# Selection Bias

EPIB 704

Gabrielle Jacob

# Let’s Play a Game!

**We are going to answer a comps question!**
**This is reflective of Part 2 of the exam.**

**Get into groups of 2-3. You will have 10 minutes to read the situation and point out potential selection bias concerns (could also include drawing a DAG). You want to identify as many concerns as possible.**

**Next, you will select one person from your group to roll die. The group that gets the highest number gets to go first.**

**Here is the catch: you cannot repeat what another group said. If your group does, then your group is out.**

# The prize is candy/snacks :)

# Any questions or concerns before we begin?!

# LEARNING EXPECTATIONS

- **DIFFERENTIATE** between type 1 and type 2 selection bias
- **IDENTIFY** methods to address selection bias
- **IDENTIFY** different types of Type 1 selection bias
- **APPLY** knowledge of Type 1 selection bias to a topical example

# Who are we really studying?

# Limited by reality

Epidemiologists often work from datasets that they did not create or are limited in the data they can collect related to their exposure and outcome. Therefore, in most epidemiological studies, there is a chance that the study has selection bias.

# In an epidemiologic study, we do three things:

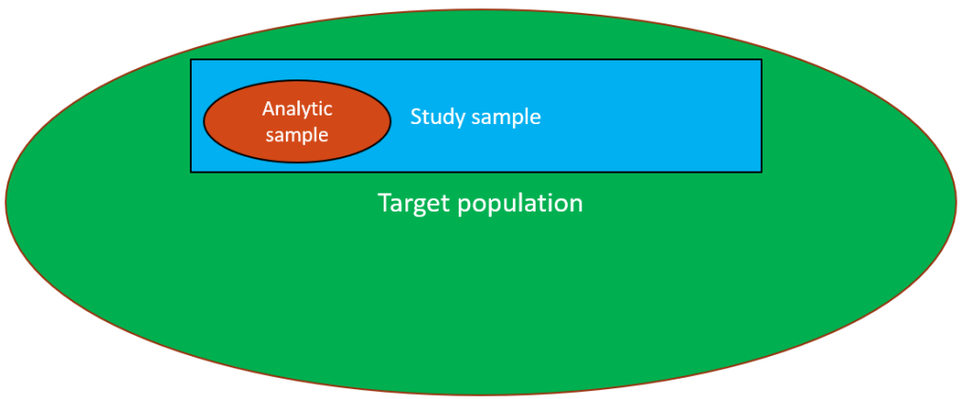
- Identify a target population
- Select the study sample from the target population
- Select the analytic sample from the study sample

# Pop off! Populations of interest

**Target population**
the population that inference is to be made about

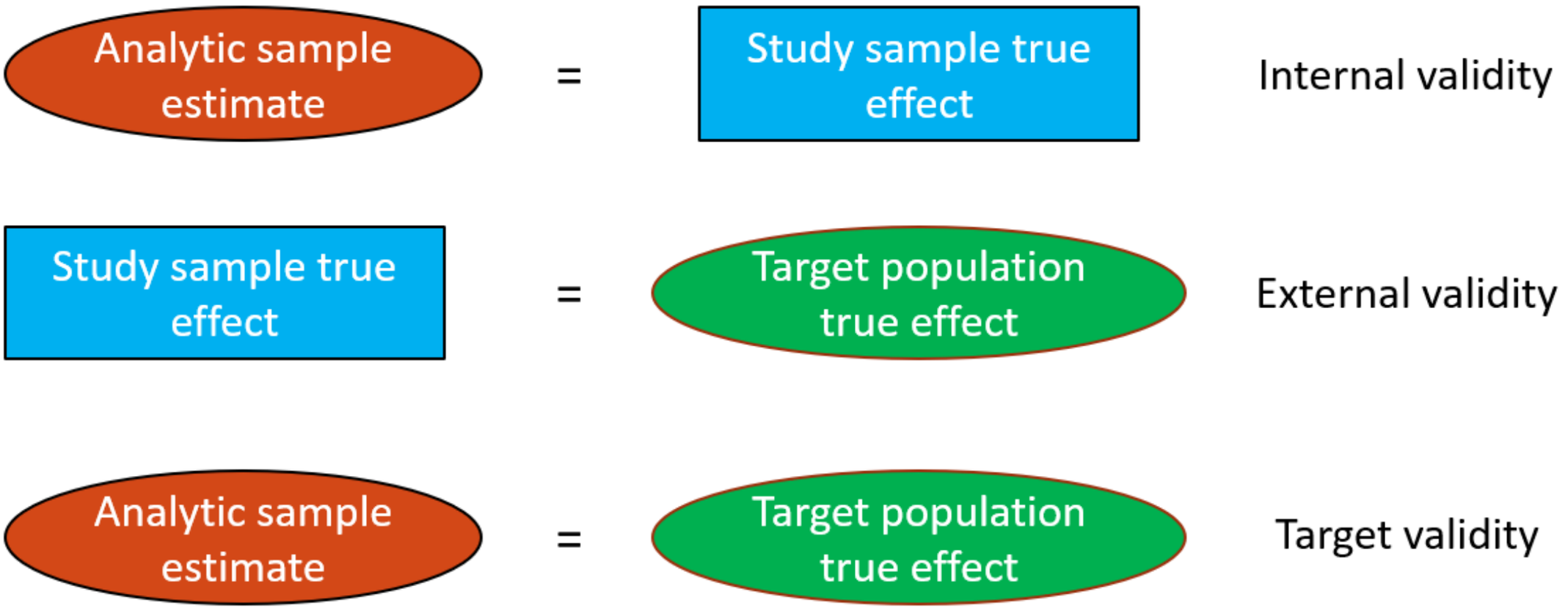
**Study sample**
the complete population that is included in the study, and it is used to make inference about the target population and may or may not be representative of the target population

**Analytic sample**
the observed portion of the study sample that is used for analysis




Image

# Internal and External Validity




Ex/In Validity Flow Chart

# Internal Validity

We achieve internal validity when the true causal effect estimated from the analytic sample is equal to the true causal effect in the study sample.

# External Validity

We achieve external validity when the true causal effect in the study sample is equal to the true causal effect in the target population.

# So what is selection bias?

Any bias away from the true causal effect in the referent population due to how the sample is selected from the referent population.

# Type 1 selection bias I

- Arises when the selection is based on a collider or its descendants which will distort the observed relationship between the exposure (cause a spurious association)
- In other words, Type 1 selection bias results from conditioning on a common effect of two causes (collider)
- Restricting to one (or more) level(s) of a collider (or a descendant of a collider) opens a noncausal backdoor path between the exposure and outcome

# Example of a collider

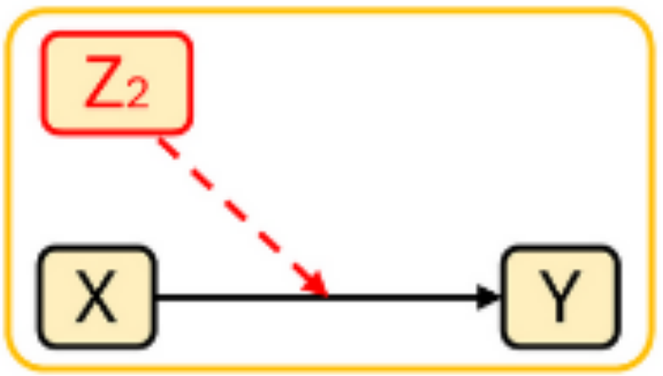
- What is a collider?
  - Variable associated with the exposure or cause of the exposure and associated with outcome or a cause of the outcome

# Type 2 Selection Bias:

- Arises when the selection is based on an effect measure modifier which will lead to an incorrect estimate of the causal effect in the selected sample.
- Scale dependent (multiplicative vs additive) since type 2 selection bias is dealing with restricting to one or more level(s) of an effect measure modifier

# Just a refresher:

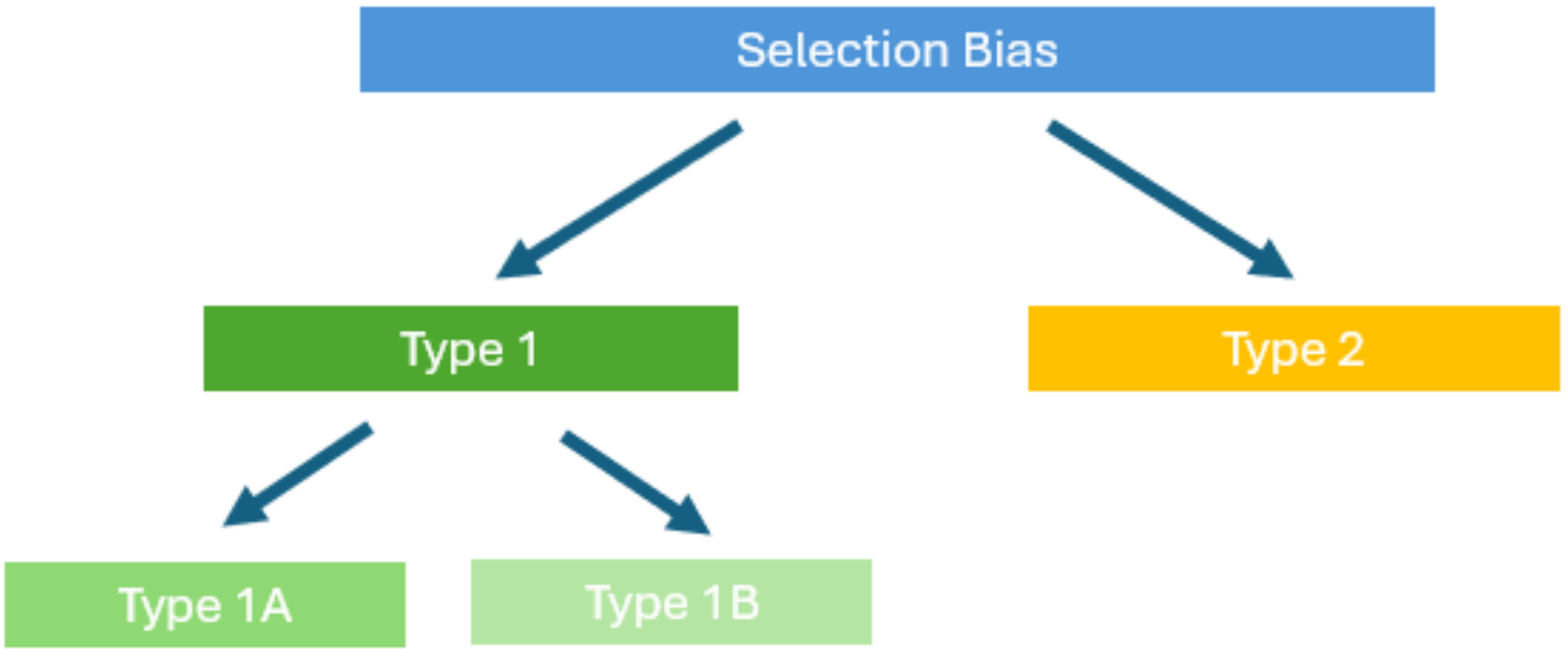
- An effect modifier is a third variable whose presence or level modifies the relationship between the exposure and outcome
- Example: the effect of smoking on heart failure might be enhanced in the subgroup of smokers who also have diabetes

| | | |
|---|---|---|
| **Effect modifier** |  | The association between smoking (X) and the risk of heart failure (Y) is different among individuals with or without diabetes ($Z_2$) |

# Type 2 Selection Bias: More on it

- Type 2 selection bias affects external validity when selecting the study sample form the target population
- In randomized and observational studies, it is rare that the study sample is randomly selected from target population, due in part to informed consent
- Generally, we cannot assume that the effect in the study is the same as in the target population

# CONCEPTUAL MAP




A nice chart!

# Let’s just focus on Type 1 Selection Bias:

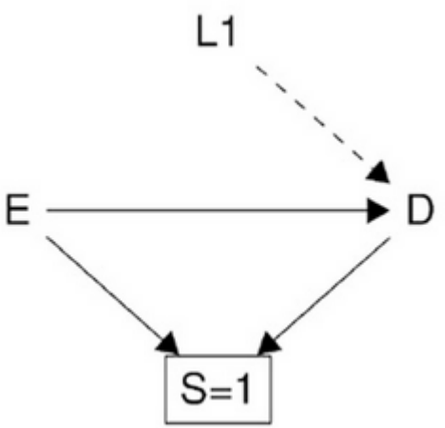
# Regular Degular

- When estimating causal effects, type 1 selection bias, is the classic selection bias we often encounter in epidemiologic literature
  - Called collider stratification bias or collider restriction bias

# Collider restriction vs collider stratification bias

- **Collider restriction bias** : Bias introduced by restricting to one level of a collider
- **Collider stratification bias** : Bias introduced by conditioning on the collider
  - Includes:
    - Bias due to restricting on a collider
    - Bias introduced through the unnecessary inclusion of collider in a regression model (analogous to restricting to more than one level of a collider)

# A Quick DAG!




Example: Berkson’s Bias

# Subtypes

Based on whether the causal effect in the referent population is identifiable or not

# Type 1A Selection Bias:

- Can be addressed
- Can find out the true causal effect by measuring and adjusting for covariates that lie on the noncausal path that is opened by restricting the collider via inverse probability weighting, g-computation, and sometimes stratification

# Type 1B Selection Bias:

- Cannot be addressed
- Occurs when there are no measured covariates that lie on the noncausal path
- The causal effect in both the selected sample and the referent population is not identifiable unless the selection probability of each combination of exposure, covariates (if any) and outcome is known, which is typically unattainable in practice

# What happens when you have Type 1B Selection bias

- Can be difficult to minimize analytically
  - Especially when selection is the direct common effect of both the exposure and the outcome,
  - and when selection is dependent on the outcome and causal effects of risk difference or risk ratio scale are desired (i.e., type 1B selection bias)

# What are some (not all!) ways we can address selection bias?

# How to correct for selection bias: Multiple imputation

- This method is useful when a complete-case analysis would lead to selection bias.
- Multiple imputation involves creating multiple imputed datasets, averaging estimates over these datasets, and then performing the analysis on the imputed data.
- We can use DAGS to understand our missing data and identify parameters of interest that are recoverable with complete-case analysis vs. those that require multiple imputation to avoid bias.

# How to correct for selection bias: IPTW

- Inverse probability of treatment weighting (IPTW) uses the propensity score (probability of treatment selection conditional on observed baseline characteristics) to obtain unbiased estimates of ATE (average treatment effect)
- By weighting subjects by the inverse probability of treatment, we can create a synthetic sample in which treatment assignment is independent of measured baseline covariates
- The causal interpretation depends on the untestable assumption that all additional factors predicting both selection and the outcome are identified.

# Different Types of Type 1 Selection Bias

Every year, at a romance book convention, Dr.Darcy sees the same trend at the convention clinic. Overwhelmingly, romance book enthusiasts come to the clinic with neck pain and severe eye strain. The doctor decides to conduct a study to assess the association between neck pain and severe eye strain.

# What is this an example of?

# What is this an example of?

Berkson’s Bias

# Why?

# Why?

The association between the exposure and outcome is not something you would see in a general population. The supposed association is based off the fact that readers may read in awkward positions and read for too long.

# Let’s visual it!

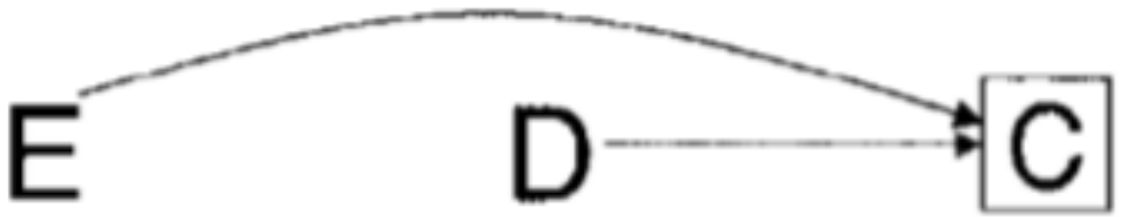



**FIGURE 3.** Conditioning on a common effect C of exposure E and outcome D.

- **The Nodes**
  - E → Neck pain
  - D →Eye strain
  - C → Patients at the convention clinic

Berkson’s Bias

A case-control study aims to study the effect of postmenopausal estrogens on the risk of myocardial infarction. The selection into the study depends on disease status. This is because cases are more likely to be included - a defining feature of case-control studies. Investigators also selected controls preferentially among women with a hip fracture. Estrogen reduces the risk of hip fracture. The study then compares estrogen use between the myocardial infarction cases and the hip-fracture controls.

# What is this an example of?

# What is this an example of?

Inappropriate Selection of Controls

# Why?

# Why?

Controls that are chosen preferentially among women with a hip fracture doesn’t seem to match the cases if the cases have different selection criteria.

# Let’s visual it!

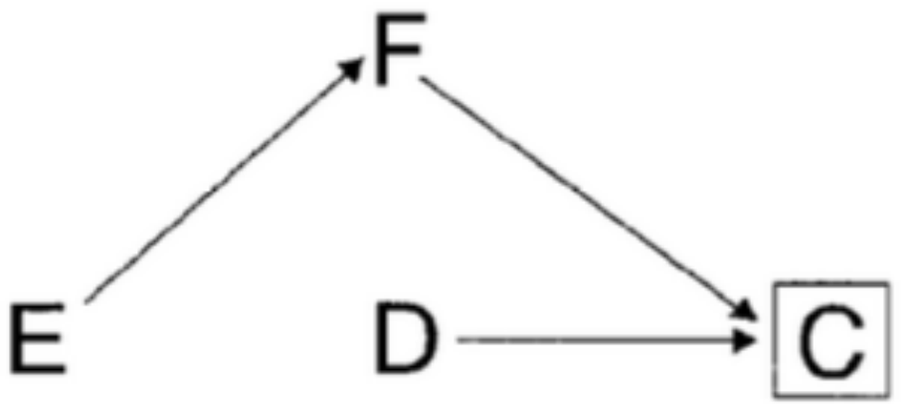



**FIGURE 5.** Selection bias in a case-control study. See text for details.

- **The Nodes**
- E → postmenopausal estrogens
- D → myocardial infarction
- C → whether a woman is selected for the case-control study
- F → women with a hip fracture

Bad controls

Investigators are assessing the impact of antiretroviral therapy on AIDS risk among HIV-infected patients using a publically available dataset. The dataset includes a positive HIV diagnosis, their initiation of antiretroviral therapy, subsequent AIDS development, and their retention status. This data set is limited and does not include data that points to their true level of immunosuppression such as symptoms, CD4 counts or viral load in plasma and analyses are limited to individuals who were retained. The greater the true level of immunosuppression, the greater the risk of AIDS. Previous studies have found that severity of AIDS initially prevents HIV-positive persons from doing every-day activities.

# What is this an example of?

# What is this an example of?

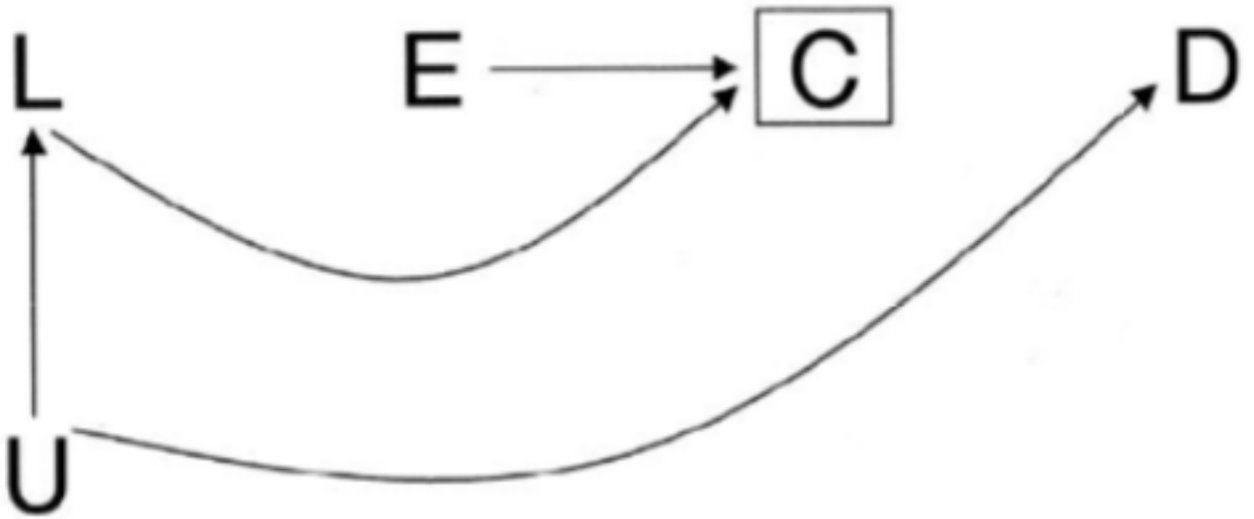
Differential Loss to Follow-up in Longitudinal Studies

# Why?

The more a person is unwell, the less likely they will be to continue to participate in the study and we won’t know what their AIDs status is.

# Let's visual it!

**a.**




- **The Nodes**
  - E → Antiretroviral therapy
  - D → AIDS
  - U → true level of immunosuppression (unmeasured)
  - C → undetermined AIDs status
  - L → presence of symptoms, CD4 count, and viral load in plasma

LTFU

# List of Other Types I

- Nonresponse Bias/Missing Data Bias
- Healthy Worker Bias
- Volunteer Bias
- Censoring by death

# List of Other Types II

- Survivor Bias
  - Occurs when those that survive in a study differ systematically from those that do not
- Index Event Bias
  - Occurs when patients are selected based on the occurrence of an index event (first clinical event)
- Survivor Bias from Competing Risks
  - Analysis is restricted to those who survive long enough to be observed, and survival depends events that preclude the occurrence of the event of interest

# TAKE AWAYS

# LEARNING EXPECTATIONS - LET'S REVIEW

- **DIFFERENTIATE** between type 1 and type 2 selection bias
- **IDENTIFY** methods to address selection bias
- **IDENTIFY** different types of Type 1 selection bias
- **APPLY** knowledge of Type 1 selection bias to a topical example

# What to remember

- Type 1 selection bias is collider stratification bias
- There might be different ways that bias is produced but the mechanism remains the same
- If you have information about the covariates that lie on the noncausal path, you can adjust for it

# How to approach Selection Bias for Comps

- think about selection bias by study design
- know methodologies to address selection bias

# Final Slide

Thank you!
Questions?