- DeCART 2019 -
Mechanistic Thinking with Models: Concepts, Examples, and Methods:
Directed Acyclic Graphs (DAGs)

- Matthew Samore, MD
- Chief, Division of Epidemiology
- HA and Edna Benning Professor of Medicine
- University of Utah
- VA Salt Lake City

Lynd Bacon, PhD MBA
Adjunct Associate Professor
Division of Epidemiology
David Eccles School of Business
University of Utah

Outline

- Structural thinking
  - The Monty Hall dilemma
  - Toy Story
- Directed acyclic graphs
- Epidemiological examples

By the end of the day you will be able to:

- Formulate causal questions
- Apply concepts such as d-separation and conditional independence
- Characterize confounding and selection bias using graphs
- Have another model toolkit at your disposal to support mechanistic thinking

Statistical methods for causal inference will be covered tomorrow

Study 1

- Causal hypothesis:
  - Coffee increases risk of pancreatic cancer
- Study design
  - Select subjects with pancreatic cancer from cancer registry in 5 state area
  - Select controls by random digit dialing in the same 5 state area
- A colleague who had "epi training" recommends that you exclude controls who had smoking or alcohol related conditions
- Should you accept his advice?

Study 2

- Causal hypothesis
  - Ibuprofen increases risk of development of empyema in children with pneumonia
- Study design
  - Divide hospitalized children with <u>community-acquired pneumonia</u> into those who have empyema and those who do not. Ascertain use of ibuprofen prior to hospitalization.
- Does this study have selection bias?
- If there is selection bias, can it be removed?

Study 3

- Causal hypothesis
  - Estrogen supplementation increases risk of uterine cancer.
- Some women experience vaginal bleeding on estrogen in the absence of uterine cancer or cancer precursors
- Vaginal bleeding leads to diagnostic evaluation for uterine cancer
- Should you limit your study of the association between estrogen and uterine cancer to women who have had diagnostic evaluation for uterine cancer?

For today, ...

- Forget about p values
- Do not worry about the names of different types of study designs
- Instead, think about prizes, tinker toys, lines and circles

I. The Monty Hall Dilemma: an introduction to structural thinking

- Behind 1 of 3 doors lies a grand prize; behind each of the other two lies a booby prize.
- Monty asks you to pick a door. You state your selection, then his assistant opens one of the doors you did not select. The door revealed always shows a booby prize. You are asked if you want to switch.

Should you stay with your original selection or switch?

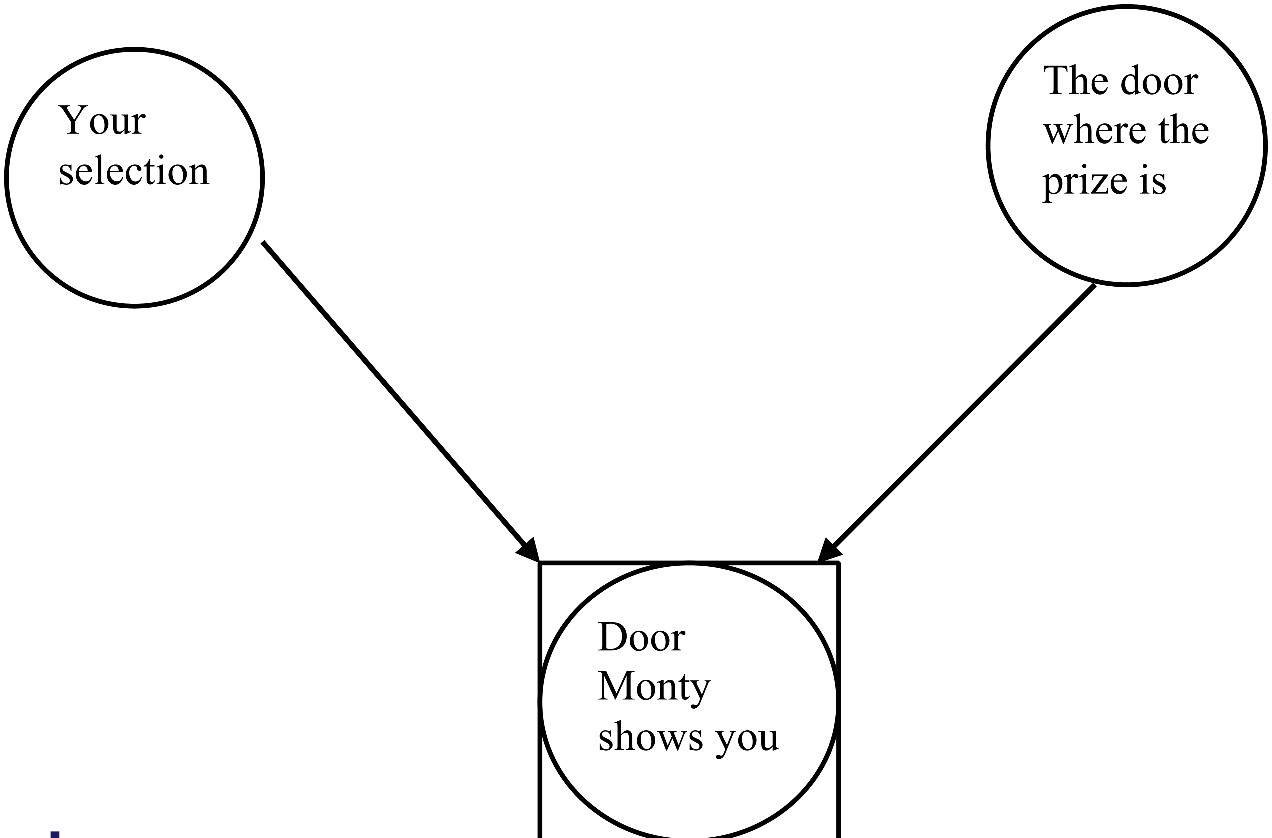
What is your answer?

- Usually, about 75% of people say it does not matter
  - Two doors left, there is an equal chance that the prize will be behind either door
- The correct answer is:
  - YOU SHOULD SWITCH

The logic expressed in words

- Your initial selection has 1/3 chance of being correct
- If your initial selection is wrong (2/3 probability), Monty's assistant has only one option
  - Then selecting the door not opened is always correct

The logic expressed in structure



Legend

variable

Indicates direction of causation

variable that is conditioned upon

If rules changed

- One of the doors showing the booby prize was opened regardless of initial selection

- Then, each of the remaining two options would have equal likelihood

Why the Monty Hall Dilemma?

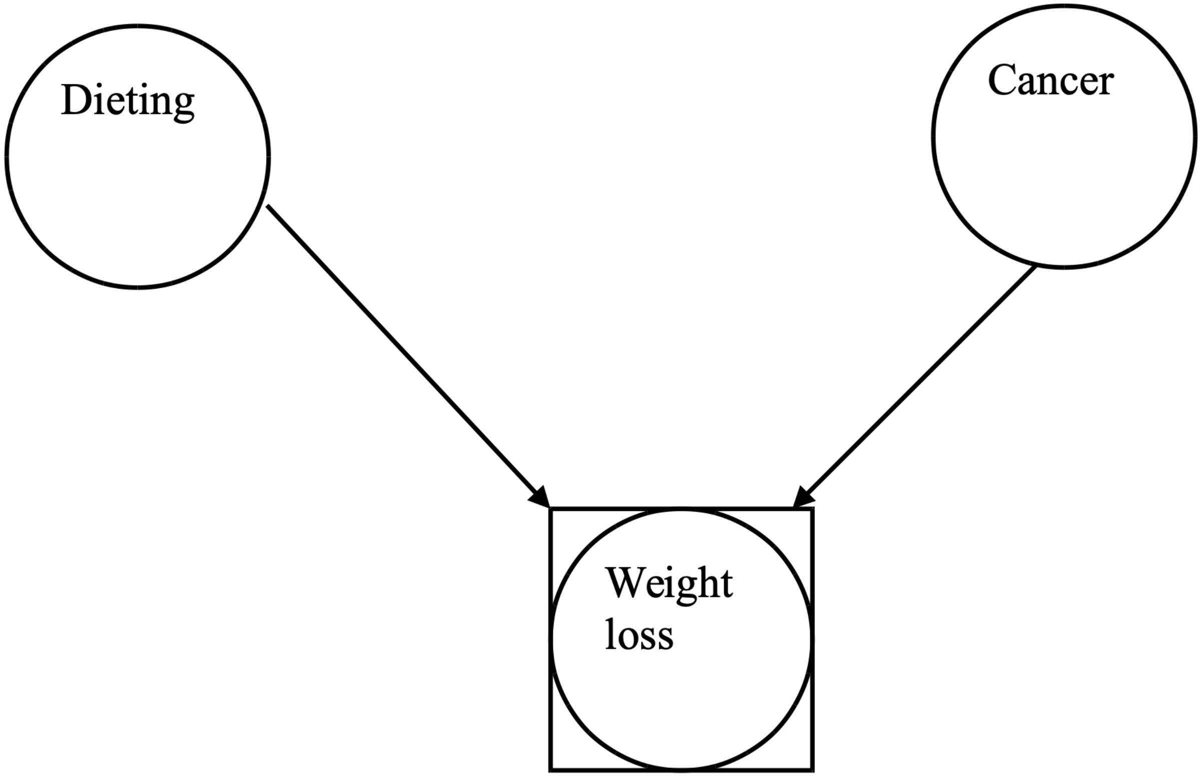
- It illustrates the collider principle: two independent causes that have a common effect are associated “conditional on” (given) the effect

- These types of problems are hard for humans

Another example of the collider principle, this one drawn from clinical medicine

- A patient presents with 30 pound weight loss
- Does information about whether the patient has been on a diet influence your suspicion that the patient has cancer?

Structural approach



- Dieting and cancer are two causes of weight loss
- They are negatively associated conditional on weight loss
- Given weight loss, if an individual has not been dieting, cancer is more likely present

How is this relevant to epidemiology and data science?

- The collider principle underlies all forms of selection bias!
- Keep your eyes open – the problem is easy to miss

Another relevant story

Begin with an imaginary, but plausible scenario:

- You're the hospital epidemiologist at a Children's Hospital
- The Vice President of Marketing initiates a gift give-away program:
  - Each day 5% of the children in the hospital will be randomly selected to receive a toy
    - Children receive at most one toy
  - Because you're the compulsive sort, you keep track of which patients received toys on which date

You are trying to to teach your fellow how to do outcomes studies

- You pose a simple question: evaluate whether receiving a toy affected a child's length of stay
- Your fellow has taken epi 101 and collects data on a cohort of patients admitted during a one month period

During the 1 month period:

- 2,207 admissions (this is a busy children's hospital)
  - Mean and median length of stay: 4.8 and 3 days (this part is based on real data!)
- 505 received a toy

  This part was simulated on the basis of a random selection process

You suggest a matched cohort study

- You give careful instructions:
- For each child that received a gift, randomly select a child who did not receive a toy and whose length of stay was at least as long as the interval from admission to receipt of the toy

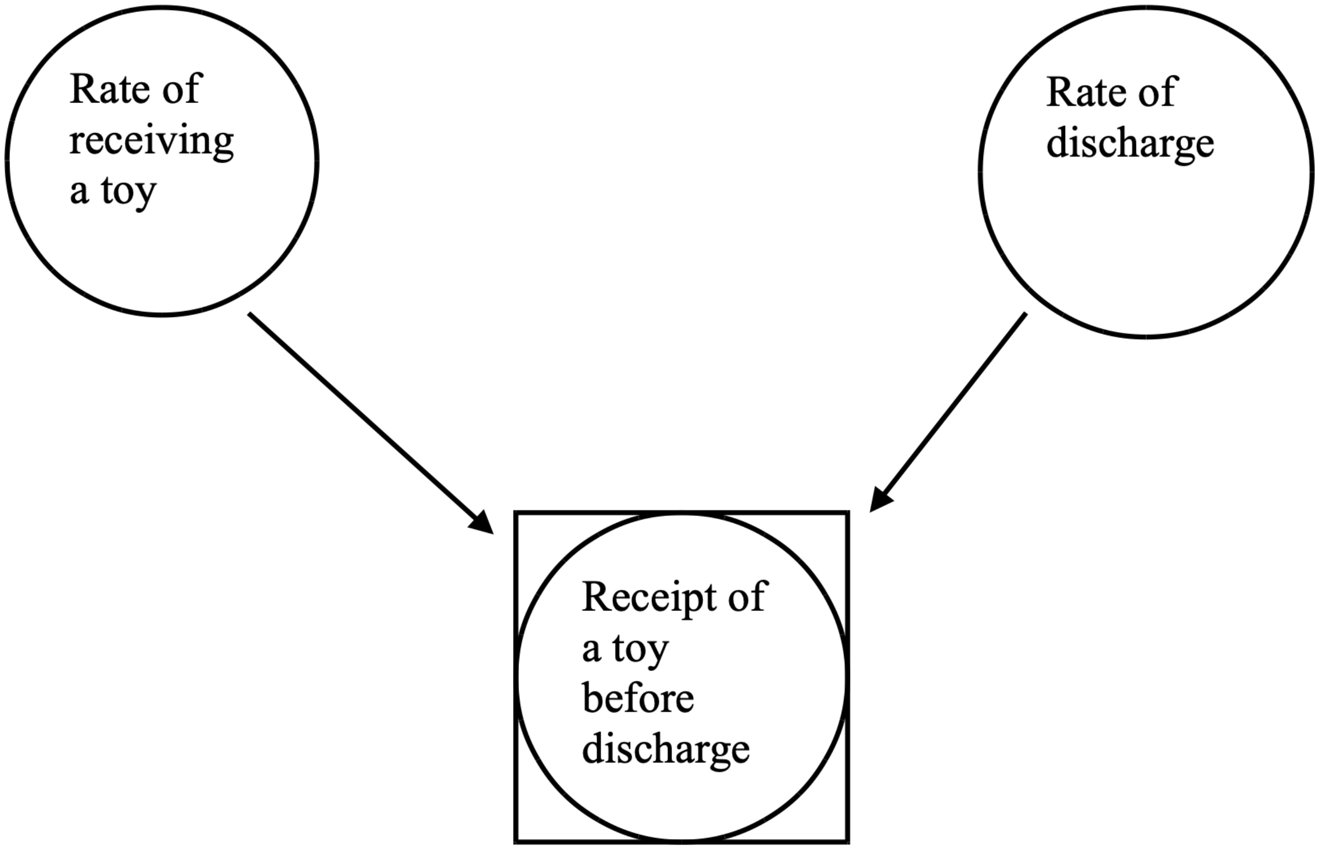
The fellow returns the following data to you:

	Toy recipients	Non-toy recipients	p value (paired t test)
Post-toy LOS study #1 (not all toy recipients matched)	6.2	3.5	p=.001
Post-toy LOS study #2 (controls selected with replacement)	6.1	4.4	p=.002

Why is this biased?

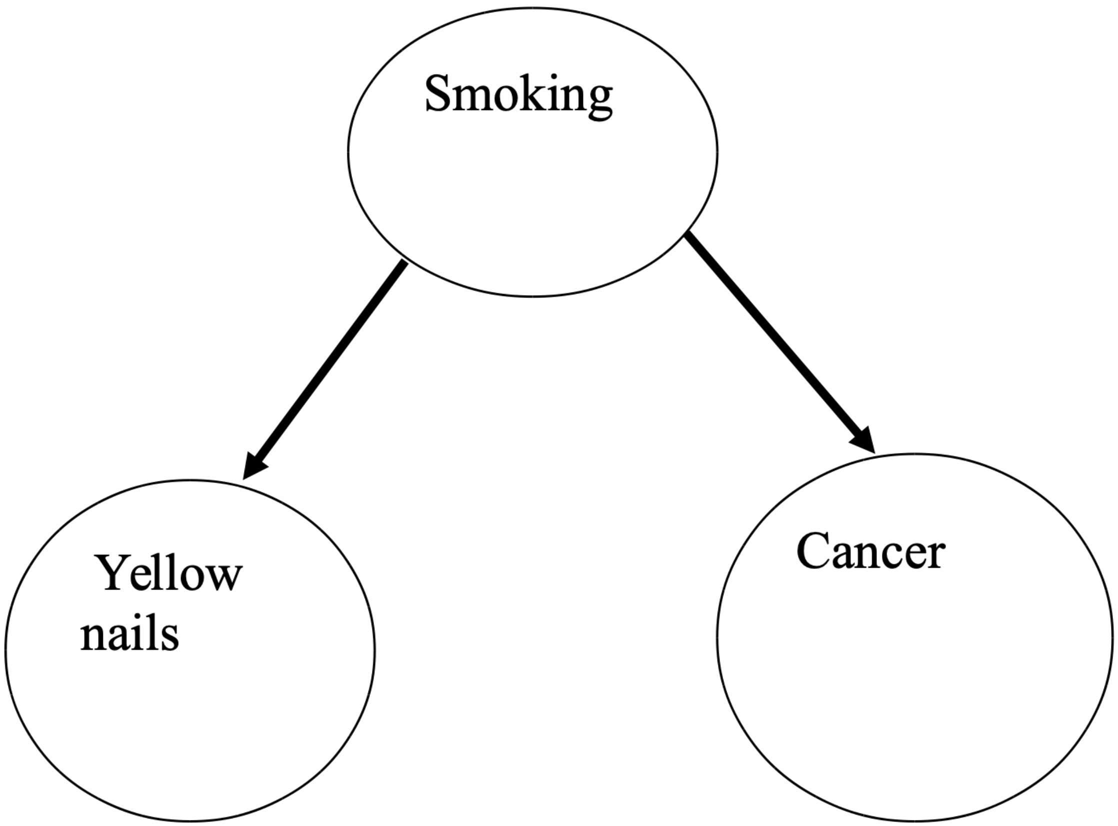
- The analysis conditions on whether a toy was received before discharge
- It gives the wrong answer because the method induces selection bias

The problem depicted structurally: the exposure and outcome collide



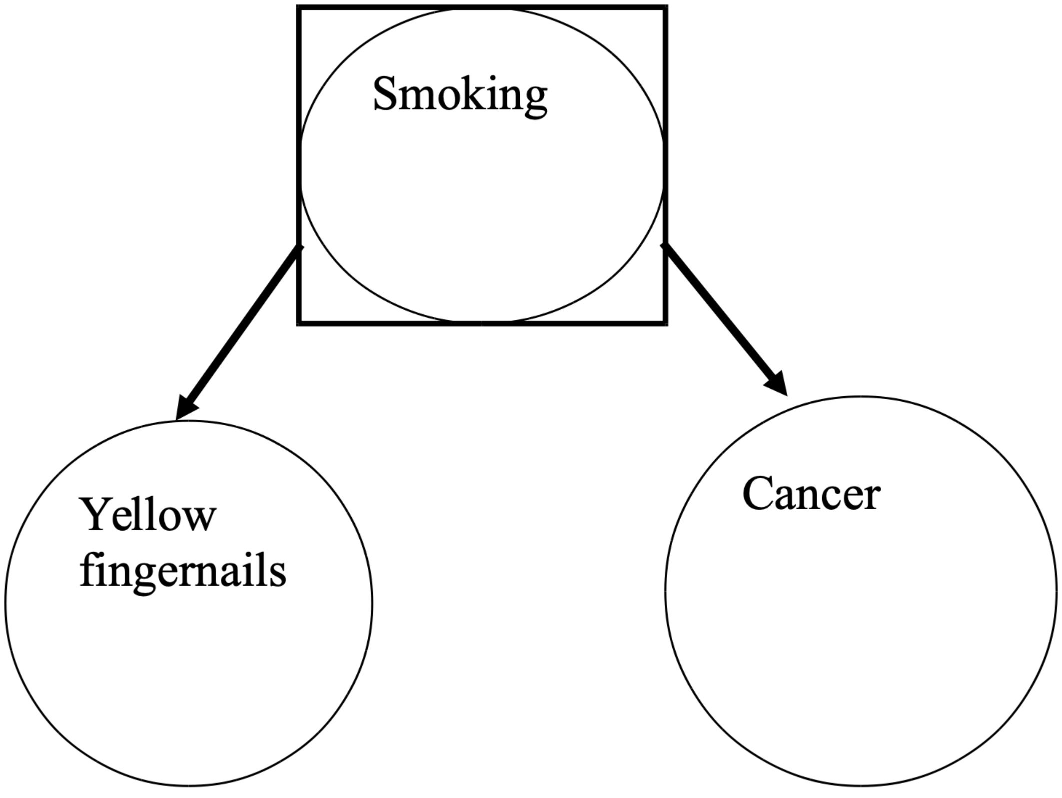
Structural thinking applied to confounding

- More intuitive than the collider principle
- Confounding is bias due to common causes
- Example: yellow fingernails are associated with cancer



To remove confounding

- Block the back-door path by conditioning on the confounder



The beautiful symmetry in epidemiology

- Biases in estimation of the causal effect of exposure on outcome
- Confounding
  - Common causes
  - Conditioning on the common cause removes confounding
- Selection bias
  - Conditioning on a common effect
  - Avoid conditioning on a common effect of exposure and outcome

Structural thinking (causal reasoning)

- Formulate causal assumptions and hypotheses
  - Do not be afraid to use your background knowledge!
  - Draw a picture or graphical model
  - Keep it simple!

Causal concepts and statistics

- Although statistical analysis can yield evidence of confounding, there is no statistical test for confounding or selection bias
- Determining whether confounding or selection bias is present relies in part on background knowledge
- "no causes in, no causes out"

Introduction to Daggity: Features

- “Whiteboard” for drawing graphs
- Naming conventions are a little different
- Graph elements can be created by clicks, clicks+keystrokes, entering simple model “code”
- “Adjustments” (conditioning, held constant) required to test causal effects are indicated
- Graphs can be exported

Daggity Introductory Tour:

www.dagitty.net

Causal DAGs

- Nodes are variables
- Directed edges: arrows indicate causal relationships
  - In the examples given here, causal relationships will be defined at a population level

Definitions for causal DAGs

- Nodes are variables
- Show all common causes of any pair of variables
- A lot of expert knowledge is encoded in the missing arrows

Why called "directed acyclic graphs"?

- Directed because arrows not just lines
- Acyclic because no arrows from descendents (effects) to ancestors (causes)
- If an "effect" variable can affect a "cause" variable it does so at a later time

  Current and past anti-HIV therapy affects current CD4 count, which in turn influences future anti-HIV therapy

- These arrows represent the postulated causal mechanisms or pathways by which exposure affects the outcome or disease.

- Causal pathways may be direct or indirect. An indirect pathway is characterized by the presence of an “intermediate variable” (I) that mediates a causal effect whereas a direct effect lacks an intermediate variable.

- Independence: lack of association between variables
    - Knowing the value of one variable provides no information about the value of another variable
- Dependence: opposite of independence
    - Synonym: correlation

Dagitty Exercise:

Chains

Forks

Colliders

Dagitty To Do's

- Start by creating a new model
  - Create a *chain*:
    - $X0 \rightarrow Z0 \rightarrow Y0$
  - Create a *fork*:
    - $X1 \leftarrow Z1 \rightarrow Y1$
  - Create a *collider*:
    - $X2 \rightarrow Z2 \leftarrow Y2$
  - Add a *descendant* Q of Z2 to your collider
- Note: None of these variables are Dagitty exposure or outcome variables.

What is an open path?

- Open pathways are nodes that are linked directly or indirectly via:

  Head-to-tail, tail-to-head, tail-to-tail connections, unless the connection is blocked

- “Conditioning” on a head-to-tail or tail-to-head or tail-to-tail (common causes) connection BLOCKS it

- Conditioning on a head-to-head connection (a collider) **UNBLOCKS** it

What is conditioning?

- “Condition” roughly means “given”
- Without conditioning on variable “C”
  - The association between “E” and “D” is the crude or unconditional association
- Conditioning on variable “C” means measuring the association between “E” and “D”, given the value of “C”

Dagitty Exercise:

Conditioning

Dagitty: Confounding

- Create the following graph, and assess what adjustment need to be made to evaluate the effect of X → Y, where X is an "exposure," and Y an "outcome:"
  - X → Y
  - Z → X Y
  - P → X Y Z
  - Q → X Y Z
  - R → Z
- What independence hypotheses are evaluable?

The concept of d-separation

- Variables that are d-separated have NO open paths between them

  d-separated variables will be independent

- Variables that are not d-separated (e.g, have open paths that connect them) will be associated unless causal effects happen to cancel each other out

Graphical rules for d-separation

- Absence of arrows connecting two variables (there is no path): **marginal or unconditional independence**

- Blocking (conditioning) on a non-collider in the path between two variables: **conditional independence**

Using DAGs to evaluate for confounding

- Draw the DAG which corresponds to the “causal null hypothesis” for the variable of interest
- The causal null hypothesis
  – The assumption that there are no indirect or direct causal pathways pointing from exposure to disease.
- Erase all arrows downstream from exposure that connect it to the disease (outcome)

After eliminating arrows pointing from exposure to disease

- All remaining open pathways between exposure and disease are non-causal pathways
- These non-causal pathways must be blocked!
  - The non-causal pathways produce spurious associations
  - Intervening on an exposure connected to disease only through non-causal pathways produces no change in disease

Dagitty: Evaluating Confounding

- Draw the the following graph. What paths need to be blocked in order to infer that X causes Y?
- X is an exposure, Y is an outcome
- $X \rightarrow Y$
- $Z \rightarrow X\ Y$
- $Z \rightarrow Q$
- $Q \rightarrow X\ Y$
- $P \rightarrow Q$

Problems & Examples

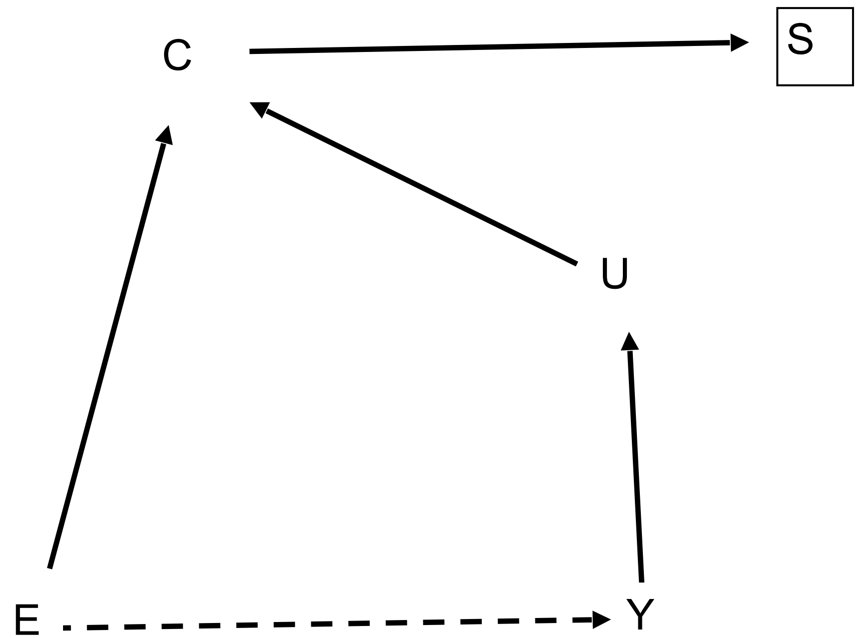
Does smoking cause coronary artery disease?

- Smoking was not associated with a “positive” cardiac catheterization among patients who were referred for cardiac cath

- Why?

Explanation of causal arrows

- E→C
  - Clinicians are more likely to refer smokers to cardiac cath than non-smokers
- E→Y depicts causal hypothesis
- C→S
  - Cardiac cath influences selection into study
- U mediates influence of disease on referral for cardiac cath—however, it is unmeasured

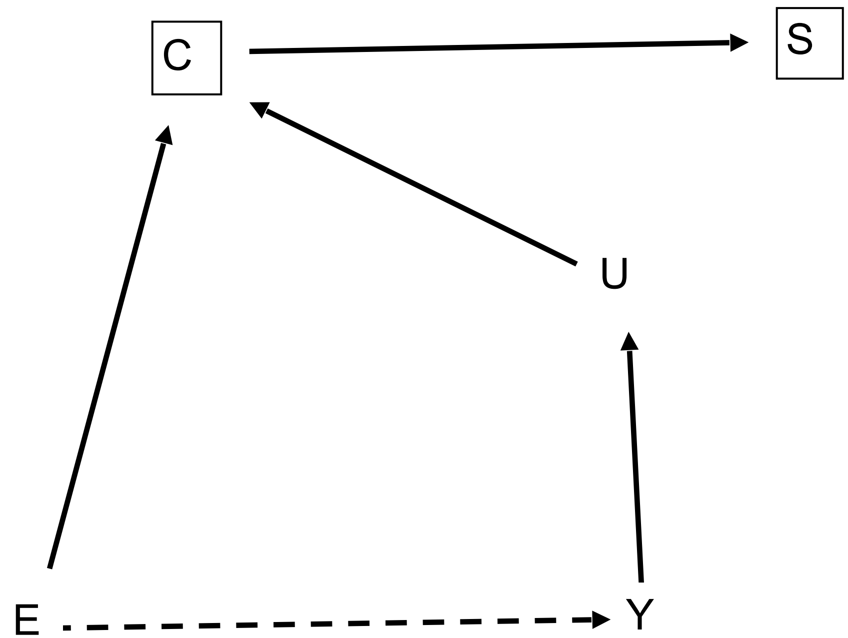


Identify the open, non-causal pathway from E to Y

E: exposure (smoking); C: cardiac catheterization; Y: coronary artery disease (outcome); U: symptoms and signs of coronary disease; S: selection into study sample

Does conditioning on C help?

- E→C
  - Clinicians are more likely to refer smokers to cardiac cath than non-smokers
- E→Y depicts causal hypothesis
- C→S
  - Cardiac cath influences selection into study
- U is unmeasured

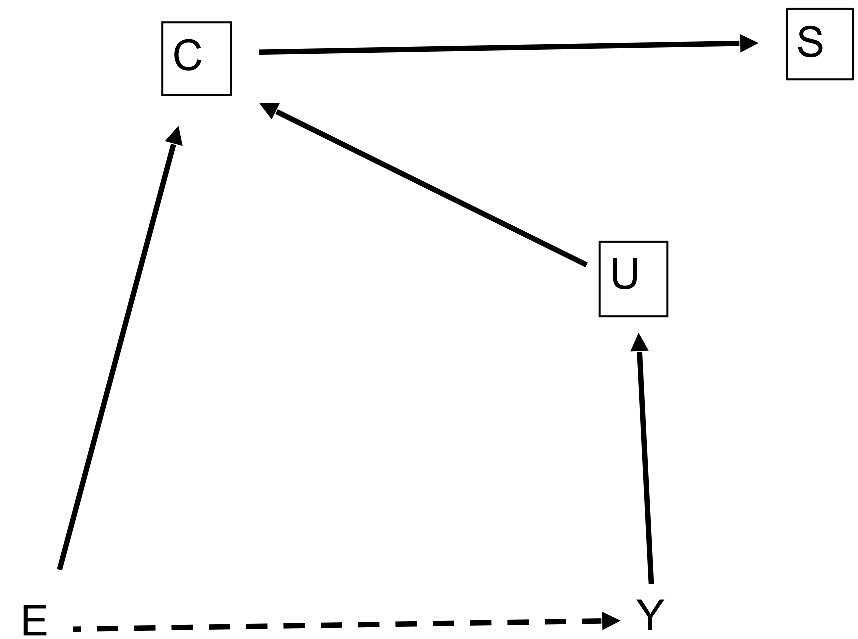


Identify the open, non-causal pathway from E to Y

E: exposure (smoking); C: cardiac catheterization; Y: coronary artery disease (outcome); U: symptoms and signs of coronary disease; S: selection into study sample

How about conditioning on U?

- E→C
  - Clinicians are more likely to refer smokers to cardiac cath than non-smokers
- E→Y depicts causal hypothesis
- C→S
  - Cardiac cath influences selection into study
- Now U is measured

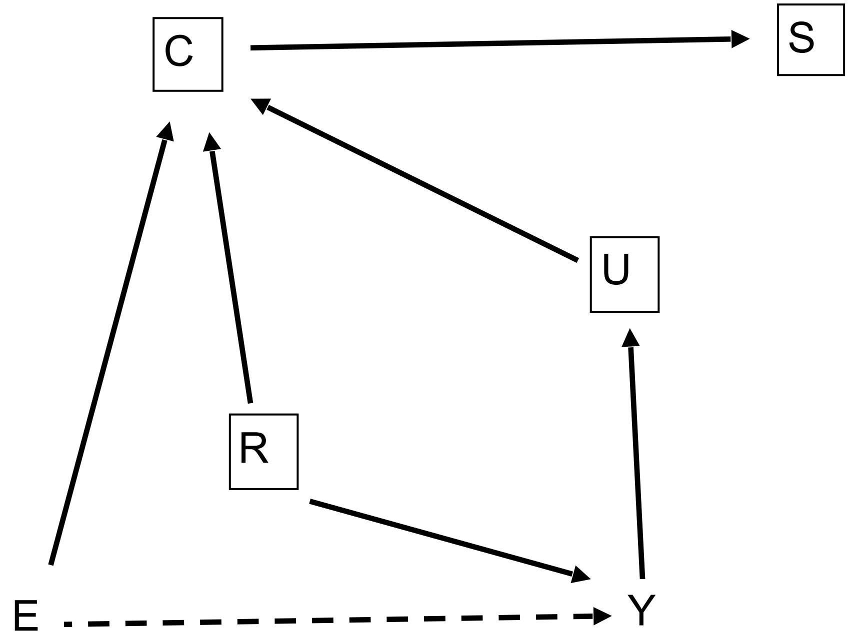


No open, non-causal pathways from E to Y!

E: exposure (smoking); C: cardiac catheterization; Y: coronary artery disease (outcome); U: symptoms and signs of coronary disease; S: selection into study sample

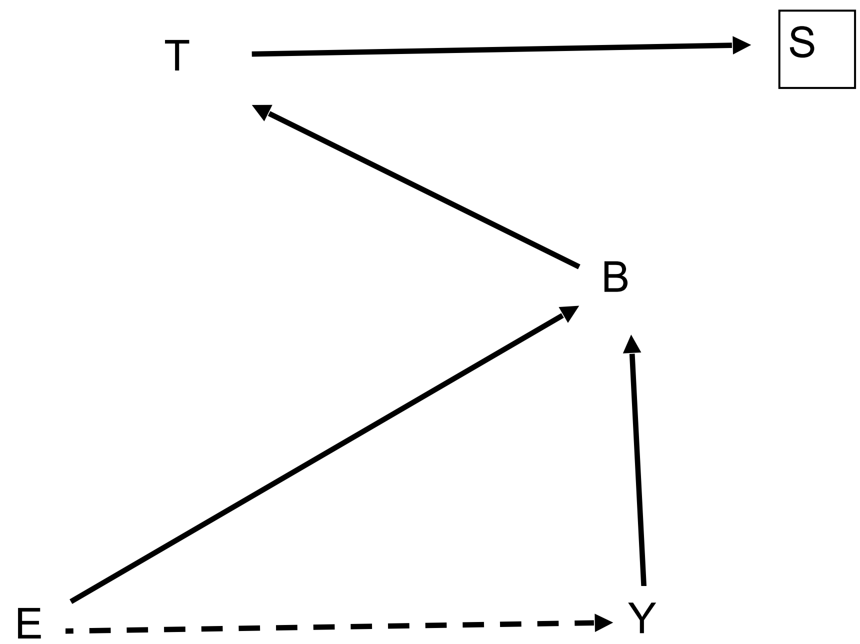
What about other risk factors for coronary artery disease?

- E→C
  - Clinicians are more likely to refer smokers to cardiac cath than non-smokers
- E→Y depicts causal hypothesis
- C→S
  - Cardiac cath influences selection into study
- Other risk factors for coronary disease that many influence referral



E: exposure (smoking); C: cardiac catheterization; Y: coronary artery disease (outcome); U: symptoms and signs of coronary disease; S: selection into study sample

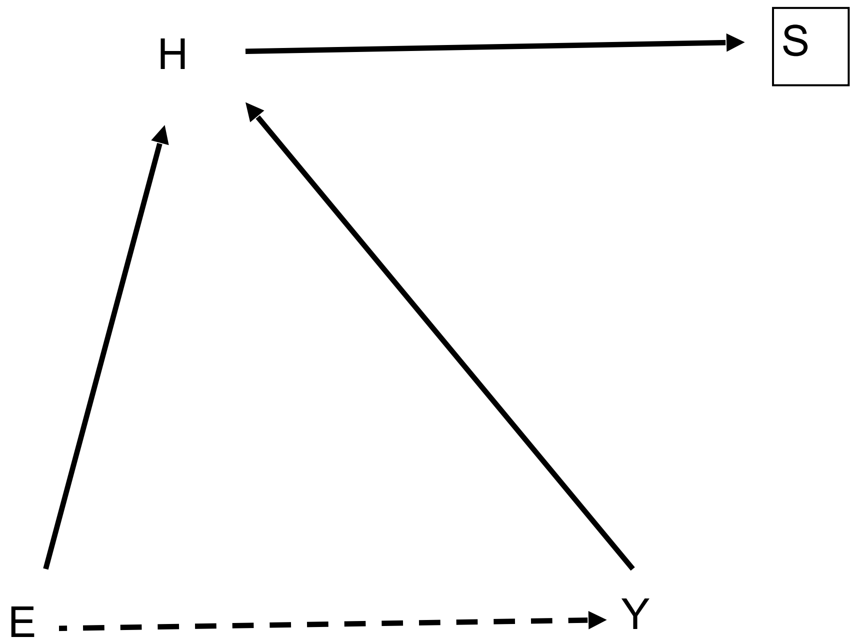
Both disease and exposure often influence diagnostic testing



E: exposure (estrogen); B: vaginal bleeding; T: diagnostic testing for uterine cancer; Y:disease (uterine cancer); S: selection into study

Ibuprofen-empyema study

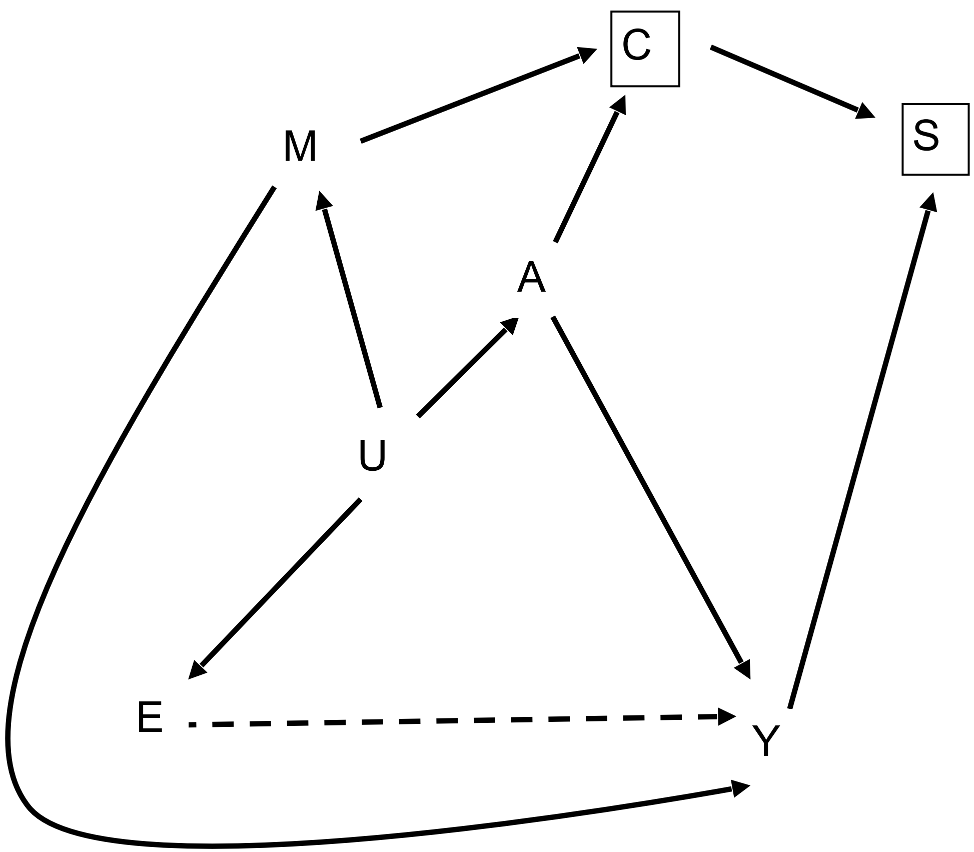
- If ibuprofen suppresses fever, it might reduce the chance of hospitalization



E: exposure (ibuprofen); Y:disease (empyema); H: hospitalization; S: selection into study

Pancreatic cancer – coffee study

- Why is there an arrow from Y to S?
- Does conditioning on M (smoking) and A (alcohol) eliminate bias?



E: exposure (coffee); U: unmeasured common causes of smoking, coffee, and EtOH (e.g, propensity for habits); M: smoking; A: EtOH (alcohol); C: conditions related to smoking & EtOH; Y: disease (pancreatic cancer); S: selection into study

Dagitty: Smoking & CAD

- Create a DAG based on the preceding slide
- Verify what “adjustments” (conditioning on confounders) is needed in order to be able to test the hypothesis of a direct causal effect of smoking on CAD.

The stillbirth question

E: folic acid supplementation
D: neural tube defects
C: stillbirth or therapeutic abortion

Daggity: Stillbirth

- Draw the preceding graph
- Verify what conditioning is needed to test for a direct effect of E on C
- How would you test for an *indirect* effect?

Low-birth weight paradox

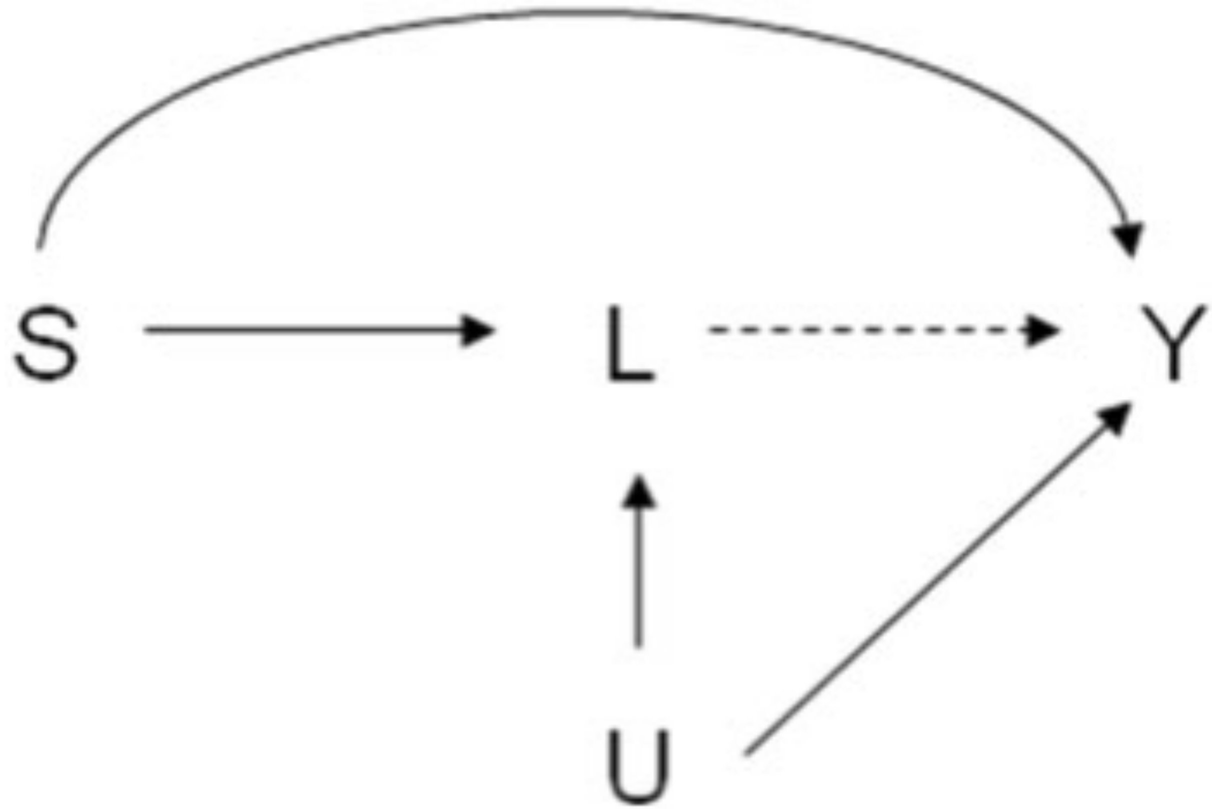
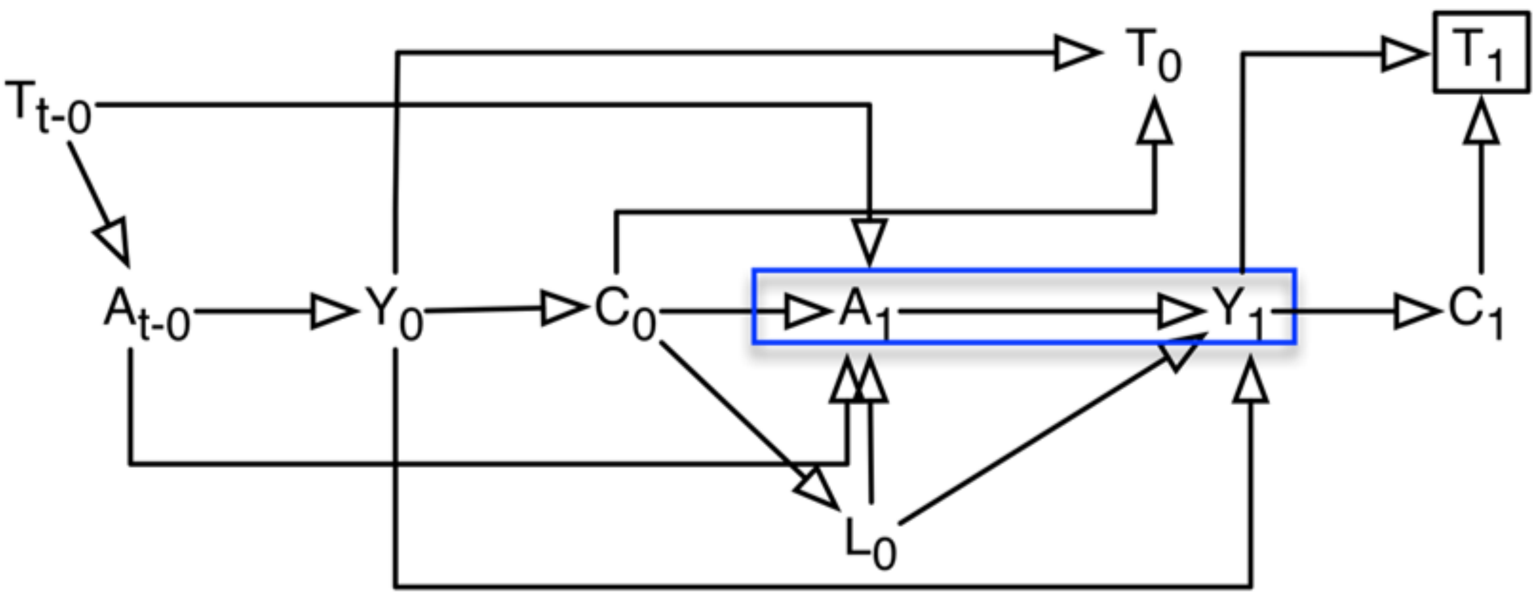


Figure 1. Unmeasured intermediate-outcome confounding as an explanation of the birthweight paradox: unmeasured common causes (U) of low birthweight (L) and infant mortality (Y), such as for instance malnutrition or birth defects, bias association between maternal smoking (S) and infant mortality.

Dagitty: Smoking and Infant Mortality

- Draw the preceding DAG
- What is the effect (if any) of conditioning on U?

Effect on antibiotics (A) on antibiotic resistant infection



Another causal question

- What is the effect of methicillin-resistant *Staphylococcus aureus* (MRSA) bloodstream infection on mortality?

The problem with this question

- What is the alternative?
  - "Compared to what?"
- Is MRSA bloodstream infection being compared to not having MRSA bloodstream infection or is it being compared to having methicillin-sensitive *Staphylococcus aureus* (MSSA) infection?

Further issues

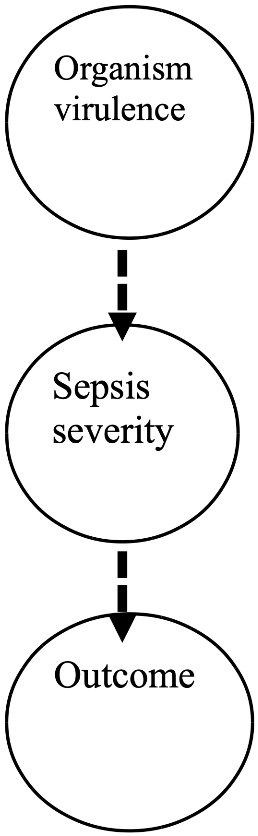
- If the comparison is MRSA versus MSSA, what is the precise question:
  - Does MRSA produce more severely symptomatic infection than MSSA?

    Or

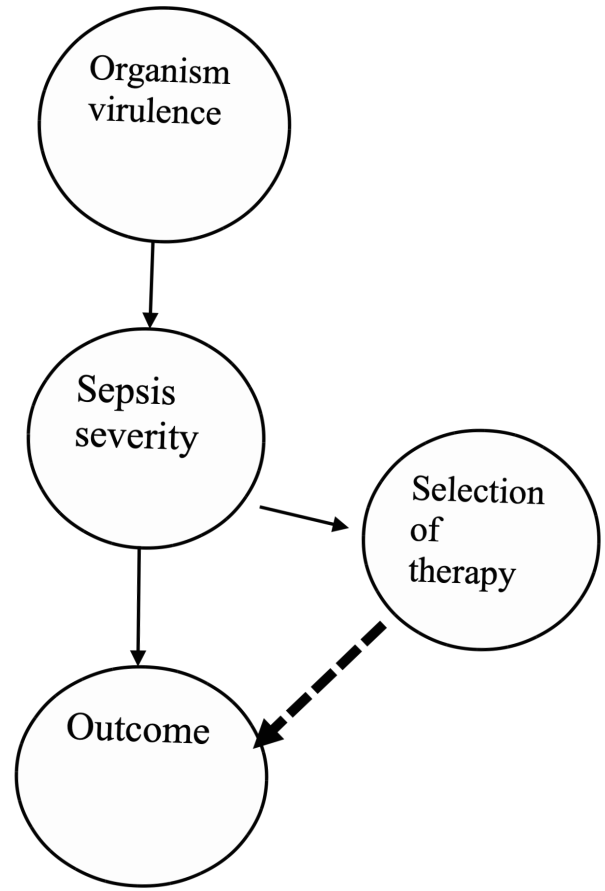
  - Is therapy against MRSA less effective than therapy against MSSA?

Structural approach

- Causal question is about virulence
- Sepsis severity is an intermediate variable



- Causal question is about therapy
- Sepsis severity is a confounder



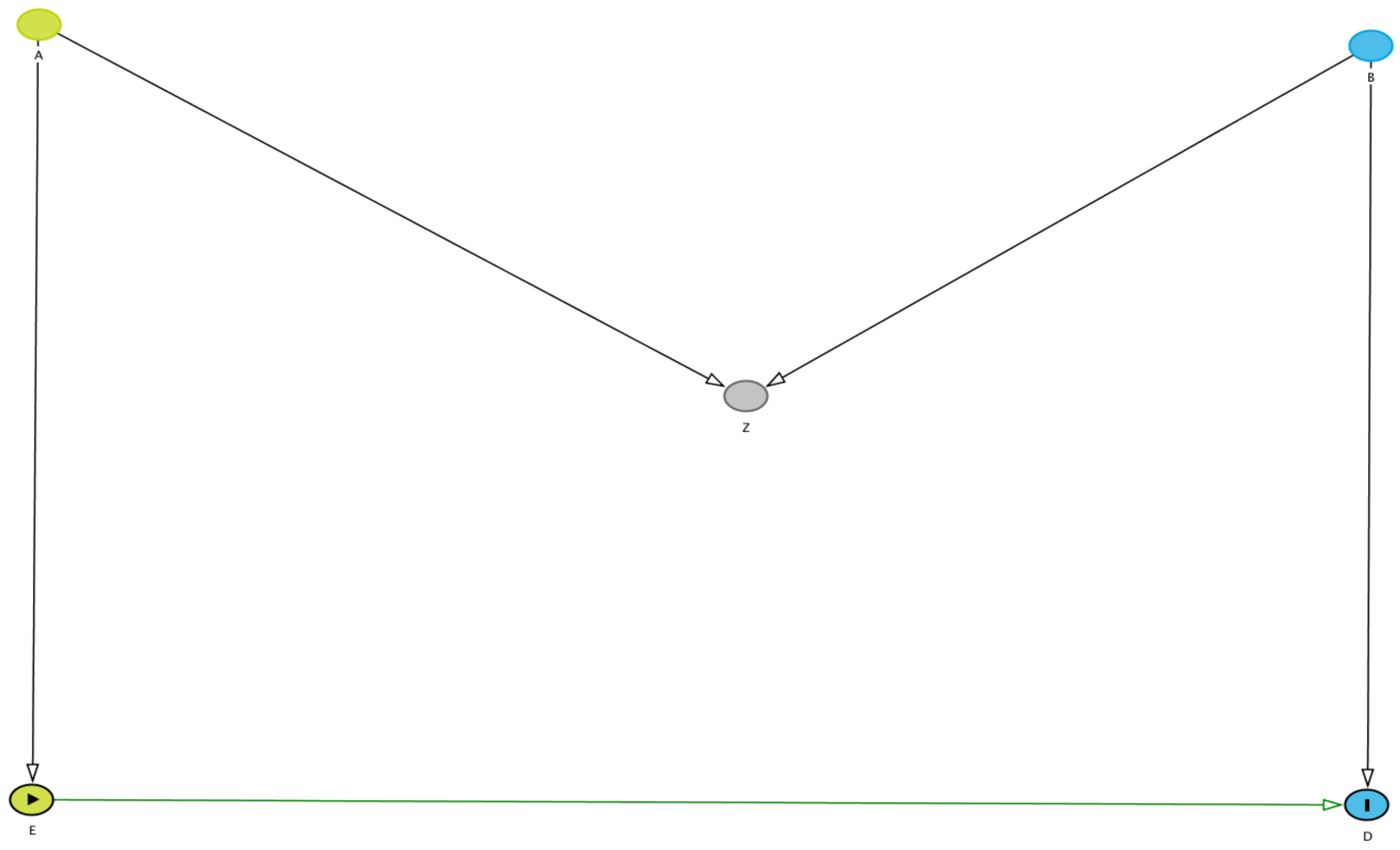
Summary I.
State the causal question explicitly

- Think counterfactually
- What would have happened?
- Would this exposed patient have experienced the same outcome if he or she had not been exposed?

Summary II: Think structurally

- Causal questions are often not directly answerable via randomization
  - Confounding and selection bias are major problems
- Use DAGs to help you encode your assumptions and knowledge, identify sources of bias, and increase the validity of your research

M Graph



Bias Graph