Causal Diagrams

Friday morning

- Introduction to graphical models
- Model definition
  - Structural equations models
  - Relationship to the counterfactual model
- Inspecting causal diagrams to discover associations.
- Detection and control of confounding

What are graphs useful for?

- **To model.** Encode assumptions of complex systems in an intuitive form. Why model?
  - **To design your study**. Help determine which variables you ***must*** measure in order to be able to estimate without bias the effect measure you care about (which L variables you must measure?)
  - **To communicate.** Communicate sources of bias with others.
  - **To understand**. To understand the limitations of your study, which causal effects can and cannot be learnt from your data

Assumption languages

- With words
  - ***Exposed and unexposed are exchangeable***
- With equations
  - **P[Y1=1|A=1] = P[Y1=1|A=0]**
  - **P[Y0=1|A=1] = P[Y0=1|A=0]**
- With graphs

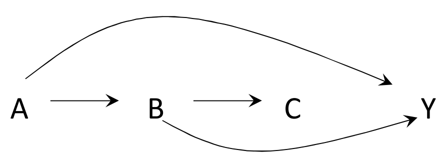
$$A \longrightarrow Y$$

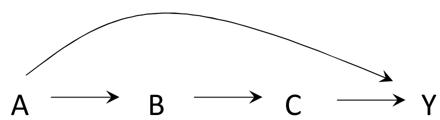
- With words
  - ***Exposed and unexposed are not exchangeable***
- With equations
  - **P[Y1=1|A=1] ≠ P[Y1=1|A=0]**
  - **P[Y0=1|A=1] ≠ P[Y0=1|A=0]**
- With graphs

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

Arrows

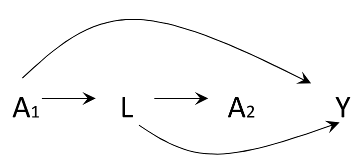
- Arrows represent potential causal effects
- **Present arrow says** you don't know causal effect
  May be there is effect, may be there is no effect
- **Absent arrows** convey the assumption of ***no causal effect***
- **Direction of arrow** conveys information of ***temporal ordering***



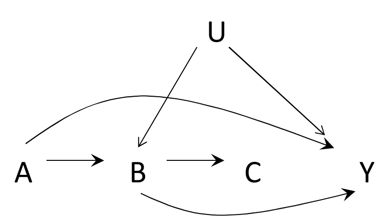


Causal Diagrams

- A diagram is called <u>*CAUSAL*</u> when it includes ***all common causes***
  - Whether measured or unmeasured

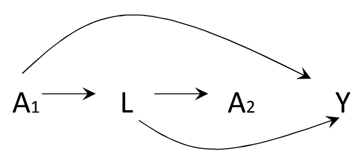


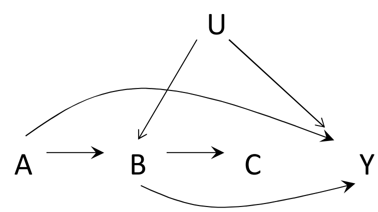
or



Causal Diagrams

- A causal diagram is a **directed acyclic graph (DAG)**
  - Directed: because edges have directions (arrows)
  - Acyclic: because no directed path crosses twice the same node (no cycles or loops)

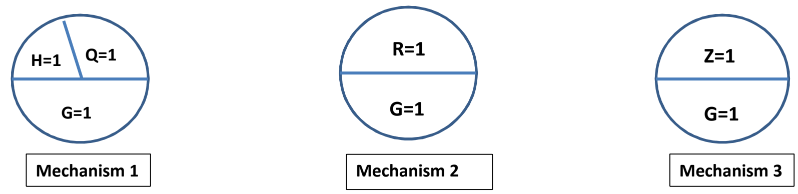




Directed acyclic graphs

- DAGs represent a number of different models about how the variables are generated in
  - the factual world, and
  - in intervened (hypothetical worlds).
- Many different causal models can be represented by a DAG.
- There is no such thing as "the causal model" represented by a DAG.
- I will now give one such model, the so called *non-parametric structural equations model*, because it is the easiest to describe. The model was proposed by Pearl (1995).

Sufficient cause model. Outcome: emphysema



Z = smoke more than 2 packs per day between age 30 and 50 : 1=yes
G = carry a protective gen : 1 = no
Q = exposed to environmental chemicals: 1 = yes
H = smoke more than 2 packs per day between age 20 and 30 : 1=yes
R = heavy passive smoker for at least 25 years : 1 = yes
D = emphysema : 1 = yes

With equations, the model implies that

$$D = 1 - (1 - H^* Q^* G) * (1 - R^* G) * (1 - Z * G)$$

Also,

$$D = f_D(Z, H, \varepsilon_D)$$ where $\varepsilon_D = \{Q, R, G\}$ are the "other causes"

Structural equations

A parametric structural equation

$$D = 1 - (1 - H^* Q^* G) * (1 - R^* G) * (1 - Z * G)$$

A non-parametric structural equation

$$D = f_D(Z, H, \varepsilon_D)$$

10

Parametric structural equations

The equation

$$D = 1 - (1 - H^* Q^* G) * (1 - R^* G) * (1 - Z * G)$$

is an example of a ***"parametric structural equation"***.

It is a ***formula*** that gives the ***prescription*** for how the components of sufficient causes combine to bring about the outcome.

11

Non-parametric structural equation.

The formula

$$D = f_D(Z, H, \mathcal{E}_D)$$

conveys also partial information about how D comes about.
It represents the assumption that if the other causes $\mathcal{E}_D$ occur at a certain level, *the outcome D would be different* if

Z and H were set (possibly contrary to fact) to some level z,h

than if

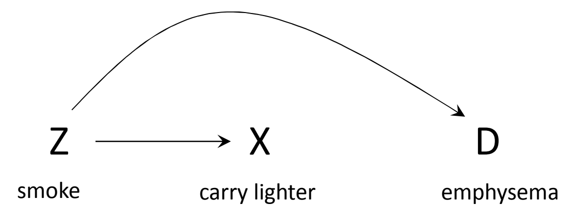
Z and H were set (possibly contrary to fact) to z,h*.*

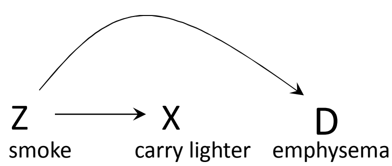
However, the formula does not convey information about

1. What are the "other causes" $\mathcal{E}_D$, or about
2. The mathematical formula for combining Z, H and $\mathcal{E}_D$

12

Causal diagrams





Non-parametric structural equations represented by the DAG

$Z = f_Z(\varepsilon_Z)$

$X = f_X(Z, \varepsilon_X)$

$D = f_D(Z, \varepsilon_D)$

The *DAG is causal* if it encodes a collection of *non-parametric structural equations*, one for each node, such that:

a) Only the "parents" of a node and the "other causes" ε intervene in the structural equation for the node

b) The "other causes" ε :
- b.1) do not include any of the other nodes in the diagram,
- b.2) are not caused by the other nodes in the diagram, or by the other ε's, and
- b.3) do not have common causes with the other ε's

Non-parametric structural equations model

- In the model proposed by Pearl (1995), assumption (b) is formalized in statistical terms with the assumption that :
  - **The Ɛ's are mutually independent random variables, i.e. unassociated.**

- When the assumption of mutually independent errors is made, the model represented by the DAG is called a

NON-PARAMETRIC STRUCTURAL EQUATIONS MODEL (NPSEM).

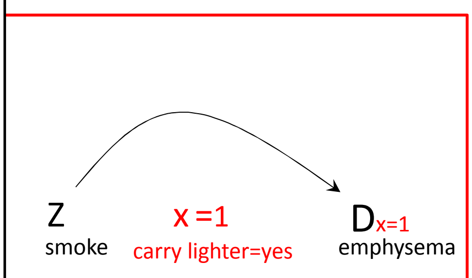
Causal Markov condition

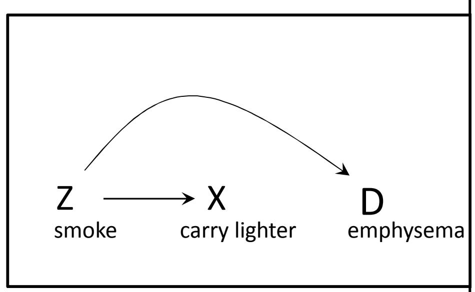
- A number of mathematical facts follow from the assumptions of an NPSEM.
- The most important one is:

- **Local Causal Markov Condition: any variable in the DAG is conditionally independent (i.e. unassociated) with its non-descendants given its parents.**

Causal Markov condition

- Statisticians and computer scientists in the last 15 years have shown that when the **Local Causal Markov Condition** holds, then one can deduce a number of results about marginal and conditional associations between the variables of a DAG, by simply inspecting the DAG and applying some simple graphical rules.
- We will learn these graphical rules today but first we will investigate the connection between the non-parametric structural equation model and the counterfactual model.





Intervened world (everybody carries a lighter)

Factual world

$Z = f_Z(\varepsilon_Z)$

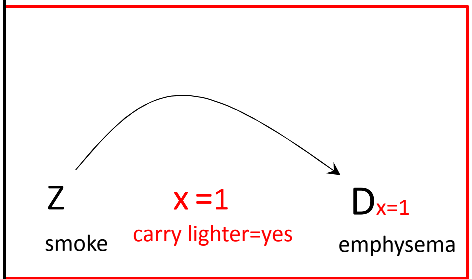
$X_{x=1} = 1$

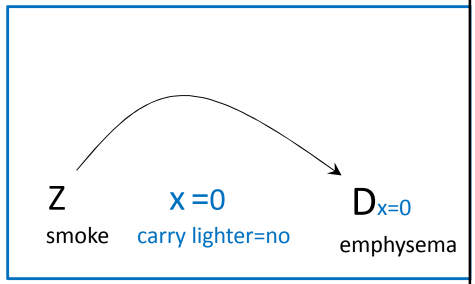
$D_{x=1} = f_D(Z,\varepsilon_D)$ same as $D = f_D(Z,\varepsilon_D)$

$Z = f_Z(\varepsilon_Z)$

$X = f_X(Z,\varepsilon_X)$

$D = f_D(Z,\varepsilon_D)$





Intervened world (everybody carries a lighter)

Intervened world (nobody carries a lighter)

$Z = f_Z(\varepsilon_Z)$

$X_{x=1} = 1$

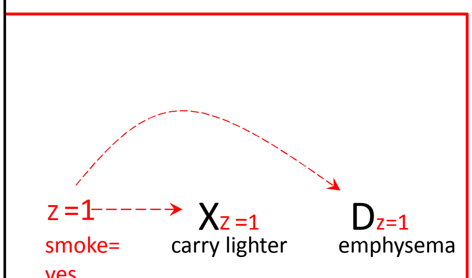
$D_{x=1} = f_D(Z,\varepsilon_D)$

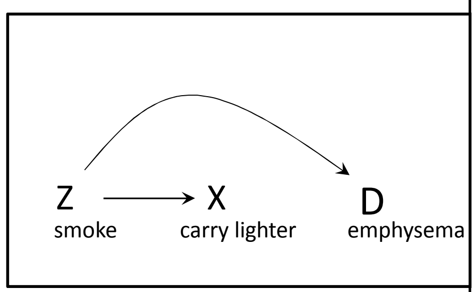
same as

$Z = f_Z(\varepsilon_Z)$

$X_{x=0} = 0$

$D_{x=0} = f_D(Z,\varepsilon_D)$





Intervened world (everybody smokes)

Factual world

$Z_{z=1} = 1$

$X_{z=1} = f_X(1,\varepsilon_X)$

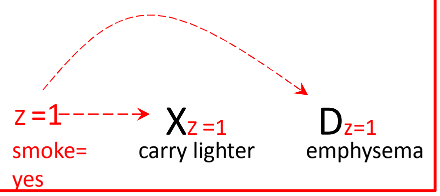
$D_{z=1} = f_D(1,\varepsilon_D)$

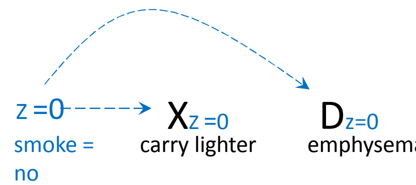
different than

$Z = f_Z(\varepsilon_Z)$

$X = f_X(Z,\varepsilon_X)$

$D = f_D(Z,\varepsilon_D)$





Intervened world (everybody smokes)

Intervened world (nobody smokes)

$Z_{z=1} = 1$

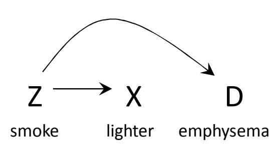
$X_{z=1} = f_X(1,\mathcal{E}_X)$

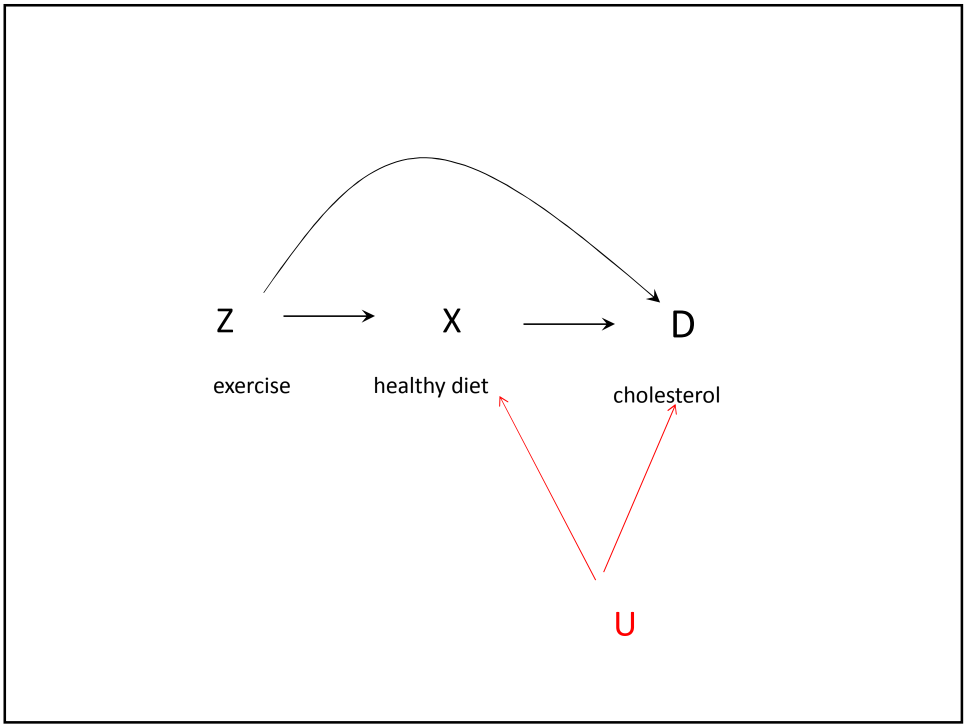
$D_{z=1} = f_D(1,\mathcal{E}_D)$ different than $D_{z=0} = f_D(0,\mathcal{E}_D)$

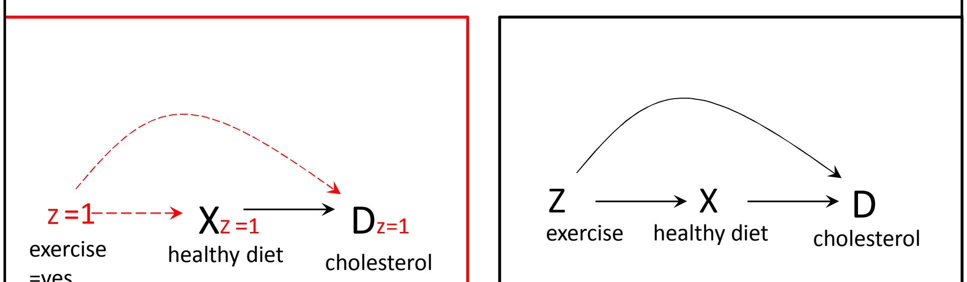
$Z_{z=0} = 0$

$X_{z=0} = f_X(0,\mathcal{E}_X)$

Patient id	Oracle's outcome when exposure combination is				Oracle's outcome when smoke is		Oracle's outcome when lighter is		Oracle's lighter when smoke is		Data analyst's table		
	Z=1 X=1	Z=0 X=1	Z=1 X=0	Z=0 X=0	Z=1	Z=0	X=1	X=0	Z=1	Z=0	Z	X	D
	D_{11} $f_D(1,\varepsilon d)$	D_{01} $f_D(0,\varepsilon d)$	D_{10} $f_D(1,\varepsilon d)$	D_{00} $f_D(0,\varepsilon d)$	$D_{z=1}$ $f_D(1,\varepsilon d)$	$D_{z=0}$ $f_D(0,\varepsilon d)$	$D_{x=1}$ $f_D(Z,\varepsilon d)$	$D_{x=0}$ $f_D(Z,\varepsilon d)$	$X_{z=1}$ $f_X(1,\varepsilon x)$	$X_{z=0}$ $f_X(0,\varepsilon x)$	$f_Z(\varepsilon z)$	$f_X(Z,\varepsilon x)$	$f_d(Z,\varepsilon d)$
1.	1	1	1	1	1	1	1	1	1	1	1	1	1
2.	1	1	1	1	1	1	1	1	0	0	1	0	1
3.	1	0	1	0	1	0	0	0	1	0	0	0	0
4.	1	0	1	0	1	0	0	0	0	0	0	0	0
5.	0	0	0	0	0	0	0	0	0	0	0	0	0
6.	0	0	0	0	0	0	0	0	1	0	1	1	0
7.	0	0	0	0	0	0	0	0	0	0	1	0	0
8.	0	0	0	0	0	0	0	0	0	0	0	0	0



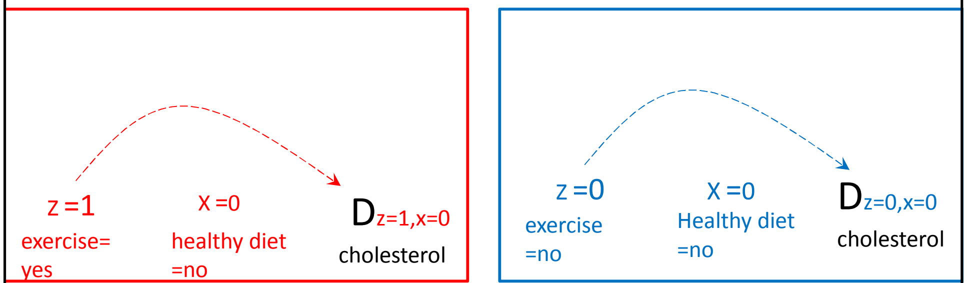




<u>Intervened world (everybody exercises)</u>

<u>Factual world</u>

Intervened world		Factual world
$Z_{z=1} = 1$		$Z = f_Z(\varepsilon_Z)$
$X_{z=1} = f_X(1, \varepsilon_X)$		$X = f_X(Z, \varepsilon_X)$
$D_{z=1} = f_D(1, X_{z=1}, \varepsilon_D)$	different than	$D = f_D(Z, X, \varepsilon_D)$



Intervened world (everybody exercises nobody has healthy diet)

Intervened world (nobody exercises nobody has healthy diet)

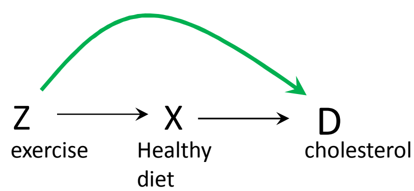
$Z_{z=1,x=0} = 1$

$X_{z=1,x=0} = 0$

$Z_{z=0,x=0} = 0$

$X_{z=0,x=0} = 0$

$D_{z=1,x=0} = f_D(1,0,\varepsilon_D)$ different than $D_{z=0,x=0} = f_D(0,0,\varepsilon_D)$



When we ***include*** a direct arrow from a variable, say Z, to another, say D, we express our belief that there ***might be*** a direct effect of Z on D ***relative to*** the other variables in the diagram.

By this we mean that there is *at least one subject and one level of the other variables in the diagram*, say X=x*, such that

$$D_{z=1,x=x^*} \neq D_{z=0,x=x^*}$$

$Z \longrightarrow X \longrightarrow D$

smoke Tar in lungs Lung cancer

$Z \longrightarrow X \longrightarrow D$

smoke Tar in lungs Lung cancer

When we ***do not include*** a direct arrow from a variable, say Z, to another, say D, we express our belief that there ***is not a direct effect*** of Z on D *relative to* the other variables in the diagram.

By this we mean that for *all subjects in the population and all the other variables in the diagram*, say X, it holds that

$$D_{z=1,x=x^*} = D_{z=0,x=x^*} \quad \text{for all } x^*$$

Patient id	Oracle's outcome when exposure combination is				Oracle's outcome when smoke is		Oracle's outcome when tar is		Oracle's tar when smoke is		Data analyst's table		
	Z=1 X=1	Z=0 X=1	Z=1 X=0	Z=0 X=0	Z=1	Z=0	X=1	X=0	Z=1	Z=0	Z	X	D
	D_{11} $f_D(1,\varepsilon_d)$	D_{01} $f_D(1,\varepsilon_d)$	D_{10} $f_D(0,\varepsilon_d)$	D_{00} $f_D(0,\varepsilon_d)$	$D_{z=1}$ $f_D(X_{z=1},\varepsilon_d)$	$D_{z=0}$ $f_D(X_{z=0},\varepsilon_d)$	$D_{x=1}$ $f_D(1,\varepsilon_d)$	$D_{x=0}$ $f_D(0,\varepsilon_d)$	$X_{z=1}$ $f_X(1,\varepsilon_x)$	$X_{z=0}$ $f_X(0,\varepsilon_x)$		$f_X(Z,\varepsilon_x)$	$f_d(X,\varepsilon_d)$
1.	1	1	1	1	1	1	1	1	1	1	1	1	1
2.	1	1	0	0	1	0	1	0	1	0	1	1	1
3.	0	0	1	1	0	0	0	1	1	1	1	1	0
4.	0	0	0	0	0	0	0	0	0	0	1	0	0
5.	1	1	0	0	1	0	1	0	1	0	0	0	0
6.	1	1	1	1	1	1	1	1	1	1	0	1	1
7.	0	0	0	0	0	0	0	0	0	0	0	0	0
8.	0	0	1	1	1	0	0	1	0	1	0	1	0

$Z \rightarrow X \rightarrow D$

smoke tar cancer

Patient id	Oracle's outcome when exposure combination is				Oracle's outcome when smoke is		Oracle's outcome when tar is		Oracle's tar when smoke is		Data analyst's table		
	Z=1 X=1	Z=0 X=1	Z=1 X=0	Z=0 X=0	Z=1	Z=0	X=1	X=0	Z=1	Z=0	Z	X	D
	D_{11} $f_D(1,\varepsilon_d)$	D_{01} $f_D(1,\varepsilon_d)$	D_{10} $f_D(0,\varepsilon_d)$	D_{00} $f_D(0,\varepsilon_d)$	$D_{z=1}$ $f_D(X_{z=1},\varepsilon_d)$	$D_{z=0}$ $f_D(X_{z=0},\varepsilon_d)$	$D_{x=1}$ $f_D(1,\varepsilon_d)$	$D_{x=0}$ $f_D(0,\varepsilon_d)$	$X_{z=1}$ $f_X(1,\varepsilon_x)$	$X_{z=0}$ $f_X(0,\varepsilon_x)$		$f_X(Z,\varepsilon_x)$	$f_d(X,\varepsilon_d)$
1.	1	1	1	1	1	1	1	1	1	1	1	1	1
2.	1	1	0	0	1	0	1	0	1	0	1	1	1
3.	0	0	1	1	0	0	0	1	1	1	1	1	0
4.	0	0	0	0	0	0	0	0	0	0	1	0	0
5.	1	1	0	0	1	0	1	0	1	0	0	0	0
6.	1	1	1	1	1	1	1	1	1	1	0	1	1
7.	0	0	0	0	0	0	0	0	0	0	0	0	0
8.	0	0	1	1	1	0	0	1	0	1	0	1	0

$D = f_D(X,\varepsilon_d)$: factual cancer status

$D_{z=1,x=1} = f_D(1,\varepsilon_d)$: counterfactl cancer. when smoke yes, tar yes

$D_{z=0,x=1} = f_D(1,\varepsilon_d)$: counterfactl cancer when smoke no, tar yes

$D_{x=1} = f_D(1,\varepsilon_d)$: counterfactl cancer when tar yes

$Z \rightarrow X \rightarrow D$

smoke tar cancer

Patient id	Oracle's outcome when exposure combination is				Oracle's outcome when smoke is		Oracle's outcome when tar is		Oracle's tar when smoke is		Data analyst's table		
	Z=1 X=1	Z=0 X=1	Z=1 X=0	Z=0 X=0	Z=1	Z=0	X=1	X=0	Z=1	Z=0	Z	X	D
	D_{11} $f_D(1,\varepsilon_d)$	D_{01} $f_D(1,\varepsilon_d)$	D_{10} $f_D(0,\varepsilon_d)$	D_{00} $f_D(0,\varepsilon_d)$	$D_{z=1}$ $f_D(X_{z=1},\varepsilon_d)$	$D_{z=0}$ $f_D(X_{z=0},\varepsilon_d)$	$D_{x=1}$ $f_D(1,\varepsilon_d)$	$D_{x=0}$ $f_D(0,\varepsilon_d)$	$X_{z=1}$ $f_X(1,\varepsilon_x)$	$X_{z=0}$ $f_X(0,\varepsilon_x)$		$f_X(Z,\varepsilon_x)$	$f_d(X,\varepsilon_d)$
1.	1	1	1	1	1	1	1	1	1	1	1	1	1
2.	1	1	0	0	1	0	1	0	1	0	1	1	1
3.	0	0	1	1	0	0	0	1	1	1	1	1	0
4.	0	0	0	0	0	0	0	0	0	0	1	0	0
5.	1	1	0	0	1	0	1	0	1	0	0	0	0
6.	1	1	1	1	1	1	1	1	1	1	0	1	1
7.	0	0	0	0	0	0	0	0	0	0	0	0	0
8.	0	0	1	1	1	0	0	1	0	1	0	1	0

$D = f_D(X,\varepsilon_d)$: factual cancer status

$D_{z=1,x=1} = f_D(1,\varepsilon_d)$: counterfactl cancer. when smoke yes, tar yes

$D_{z=0,x=1} = f_D(1,\varepsilon_d)$: counterfactl cancer when smoke no, tar yes

$D_{x=1} = f_D(1,\varepsilon_d)$: counterfactl cancer when tar yes

$\underset{\text{smoke}}{Z} \rightarrow \underset{\text{tar}}{X} \rightarrow \underset{\text{cancer}}{D}$

Lack of a direct arrow from Z to D implies $D_{z=1,x=1} = D_{z=0,x=1} = D_{x=1}$

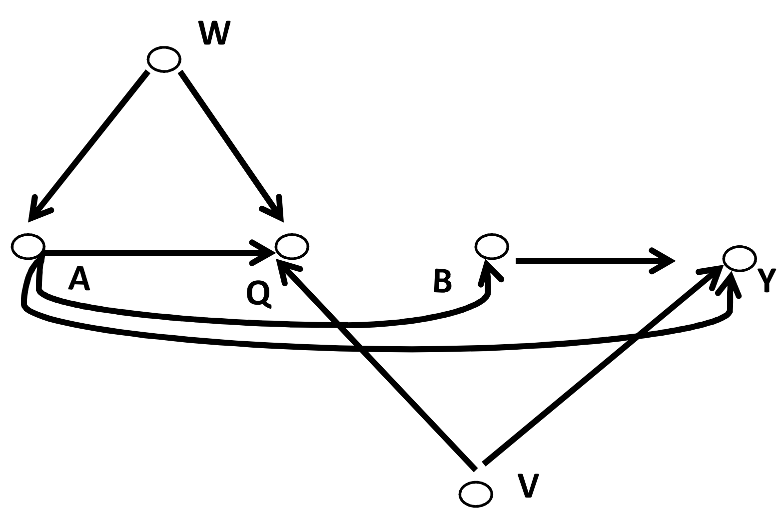


Diagram implies that A may have a causal effect on Y, that is, $P(Y_{a=0}=1) \neq P(Y_{a=1}=1)$

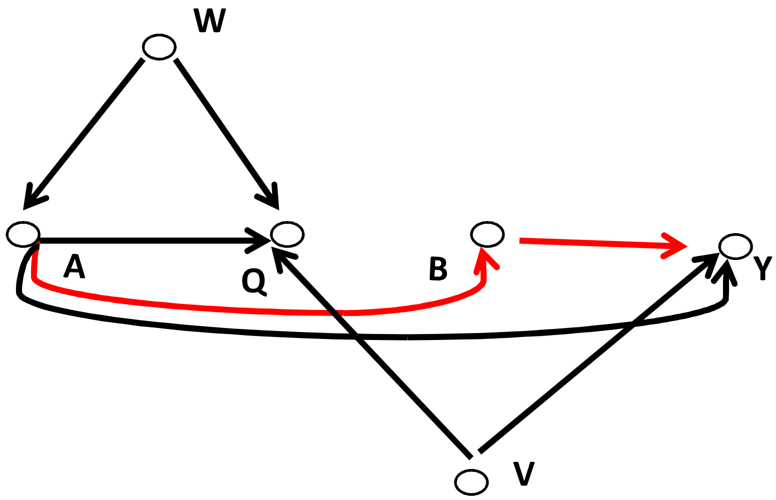


Diagram implies that A may have a causal effect on Y, that is,
$P(Y_{a=0}=1) \neq P(Y_{a=1}=1)$

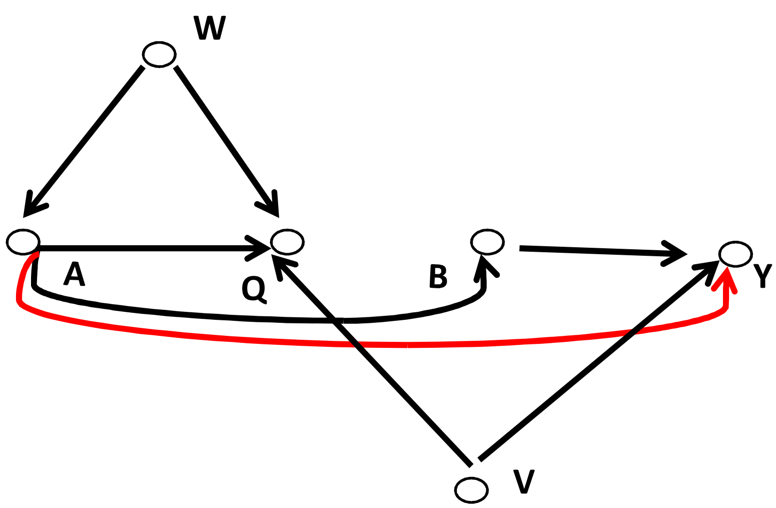
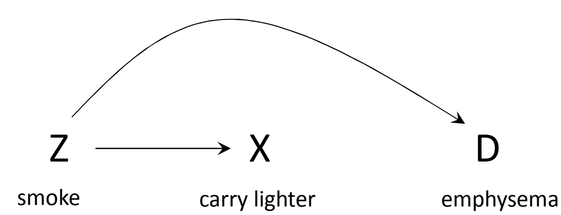


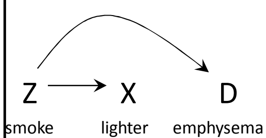
Diagram implies that A may have a causal effect on Y, that is,
$P(Y_{a=0}=1) \neq P(Y_{a=1}=1)$

Inspecting causal diagrams to discover associations.

Back to our first example



Patient id	Oracle's outcome when exposure combination is				Oracle's outcome when smoke is		Oracle's outcome when lighter is		Oracle's lighter when smoke is		Data analyst's table		
	Z=1 X=1	Z=0 X=1	Z=1 X=0	Z=0 X=0	Z=1	Z=0	X=1	X=0	Z=1	Z=0	Z	X	D
	D_{11} $f_D(1,\varepsilon d)$	D_{01} $f_D(0,\varepsilon d)$	D_{10} $f_D(1,\varepsilon d)$	D_{00} $f_D(0,\varepsilon d)$	$D_{z=1}$ $f_D(1,\varepsilon d)$	$D_{z=0}$ $f_D(0,\varepsilon d)$	$D_{x=1}$ $f_D(Z,\varepsilon d)$	$D_{x=0}$ $f_D(Z,\varepsilon d)$	$X_{z=1}$ $f_X(1,\varepsilon x)$	$X_{z=0}$ $f_X(0,\varepsilon x)$	$f_z(\varepsilon z)$	$f_x(Z,\varepsilon x)$	$f_d(Z,\varepsilon d)$
1.	1	1	1	1	1	1	1	1	1	1	1	1	1
2.	1	1	1	1	1	1	1	1	0	0	1	0	1
3.	1	0	1	0	1	0	0	0	1	0	0	0	0
4.	1	0	1	0	1	0	0	0	0	0	0	0	0
5.	0	0	0	0	0	0	0	0	0	0	0	0	0
6.	0	0	0	0	0	0	0	0	1	0	1	1	0
7.	0	0	0	0	0	0	0	0	0	0	1	0	0
8.	0	0	0	0	0	0	0	0	0	0	0	0	0



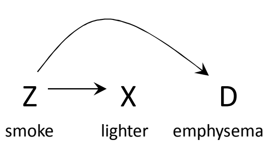
P(D=1|Z=1) =1/2

different from

P(D=1|Z=0) = 0

D and Z are (marginally) associated, i.e. dependent

Patient id	Oracle's outcome when exposure combination is				Oracle's outcome when smoke is		Oracle's outcome when lighter is		Oracle's lighter when smoke is		Data analyst's table		
	Z=1 X=1	Z=0 X=1	Z=1 X=0	Z=0 X=0	Z=1	Z=0	X=1	X=0	Z=1	Z=0	Z	X	D
	D_{11} $f_D(1,\varepsilon d)$	D_{01} $f_D(0,\varepsilon d)$	D_{10} $f_D(1,\varepsilon d)$	D_{00} $f_D(0,\varepsilon d)$	$D_{z=1}$ $f_D(1,\varepsilon d)$	$D_{z=0}$ $f_D(0,\varepsilon d)$	$D_{x=1}$ $f_D(Z,\varepsilon d)$	$D_{x=0}$ $f_D(Z,\varepsilon d)$	$X_{z=1}$ $f_X(1,\varepsilon x)$	$X_{z=0}$ $f_X(0,\varepsilon x)$	$f_z(\varepsilon z)$	$f_x(Z,\varepsilon x)$	$f_d(Z,\varepsilon d)$
1.	1	1	1	1	1	1	1	1	1	1	1	1	1
2.	1	1	1	1	1	1	1	1	0	0	1	0	1
3.	1	0	1	0	1	0	0	0	1	0	0	0	0
4.	1	0	1	0	1	0	0	0	0	0	0	0	0
5.	0	0	0	0	0	0	0	0	0	0	0	0	0
6.	0	0	0	0	0	0	0	0	1	0	1	1	0
7.	0	0	0	0	0	0	0	0	0	0	1	0	0
8.	0	0	0	0	0	0	0	0	0	0	0	0	0



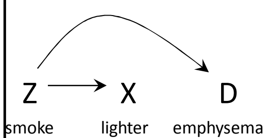
P(D=1|X=1) =1/2

different from

P(D=1|X=0) = 1/6

D and X are (marginally) associated i.e. dependent

	Oracle's outcome when exposure combination is				Oracle's outcome when smoke is		Oracle's outcome when lighter is		Oracle's lighter when smoke is		Data analyst's table		
Patient id	Z=1 X=1	Z=0 X=1	Z=1 X=0	Z=0 X=0	Z=1	Z=0	X=1	X=0	Z=1	Z=0	Z	X	D
	D_{11} $f_D(1,\varepsilon_d)$	D_{01} $f_D(0,\varepsilon_d)$	D_{10} $f_D(1,\varepsilon_d)$	D_{00} $f_D(0,\varepsilon_d)$	$D_{z=1}$ $f_D(1,\varepsilon_d)$	$D_{z=0}$ $f_D(0,\varepsilon_d)$	$D_{x=1}$ $f_D(Z,\varepsilon_d)$	$D_{x=0}$ $f_D(Z,\varepsilon_d)$	$X_{z=1}$ $f_X(1,\varepsilon_x)$	$X_{z=0}$ $f_X(0,\varepsilon_x)$	$f_z(\varepsilon_z)$	$f_x(Z,\varepsilon_x)$	$f_d(Z,\varepsilon_d)$
1.	1	1	1	1	1	1	1	1	1	1	1	1	1
2.	1	1	1	1	1	1	1	1	0	0	1	0	1
3.	1	0	1	0	1	0	0	0	1	0	0	0	0
4.	1	0	1	0	1	0	0	0	0	0	0	0	0
5.	0	0	0	0	0	0	0	0	0	0	0	0	0
6.	0	0	0	0	0	0	0	0	1	0	1	1	0
7.	0	0	0	0	0	0	0	0	0	0	1	0	0
8.	0	0	0	0	0	0	0	0	0	0	0	0	0



P(D=1|X=1,Z=1) =1/2

Equal to

P(D=1|X=0,Z=1) = 1/2

D and X are not associated, conditionally on Z=1, i.e. conditionally Independent given Z=1

Three basic graphical structures

Chain

$$A \longrightarrow B \longrightarrow Y$$

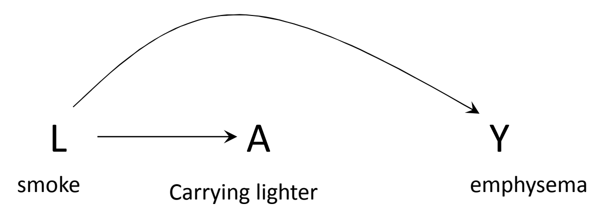
smoke   Tar in lungs   Lung cancer

Collider

$$\begin{tikzcd} A & C \arrow[r] & B \arrow[from=1-1, to=1-3, bend left] \end{tikzcd}$$

gene   smoke   Lung cancer

Forks



Causal DAGs and associations

<u>Causal Dags are useful to uncover associations</u>

There are four sources of associations

In causal language

1. Cause and effect
2. Common causes
3. Conditioning on a common effect
4. Conditioning on a consequence of a common effect

In graphical language

1. Direct path
2. Back door path
3. Conditioning on a collider
4. Conditioning on a descendant of a collider

Causal DAGs and associations

Causal Dags are useful to uncover associations
There are four sources of associations

In causal language

1. **Cause and effect**
2. Common causes
3. Conditioning on a common effect
4. Conditioning on a consequence of a common effect

In graphical language

1. **Direct path**
2. Back door path
3. Conditioning on a collider
4. Conditioning on a descendant of a collider

Causal DAGs and associations

Causal Dags are useful to uncover associations
There are four sources of associations

In causal language

1. **Cause and effect**

In graphical language

1. **Open direct path**

$$A \longrightarrow B \longrightarrow Y$$

smoke | Tar in lungs | Lung cancer

$$P(Y=1|A=1) \neq P(Y=1|A=0)$$

Y and A are associated

Causal DAGs and associations

Causal Dags are useful to uncover associations

There are four sources of associations

In causal language

1. Cause and effect
2. **Common causes**
3. Conditioning on a common effect
4. Conditioning on a consequence of a common effect

In graphical language

1. Direct path
2. **Back door path**
3. Conditioning on a collider
4. Conditioning on a descendant of a collider

Causal DAGs and associations

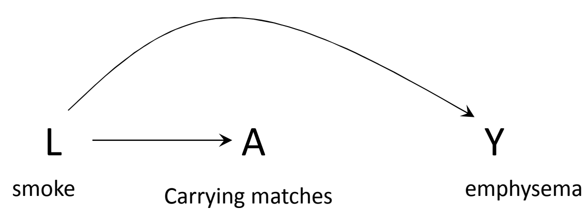
Causal Dags are useful to uncover associations

In causal language

2. **Common causes**

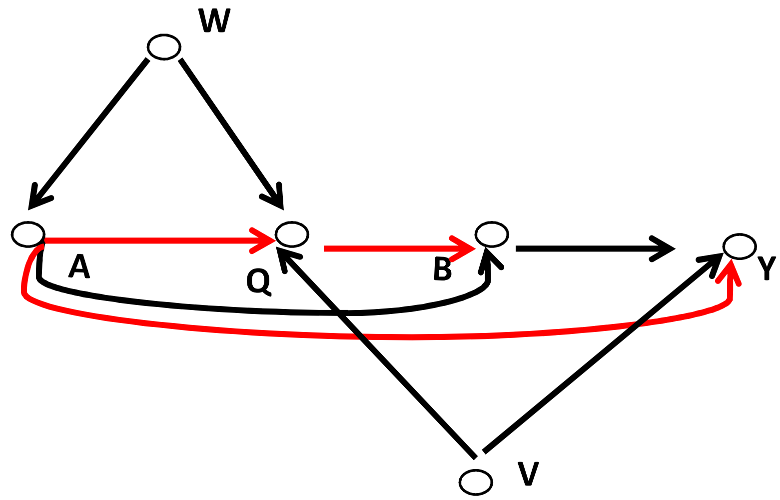
In graphical language

2. **Open back door path**

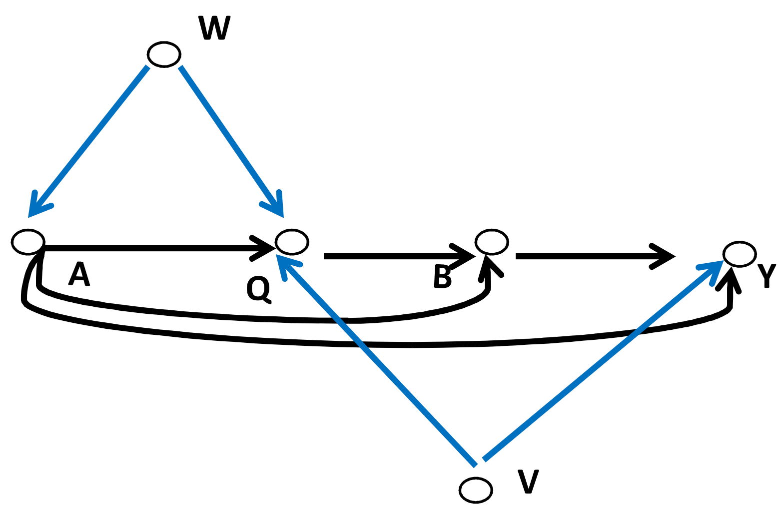


$P(Y=1|A=1) \neq P(Y=1|A=0)$ even though $P(Y_1=1) = P(Y_0=1)$

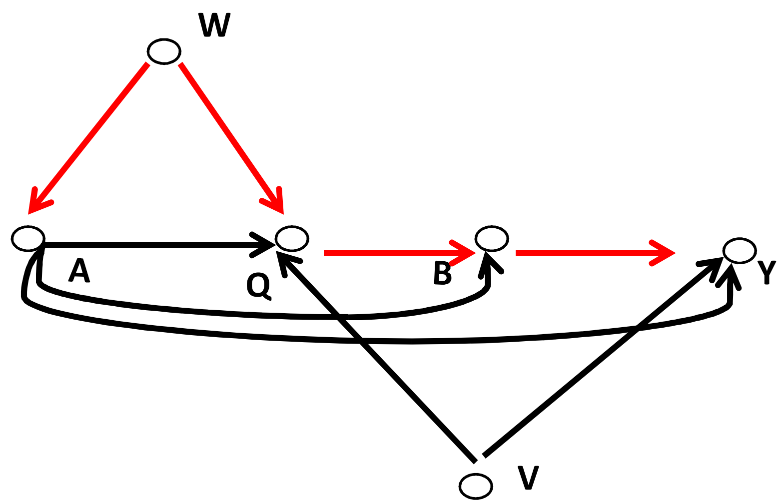
Association without causation because of common causes



Path in red is an **open back door path** between B and Y



Path in blue is a **blocked back door path** between A and Y because it hits ***the collider Q.***



Path in red is an **open back door path** between A and Y

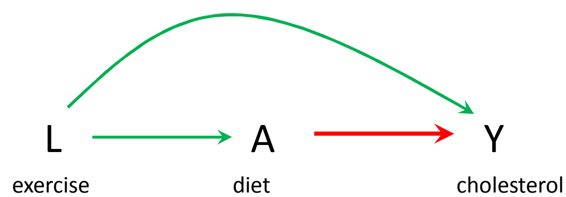
Open back door path and association

- **Result**: if in a causal DAG there is *at least one* **open back door path** between a node A and a node Y, then ("*most likely*"), Y and A are **marginally associated**.

Causal DAGs and associations

Two simultaneous sources of association

- **Cause and effect and Common causes**
- **Open direct path and back door path**



$P(Y=1|A=1) - P(Y=1|A=0)$ different from $P(Y_1=1) - P(Y_0=1)$

Association parameter different from causation parameter because of common causes

Causal DAGs and associations

Causal Dags are useful to uncover associations

There are four sources of associations

In causal language	**In graphical language**
1. Cause and effect	1. Direct path
2. Common causes	2. Back door path
3. Conditioning on a common effect	3. Conditioning on a collider
4. Conditioning on a consequence of a common effect	4. Conditioning on a descendant of a collider

Postpone temporarily the discussion of this

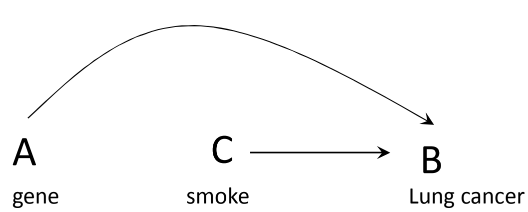
- We have seen two structures that generate association between A and Y

- We are now going to learn a structure that **DOES NOT** generate association between A and Y

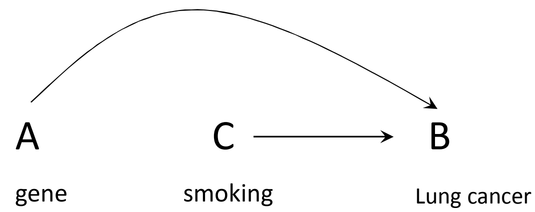
Common effects: colliders

- Paths that hit colliders do not generate associations

- **A and C are not associated**: only path between them is through a collider



Colliders: an example



Smoking and carrying the gene are not associated in the entire population regardless of the fact that they both cause lung cancer

Pr(smoke| **gene**) = Pr(smoke | **no gene**)
= Pr(smoke)

Knowing that someone carries the gene gives no information about his/her smoking status, i.e. does not change the chance that the person is a smoker

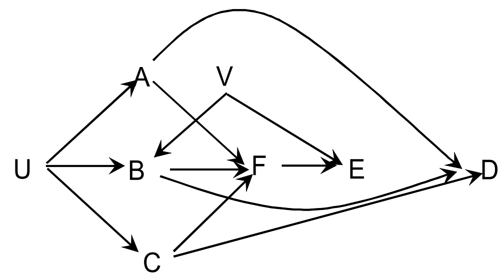
Lack of marginal association

- Two variables **are not (marginally)** associated when either

1. There exists no path between them, or
2. All paths between them **hit colliders**

Exercise

Which of the indicated paths are blocked and which are open?



		open	blocked
1	A F B		
2	A U B		
3	A F B D		
4	E F C D		
5	A U C D		
6	C F A D		
7	E V B U A D		
8	C U B V E		
9	E V B U C D		
10	E V B D		
11	**E F B D**		

Causal DAGs and associations

Causal Dags are useful to uncover associations

There are four sources of associations

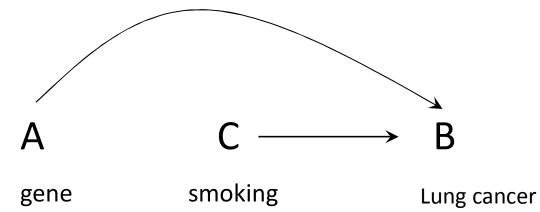
In causal language

1. Cause and effect
2. Common causes
3. **Conditioning on a common effect**
4. Conditioning on a consequence of a common effect

In graphical language

1. Direct path
2. Back door path
3. **Conditioning on a collider**
4. Conditioning on a descendant of a collider

Association generated by conditioning on a common effect: an example



Smoking and carrying the gene are not associated in the entire population

Pr(smoke| **gene**) = Pr(smoke | **no gene**) = Pr(smoke)

Knowing that someone carries the gene gives no information about his/her smoking status, i.e does not change the chance that the person is a smoker

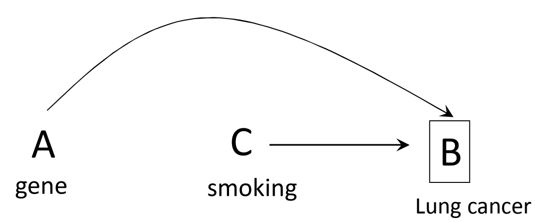
Causal DAGs and associations

In causal language

3. Conditioning on a common effect

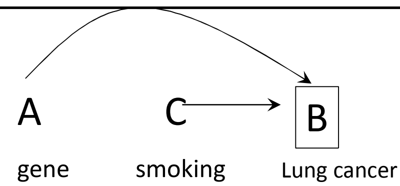
In graphical language

3. Open path through a conditioned collider



P(C=1|A=1,**B=1**) ≠ P(C=1|A=0,**B=1**)
or
P(C=1|A=1,**B=0**) ≠ P(C=1|A=0,**B=0**)

A and C are associated conditionally on B, i.e. within, at least one, B stratum



Smoking and carrying the gene are **negatively associated** in subjects with lung cancer.

<u>Extreme example.</u>

Suppose Pr(**smoke**)=.50, P(**gene**)= .01 , independent (marginally).

B = 1-(1-A)*(1-C)

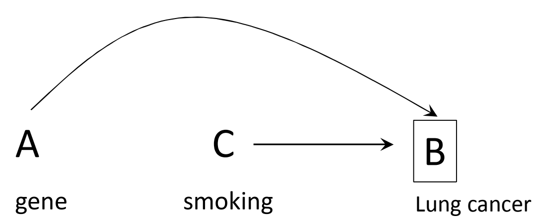
<u>(lung cancer if and only if have gene or smoke)</u>

Then, if someone has lung cancer and I know that she does not have the gene, *<u>she must smoke</u>*

Pr(**smoke**| **no gene, lung cancer**) = **1** which is greater than

Pr(**smoke**| **gene, lung cancer) = 0.5**

Association generated by conditioning on a common effect: an example



Smoking and carrying the gene are **negatively associated** in the subjects with lung cancer

Pr(smoke| **no gene, lung cancer**) **is higher than**
Pr(smoke | **gene, lung cancer**)

Knowing that a person with lung cancer does not have the gene provides some information about his smoking status because, in the absence of the gene, it is more likely that another cause of lung cancer such as smoking is present.

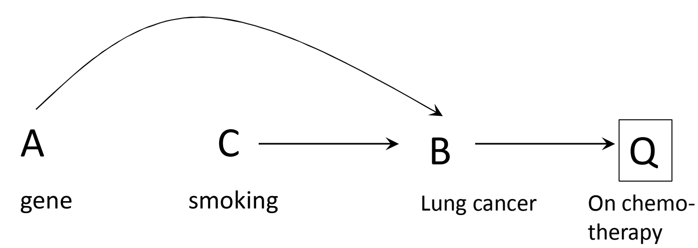
Causal DAGs and associations

Causal Dags are useful to uncover associations
There are four sources of associations

In causal language
1. Cause and effect
2. Common causes
3. Conditioning on a common effect
4. **Conditioning on a consequence of a common effect**

In graphical language
1. Direct path
2. Back door path
3. Conditioning on a collider
4. **Conditioning on a descendant of a collider**

Association generated by conditioning on a common effect: an example



Smoking and carrying the gene are **negatively associated** in the subjects with lung cancer

Pr(smoke| **no gene, on chemo**) **is higher than**
Pr(smoke | **gene, on chemo**)

Knowing that a person on chemo does not have the gene provides some information about his smoking status because, in the absence of the gene, it is more likely that another cause resulting in chemotherapy such as smoking is present.

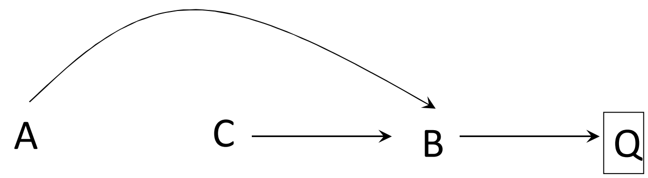
Causal DAGs and associations

In causal language

3. Conditioning on a consequence of a common effect

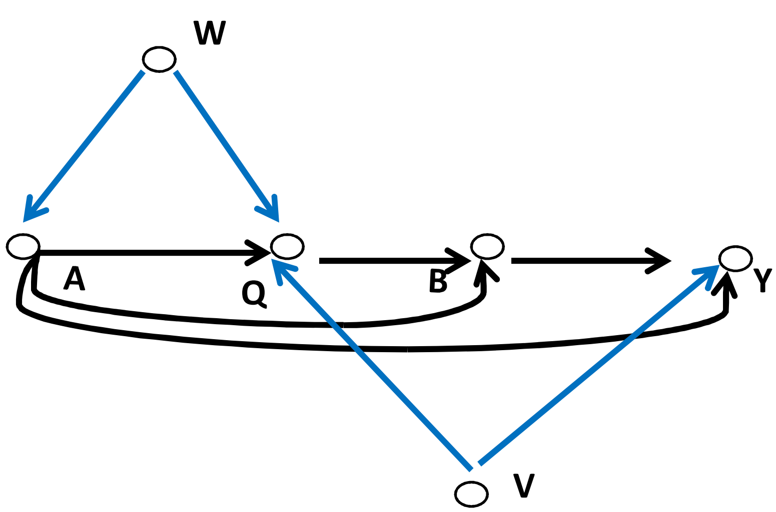
In graphical language

3. Open path through conditioning on a descendant of a collider

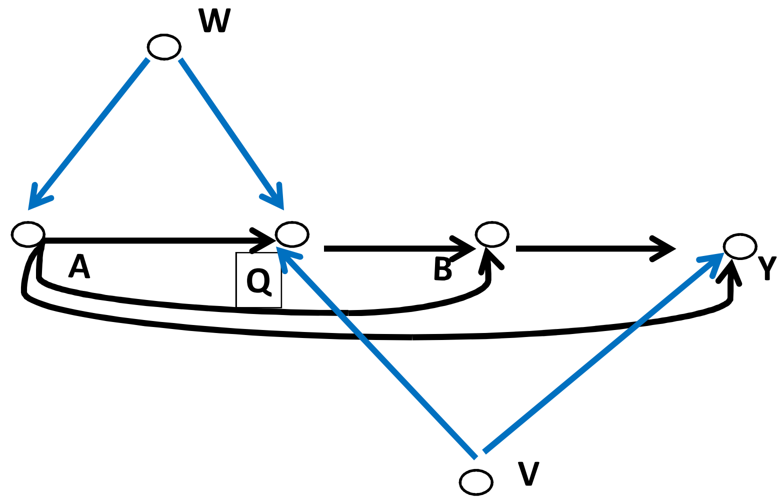


$P(C=1|A=1,Q=1) \neq P(C=1|A=0,Q=1)$

Y and A are associated conditionally on Q, i.e. within each Q stratum



Path in blue is a **blocked back door path** between A and Y because it hits ***the collider Q.***

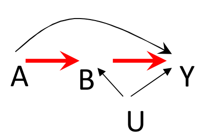


Path in blue is an **unblocked back door path** between A and Y because we have conditioned on ***<u>the collider Q.</u>***

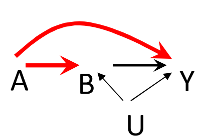
Open paths: summary

Marginal associations

- Direct path: **between A and Y**

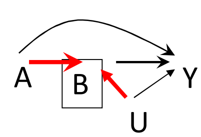


- Back door path: **between B and Y**

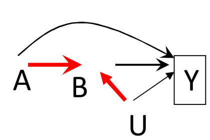


Conditional associations

- Path through a conditioned collider: **between A and U conditioned on B**



- Path through conditioning on a descendant of a collider: **between A and U conditioned on Y**



Lack of associations. Blocked paths

In causal language | **In graphical language**

Marginal lack of associations

In causal language	In graphical language
1. Common effect	1. Collider

Conditional associations

In causal language	In graphical language
2. Conditioning on a common cause	**2. Blocking** a back door path **by** a common ancestor
3. Conditioning on a variable on the causal path	**3. Blocking** a direct path **by** a variable in the path

Lack of associations. Block paths

In causal language | **In graphical language**

Marginal associations

In causal language	In graphical language
1. Common effect	1. Collider

Conditional associations

In causal language	In graphical language
2. Conditioning on a common cause	**2. Blocking a back door path by a common ancestor**
3. Conditioning on a variable on the causal path	**3. Blocking** a direct path **by** a variable in the path

Causal DAGs and associations

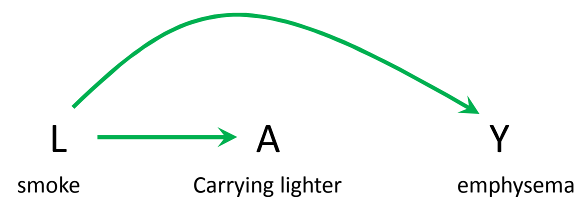
Causal Dags are useful to uncover associations

In causal language

2. Common causes

In graphical language

2. Open back door path



$P(Y=1|A=1) \neq P(Y=1|A=0)$ even though $P(Y_1=1) = P(Y_0=1)$

Association without causation because of common causes

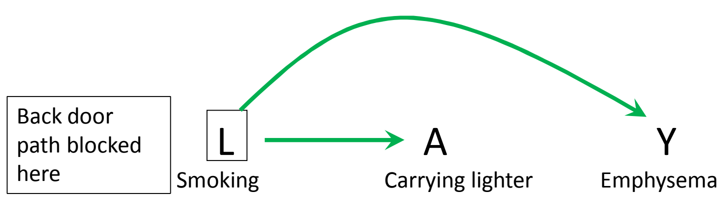
But conditioning on common causes destroys associations, closes or blocks paths

In causal language

2. Conditioning on a common causes

In graphical language

2. Blocking by conditioning on a common ancestor



$P(Y=1|A=1,L=1) = P(Y=1|A=0,L=1)$
and
$P(Y=1|A=1,L=0) = P(Y=1|A=0,L=0)$

Association prevented by conditioning on the common cause

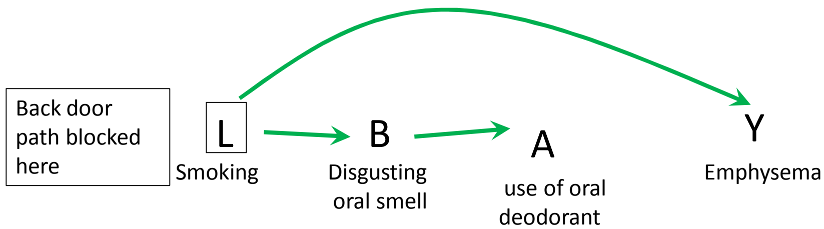
Another example of a closed or blocked path

In causal language

2. Conditioning on a common causes

In graphical language

2. Blocking by conditioning on a common ancestor



P(Y=1|A=1,**L=1**) = P(Y=1|A=0,**L=1**)
and
P(Y=1|A=1,**L=0**) = P(Y=1|A=0,**L=0**)

Association prevented by conditioning on the common cause

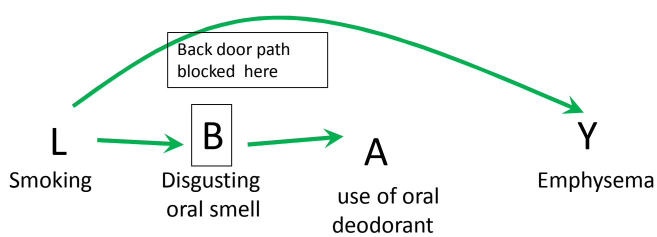
Another example of a closed or blocked path

In causal language

2. Conditioning on a common causes

In graphical language

2. Blocking by conditioning on a variable between trx and the common ancestor



P(Y=1|A=1,**B=1**) = P(Y=1|A=0,**B=1**)
and
P(Y=1|A=1,**B=0**) = P(Y=1|A=0,**B=0**)

Association prevented by conditioning on a variable in the Path between trx and the common cause of the outcome

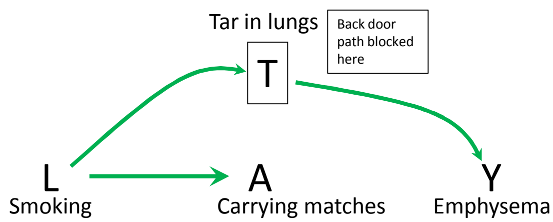
Another example of a closed or blocked path

In causal language

2. **Conditioning on a common causes**

In graphical language

2. **Blocking by conditioning on a variable between the outcome and the common ancestor**



P(Y=1|A=1,**T=1**) = P(Y=1|A=0,**T=1**)

and

P(Y=1|A=1,**T=0**) = P(Y=1|A=0,**T=0**)

Association prevented by conditioning on a variable in the path between the outcome and the common cause of the trx

Lack of associations that **close** paths

In causal language	In graphical language
Marginal associations	
1. Common effect	1. Collider
Conditional associations	
2. **Conditioning** on a common cause	2. **Blocking** a back door path **by** a common ancestor
3. **Conditioning on a variable on the causal path**	3. **Blocking a direct path by a variable in the path**

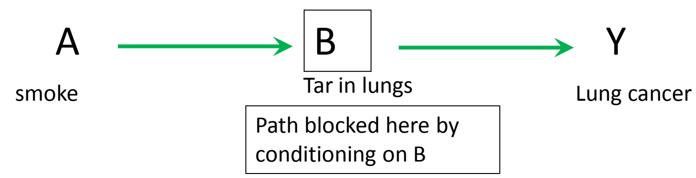
But conditioning on common causes destroys associations, blocks or closes paths

In causal language

3. **Conditioning on a variable on the causal path**

In graphical language

3. **Blocking a direct path by conditioning on a variable in the path**



$P(Y=1|A=1,B=1) = P(Y=1|A=0,B=1)$

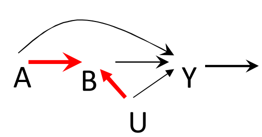
and

$P(Y=1|A=1,B=1) = P(Y=1|A=0,B=1)$

Closed or blocked paths: summary

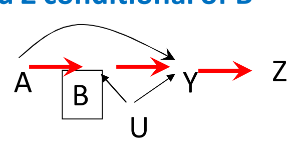
Marginal lack of association

- Path that hits a collider: **between A and U**



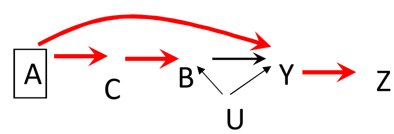
Conditional lack of association

- Conditioning on a node in a directed path: **between A and Z conditional of B**

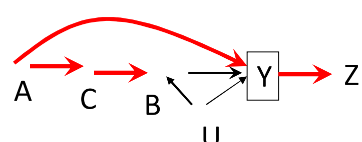


Conditional lack of association

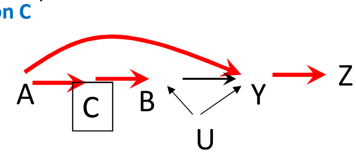
- Conditioning on a common ancestor: **between B and Z conditional of A**



- Conditioning on a variable on the back door path : **between B and Z conditioned on Y**

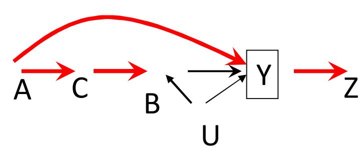


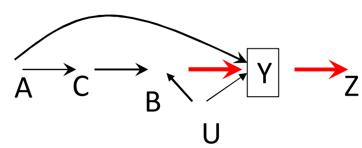
- Conditioning on a variable on the back door path: **between B and Z conditioned on C**

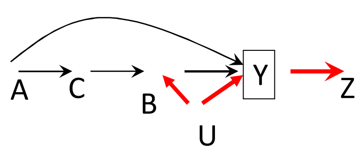


D-separation

- Two nodes B and Z of a DAG are **d-separated** by a set of nodes Y if all paths between B and Z are blocked by conditioning on Y





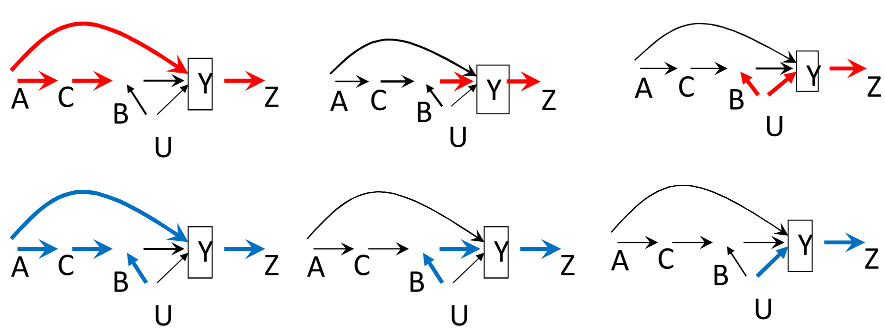


There are three paths between B and Z. All three are blocked by conditioning on Y.

B and Z are d-separated given Y
B and Z are d-separated conditional on Y

D-separation of sets of nodes

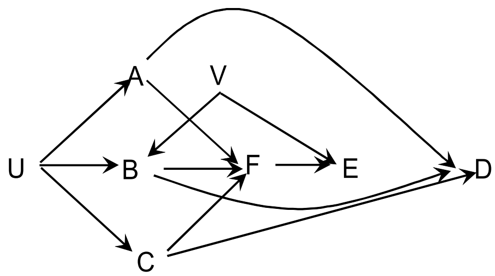
- Two sets of nodes D and E of a DAG are **d-separated** by a set of nodes G if all paths between any node of D and any node of E are blocked by conditioning on G



There are three paths between B and Z (in red). All three are blocked by conditioning on Y.
There are three paths between U and Z (in blue). All three are blocked by conditioning on Y.
The set of variables {B,U} and Z are **d-separated given** Y. The variables {B,U} **are d-separated conditional on Y.**

Exercise

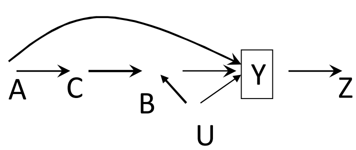
Which of the indicated pair of variables or sets of variables are d-separated by the third set of variables



		by set	D-separated	
			YES	NO
1	A & C	{ F}		
2	A & C	{F, U}		
3	A & C	{U}		
4	A & D	{C}		
5	A & E	{F}		
6	A & E	{F, B}		
7	E & D	{A, B, C}		
8	E & D	{F, A,C}		
9	E & D	{F, V}		
10	{A,C} & {B,V}	Empty set		
11	{A,C} & {B,V}	{U}		
12	{C,D} & V	Empty set		
13	{C,D} & V	{B }		
14	{C,D} & V	{B, U}		

D-separation theorem

- **Theorem**. (Verma and Pearl, 1988). If in a causal diagram two sets of nodes ***A and B are d-separated conditional on a set of nodes C***, then *A and B are statistically independent given C*
- *Note: Statistically independent is the same as "not associated"*



B and Z are d-separated given Y, then

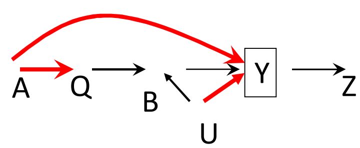
$P(Z=1|B=1,Y=1) = P(Z=1|B=0,Y=1)$

and

$P(Z=1|B=1,Y=0) = P(Z=1|B=0,Y=0)$

D-connection

- Two nodes A and B are d-connected conditional on a set of nodes C if A and B are not d-separated given C, i.e. if there exists at least one open path between A and B when we condition on C.

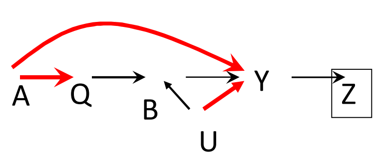


Q and U are d-connected given Y because the red path is open by conditioning on Y. (Y is a collider on the red path)

Is there any other open path between Q and U conditional on Y?

D-connection of sets of nodes

- Two **sets of nodes** D and E are d-connected conditional on another set of nodes C if there exists at least one node in D and one node in E that are not d-separated given C.

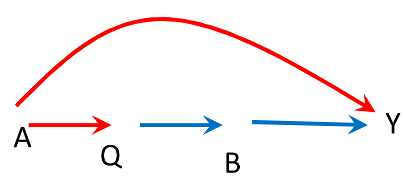


{A,Q} is d-connected with U given Z because the red path is open by conditioning on Z. (Z is a descendant of a collider on the red path)

D-connection and association

- **D-separation implies lack of association** (recall Verma and Pearl, Theorem from last class).
- *Does d-connection imply association?*
- **Answer:** it is **not true** that d-connection between two sets of nodes, say D and E, given G, implies association between D and E in at least one stratum of G.
- The reason is because there can exist associations in different directions that balance out and cancel with each other

D-connection and association



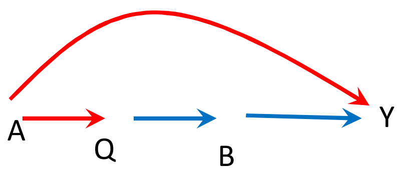
Q and Y are d-connected (given nothing). Two open paths: the red and the blue paths.

It may happen that the **red path** generates, say, a **negative association** between Q and Y, and the **blue path** generates, say, a **positive association** between Q and Y and that these associations cancel with each other to yield **NO association between Q and Y**

Faithfulness assumption

- **Faithfulness assumption**. If in a causal diagram two nodes ***A and B are d-connected conditional on a set of variables C***, then *A and B are statistically dependent given C*

- *Note: Statistically dependent is the same as "associated"*



Under the faithfulness assumption: Q and Y are marginally associated (because Q and Y are d-connected conditional on nothing)

P(Y=1|**Q=1**) ≠ P(Y=1|**Q=0**)

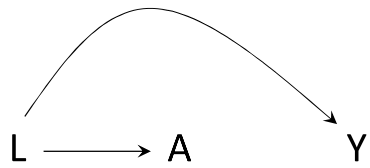
Faithfulness

- **Faithfulness assumption in plain English**: positive and negative associations along different connected paths never perfectly offset one another.

- Faithfulness does not hold when there exist perfect cancellations. Though some exceptions exist (see Robins and Wasserman, 1999), in most epidemiological studies these perfect cancellations will be rare. **So, from now on, we will make the faithfulness assumption**.

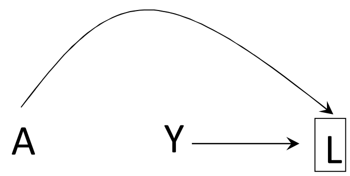
Two potential sources of bias in your causal analysis

Open back door path: lack of adjustment for variables that block a back door path that exists because of a common cause.



Confounding bias

Open path through adjustment on a collider. Open a path by adjusting on a common effect



Selection bias,
Berkson's bias

Confounding

- Confounding
  - Definition
  - How to check for confounding bias
  - How to adjust to prevent confounding bias
    - Which variables suffice to adjust for in order to prevent confounding bias?
    - Which variables we cannot adjust for because we will create artificial confounding bias?
- Confounders
  - Why are usual definitions wrong?
  - What is a sufficient set of confounders?

Confounding

- Confounding
  - Definition
  - How to check for confounding bias
  - How to adjust to prevent confounding bias
    - Which variables suffice to adjust for in order to prevent confounding bias?
    - Which variables we cannot adjust for because we will create artificial confounding bias?
- Confounders
  - Why are usual definitions wrong?
  - What is a sufficient set of confounders?

Confounding bias

- ***Confounding definition***: (Miettinen and Cook, 1981). Confounding bias (or simply, confounding) will be present if and only if exposure would remain associated with disease even if all exposure effects were removed, prevented or blocked.

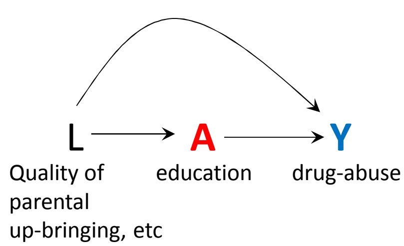
- If we interpret "***all exposure effects***" as effects of the exposure on "***all future events***" not just on disease, then ***graphically*** all exposure effects are prevented when we ***remove*** from the graph ***all arrows that emanate from the exposure***.

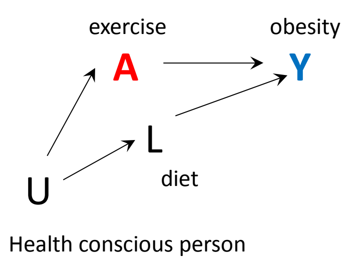
Confounding bias: graphical definition

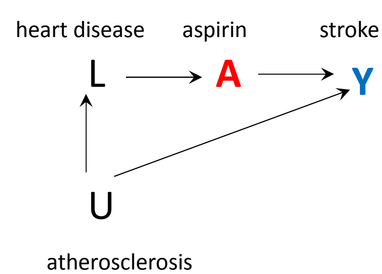
- ***Graphical definition of confounding bias***: Confounding bias (or simply confounding) will be present if and only if there exists at least one open back door path between exposure and outcome.

Confounding bias due to the presence of common causes

A : exposure Y: outcome



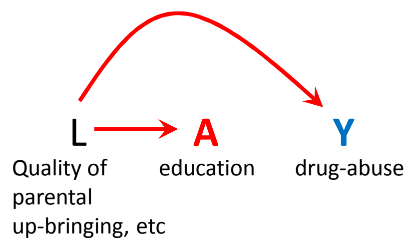


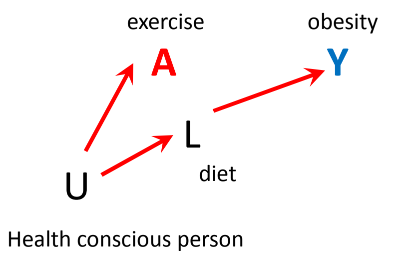


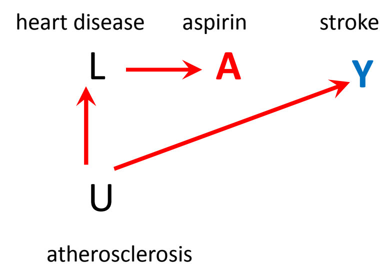
Confounding bias due to the presence of common causes

A : exposure Y: outcome

A and Y are still d-connected when we *remove all arrows emanating* from A.
=> *Confounding bias*







Confounding

- Confounding
  - Definition
  - How to check for confounding bias
  - How to adjust to prevent confounding bias
    - Which variables suffice to adjust for in order to prevent confounding bias?
    - Which variables we cannot adjust for because we will create artificial confounding bias?
- Confounders
  - Why are usual definitions wrong?
  - What is a sufficient set of confounders?

Confounding bias prevention

- **Back-door theorem** (Pearl, 1995). Suppose that

1. L is a set of nodes that are ***non-descendants*** of A
2. all ***back door paths*** between A and Y are ***blocked*** when we ***condition on L***

then, **within** each strata defined by the cross-classification of the **levels of L**, **the treated and the untreated are exchangeable**

Confounding bias prevention

- **Back-door theorem** (Pearl, 1995). Suppose that

1. L is a set of nodes that are ***non-descendants*** of A
2. all ***back door paths*** between A and Y are ***blocked*** when we ***condition on L***

then, *in each stratum defined by cross-classification of all variables in L*,

$$P(Y=1|A=1,L=l) = P(Y_1=1|L=l)$$

crude rate in the exposed in stratum l = **counterfactual rate if everyone in stratum l had been exposed**

and

$$P(Y=1|A=0,L=l) = P(Y_0=1|L=l)$$

crude rate in the unexposed in stratum l = **counterfactual rate if everyone in stratum l had not been exposed**

Confounding bias prevention

- **Back-door theorem** (Pearl, 1995) (Y continuous). Suppose that

1. L is a set of nodes that are ***non-descendants*** of A
2. all ***back door paths*** between A and Y are ***blocked*** when we ***condition on L***

then, *in each stratum defined by cross-classification of all variables in L*,

$$E(Y|A=1,L=l) = E(Y_1|L=l)$$

crude mean in the exposed in stratum l = counterfl mean if everyone in stratum l had been exposed

and

$$E(Y|A=0,L=l) = E(Y_0|L=l)$$

crude mean in the unexposed in stratum l = counterfl mean if everyone in stratum l had not been exposed

Confounding bias prevention

- **Back-door theorem** (Pearl, 1995). Suppose that

1. L is a set of nodes that are ***non-descendants*** of A
2. all ***back door paths*** between A and Y are ***blocked*** when we ***condition on L***

then, *standardized/ ipw rates adjusting for L are equal to the counterfactual rates in the entire population*,

$$\sum P(Y=1|A=1,L=l)P(L=l) = P(Y_1=1)$$

standardized rate = counterfactual rate if everyone had been exposed

and

$$\sum P(Y=1|A=0,L=l)P(L=l) = P(Y_0=1)$$

standardized rate = counterfactual rate if everyone had not been exposed

Confounding bias prevention

- **Back-door theorem** (Pearl, 1995). (Y continuous) Suppose that

  1. L is a set of nodes that are ***non-descendants*** of A
  2. all ***back door paths*** between A and Y are ***blocked*** when we ***condition on L***

then, ***standardized/ ipw means adjusting for L are equal to the counterfactual rates in the entire population***,

$$\sum E(Y|A=1,L=l)P(L=l) = E(Y_1)$$

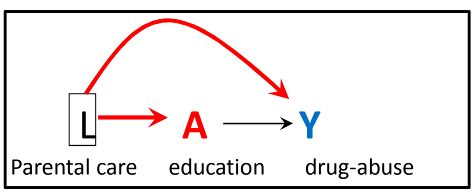
standardized mean = **counterfactual mean if everyone had been exposed**

and

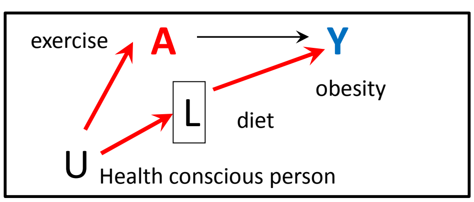
$$\sum E(Y|A=0,L=l)P(L=l) = E(Y_0)$$

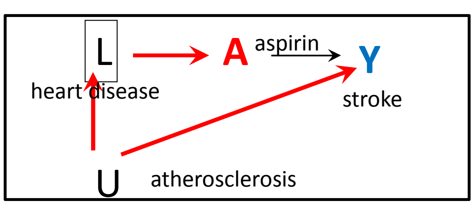
standardized mean = **counterfactual mean if everyone had not been exposed**

Stratification to prevent confounding bias



A : exposure Y: outcome
L = measured covariate
U = unmeasured covariate



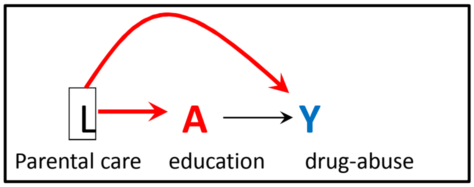


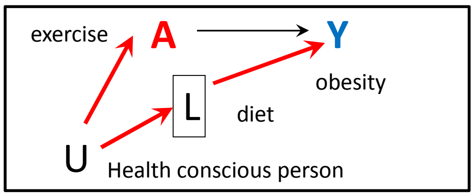
L blocks all back door paths between A and Y

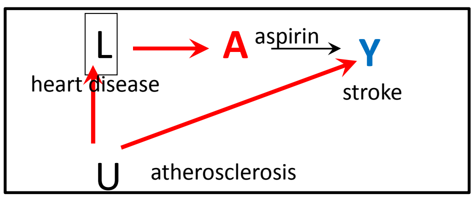
$\Rightarrow$within strata of L, no confounding bias

$\Rightarrow$ within strata of L crude rates = counterfcl rates

Standardization/IPW to prevent confounding bias







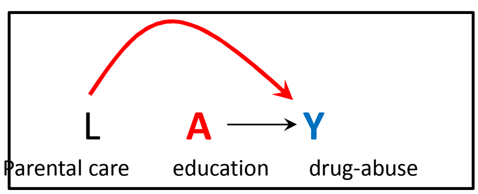
A : exposure Y: outcome
L = measured covariate
U = unmeasured covariate

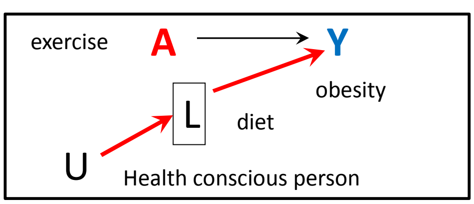
L blocks all back door paths between A and Y

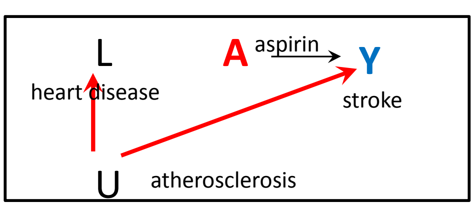
⇒

Standardized/IPW rates adjusted for L
=
counterfactual rates in the entire population

Graphs in pseudo-population after Inverse Probability Weighting







L blocks all back door paths between A and Y

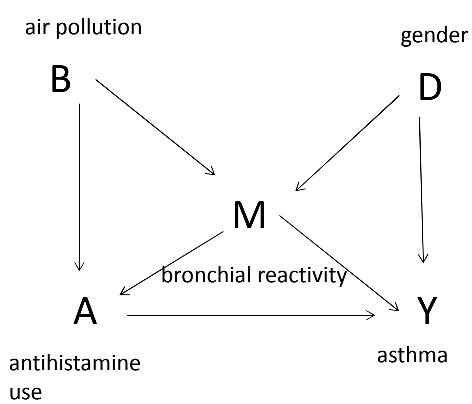
⇒ IPW (adjusting for L) creates a pseudo-population with two *clones of each person*: one person is treated and another is not treated.

⇒ in pseudo population there is exchangeability, *A has been randomized*

⇒In pseudo-population
Crude rates = counterfl rates

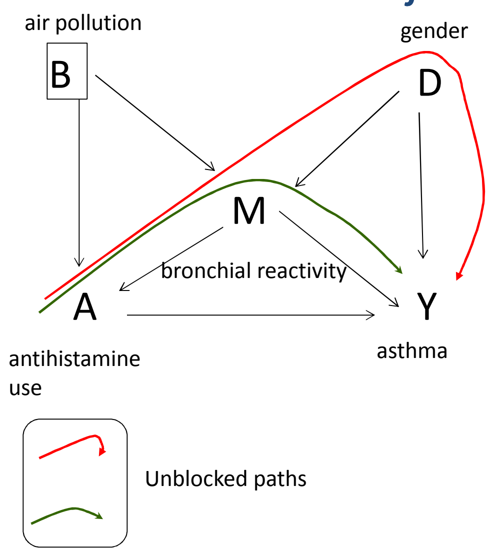
Sometimes we have more than one option to choose the variables that control confounding

Confounding bias: which variables should we adjust for?



Example taken from Greenland, Pearl and Robins, 1999.

Confounding bias: which variables should we adjust for?

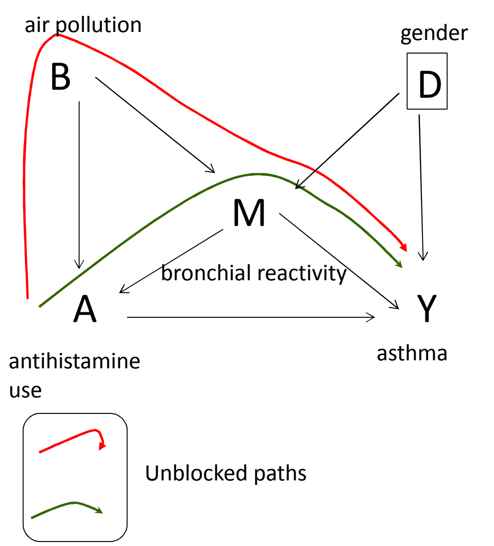


Design: measuring B alone is not enough to control confounding bias

Analysis:
Standardization/ipw
by B alone yields
biased estimates of
Avg causal effects in
the entire population.

Stratification within levels of B
Crude rates differences or ratios
are not the same as causal rate
differences or ratios

Confounding bias: which variables should we adjust for?

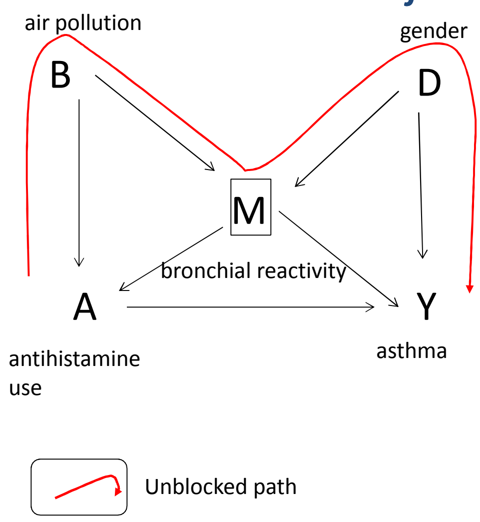


Design: measuring D alone is not enough to control confounding bias

Analysis:
Standardization/ipw
by D alone yields
biased estimates of
Avg causal effects in
the entire population.

Stratification within levels of D
Crude rates differences or ratios
are not the same as causal rate
differences or ratios

Confounding bias: which variables should we adjust for?

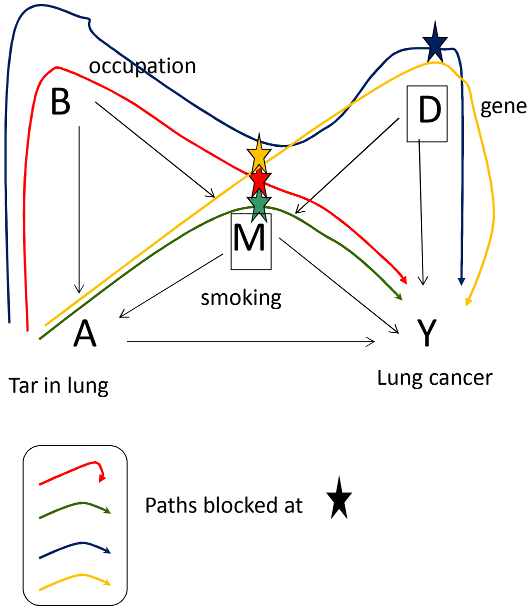


Design: measuring M alone is not enough to control confounding bias

Analysis:
Standardization/ipw by M alone yields biased estimates of Avg causal effects in the entire population.

Stratification within levels of M
Crude rates differences or ratios are not the same as causal rate differences or ratios

Confounding bias: which variables should we adjust for?

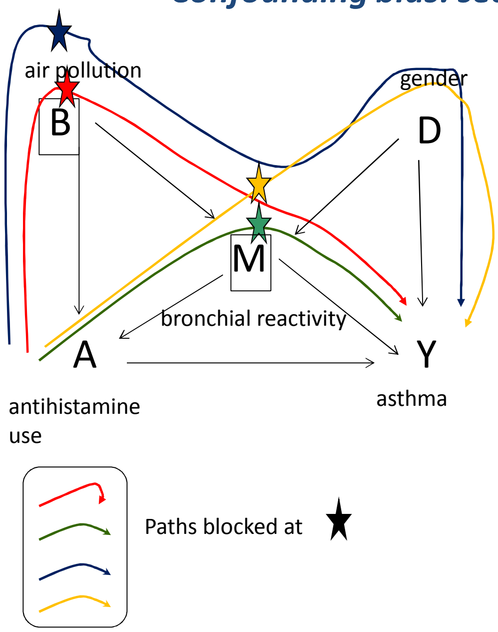


Design: measuring M and D enough to control confounding bias

Analysis:
Standardization/ipw by M and D yields valid estimates of Avg causal effects in the entire population.

Stratification within cross-classified levels of (M,D)
Crude rates differences or ratios are the same as causal rate differences or ratios

Confounding bias: second example

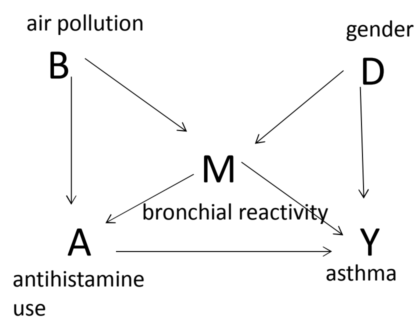


Design: measuring M and B enough to control confounding bias

Analysis:
Standardization/ipw by M and B yields valid estimates of Avg causal effects in the entire population.

Stratification within cross-classified levels of (M,B)
Crude rates differences or ratios are the same as causal rate differences or ratios

Confounding bias: Conclusions



- To control for confounding bias need to measure either (B,M) or (D,M)
- No need to measure all three B,M,D unless you want to estimate causal effect conditional on all three

- {B,M} are **minimal sufficient variables to control for confounding.**
- {D,M} are **minimal sufficient variables to control for confounding.**
- **{B,D,M} are sufficient but not minimal suff to control for confounding**

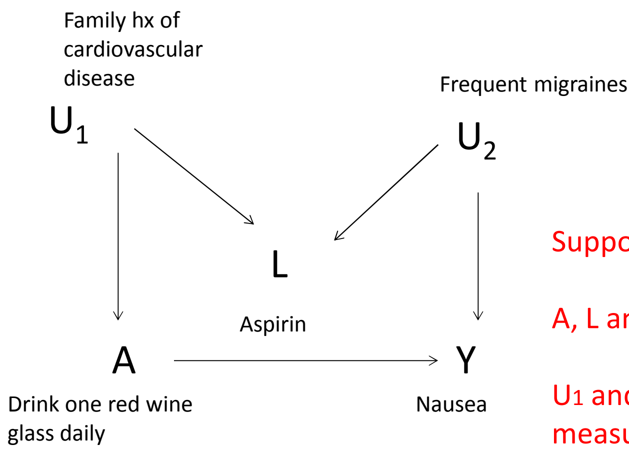
Confounding

- Confounding
  - Definition
  - How to check for confounding bias
  - How to adjust to prevent confounding bias
    - Which variables suffice to adjust for in order to prevent confounding bias?
    - Which variables we cannot adjust for because we will create artificial confounding bias?
- Confounders
  - Why are usual definitions wrong?
  - What is a sufficient set of confounders?

Which L's should I include in my analysis?

- We have seen that to prevent confounding bias you must measure and adjust (either stratify on or standardize by) a set of variables L that blocks all back door paths between exposure and outcome.
- Most standard books call these L's "confounders" and give you rules for when to include them in your analysis
- We will now see that following these rules unblindly is wrong because
  - Sometimes they lead you to collect more covariates than needed, and thus to a costly study, and/or
  - Sometimes they actually lead you to artificially create confounding bias in your analysis when in fact no confounding bias existed!!!

M-structures

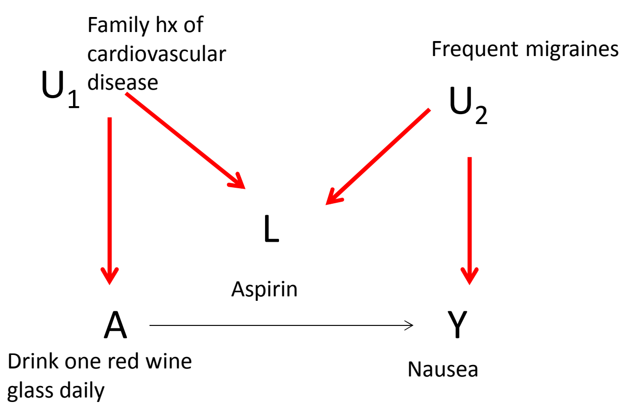


Suppose

A, L and Y are measured

U_1 and U_2 are not measured

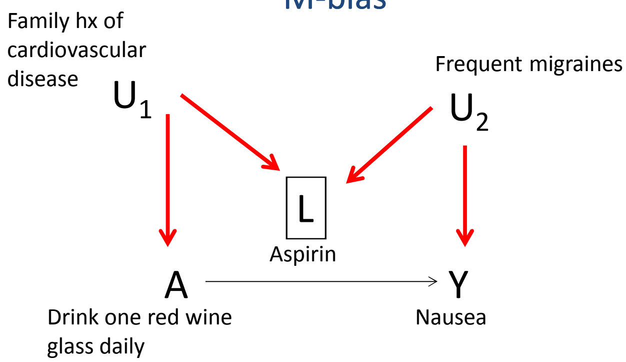
M-structures



Only back door path between A and Y is **blocked** because it goes through the collider L

No confounding bias for the effect of A on Y : Crude A-Y association measures are **equal to** causal effect measures

M-bias

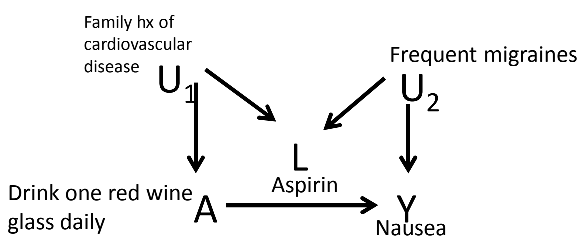


Conditional A-Y association given L is **not causal** .
Back-door path between A and Y is opened if we condition on L.

Artificial confounding by analysis: by adjusting on L we have artificially opened a back door path and created confounding

Standardized/ipw adjusting for L yields **invalid** marginal causal effects
Stratification on L yields **invalid** conditional, i.e. stratum specific, causal effects

M-bias

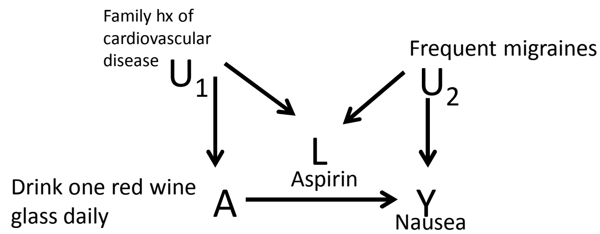


But L meets all three conditions that some authors use to define "confounder":

1. L associated with A (path A-U_1-L is unblocked)
2. L is associated with Y conditional on A (path Y-U_2-L unblocked when we condition on A)
3. L not an effect of A or of Y

So usual confounder rule would have lead us to the wrong analysis!!!

M-bias

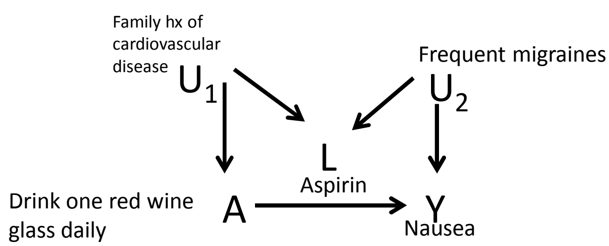


But L meets all three conditions that some authors use to define "confounder":

1. L associated with A (path A-U_1-L is unblocked)
2. L is associated with Y conditional on A (path Y-U_2-L unblocked when we condition on A)
3. L not an effect of A or of Y

So usual confounder rule would have lead us to the wrong analysis!!!

Problems with usual definitions of confounder

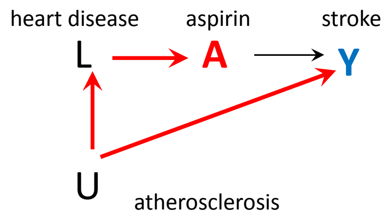


Some authors change condition 2 to and define L as a "confounder" if it satisfies:

1. L associated with A (path A-U_1-L is unblocked)
2. L is an "independent" cause of Y
3. L not an effect of A or of Y

L fails condition 2, so according to this definition we would not adjust for L, however,…

Problems with the usual definitions of confounders



A : exposure Y: outcome
L = measured
U = unmeasured

Some authors change condition 2 to and define L as a "confounder" if it satisfies:

1. L associated with A (path A-U_1-L is unblocked)
2. L is a "independent" cause of Y
3. L not an effect of A or of Y

L fails condition 2, so this definition would have incorrectly told us not to adjust for L!!!

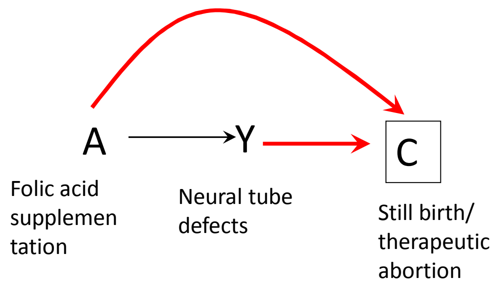
Problems with definitions of confounders based on change in estimated parameters

- Data adapted from Slone epidemiology birth defects study

	C=1		C=0	
	Y=1	Y=0	Y=1	Y=0
A=1	19	8	24	231
A=0	100	46	94	658

- Y: mother of infant with(1) /without(0) neural tube defects
- A: presence(1)/absence(0) of daily supplementation with folic acid during first 2 months after last menstrual period
- C: stillbirth or therapeutic abortion (1/0)
- **Goal:** compute causal odds ratio of A on Y.
- Assume A was randomized.
- **Analysis question**: which estimator is right? Odds Ratio of standardized/ipw rates adjusting for C? or Odds Ratio of crude rates?

- C a common consequence, not a common cause, of A and Y.



Analysis
- *Standardization/ipw* adjustment by C wrong!.
- *Stratified* analysis by C wrong!
- *Crude rates* ok.
- In our data crude OR = 0.65 95% CI (0.45; 0.94)

Problems with definitions of confounders based on change in estimated parameters

- Let's analyze what some of the common rules would have told us to do …
- **Rule 1**: variable selection
  1. Fit a (logistic) regression model of outcome Y with A and C as covariates.
  2. Keep C in the model if p value, say < 0.05, (or 0.010)
  3. Report exp(coefficient of A) as the causal odds ratio estimate
- In step (2) p < 0.001
- In step (3) exp(coeff of A) = **0.8** (95% CI (0.53; 1.20))
- Compare this with correct analysis
  crude odds ratio = **0.65** (95% CI (0.45; 0.94))
- **Conclusion:** Don't follow blindly rule 1!

Problems with definitions of confounders based on change in estimated parameters

- **Rule 2**: relative change in parameter estimators
  - leave C in logistic regression model if relative change in estimates before versus after including C in your logistic regression model is greater than 10%
- In our data
  - OR in model that included C was 0.8,
  - OR in model that did not include C was 0.65
- relative change is

  $$(0.8 - 0.65) / 0.65 = 0.23$$

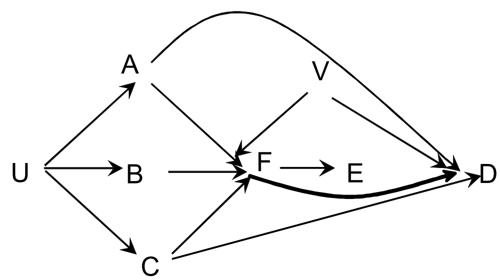
- This rule again tell us to report the adjusted (wrong) odds ratio estimate of 0.8
- Conclusion: don't follow blindly rule 2!

Preventing confounding bias: summary

- **Confounding bias**: spurious association that arises from open back door paths between exposure and outcome
- **Confounding bias prevention strategy**:
  - **Design**:
    - Draw causal dag
    - Find a subset of variables sufficient to block all back-door paths (there may be many subsets, choose one)
  - **Analysis**:
    - Decide whether you care about conditional effects or marginal effects. Choose your analytic method consequently.

Exercise

Suppose you are interested in estimating the **causal effect of C on D**. If you stratify by the cross-classification of the indicated variables, will the difference of the stratified crude means be equal to the difference of the stratified counterfactual means?



	Stratify by	Association = causation?	
		YES	NO
1	A		
2	B		
3	A & B		
4	A & F		
5	A & B & F		
6	A & B & F & V		
7	U		