DIRECTED ACYCLIC GRAPHS

IDENTIFYING STRUCTURAL SOURCES OF BIAS

March 2, 2021

Joy Shi
Department of Epidemiology
Harvard T.H. Chan School of Public Health

LEARNING OBJECTIVES.

After this session, you should be able to:

1. Identify features of a DAG.
2. Understand the rules of d-separation.
3. Use a causal DAG to identify bias due to confounding and selection bias.
4. Identify control strategies to account for bias due to confounding and selection bias.

INTRODUCTION: POLL QUESTION.

How familiar are you with DAGs?

A. Not at all familiar
B. Slightly familiar
C. Somewhat familiar
D. Moderately/extremely familiar

INTRODUCTION TO DAGs

WHAT IS A DAG?

Directed
Acyclic
Graphs

WHAT IS A DAG?

Visual representation of one's assumptions about the relationship between variables

USES

- Making assumptions explicit
- Identifying sources of structural bias
- Informing study design and analytical strategy

WHAT DAGs DON'T TELL YOU…

- Strength/direction of relationships
- Sampling variability
- Scale (i.e. additive vs. multiplicative)
- True state of nature

COMPONENTS OF A DAG.

Here is a simple DAG:

$A \longrightarrow Y$

There are three key components/characteristics of a DAG:

1. Nodes: variables (often represented by letters)
   A: exposure
   Y: outcome
   Optional: nodes are placed temporally from left to right
2. Edges: arrows, representing the direction of causality
   A causes Y
   Note: you would include an arrow from A to Y if A causes Y for at least one person in your population; therefore, the absence of an arrow is a stronger assumption than the presence of one
3. Acyclic: no cycles or loops; i.e., a variable cannot cause itself, either directly or through another variable

FLOW OF ASSOCIATION.

Here is a simple DAG:

$A \rightarrow Y$

Associations ignore the direction of the arrows.

A is associated with Y.

Y is associated with A.

Causality follows the direction of the arrows.

A causes Y.

Y does not cause A.

A path is a sequence of edges (i.e., arrows) connecting two variables on the graph

THREE KEY DAG STRUCTURES.

(1) MEDIATOR

$A \rightarrow M \rightarrow Y$

(2) COMMON CAUSE

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A & Y \end{tikzcd}$$

(3) COMMON EFFECT

$$\begin{tikzcd} A \arrow[rr, bend left] & Y \arrow[r] & L \end{tikzcd}$$

In each structure, we can identify a path from A to Y (either through M or through L). Let's consider each of these paths from A to Y in more detail.

MEDIATOR.

$A \longrightarrow M \longrightarrow Y$

- M is a mediator for the effect of the exposure (A) on the outcome (Y)
- A causes M, which in turn causes Y
- Example:

Maternal air pollution exposure $\longrightarrow$ Preterm birth $\longrightarrow$ Childhood academic achievement

- The path from exposure to outcome is A to M to Y
- This path is open
  - Association flows along this path
  - A and Y are associated

CONDITIONING ON A MEDIATOR.

$A \longrightarrow \boxed{M} \longrightarrow Y$

- Conditioning on a variable means to stratify/restrict on that variable (or adjusting for that variable in a regression model)
- In a DAG, conditioning on a DAG is represented by drawing a box around that variable
- Conditioning on M blocks the path that is A to M to Y
- For dichotomous M:
  - Among people with M = 1, A and Y are independent
  - Among people with M = 0, A and Y are independent

Important note:

Open path from A to Y = A and Y are associated = A and Y are *not* independent

All paths from A to Y are blocked = A and Y are *not* associated = A and Y are independent

COMMON CAUSE.

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A & Y \end{tikzcd}$$

- L is a cause of both A and Y
- The path from exposure to outcome is A to L to Y
- This path is open
  - Association flows along this path (even though we are not following the directionality of the arrows)
  - A and Y are associated (even though A does not cause Y)
  - This is the structure for confounding (i.e. the effect of A on Y is confounded by L)

- Conditioning on L blocks this path:

$$\begin{tikzcd} \boxed{L} \arrow[r] \arrow[rr, bend left] & A & Y \end{tikzcd}$$

- A and Y are independent, conditional on L

COMMON EFFECT.

$$\begin{tikzcd} A \arrow[rr, bend left] & Y \arrow[r] & L \end{tikzcd}$$

- A and Y both cause L (i.e. L is an effect of A and Y)
- L is a collider because there are two arrowheads colliding on that variable
- The path from exposure to outcome is A to L to Y
- This path is closed
  - Colliders block the flow of association
  - A and Y are independent

- Conditioning on L opens this path:

$$\begin{tikzcd} A \arrow[rr, bend left] & Y \arrow[r] & \boxed{L} \end{tikzcd}$$

- A and Y are associated, conditional on L (even though A does not cause Y)
- This is the structure for selection bias

D-SEPARATION.

A set of rules that allow us determine whether two variables on a DAG are associated (i.e. whether the path between them is open or blocked)

1. If there are no variables being conditioned on, a path is blocked if two arrowheads on a path collide at some variable on the path.
2. A path that contains a non-collider that is conditioned on is blocked.
3. A collider that has been conditioned on does not block a path.
4. A collider that has a descendant that has been conditioned on does not block a path.

TLDR version

1. *Colliders block paths*
2. *Conditioning on a mediator or a common cause blocks a path*
3. *Conditioning on a collider opens a path*
4. *Conditioning on a descendant of a collider opens a path*

D-SEPARATION: POLL QUESTION 1.

$$A \rightarrow \boxed{M} \rightarrow Y$$

In the DAG above:

- A and Y are independent, conditional on M
- A and Y are not associated, conditional on M
- The path from A to Y is blocked
- A and Y are d-separated

Which d-separation rule tells us this?

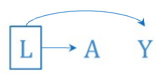
A. If there are no variables being conditioned on, a path is blocked if two arrowheads on a path collide at some variable on the path.

B. A path that contains a non-collider that is conditioned on is blocked.

C. A collider that has been conditioned on does not block a path.

D. A collider that has a descendant that has been conditioned on does not block a path.

D-SEPARATION: POLL QUESTION 2.



In the DAG above:
- A and Y are independent, conditional on L
- A and Y are not associated, conditional on L
- The path from A to Y is blocked
- A and Y are d-separated

Which d-separation rule tells us this?

A. If there are no variables being conditioned on, a path is blocked if two arrowheads on a path collide at some variable on the path.

B. A path that contains a non-collider that is conditioned on is blocked.

C. A collider that has been conditioned on does not block a path.

D. A collider that has a descendant that has been conditioned on does not block a path.

D-SEPARATION: POLL QUESTION 3.

In the DAG above:
- A and Y are marginally independent
- A and Y are not associated, marginally
- The path from A to Y is blocked
- A and Y are d-separated

Which d-separation rule tells us this?

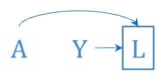
A. If there are no variables being conditioned on, a path is blocked if two arrowheads on a path collide at some variable on the path.

B. A path that contains a non-collider that is conditioned on is blocked.

C. A collider that has been conditioned on does not block a path.

D. A collider that has a descendant that has been conditioned on does not block a path.

D-SEPARATION: POLL QUESTION 4.



In the DAG above:

- A and Y are not independent, conditional on L
- A and Y are associated, conditional on L
- The path from A to Y is open
- A and Y are not d-separated

Which d-separation rule tells us this?

A. If there are no variables being conditioned on, a path is blocked if two arrowheads on a path collide at some variable on the path.

B. A path that contains a non-collider that is conditioned on is blocked.

C. A collider that has been conditioned on does not block a path.

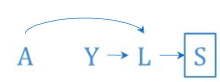
D. A collider that has a descendant that has been conditioned on does not block a path.

D-SEPARATION: POLL QUESTION 5.

D-separation rules:

1. If there are no variables being conditioned on, a path is blocked if two arrowheads on a path collide at some variable on the path.
2. A path that contains a non-collider that is conditioned on is blocked.
3. A collider that has been conditioned on does not block a path.
4. A collider that has a descendant that has been conditioned on does not block a path.

Consider the following DAG:



Are A and Y associated (i.e. is there an open path from A to Y)?

A. Yes

B. No

SUMMARY OF DAG STRUCTURES.

	DAG	Are we conditioning on anything?	Are A and Y associated?	Conclusion
Mediator	A → M → Y	No	Yes	A and Y are marginally associated
	A → $\boxed{M}$ → Y			
Common cause	L → A Y	No	Yes	A and Y are marginally associated
	L → $\boxed{A}$ Y	Yes	No	A and Y are independent, conditional on L
Common effect	A Y → L	No	No	A and Y are marginally independent
	A Y → $\boxed{L}$	Yes	Yes	A and Y are associated, conditional on L
	A Y → L → $\boxed{S}$	Yes	Yes	A and Y are associated, conditional on S

WHERE DOES BIAS COME IN?

Systematic bias: structural association between exposure and outcome that is not the result of the causal effect of exposure on outcome

Confounding: common cause of the exposure or outcome

L → A Y

Selection bias (collider stratification bias): common effect of the exposure or outcome

A Y → $\boxed{L}$

A Y → L → $\boxed{S}$

CONFOUNDING

EXAMPLE OF CONFOUNDING.

In a randomized trial, we expect the following DAG:

$$A \longrightarrow Y$$

A: alcohol intake
Y: mortality

- No causes of A because we randomize exposure
- No common causes of A and Y
- No confounding

In an observational study:

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

A: alcohol intake
Y: mortality
L: age

- There are variables which affect both the exposure and the outcome
- There is confounding

STRUCTURAL DEFINITION OF CONFOUNDING.

Confounding: presence of a backdoor path from the exposure to the outcome

- Backdoor paths are non-causal
- Backdoor paths consist of an arrow going into the exposure (A)

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

If we condition on L in the above DAG, we close the backdoor path

$$\begin{tikzcd} \boxed{L} \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

Any variable that closes a backdoor path once you condition on it is a confounder.

CONFOUNDING: POLL QUESTION 1.

Which of the following DAGs show a backdoor path between access to mental healthcare services and depression?

A. $\begin{tikzcd} \text{Access to mental healthcare services} \arrow[r] & \text{Depression} \end{tikzcd}$

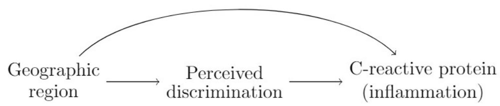
B. $\begin{tikzcd} \text{Access to mental healthcare services} \arrow[r] & \text{Psychotherapy} \arrow[r] & \text{Depression} \end{tikzcd}$

C. $\begin{tikzcd} \text{Access to mental healthcare services} \arrow[r] \arrow[rr, bend left] & \text{Depression} \arrow[r] & \text{Antidepressant use} \end{tikzcd}$

D. $\begin{tikzcd} \text{SES} \arrow[r] \arrow[rr, bend left] & \text{Access to mental healthcare services} \arrow[r] & \text{Depression} \end{tikzcd}$

CONFOUNDING: POLL QUESTION 2.

Consider the following DAG:



Which of the following statements is true?

A. There is an open backdoor path from perceived discrimination to C-reactive protein.

B. Geographic region is a confounder for the relationship between perceived discrimination and C-reactive protein.

C. If we do not adjust for geographic region, the association between perceived discrimination and C-reactive protein is a biased estimate of the causal effect of perceived discrimination and C-reactive protein.

D. All of the above.

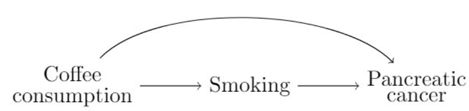
CONFOUNDING: POLL QUESTION 3.

Suppose you are interested in the relationship between:

A: coffee consumption
Y: pancreatic cancer

- You know that smoking and coffee consumption are highly correlated
- You know that smoking causes pancreatic cancer

How would you add smoking into your DAG?

A.


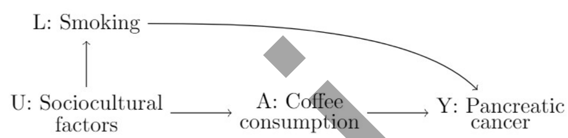
B.


C. None of the above

ANOTHER STRUCTURE FOR CONFOUNDING.

We may not be convinced that drinking coffee *causes* one to smoke, or vice versa

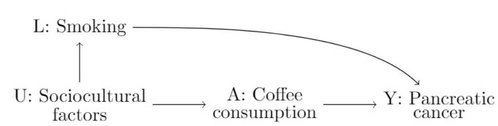
Rather, there may be some unknown/unidentified factor that is more likely to cause someone to both drink coffee and smoke, e.g.



U is often used to indicate an unknown/unmeasured variable

CONFOUNDING: POLL QUESTION 4.

Given the DAG below, which of the following statements is true?



A. There is no open backdoor path from coffee consumption to pancreatic cancer.
B. There is no way to eliminate confounding because we have unmeasured sociocultural factors.
C. We can adjust for smoking to eliminate confounding.
D. All of the above.

POSSIBLE CONFOUNDING STRUCTURES.

There are many possible DAG structures that can correspond to the presence of confounding, e.g.:

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

$$\begin{tikzcd} L \arrow[rrd, bend left] & & \\ U \arrow[u] \arrow[r] & A \arrow[r] & Y \end{tikzcd} \qquad \begin{tikzcd} U \arrow[d] \arrow[rrd, bend left] & & \\ L \arrow[r] & A \arrow[r] & Y \end{tikzcd}$$

In all three of these DAGs, L is a confounder because conditioning on it will block the backdoor path from A to Y.

HISTORICAL (NON-STRUCTURAL) DEFINITIONS OF A CONFOUNDER.

You may have previously encountered alternate criteria for identifying confounders

1. Change-in-estimate: a variable is a confounder if the magnitude of the association between the exposure and outcome changes (e.g. by 10%) once you condition on that variable
2. Conventional definition: a variable is a confounder if it meets three conditions
   a. It is associated with the exposure.
   b. It is associated with the outcome within levels of the exposure.
   c. It is not on the causal pathway from treatment to outcome.

What is wrong with using these criteria?

CHANGE-IN-ESTIMATE APPROACH.

According to the change-in-estimate approach, a variable is a confounder if the magnitude of the association between the exposure and outcome changes (e.g. by 10%) once you condition on that variable

Consider the following DAG:

$$\begin{tikzcd} A \arrow[rr, bend left] & Y \arrow[r] & L \end{tikzcd}$$

By conditioning on L:
- Open the path from A to Y to L
- Introduce collider-stratification bias (L is a collider)
- The magnitude of the association between exposure and outcome will change (because we've introduced bias)

Here, L is not a confounder and should not be conditioned on.

CONVENTIONAL DEFINITION OF A CONFOUNDER.

Consider again the following DAGs for confounding:

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

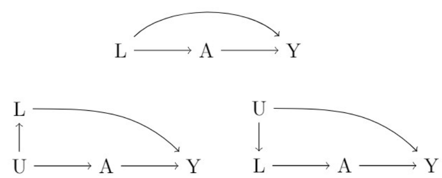
$$\begin{tikzcd} L \arrow[rr, bend left] & & \\ U \arrow[u] \arrow[r] & A \arrow[r] & Y \end{tikzcd} \qquad \begin{tikzcd} U \arrow[d] \arrow[rr, bend left] & & \\ L \arrow[r] & A \arrow[r] & Y \end{tikzcd}$$

In each of those DAGs, does L meet each of the following three criteria?
- It is associated with the exposure.
- It is associated with the outcome within levels of the exposure.
- It is not on the causal pathway from treatment to outcome.

A. Yes
B. No
C. Sometimes

CONVENTIONAL DEFINITION OF A CONFOUNDER.

Consider again the following DAGs for confounding:

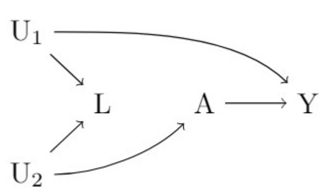


- In all three DAGs, L meets the three conventional criteria for being a confounder
- The structural and conventional definitions both identify L as a confounder

- Are there scenarios where the structural and conventional definitions of confounding contradict each other?

M-BIAS: POLL QUESTION 1.

Consider the following DAG:

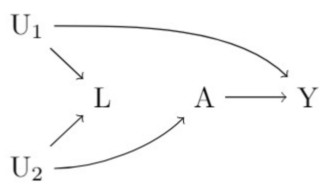


Is L associated with A?

A. Yes
B. No

M-BIAS: POLL QUESTION 2.

Consider the following DAG:

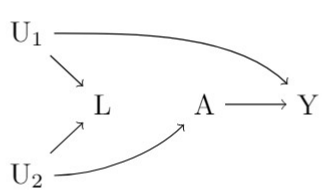


Is L associated with Y (not through A)?

A. Yes
B. No

M-BIAS: POLL QUESTION 3.

Consider the following DAG:

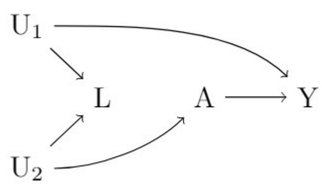


Is L on the causal pathway from A to Y?

A. Yes
B. No

M-BIAS:
POLL QUESTION 4.

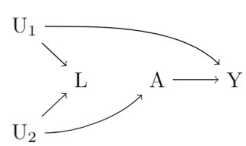
Consider the following DAG:



L meets the three criteria for the traditional definition of a confounder. However, what happens if we condition on L in this DAG?

A. We eliminate bias by closing the backdoor path from A to U_2 to L to U_1 to Y
B. We introduce bias by opening a backdoor path from A to U_2 to L to U_1 to Y
C. Nothing – the path from A to U_2 to L to U_1 to Y remains open
D. Nothing – the path from A to U_2 to L to U_1 to Y remains closed

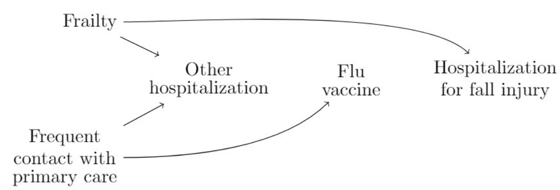
M-BIAS: WHERE TRADITIONAL DEFINITIONS FAIL.



This DAG structure (referred to as M-bias) is an example of when:

- The traditional definitions identify L as a confounder, but
- The structural definition tells us not to condition on L (and doing so will introduce bias)

M-BIAS: EXAMPLE.



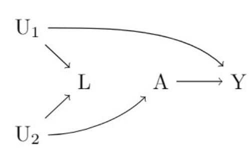
Suppose flu vaccine has no effect on being hospitalized for a fall injury.

Among people who have had other hospitalizations:

- Frailty and frequent contact with primary care are inversely related
- People who have had flu vaccine are:
  - More likely to have frequent contact with primary care
  - Less likely to be frail
  - Less likely to be hospitalized for a fall injury

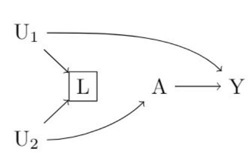
CONFOUNDERS ARE A RELATIVE CONCEPT.

Whether or not a variable is a confounder depends on what other variables in the DAG are (or are not) being conditioned on



In this DAG:

- No open backdoor paths from A to Y
- No confounding
- No confounders



In this DAG:

- Open backdoor path from A to Y
- U_1 and U_2 are confounders; both of these variables (if measured) can block the open path

SURROGATE CONFOUNDERS.

Sometimes, we don't have data on a confounder itself (U), but we have collected data on a proxy or surrogate confounder (L)

Conditioning on this variable will reduce some (but not all) of the bias

$$\begin{tikzcd}
U \arrow[r] \arrow[rr, bend left] \arrow[dr] & A \arrow[r] & Y \\
& \boxed{L} &
\end{tikzcd}$$

CONFOUNDING: TAKEAWAYS.

Confounding bias arises from an open backdoor path from A to Y, or when there is a variable that is a common cause of A and Y.

Confounders are variables that will block an open backdoor path when conditioned on.

Using non-structural definitions of confounders can potentially introduce bias (by identifying colliders as confounders).

SELECTION BIAS

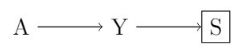
WHAT IS SELECTION BIAS?

Many different forms of selection bias:
- Berkson's bias
- Loss to follow-up
- Non-response bias
- Volunteer bias
- Missing data bias
- Etc.

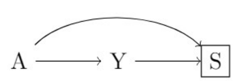
Not all forms of selection result in selection bias

Arises through the selection of participants into a study or analysis
- Conditioning on a common effect of treatment (or a cause of treatment) and outcome (or a cause of the outcome)
- Also referred to as collider-stratification bias (because the structure of selection bias is stratifying on a collider)

SELECTION BIAS IN CASE-CONTROL STUDIES.



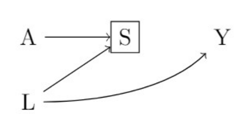
- Case-control studies selects individuals based on their outcome
  - Individuals who develop the outcome are oversampled in the study population
- In a DAG, indicated by drawing an arrow from the outcome (Y) to selection (S)
- We draw a box around selection (S) because our analysis would necessarily be restricted to individuals selected into the study



- If selection of controls is related to exposure, we introduce selection bias

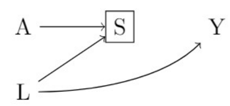
SELECTION BIAS IN FOLLOW-UP STUDIES.

- Selection bias can arise from:
  - Selection into the study
  - Loss to follow-up
- Example:



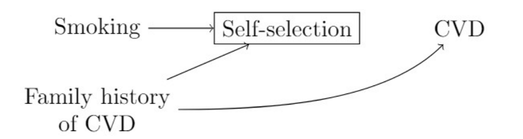
- Bias due to conditioning on a collider S
- S is a common effect of A (exposure) and L (cause of the outcome)

SELECTION BIAS IN OBSERVATIONAL FOLLOW-UP STUDIES.

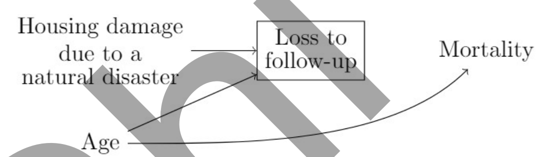


In an observational follow-up study, similar DAG structures could apply to both selection into the study and loss to follow-up

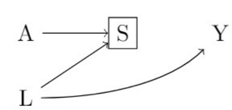
Selection into the study



Loss to follow-up



SELECTION BIAS IN RANDOMIZED TRIALS: POLL QUESTION 1.

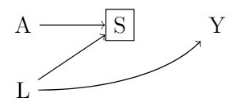


In randomized trials, can you have selection bias due to loss to follow-up?

A. Yes

B. No

SELECTION BIAS IN RANDOMIZED TRIALS: POLL QUESTION 2.

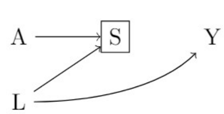


In randomized trials, can you have selection bias due to selection into the study?

A. Yes

B. No

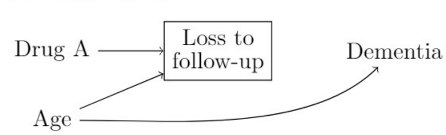
SELECTION BIAS IN RANDOMIZED TRIALS.



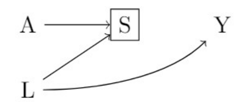
Selection into the study

- Selection bias does not arise from selection into a randomized trial
- Treatment is assigned *after* being selected into the study
- Selection, not selection bias
- Internal validity, but not necessarily external validity

Loss to follow-up



CONTROLLING FOR SELECTION BIAS: POLL QUESTION.



The bias arises from conditioning on S which opens the path from A to S to L to Y.

The best way to address selection bias is to prevent it from happening in the first place:

- Carefully evaluate inclusion/exclusion criteria
- Minimize loss to follow-up and missing data

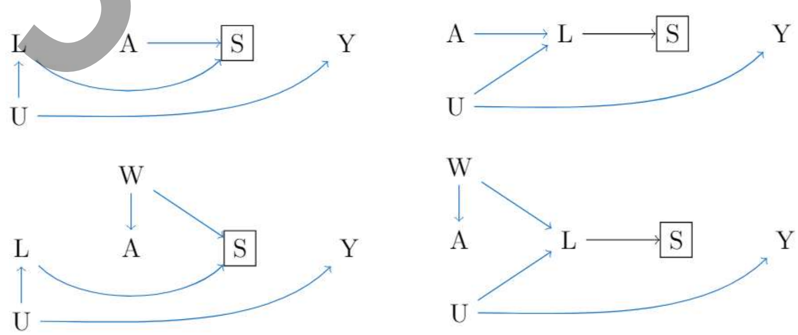
Given the DAG above, is there any way to address selection bias in the analysis?

A. No, it's a hopeless cause.

B. Yes, condition on L, which blocks the path from A to S to L to Y.

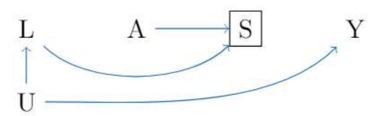
OTHER CAUSAL STRUCTURES FOR SELECTION BIAS.

! Don't forget that conditioning on a descendant of a collider can also open a path.



Modified from What If (Hernán and Robins, 2020)

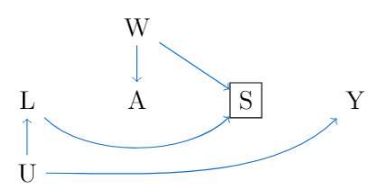
SELECTION BIAS STRUCTURES: POLL QUESTION 1.



In the DAG above, can we estimate the causal effect of A on Y by conditioning on L?

A. Yes
B. No

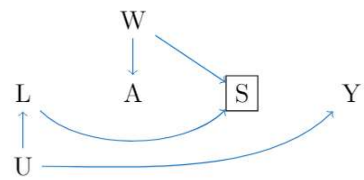
SELECTION BIAS STRUCTURES: POLL QUESTION 2.



In the DAG above, can we estimate the causal effect of A on Y by conditioning on L?

A. Yes
B. No

SELECTION BIAS STRUCTURES: POLL QUESTION 3.

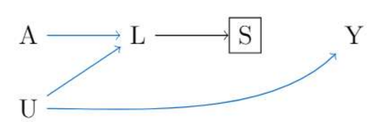


Is the above DAG possible in a randomized trial?

A. Yes

B. No

SELECTION BIAS STRUCTURES: POLL QUESTION 4.

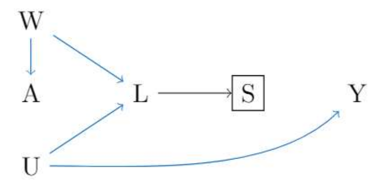


In the DAG above, can we estimate the causal effect of A on Y by conditioning on L?

A. Yes

B. No

SELECTION BIAS STRUCTURES: POLL QUESTION 5.

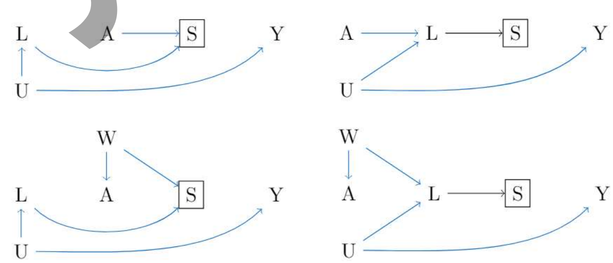


In the DAG above, can we estimate the causal effect of A on Y by conditioning on L?

A. Yes

B. No

SELECTION BIAS CONTROL.



- Conditioning (i.e. stratifying) on variables doesn't always succeed in addressing selection bias
- In fact, can sometimes exacerbate the bias (see the two DAGs on the right)
- Different analytical strategy must be used to address bias in these cases
  - Inverse probability weighting
  - G-formula
  - For more information, refer to What If (Hernán and Robins, 2020)

SELECTION BIAS: TAKEAWAYS.

Selection bias arises from conditioning on a common effect of treatment (or a cause of treatment) and outcome (or a cause of the outcome)

Different study designs are more prone to certain types of selection bias than others.

Note: Case-control studies sample from an underlying cohort for efficiency. This underlying cohort is vulnerable to biases from selection into the cohort and loss to follow-up. These biases will carry into the case-control study as well.

Stratification-based methods don't always work to address selection bias (but inverse probability weighting and g-formula always work).

BREAKOUT ROOM: DAG EXERCISES.