



Confounder selection principles:

Epidemiological and statistical considerations

ehsan.karim@ubc.ca



🗨️ When poll is active, respond at Pollev.com/ehsank878

📱 Text **EHSANK878** to **22333** once to join

What was the most difficult part of week 1

Following video materials

Using wall of confusion

Coming to office hour

Quiz questions

Concept questions

Lab exercises

Thinking about final project

2nd wave of pandemic is coming?



When poll is active, respond at Pollev.com/ehsank878

Text **EHSANK878** to **22333** once to join

Have you selected your final project data source yet?

Contemplating multiple available options

CCHS

NHANES

PopData

Other sources, where I have an ethics certificate, or I am on it

Other open sources

No idea yet

What final project?



When poll is active, respond at Pollev.com/ehsank878

Text **EHSANK878** to **22333** once to join

If you had a question in week 1, how did you get answer to that (in general)?

I read the suggested text/online materials, and found the answer.

I had a question, but did not ask

wall of confusion

office hour

Emailed the instructor directly

Communicating with classmates

Communicating with mentors outside of this class

I asked the question, but did not get answer from instructor/TA

I didn't have a question: things were pretty straightforward



Ref

[HTML] Principles of confounder selection

[TJ VanderWeele](#) - *European journal of epidemiology*, 2019 - Springer

Selecting an appropriate set of confounders for which to control is critical for reliable causal inference. Recent theoretical and methodological developments have helped clarify a number of principles of confounder selection. When complete knowledge of a causal ...

☆ [🔗](#) Cited by 10 [Related articles](#) [All 4 versions](#) [Import into BibTeX](#)

- will breakdown this paper.
- will re-use some of the graphs as well.

Content / parts

1. Counterfactual framework
2. directed acyclic graph (DAG)
3. Identifying confounders: 4 empirical criteria
4. Identifying confounders: Modelling criteria

Part 1

Counterfactual framework

A basic framework to define
causal effect

Notations

A: Exposure status

Y: Outcome

L: measured variable (Confounder)

U: unmeasured variable



Notations



A: Exposure status

1 = takes Rosuvastatin

0 = does not take rosuvastatin



Y: Outcome: Total cholesterol levels

- $Y(A=1)$ = potential outcome when exposed
- $Y(A=0)$ = potential outcome when not exposed
- $Y|A=1$ = observed outcome when exposed
- $Y|A=0$ = observed outcome when not exposed

Demographic	Total Cholesterol
Age 19 or younger	Less than 170 mg/dL
Men age 20 or older	125 to 200 mg/dL
Women age 20 or older	125 to 200 mg/dL

Notations: our interests

When assessing the effect of an exposure on an outcome, we are interested about the following estimands

1. treatment effect for an individual (TE)
2. average treatment effect (ATE)

Notations: TE

Counterfactual!

Scenario 1

- John takes Rosuvastatin ($A=1$) and his total cholesterol level is = $Y(A=1) = 195$ mg/dL (milligrams per deciliter) after 3 months

Scenario 2

- John does not take Rosuvastatin ($A=0$) and his total cholesterol level is = $Y(A=0) = 245$ mg/dL after 3 months

- Effect of Rosuvastatin on John is =

$$TE = Y(A=1) - Y(A=0) = 195 - 245 = - 50$$

Notations: ATE

Counterfactual!

Person	$Y(A=1)$	$Y(A=0)$	TE
John	195	245	- 50
Jim	100	160	- 60
Jake	210	270	- 60
Cody	155	210	- 55
Luke	165	230	- 65

$$ATE = E[Y(A=1) - Y(A=0)] = -(50+60+60+55+65)/5 = - 58$$

Real-world Problem

Observed!

Person	$Y(A=1)$	$Y(A=0)$	TE
John	195		- ?
Jim		160	- ?
Jake		270	- ?
Cody	155		- ?
Luke		230	- ?

$$ATE = E[Y(A=1) - Y(A=0)] = (?+?+?+?+?)/5 = ?$$

Real-world Problem

Outcomes under both treatments

- $Y(A=1)$ and
- $Y(A=0)$

are not possible to measure for the same subject (at the same time/condition).

Therefore, estimating TE for each person not possible.

Real-world Problem

Observed!

Person	$Y(A=1)$	$Y(A=0)$	TE
John	195		- ?
Jim		160	- ?
Jake		270	- ?
Cody	155		- ?
Luke		230	- ?

$$\text{Diff} = E[Y|A=1] - E[Y|A=0] = \frac{(195+155)}{2} - \frac{(160+270+230)}{3} = -45 \text{ (not } -58)$$

Real-world Solution

Mean outcomes under both treatments

- $E[Y|A=1]$, where $A=1$ is the treated group
- $E[Y|A=0]$, where $A=0$ is the control group

are possible to measure for 2 groups (treated and control groups, who are comparable / exchangeable / ignorable tx assignment through **randomization/RCT** [adequate n?]).

Therefore, estimating $ATE = E[Y|A=1] - E[Y|A=0]$ is possible in an RCT (no systematic difference in groups).

Real-world Solution in observational setting

In absence of randomization,

- $E[Y|A=1] - E[Y|A=0]$

Includes

1. Treatment effect
2. Systematic differences in 2 groups ('confounding')
 - Doctors may prescribe tx more to frail/older age patients).
 - In here, **L = age** is a confounder.
3. **Valid causal inference requires addressing component 2.**

Real-world Solution in observational setting

In absence of randomization, **if age is a known issue**

- Causal effect for young

$$E[Y|A=1, L = \text{younger age}] - E[Y|A=0, L = \text{younger age}]$$

- Causal effect for old

$$E[Y|A=1, L = \text{older age}] - E[Y|A=0, L = \text{older age}]$$

- **Conditional exchangeability**; only works if L is measured

Part 2

directed acyclic graph

(DAG)

A tool to identify “confounding”
(much more than a confounder)

Graphical models

- Wright, S. (1921). Correlation and causation. *J. agric. Res.*, 20, 557-580. ([link](#))
- Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 37-48. ([link](#))

DAG

[Graphical presentation of confounding in directed acyclic graphs](#)

MM Suttorp, B Siegerink, KJ Jager... - Nephrology Dialysis ..., 2014 - academic.oup.com

Since confounding obscures the real effect of the exposure, it is important to adequately address confounding for making valid causal inferences from observational data. Directed acyclic graphs (DAGs) are visual representations of causal assumptions that are increasingly used in modern epidemiology. They can help to identify the presence of confounding for the causal question at hand. This structured approach serves as a visual aid in the scientific discussion by making underlying relations explicit. This article explains the ...


☆ 🔖 Cited by 42 Related articles All 7 versions Import into BibTeX


“the arrows represent the direction of the causal relationship”

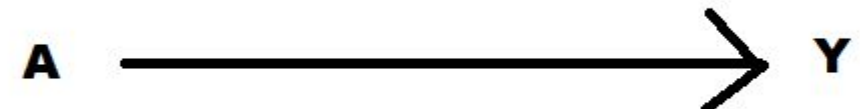
“**Directed:** the factors in the graph are connected with arrows”

“**Acyclic:** no directed path can form a closed loop, as a factor cannot cause itself”

DAG

- Not a DAG/bidirectional (path)


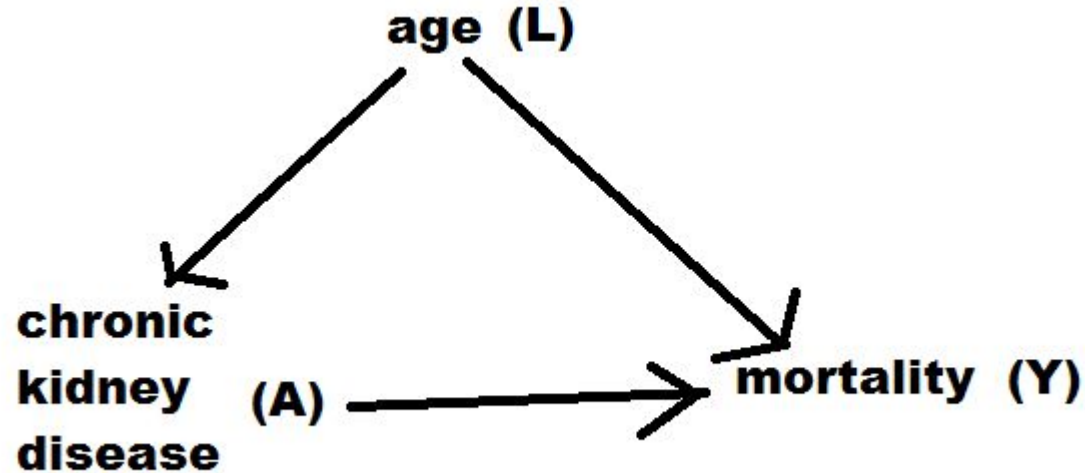
A ←→ Y
- A does not cause Y (no arrow/ absence of arrow means no relationship)


A Y
- A causes Y (directional/ DAG/ no loop/ A is a parent (or cause or ancestor) of Y/ Y is and child (or effect or descendent) of A/ cause could mean small or large effect)


A → Y

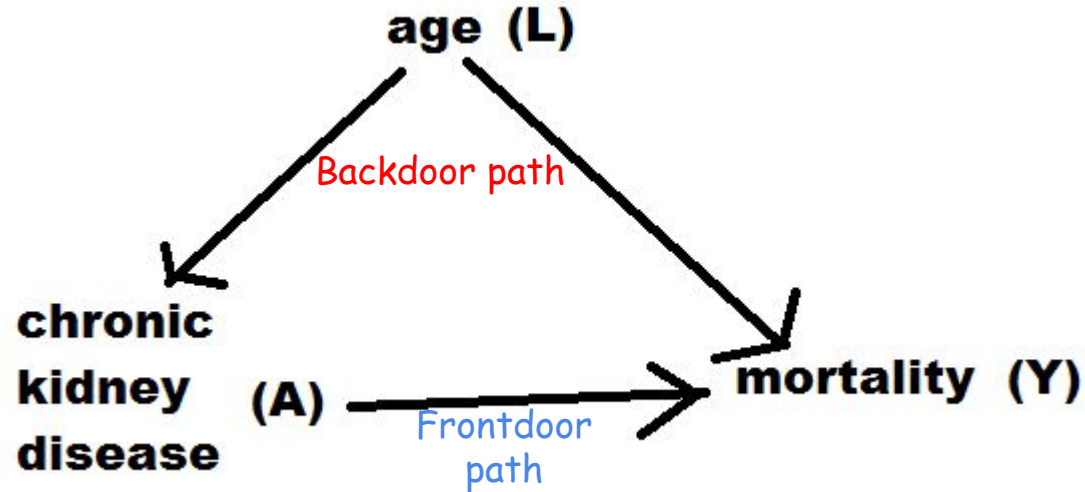
DAG representation:

Confounding in observational setting



DAG representation:

Confounding in observational setting

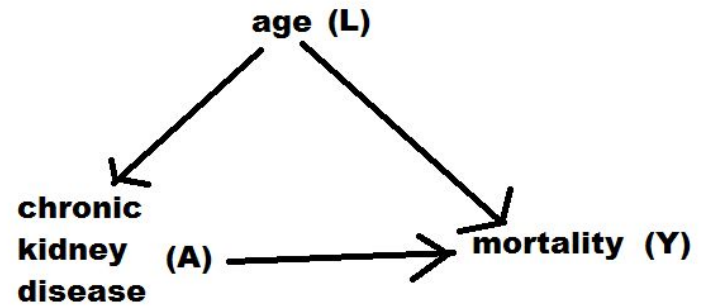


Backdoor path: non-causal path from A to Y (this path includes an arrow into A)

Frontdoor path: causal path from A to Y (this path includes an arrow from A)

What is a confounder?

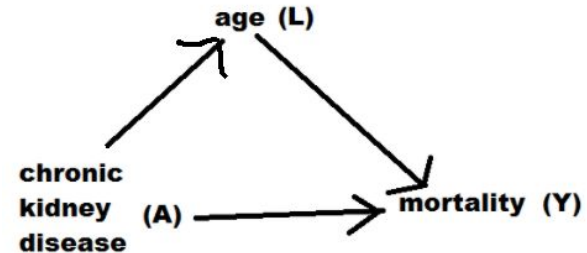
Definition: ([Rothman, 2012](#) Epi textbook, ch 7, page 141)



1. "Must be associated with the disease (either as a cause or a proxy for a cause, but not as an effect of the disease)"
2. "Must be associated with the exposure"

Mediator will also follow the above definition!

3. Not an "effect of the exposure"



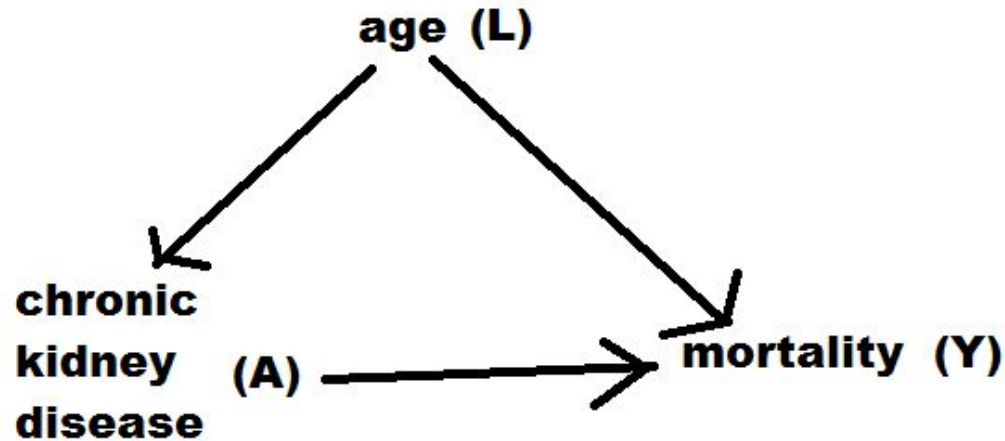
Not in "an intermediate step in the causal pathway from exposure to disease"

* The association with outcome is not just through exposure ("instrument")

DAG representation:

Confounder (L): common cause of A & Y:

control!

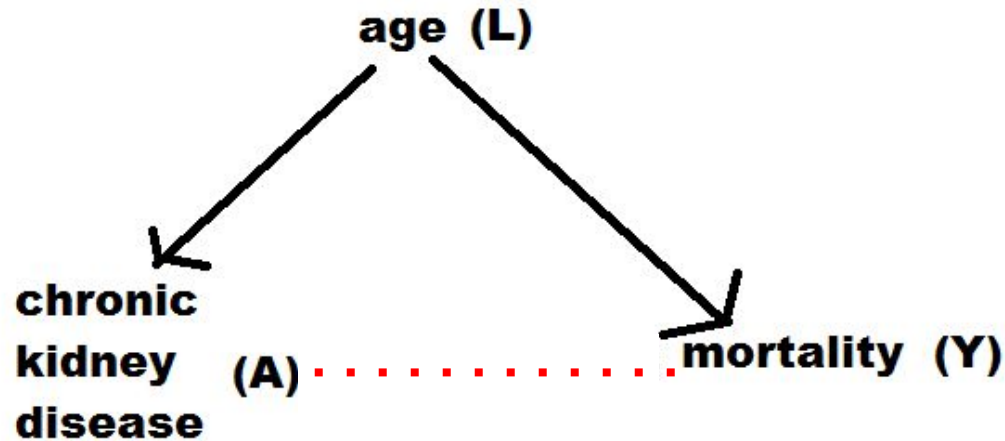


DAG representation:

..... Correlated = open path
Not correlated = blocked path

Confounder (L): common cause of A & Y:

control!



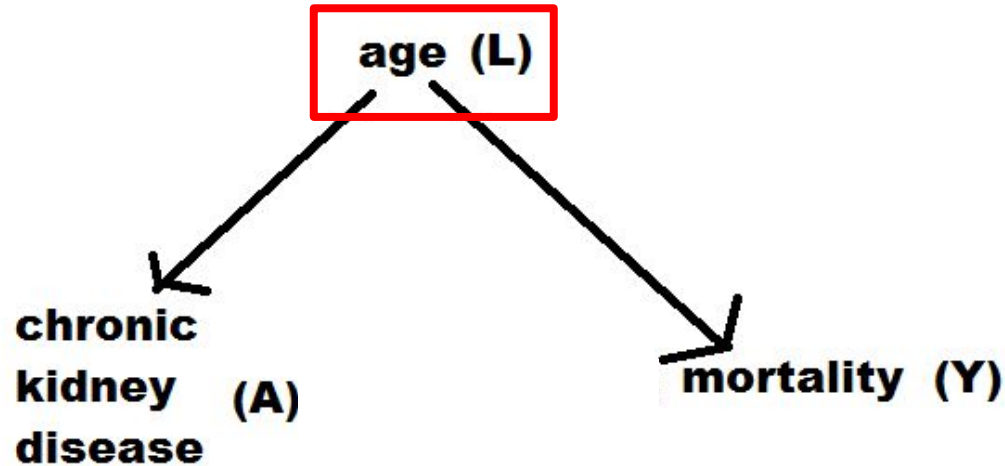
A and Y would be correlated (open path) even if the arrow between A and Y is deleted. Why? This would be an example of non-causal/biased relationship (correlation exists, but not causal).

DAG representation:

..... Correlated = open path
Not correlated = blocked path

Confounder (L): common cause of A & Y:

control!



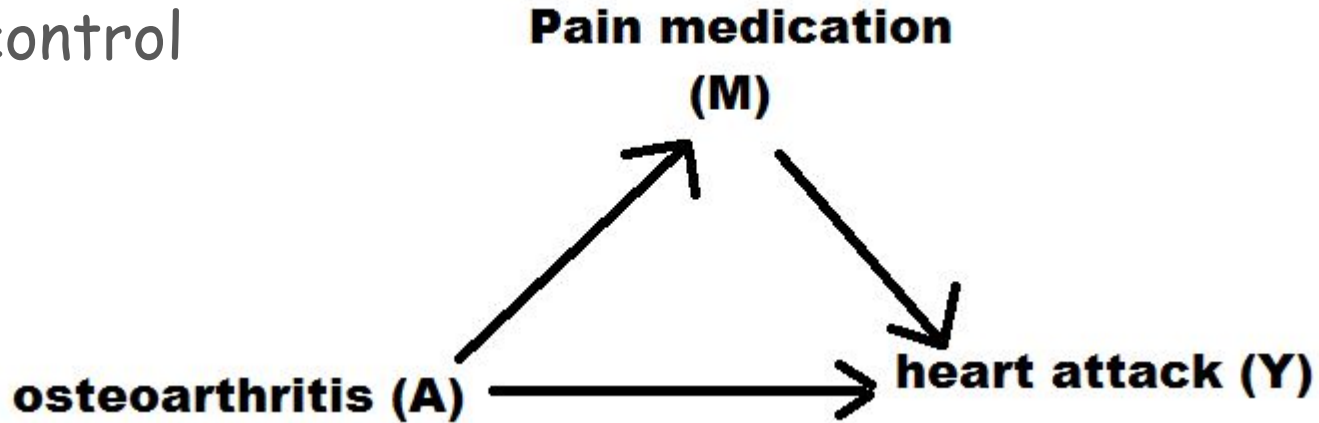
Adjusting for L would result in no correlation between A & Y (assuming enough sample size).
If we are only dealing with L = 20, irrespective of the status of A (0 or 1), will A & Y be correlated?

DAG representation:

..... Correlated = open path
Not correlated = blocked path

Mediator (in the causal pathway between A & Y):

do not control

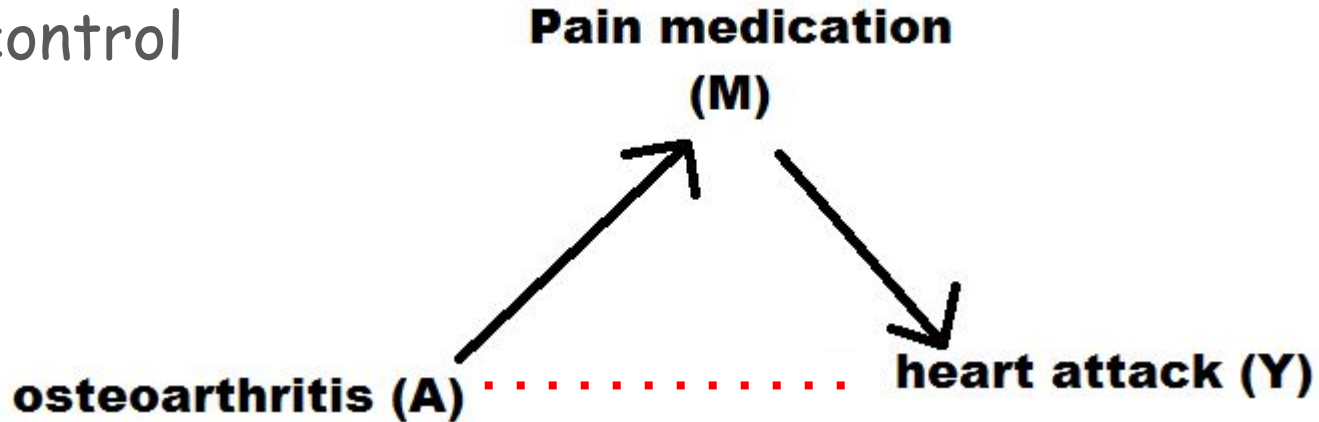


DAG representation:

..... Correlated = open path
Not correlated = blocked path

Mediator (in the causal pathway between A & Y):

do not control



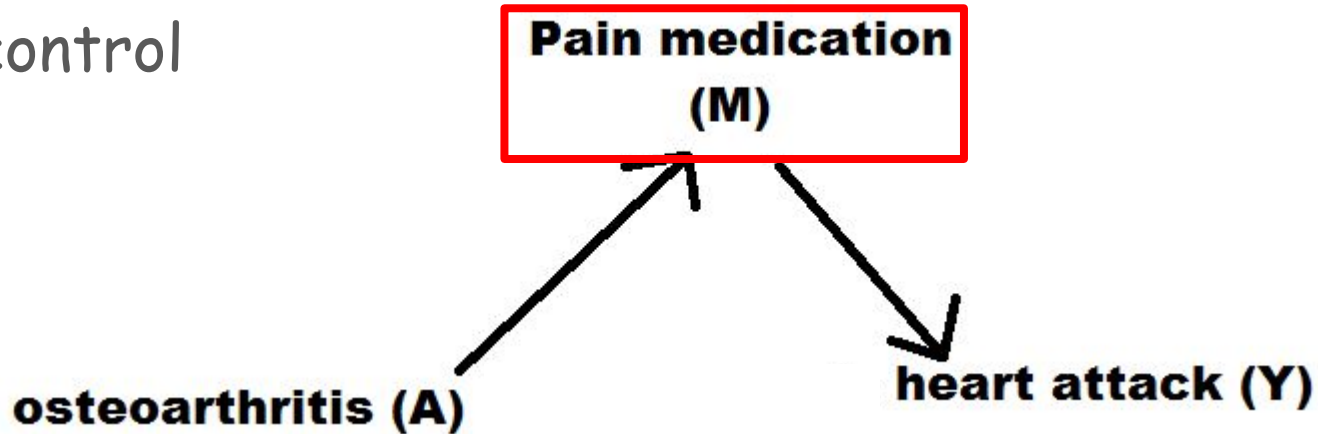
A and Y would be correlated (open path) even if the direct arrow between A and Y is deleted. Why?

DAG representation:

..... Correlated = open path
Not correlated = blocked path

Mediator (in the causal pathway between A & Y):

do not control

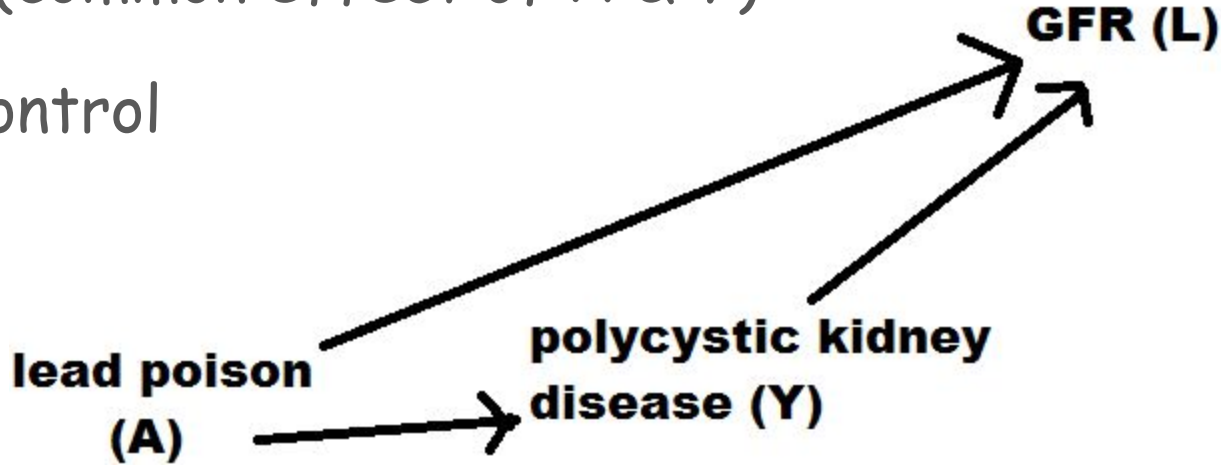


Adjusting for M would block the path, i.e., relationship between A & Y (there will not be any correlation). If we are only dealing with $M = 0$, irrespective of the status of A (0 or 1), will A & Y be correlated?

DAG representation:

Collider (common effect of A & Y):

do not control



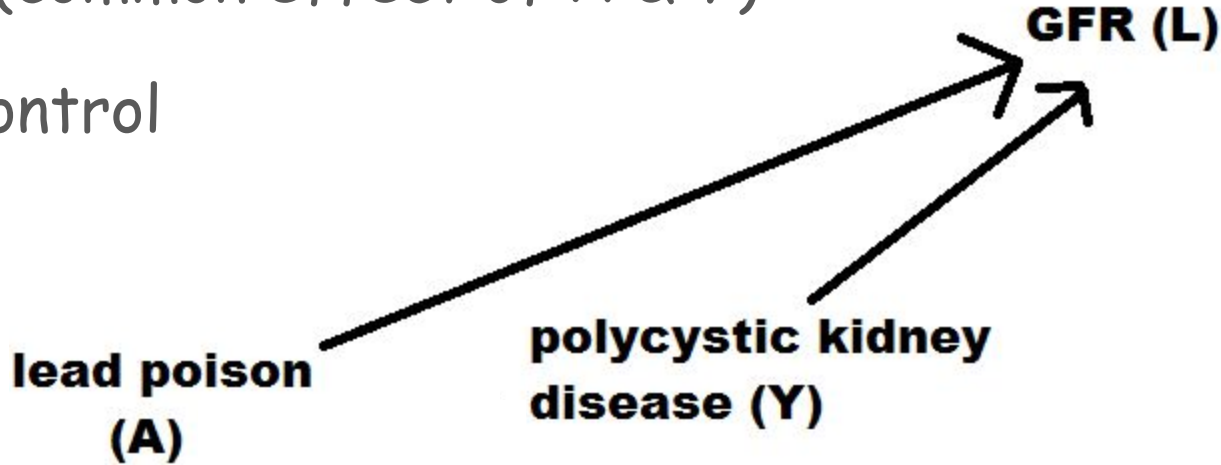
..... Correlated = open path
Not correlated = blocked path

DAG representation:

..... Correlated = open path
Not correlated = blocked path

Collider (common effect of A & Y):

do not control

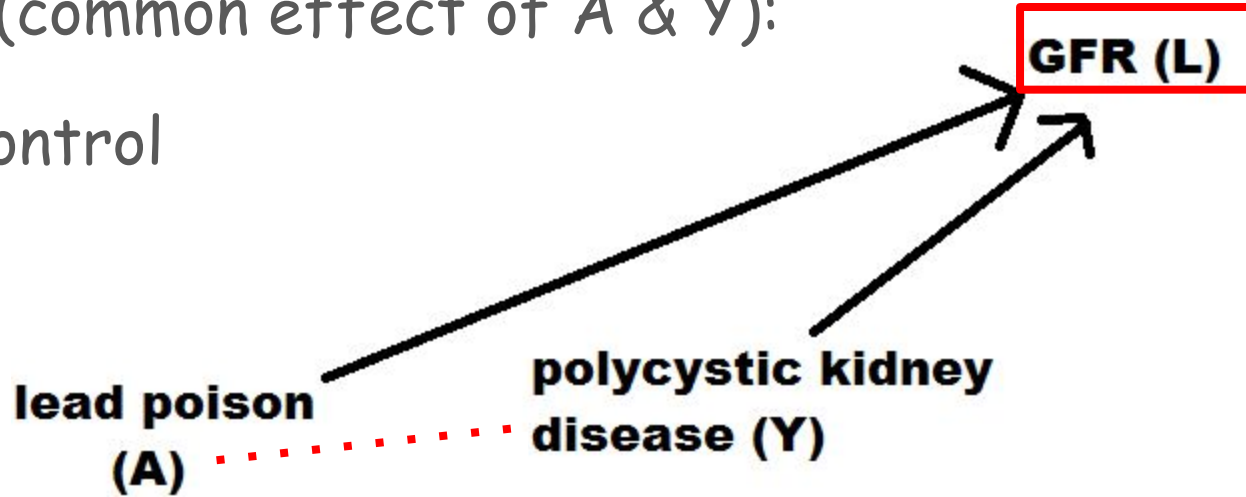


A and Y would not be correlated (blocked path) if the direct arrow between A and Y is deleted. Why? (this is often confusing!!)

DAG representation:

Collider (common effect of A & Y):

do not control

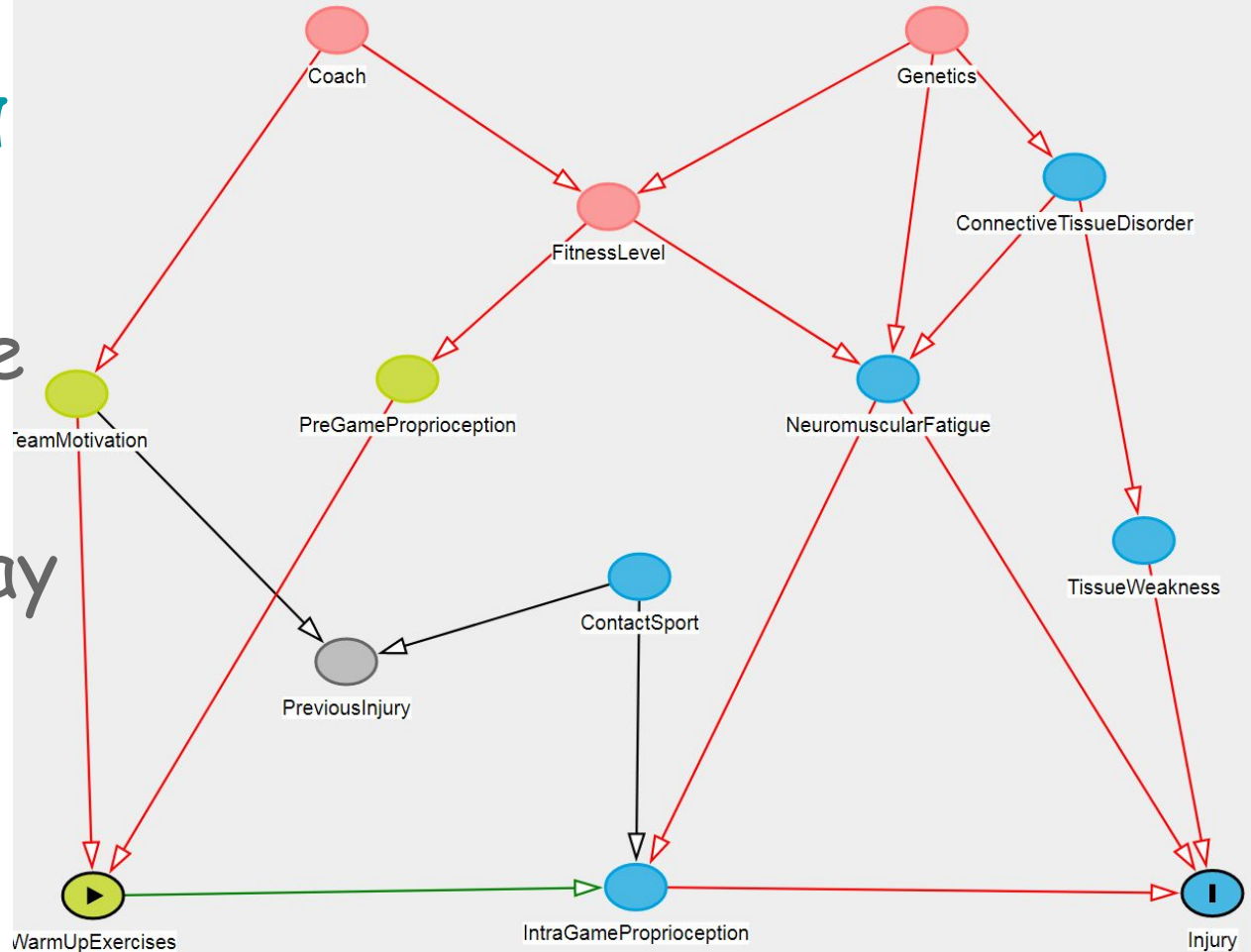


..... Correlated = open path
Not correlated = blocked path

Adjusting for L would open the path, i.e., induce a relationship between A & Y (there will be correlation). If we are only dealing with $L = 0$, irrespective of the status of A (0 or 1), will A & Y be correlated?

Complex DAG

- Could be subjective
- Other expert may not agree

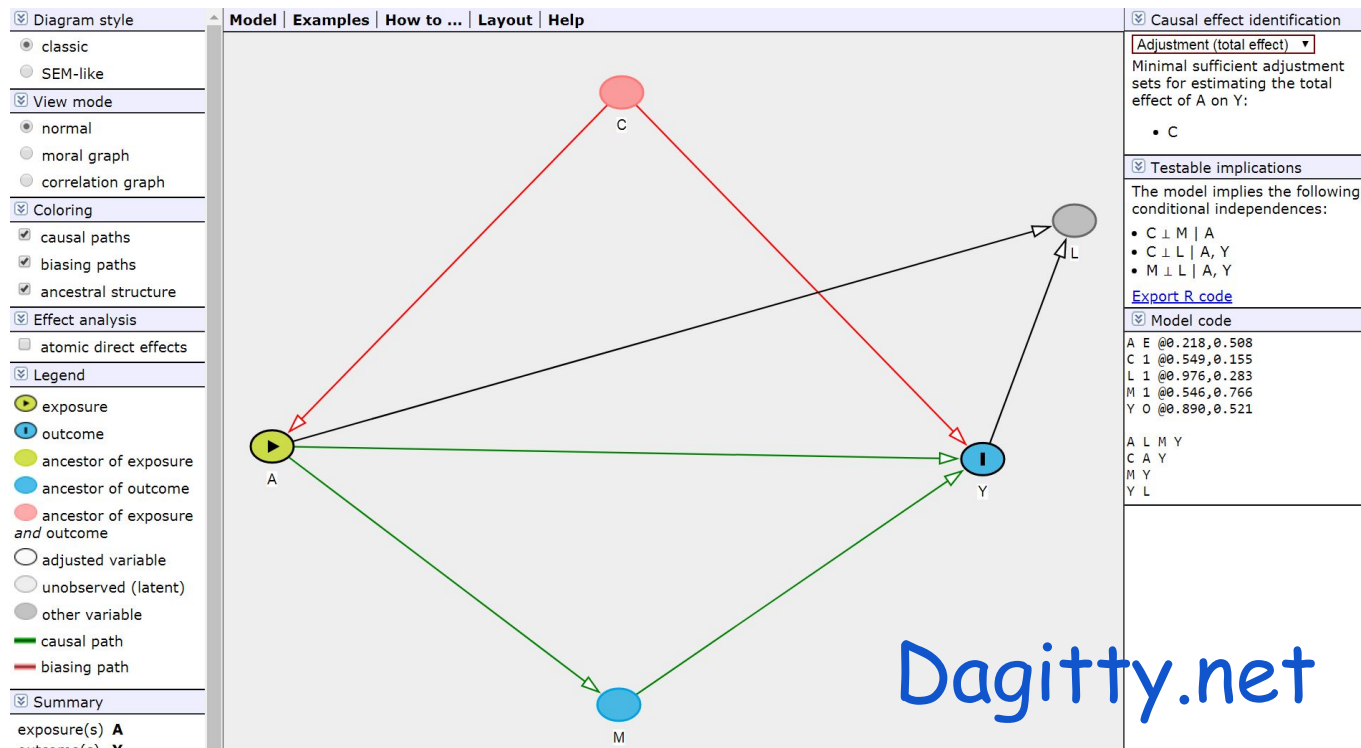


Dagitty

Backdoor path criterion

Red path =
back door

(overly simplified)



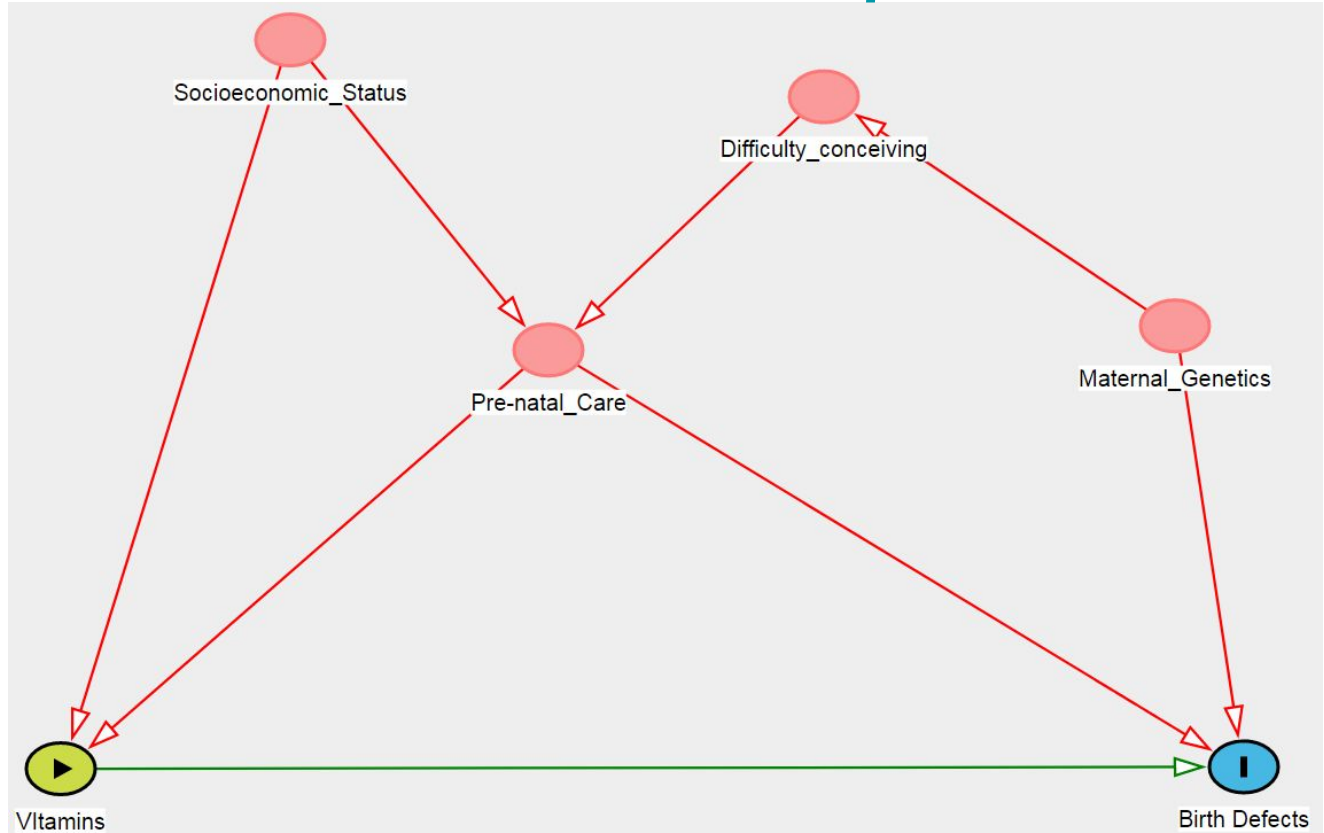
Dagitty.net

To get unbiased estimate, how many to control?

Vittinghoff
textbook
example:

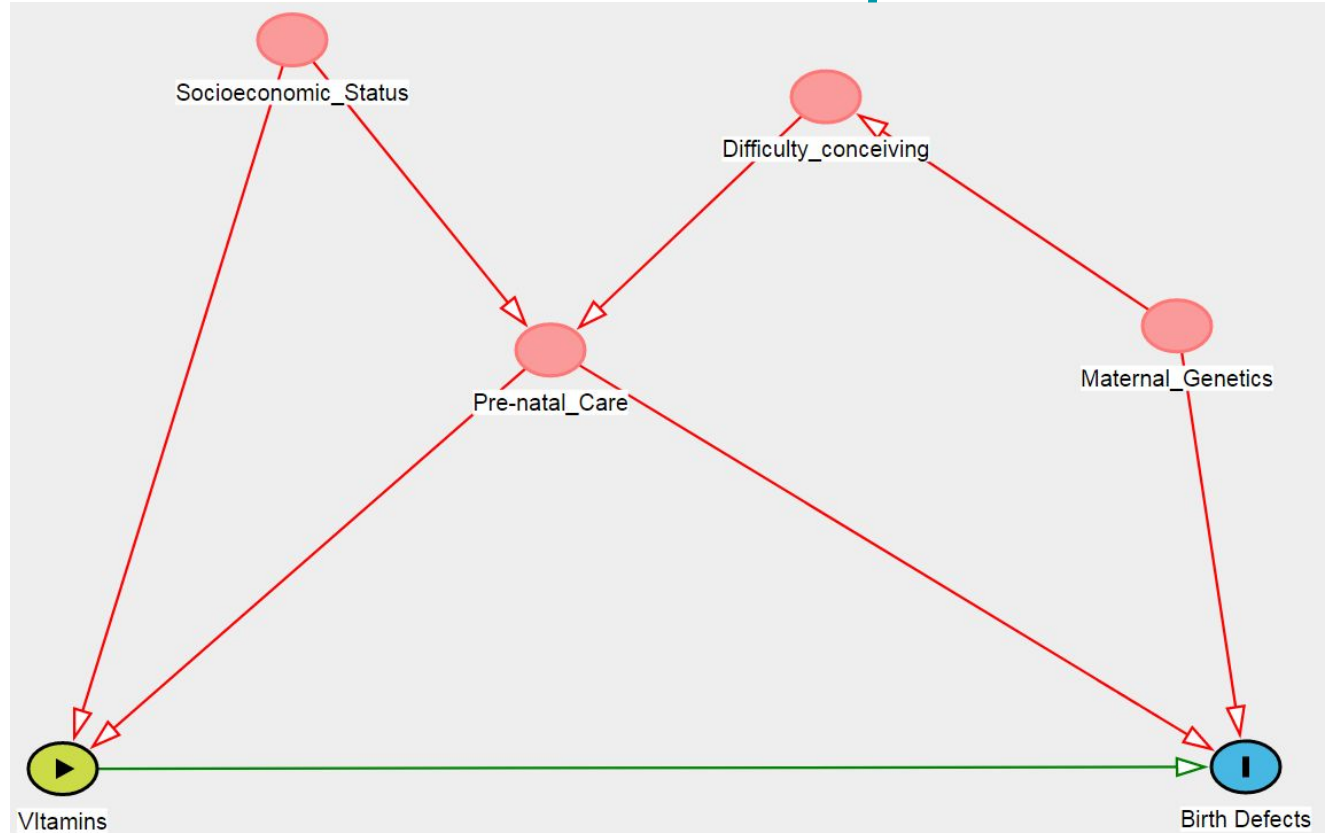
Red node =
unadjusted

Red path =
back door /
biasing path



To get unbiased estimate, how many to control?

Adjusting for **Pre-natal care** enough to get causal effect of vitamins on birth defects?

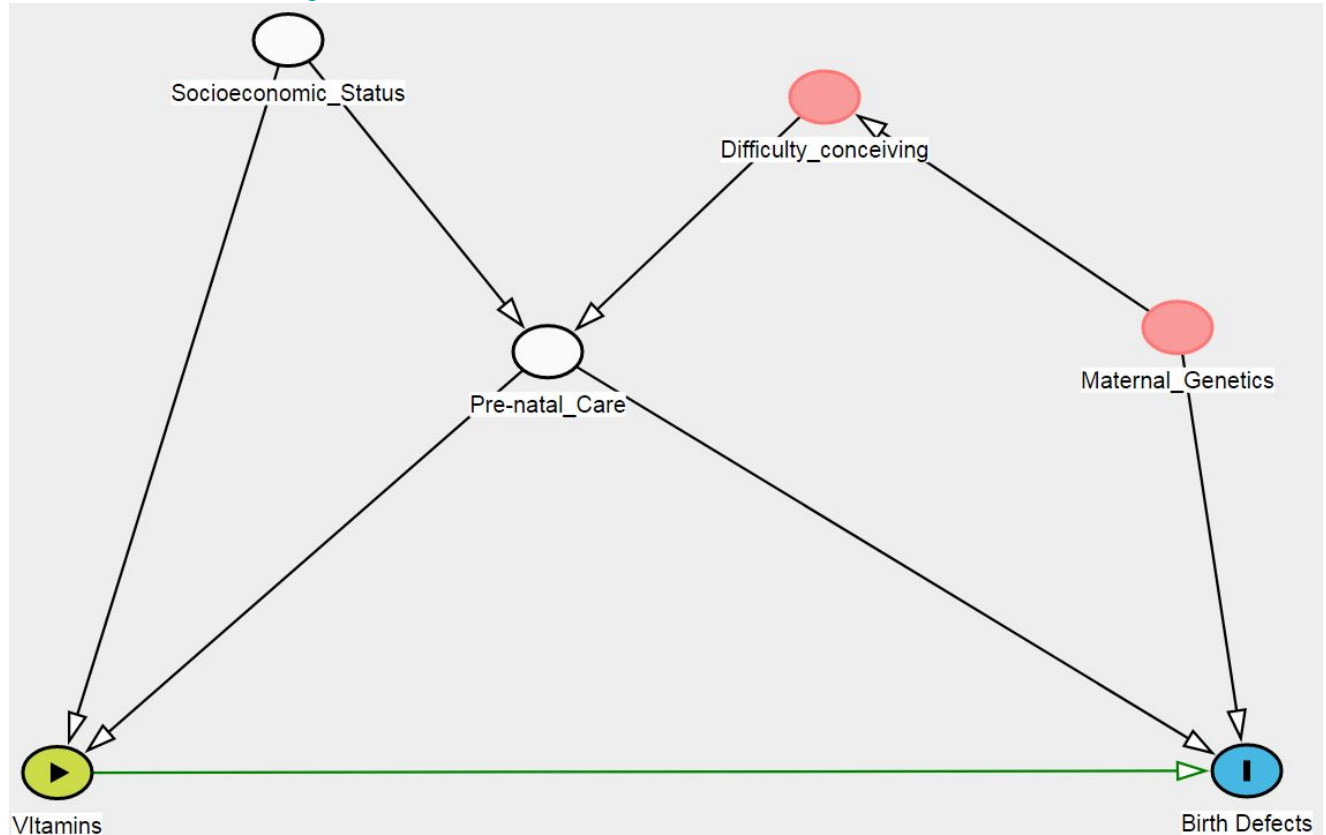


Minimal sufficient adjustment set - 1st set

Vittinghoff
textbook
example:

**white node =
adjusted**

No red path
= no back
door

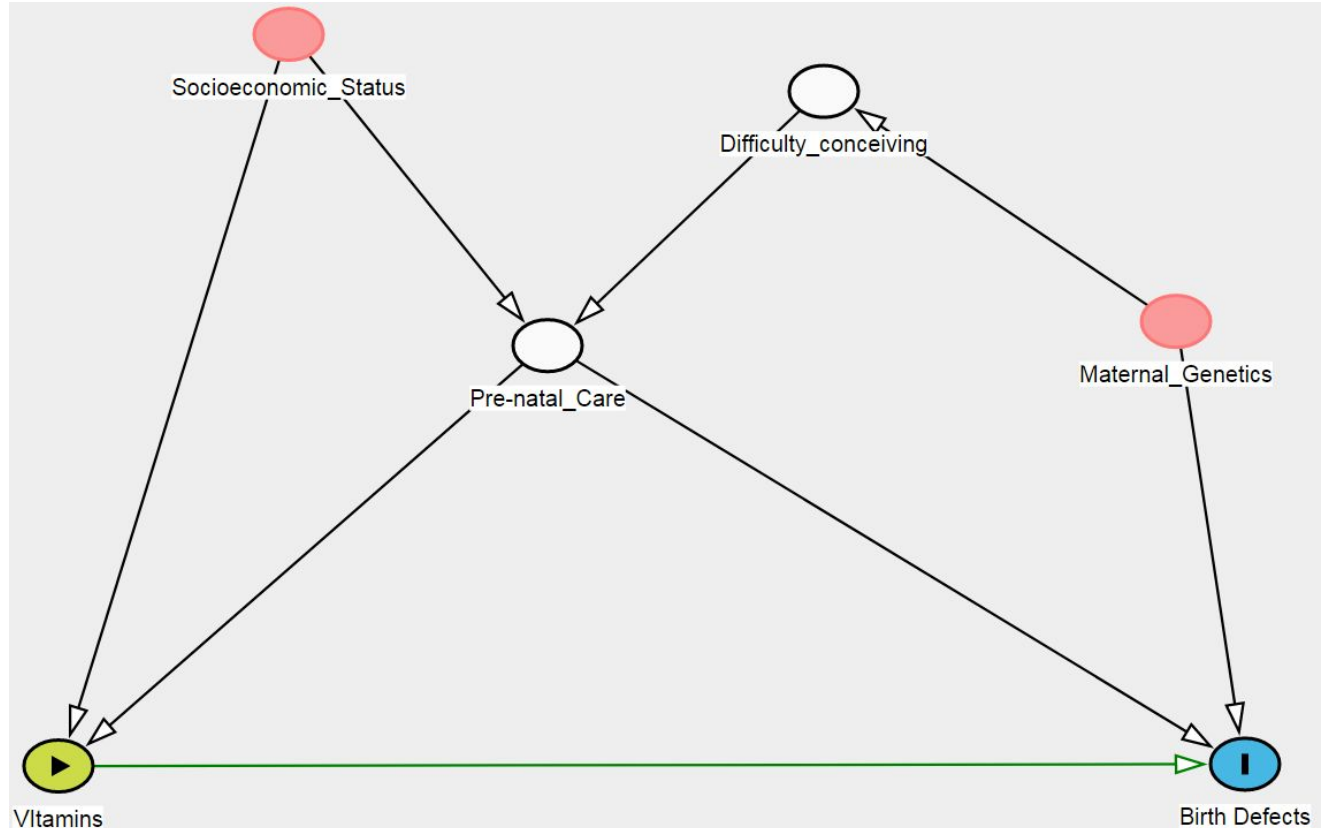


Minimal sufficient adjustment set - 2nd set

Vittinghoff
textbook
example:

white node =
adjusted

No red path
= no back
door

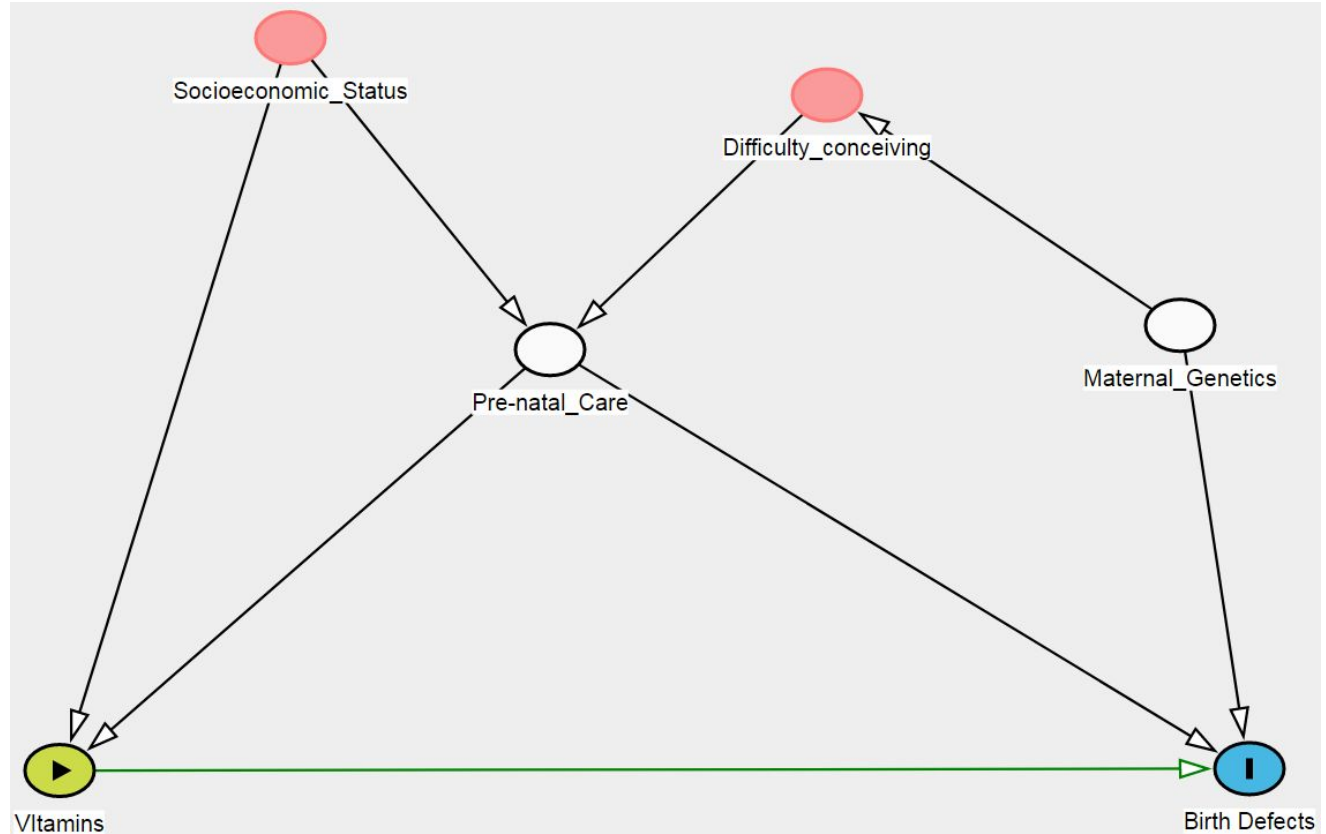


Minimal sufficient adjustment set - 3rd set

Vittinghoff
textbook
example:

white node =
adjusted

No red path
= no back
door



Identifying confounders

- DAGs can be helpful in identifying suitable variables (subject-area knowledge) [[can be subjective!](#)]
- Variables of clinical significance
- Textbook's suggestions are clear, but hard to formalize in 2ndary / observational data.
- More problematic in non-experimental studies, as variables do not come with labels (confounder/IV/Risk Factor/collider).

Part 3

Identifying confounders:

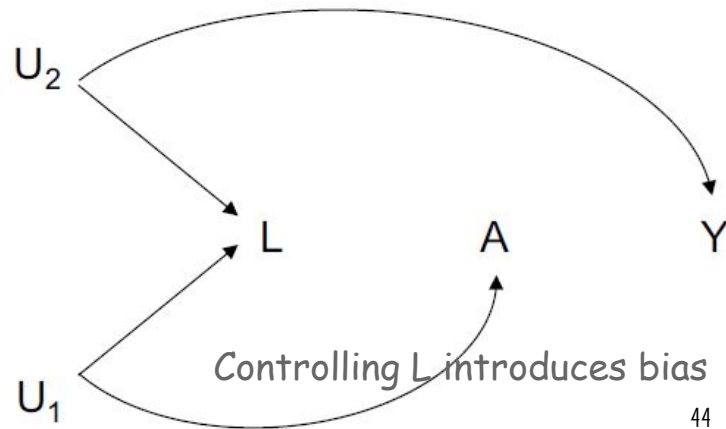
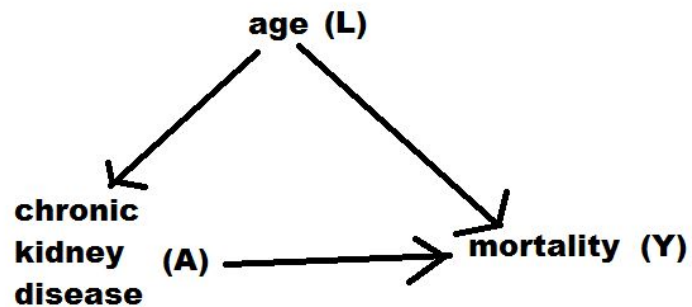
5 empirical criteria

Does not require "full knowledge"
of the DAGs (still requires some)!

Empirical criteria: I

- Pre-treatment criterion
 - Control any variable L prior to the treatment/exposure A
 - Fails due to collider bias or M -bias
 - Can be viewed as too liberal
 - Here u is unmeasured

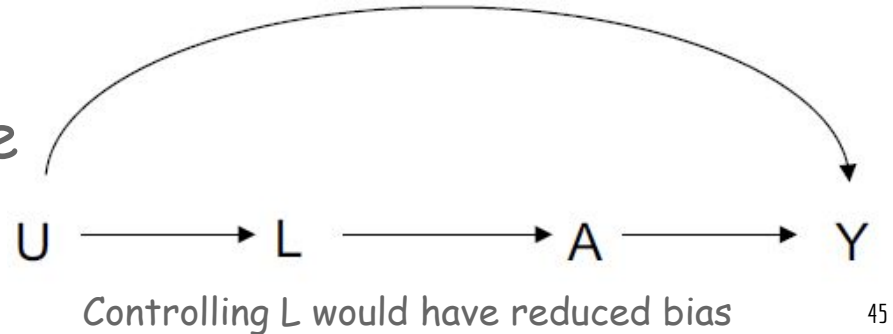
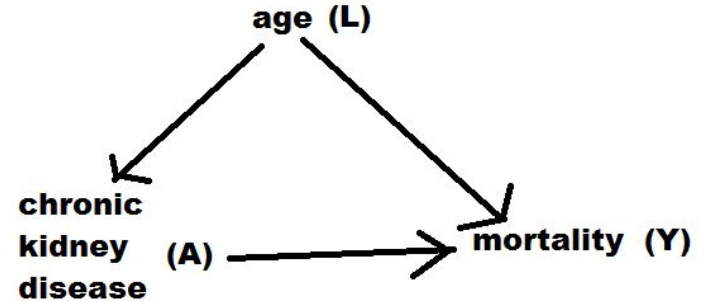
Works / fails?



Empirical criteria: 2

- Common cause criterion
 - Adjust for pre-exposure covariates that are common causes of exposure and outcome
 - Somewhat restrictive/conservative than pre-treatment criterion

Works / fails?

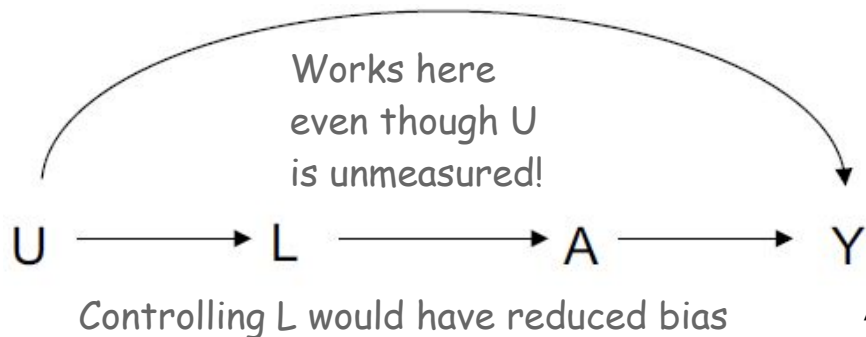
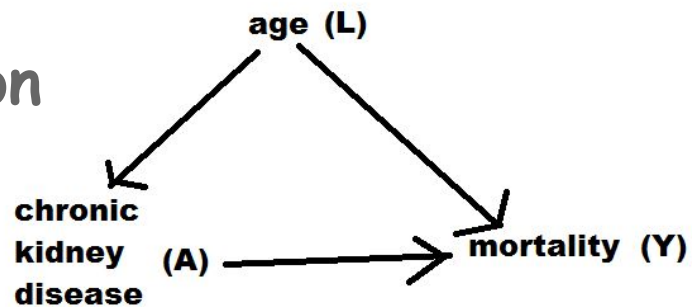


Empirical criteria: 3

- Extended Common cause criterion

- Adjust for any variable L that is
 - i. either a common cause (U here) of the exposure (A) and outcome (Y), or
 - ii. that is on the pathway from such a common cause (U) to exposure (A) or outcome (Y)

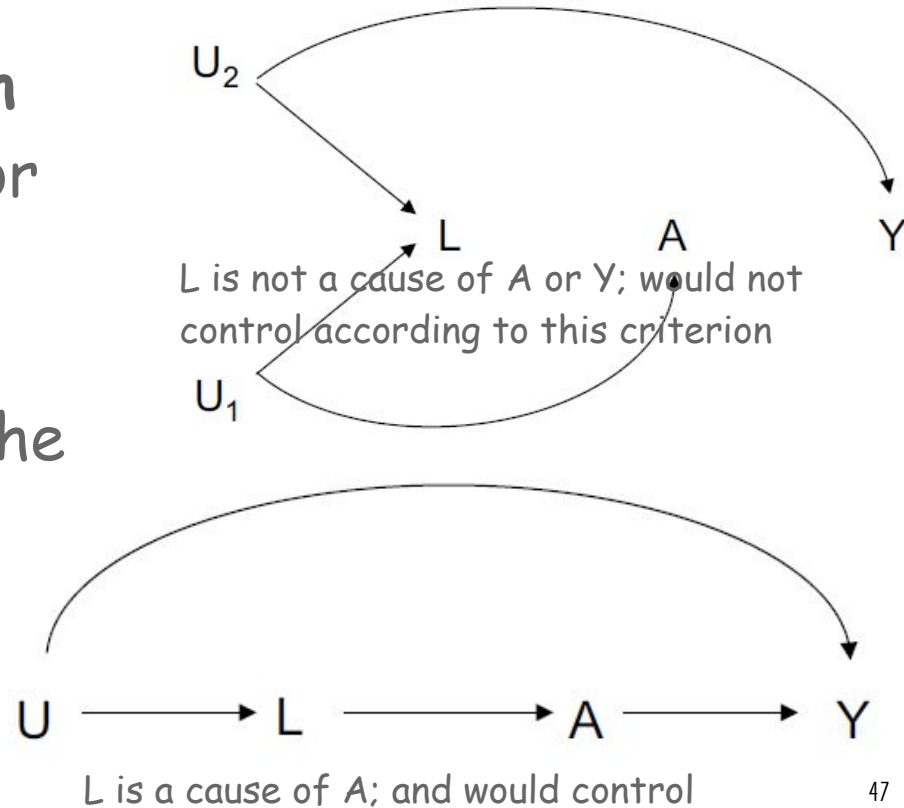
Works / fails?



Empirical criteria: 4

- Disjunctive cause criterion
 - Cause of treatment A, or outcome Y, or both
 - Can be viewed as intermediate between the above 2 criteria

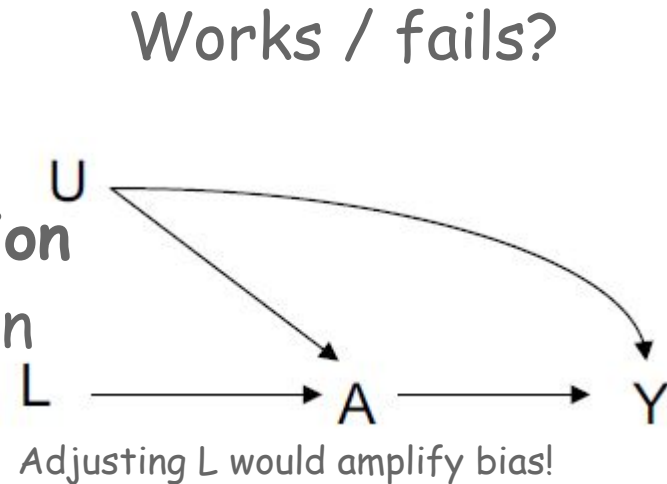
Works / fails?



Empirical criteria: 5

- **Modified Disjunctive cause criterion**

- *All of* Disjunctive cause criterion
- + *Part A*:

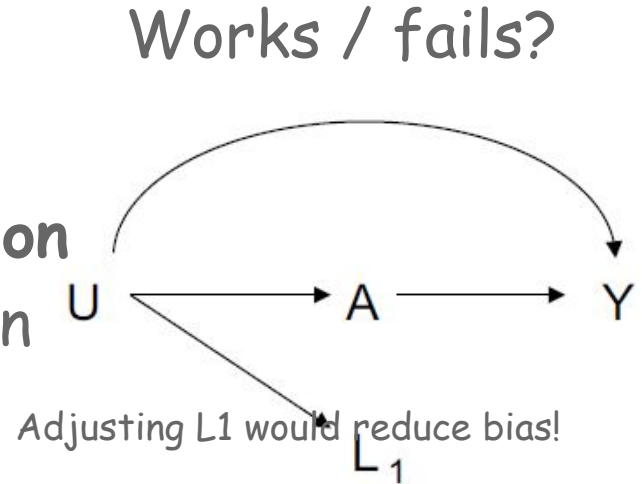


- Avoid adjusting for known instruments L. Controlling for L (in the presence of U) will amplify bias, Z-bias. Theoretically eliminating IV is good, in practice justifying IV is rare (except policy variables).

Empirical criteria: 5

- **Modified Disjunctive cause criterion**

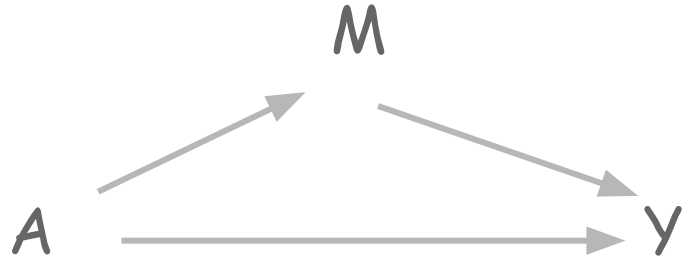
- *All of Disjunctive cause criterion*
- *+ Part A + Part B:*



- Proxy variables are viewed as confounders L with measurement errors. Should reduce bias, but not guaranteed (direction is uncertain). Best to adjust for proxies for L that are common cause of A & Y .

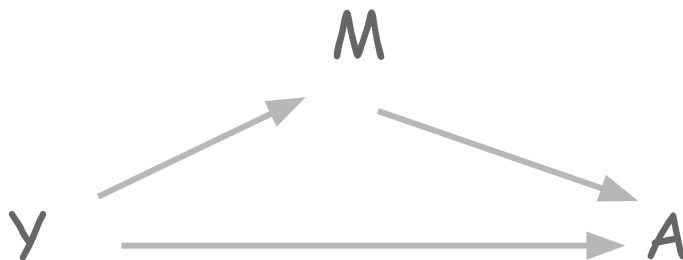
Additional consideration

- **Modified Modified (!)** Disjunctive cause criterion
 - *All of* Disjunctive cause criterion
 - + Part A + Part B:
 - Also eliminate any known mediator!
 - How?



Additional consideration

- Establishing temporality
 - Relevant for time-varying covariates in longitudinal scenario
 - Also useful for point-treatment scenario
 - Help us rule-out
 - reverse-causality /
 - mediator
 - Time-dependent confounding that affects future treatment: MSM



Part 4

Identifying confounders:

Modelling criteria

Might be somewhat useless to fit a model
blindly; but great for new discovery

Why Modelling?

- Controlling more confounder is good?
- Depends on the goal.
 - Causality? Primary outcome of interest vs primary exposure of interest?
 - Prediction of primary outcome of interest?
- Statistical model might show convergence / multicollinearity issues with too many covariates.
- Manually going through a large number of covariates is not always feasible.

Popular Statistical approaches

- Statistical covariate selection process are popular in getting parsimonious model.
 - Univariate selection of covariates.
 - Backward/forward selection is most popular.
 - Small p-value based selection is not encouraged.
 - Hold important variables fixed, then explore
 - AIC vs p-value based selection?
 - Post-selection is an issue (SE /CI not valid).
 - Data splitting /cross-validation (resulting in larger SE!). Alternatively bootstrap (majority rule)!

Popular Epidemiologic approaches

- Change in estimate (e.g., 10%) is also popular, but
 - Basic assumption is that we have a set of confounders; not mediators / colliders.
 - Does not work for non-collapsible measure (HR/OR).
 - That 10% may still could be due to chance variation.

Popular Epidemiologic approaches

- hdPS
 - Takes into account the magnitude of association with Y and A .
 - Not much theoretical properties studied so far.
 - Selects each proxies one by one; no multivariate adjustment, could be subject to multicollinearity / noise.
 - Post-selection is still an issue.

Popular Machine learning approaches

- Identifying important variables
 - LASSO, addresses multicollinearity
 - Elasticnet, middle ground between lasso (somewhat unstable) and ridge (stable)
 - Random forest; variable importance flexible model fitting
 - Causal discovery methods / create DAG empirically?
- SE calculation may be problematic
 - Data splitting may still be a reasonable solution
- Double Robust/TMLE approaches

FAQ

- Machine learning = prediction (generally speaking)
- Why we are talking about predictors, and equating them with confounders?
 - Because of PS
 - PS is basically a prediction model
 - Whatever model gets prediction of PS right is appropriate as long as balance is there.
 - These prediction methods are not suitable for outcome model (generally speaking) if the goal is establishing causality!

Which of the following methods take causality into consideration while performing variable selection?

Backward selection based on AIC

P-value based selection

Disjunctive cause criterion

LASSO

Change in estimate

None of the above



Thanks!

ehsan.karim@ubc.ca

www.ehsank.com