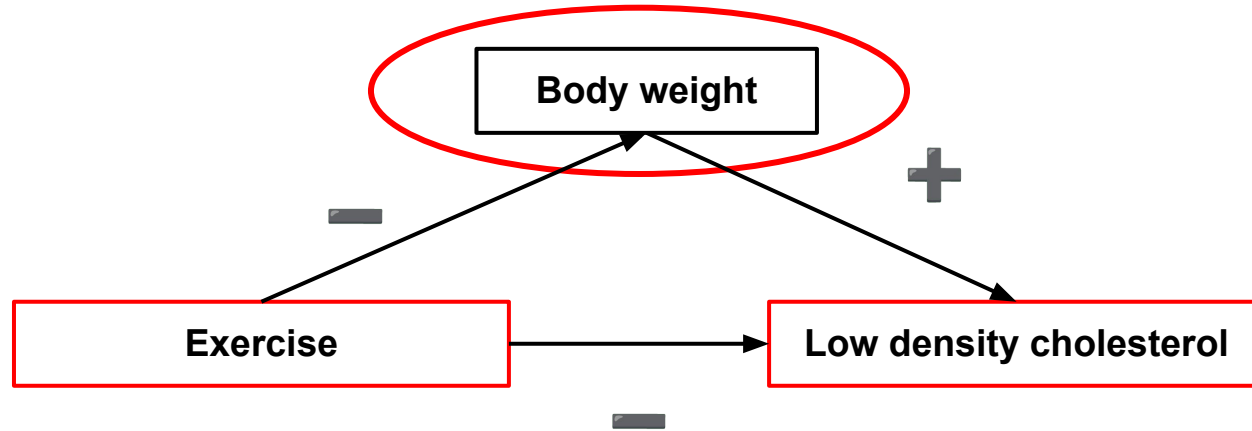
nist.com May 2nd 2020 Share

What is really happening here?

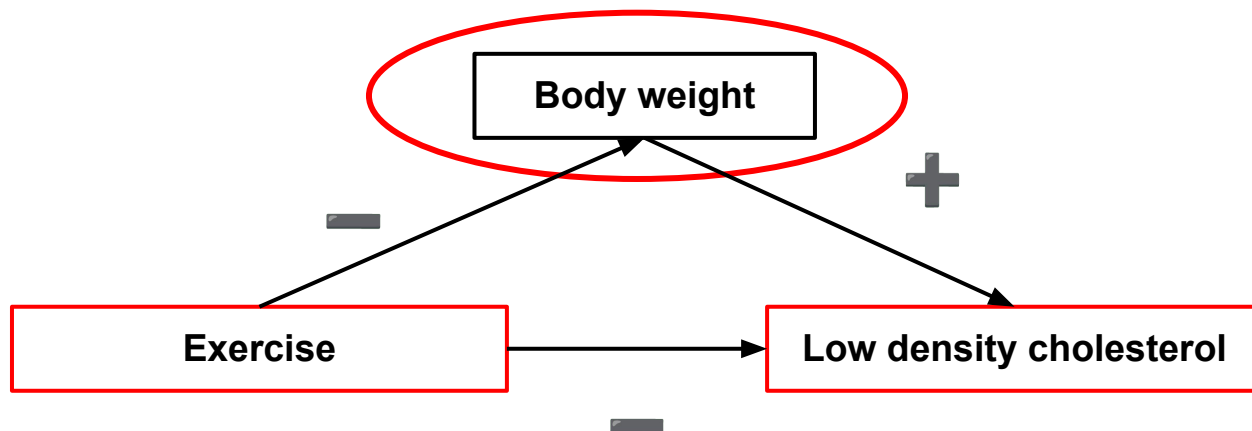
- **Inclusion to the study was based on Covid-19 testing**
- Conditional on Covid-19 testing, smoking is associated with Covid-19 severity
 - Backdoor path opened by conditioning on collider
- **Which variables should we condition on** to assess the relationship between smoking and covid-19 severity?



Should we condition on mediators?

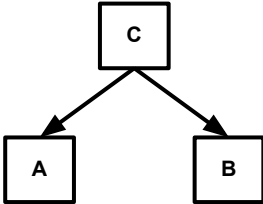
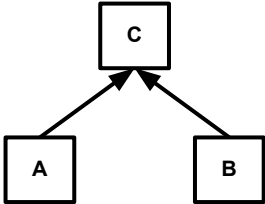
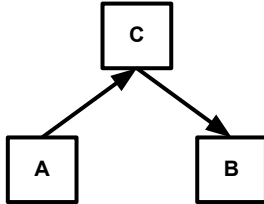


Should we condition on mediators?

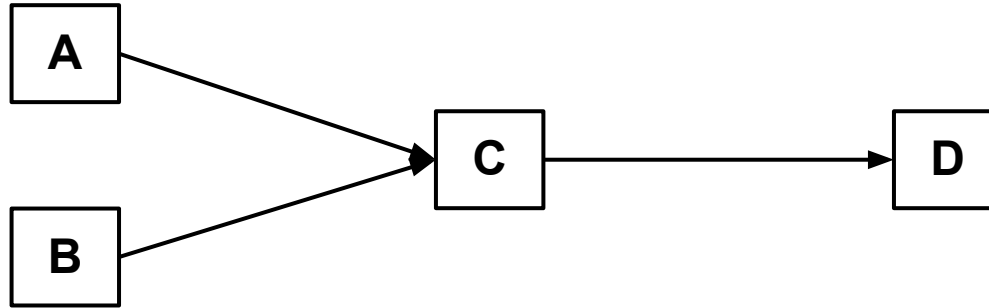


- Are we interested in the total effect of exercise on cholesterol?
 - We should not condition on body weight
- Are we interested in the **direct effect** of exercise on cholesterol, **outside of its effects on body weight**?
 - We should condition on body weight

Summary of concepts

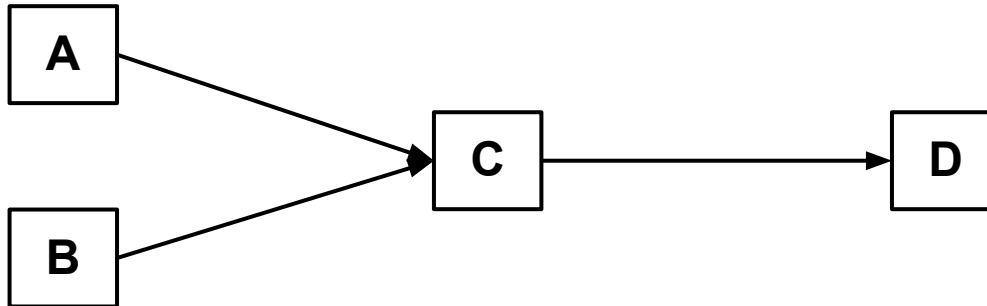
	Confounder	Collider	Mediator
DAG	 <pre>graph TD; C[C] --> A[A]; C[C] --> B[B];</pre>	 <pre>graph TD; A[A] --> C[C]; B[B] --> C[C];</pre>	 <pre>graph TD; A[A] --> C[C]; C[C] --> B[B];</pre>
Path	Open	Closed	Open
Condition?	Yes	No	Depends
Consequence of conditioning	Reduces bias from C	Causes selection bias	Isolates direct effect of A on B

What information is encoded in the DAG?



Causal assumptions	Marginal associations	Independencies	Conditional Associations
<ul style="list-style-type: none">• A is an direct cause of C• B is an direct cause of C• C is a direct cause of D• A is an indirect cause of D• B is an indirect cause of D• No common causes have been omitted	<ul style="list-style-type: none">• A and C• A and D• B and C• B and D• C and D	<ul style="list-style-type: none">• _ and _ are unconditionally independent• A and D are independent conditional on _• _ and D are independent conditional on C	<ul style="list-style-type: none">• _ and _ are associated conditional on C• A and _ are associated conditional on D

What information is encoded in the DAG?



Causal assumptions	Marginal associations	Independencies	Conditional Associations
<ul style="list-style-type: none">• A is an direct cause of C• B is an direct cause of C• C is a direct cause of D• A is an indirect cause of D• B is an indirect cause of D• No common causes have been omitted	<ul style="list-style-type: none">• A and C• A and D• B and C• B and D• C and D	<ul style="list-style-type: none">• A and B are unconditionally independent• A and D are independent conditional on C• B and D are independent conditional on C	<ul style="list-style-type: none">• A and B are associated conditional on C (collider)• A and B are associated conditional on D (descendant of a collider)

How to construct a DAG?

1. State the research question
2. State the exposure and outcome of interest
3. Consider important variables (mediators, effect modifiers, confounders including unmeasurable confounders)
4. Make selection variables explicit

Daggity

Online tool to construct DAGs

<https://www.dagitty.net/dags.html>

Construct your own DAG - example 1.

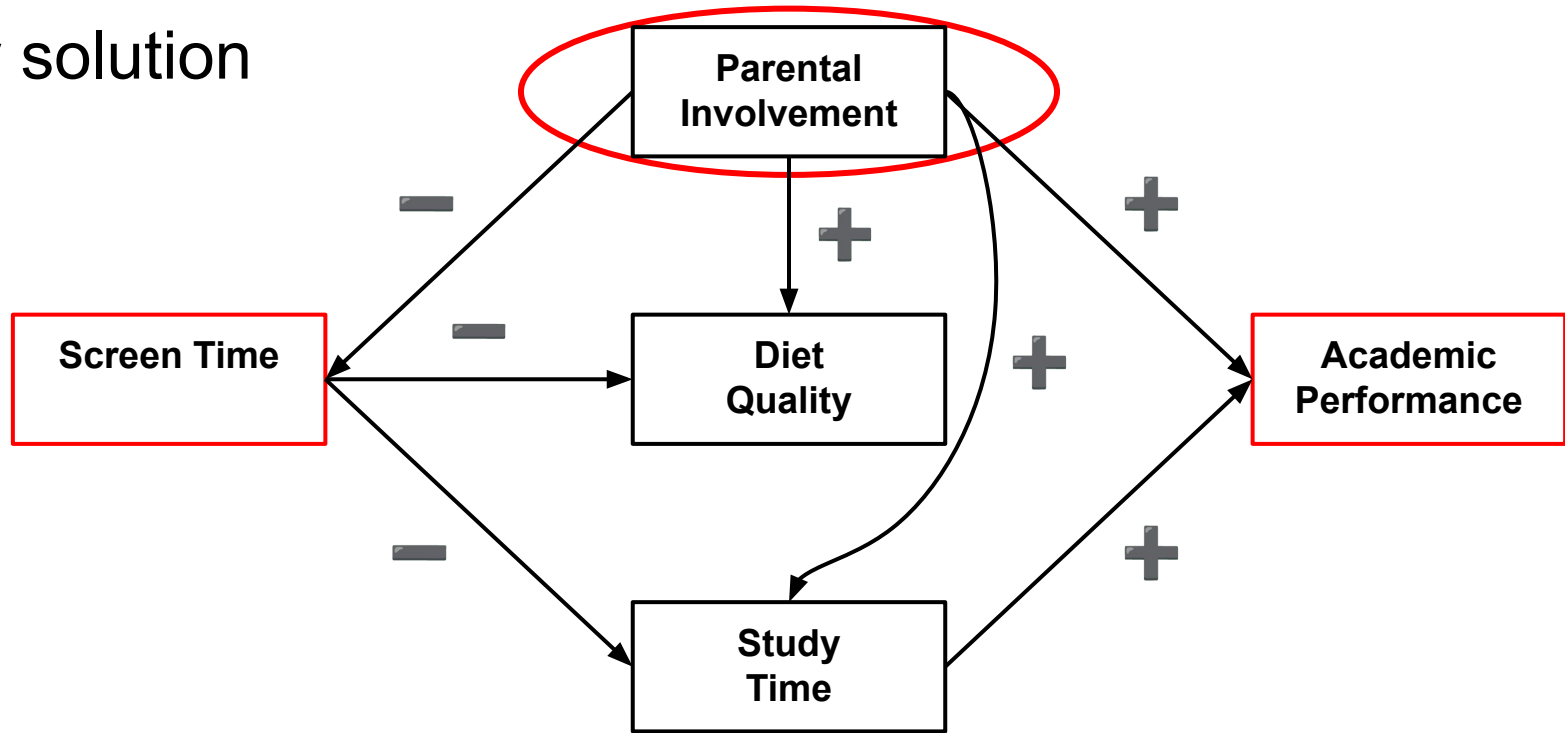
We are interested in studying the relationship between daily screen time and student's academic performance. Assume that the only variables to consider are

- S - daily screen time
- A - academic performance
- D - diet quality
- M - daily study time
- P - parental involvement

Construct a DAG based on these variables.

1. Identify any confounders, mediators, and colliders.
2. Which variables should be controlled for and why?
3. Hypothesise the direction of the confounding bias had these variables not been controlled for.

My solution



1. Diet is a collider, study time is a mediator, and parental involvement is a confounder of the relationship between screen time and academic performance
2. Parental involvement should be controlled for.
3. If parental involvement is not controlled for, the effect of screen time on academic performance will be negatively confounded.

Construct your own DAG - example 2.

We are interested in studying the relationship between red meat consumption and stomach cancer. Assume that the only variables to consider are

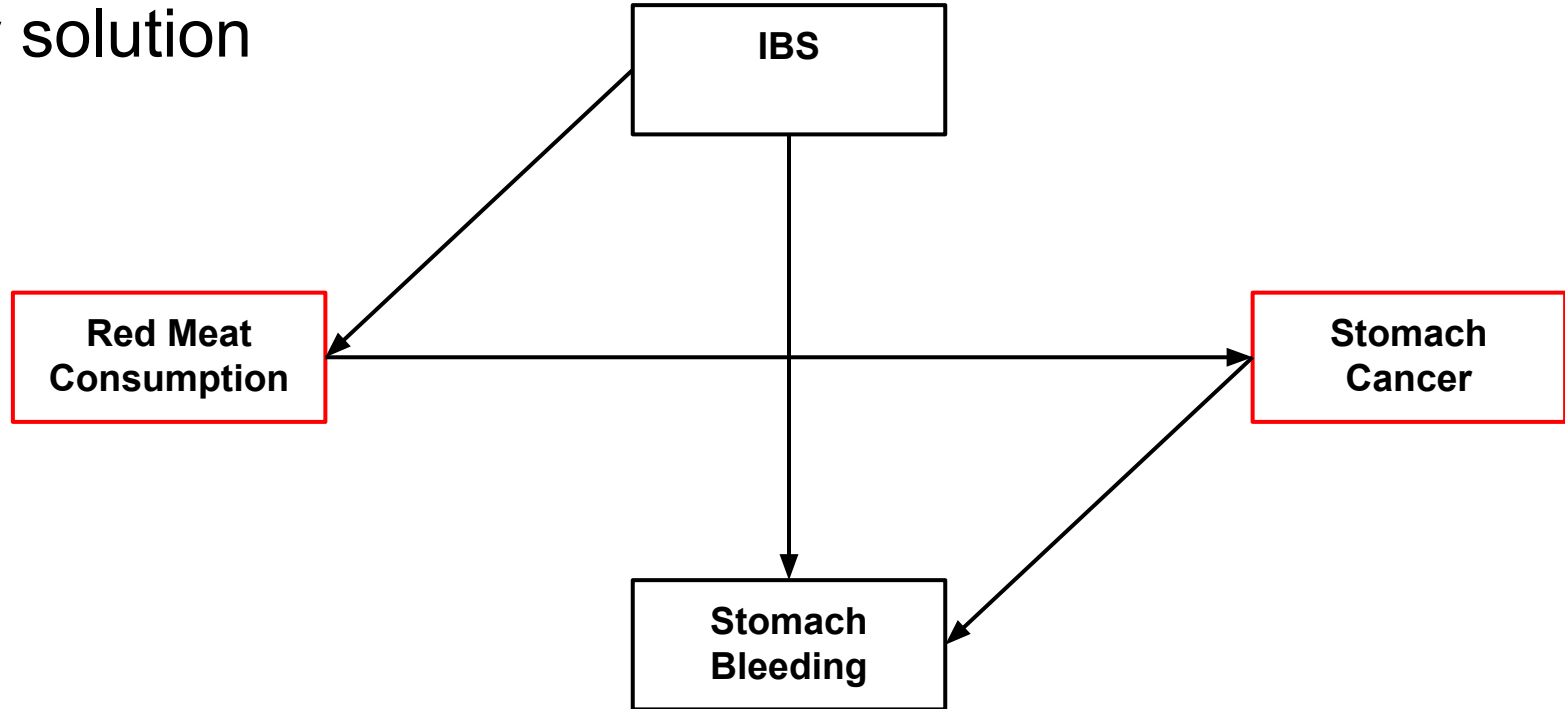
- R - average red meat consumption
- C - stomach cancer
- I - Inflammatory bowel syndrome
- B - stomach bleeding

You may assume that people with inflammatory bowel syndrome eat less red meat. You may also assume that IBS and stomach cancer are associated with stomach bleeding as a symptom, and that IBS is not associated with stomach cancer.

Construct a DAG based on these variables.

1. Should we control for stomach bleeding?

My solution



1. Stomach bleeding is a collider and should not be controlled for.

Limitations of DAGs

- DAGs are **non-parametric** and **qualitative** : they do not represent the form of the relations, nor the size or direction of the associations.
- DAGs assume every relevant variable is included : therefore **expert knowledge** is required to construct DAGs!
- DAGs can be too complex or too subjective
- DAGs cannot depict sources of **random error**

Summary

- Good epidemiological research requires **careful thinking about causal relationships** in order to avoid systematic errors, which come in many forms.
- Directed acyclic graphs depict hypothesised causal relationships between variables in a system, allowing us to better **communicate the causal question of interest**, identify **conditional dependencies**, and choose **sets of variables to control for** in order to reduce bias.

Reminder on ways to remove confounder bias

Study design

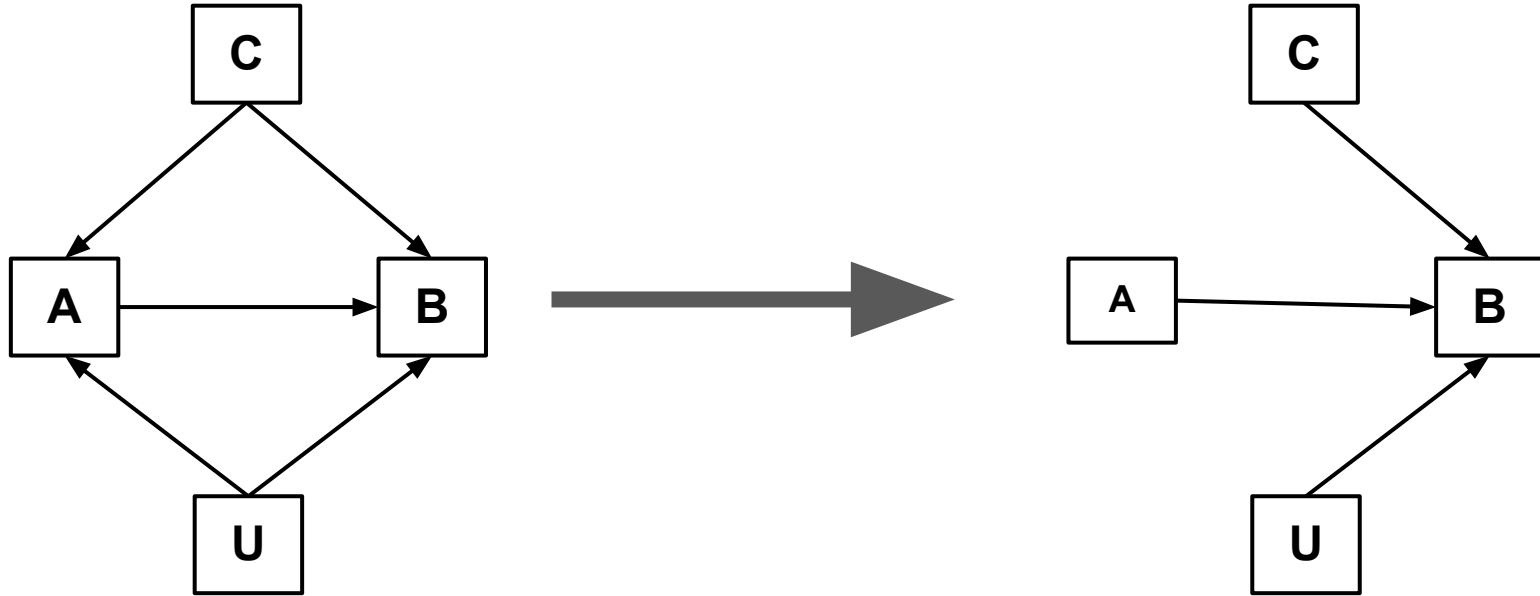
- Randomisation
- Matching
- Restriction

Analytically

- Regression
- Stratification
- Inverse probability of treatment weighting (IPW)
- G-computation

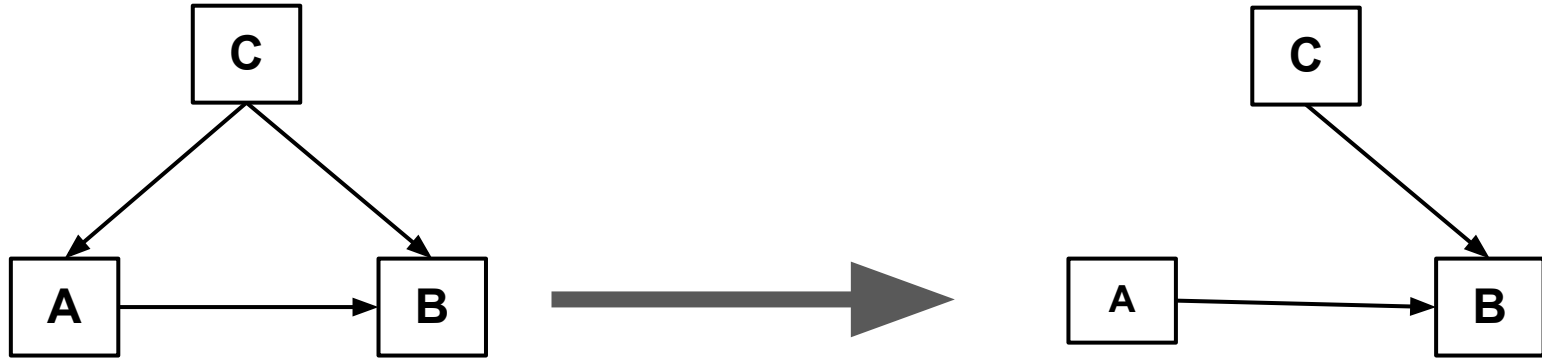
Can we represent what these methods do in terms of DAGs?

Randomisation



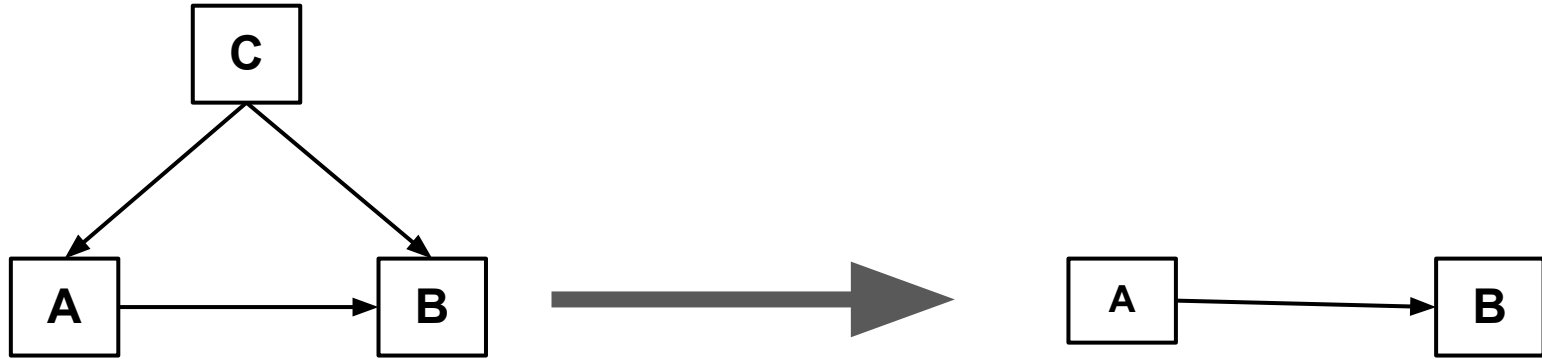
- Breaks the link between exposure A and both unknown and known confounders
- Not always ethical or practical to randomise exposure

Matching (cohort studies)



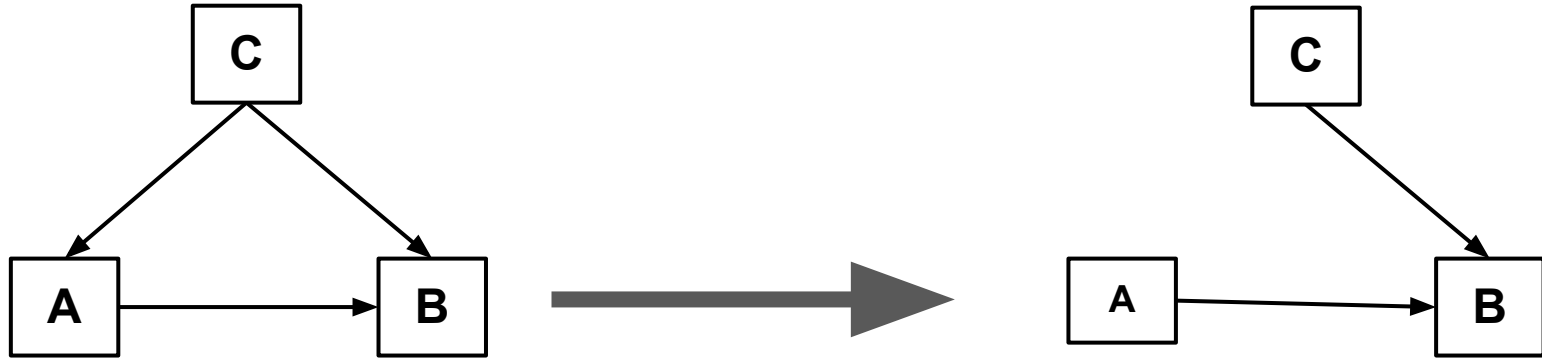
- For each individual with $A = 1$, a participant with $A = 0$ and the same value of C will be selected.
- Breaks the links between both exposure A and confounder C .

Restriction



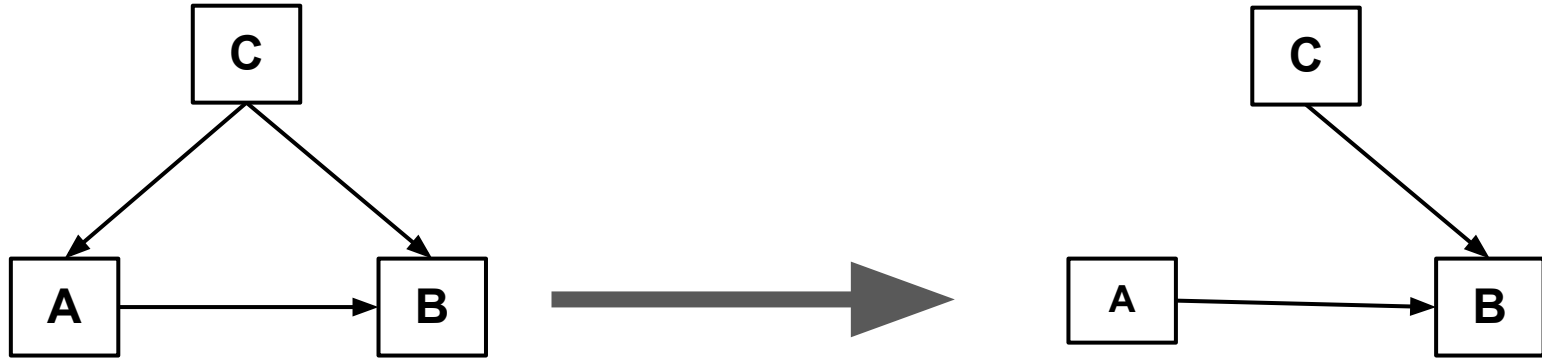
- Breaks the links between both exposure A and confounder C as well as outcome B and confounder C.
- **Problem** : Generalisability, can only control for small amounts of confounders

Inverse probability of treatment weighting



- Creates pseudo-population where exposure A and confounder C are independent.
- Breaks the links between both exposure A and confounder C.

Multivariate regression



- Examine effect of A on B *within levels* of confounder C.
- Breaks the links between both exposure A and confounder C.