Chapter 8: Selection bias

Abdullah Abdelaziz, BPharm, MSc

Important disclaimer

- The term "Selection bias" can refer to different things depending on the discipline
    - Economists' selection bias (selection on observables) is actually epidemiologists' confounding bias.
    - Survey statisticians use selection bias term to sample selection from population, which can lead to biased conclusions in descriptive research.
- Epidemiologists excelled in conceptualizing and making the distinction between confounding and selection.
- It's a good habit to make sure that your collaborators are on the same page regarding terminology to avoid confusions.

General guidelines

General guidelines

- Causal effects are linked to specific **populations**
- In many epidemiologic studies, you end up analyzing a cohort that's different from your original cohort
    - Survival analysis: we analyze uncensored individuals (we don't see the outcome in censored individuals).
    - Case-control studies: we analyze individuals who got the outcome and a sample of patients who did not get the outcome.

General guidelines

- The hope is that the estimate we get in this subset is the same as the one we would've had if we did the estimation on the original cohort.
- If this is not the case   Selection bias
- In all the coming DAGs, selection is represented as a variable with a square around it.
- The square here does not mean statistical adjustment. It means that the analysis is done on one stratum of the selection variable.
  - DAGs are not just useful in helping identifying adjustment covariates. They can help in the design as well.

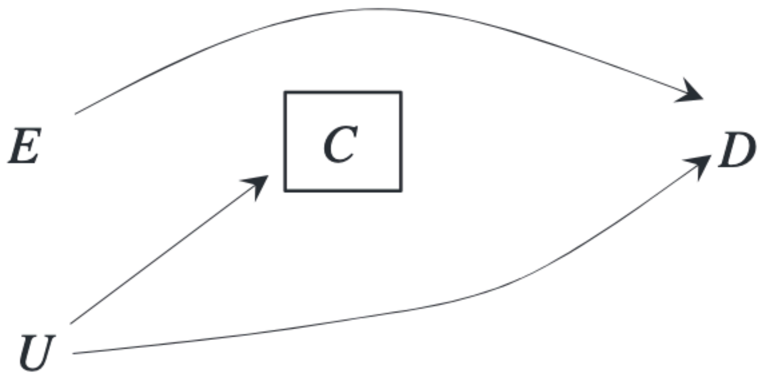
Types of selection bias

Selection bias under the null

- This type of selection bias can arise regardless of the true treatment effect being null or not.
- Conditioning on a collider (or a descendant of a collider) is necessary for this bias to happen.
- This is the topic of this book and Hernan's famous paper in 2004 (Hernán, Hernández-Díaz, and Robins 2004)

Selection bias off the null

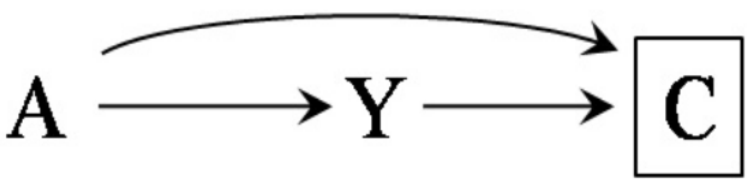
- This type of selection bias **cannot** happen under the null.
- This type of selection bias **can** happen even without conditioning on colliders.



- This bias is not covered in the chapter.

Selection bias in cross-sectional studies

Example 1



- A : Treatment
- Y : Fetal malformation
- C : Live birth

  We don't have Y for dead fetuses, so we essentially **restricting our analysis to** living fetuses.

Example 1

- Our regression will give us this quantity

$$\frac{Pr[Y=1|A=1,C=0]}{Pr[Y=1|A=0,C=0]}$$

- Is this a valid estimate for our estimand?

$$\frac{Pr[Y^{a=1}=1]}{Pr[Y^{a=0}=1]}$$

  The answer is no, because we have association transmitted through the path $A \rightarrow C \leftarrow Y$

Example 2

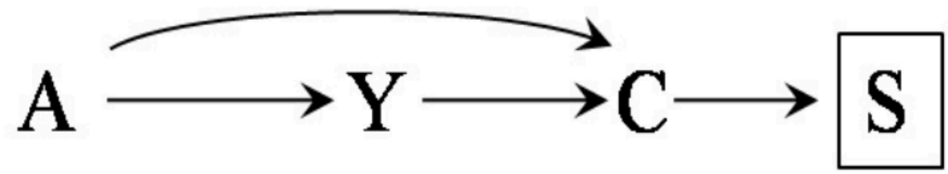


Figure 8.2

- A : Treatment
- Y : Fetal malformation
- C : Live birth
- S: Parental grief

A descendant of a collider is as dangerous as the collider itself.

Selection bias in cohort studies

Including randomized trials

Example

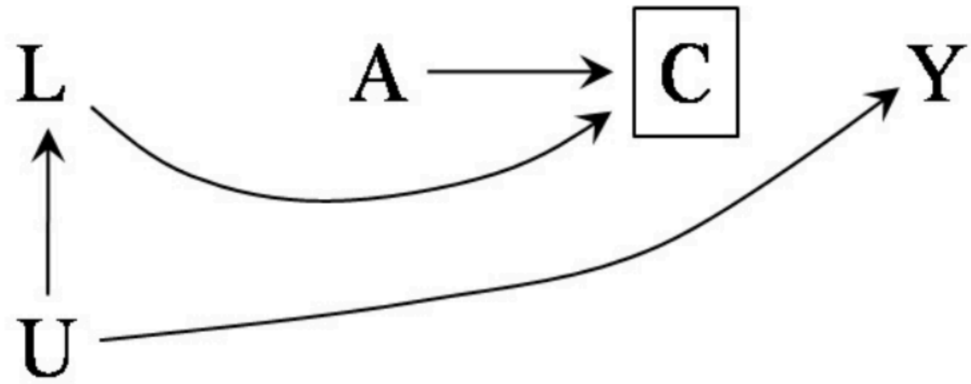


Figure 8.3

- A : Antiretroviral treatment
- Y: Death
- L: Disease severity
- U: High level of immunosuppression
- C: Loss to follow-up

Example

- Remember, C is not a variable we put in a regression model. It's a part of how your **analyzed data was formed**.
- In this example, A can show favorable result not because it's actually effective in reducing mortality, but because it caused sick people to leave the study. Although in reality, A and Y are not associated.
- The previous DAG is an example of selection bias due to **differential loss-to-follow-up** or **informative censoring.**

Similar DAGs for differential loss-to-follow-up

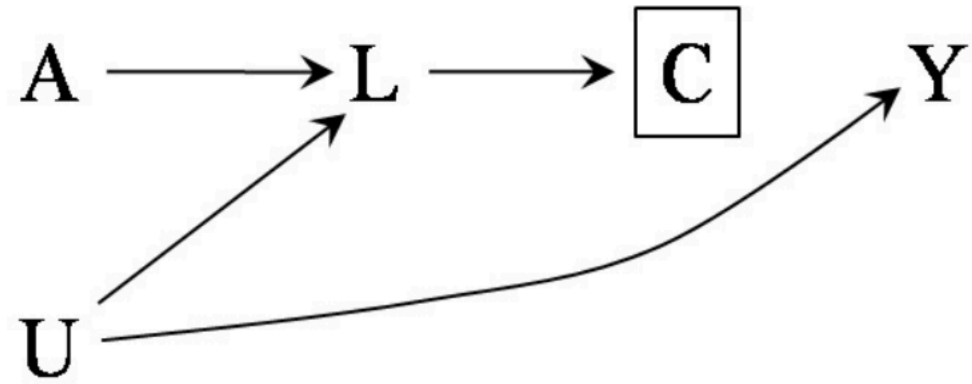


Figure 8.4

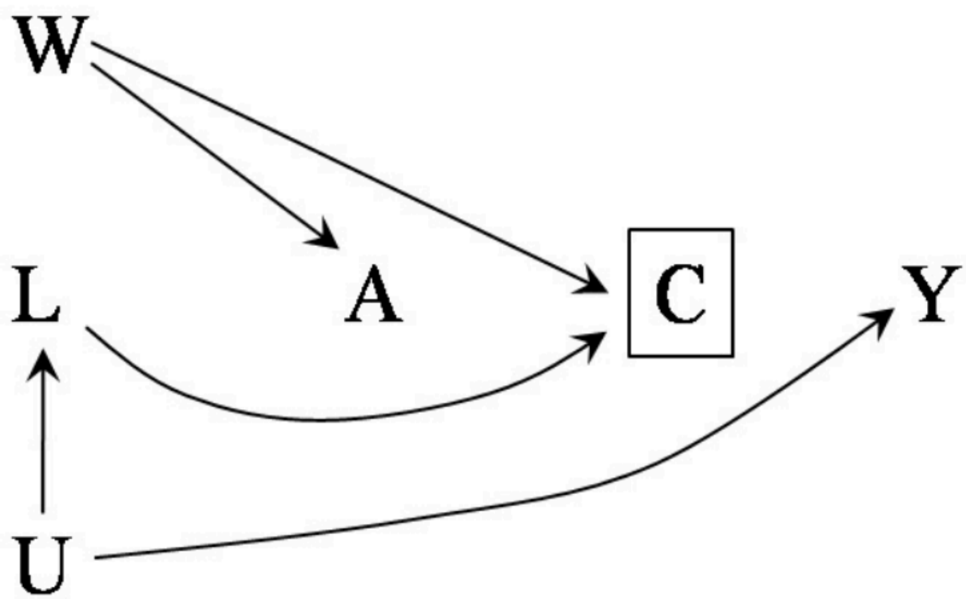


Figure 8.5

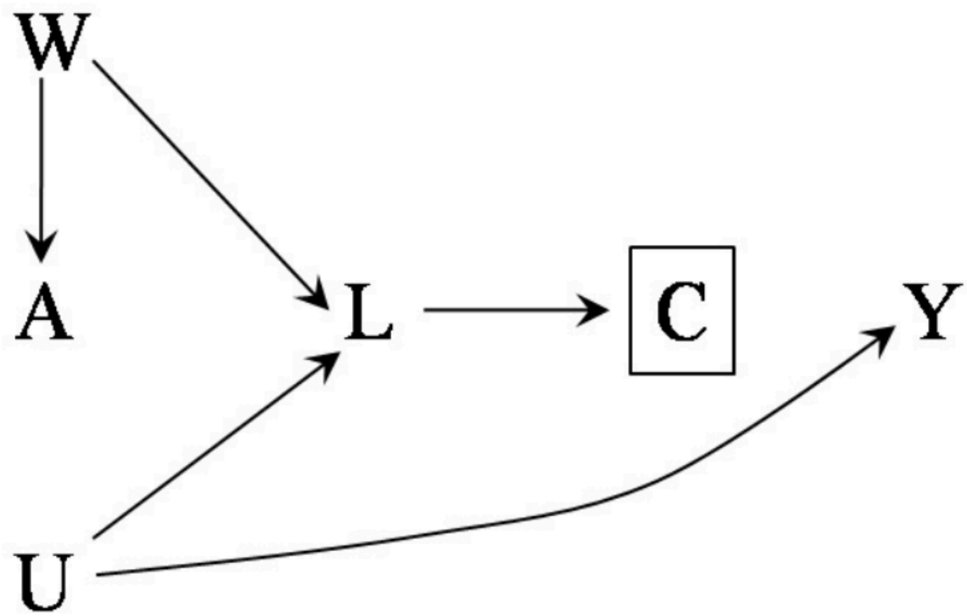
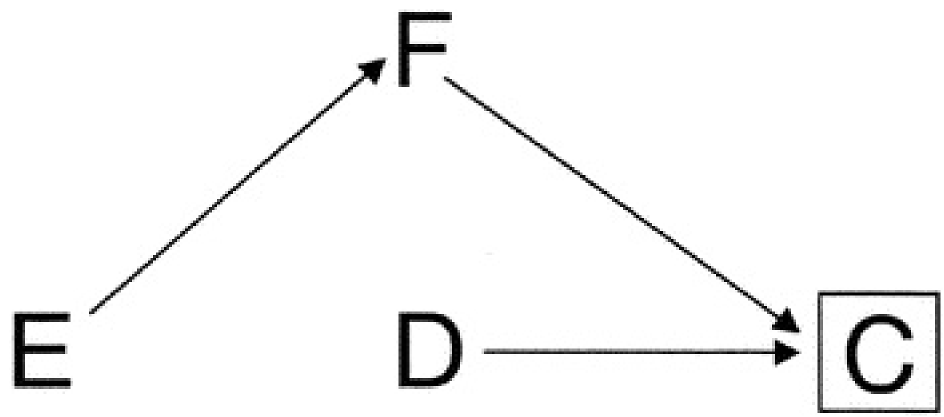


Figure 8.6

- These DAGs are modified versions of Figure 8.3
  - For instance, in figure 8.4, the association between A and L is represented by mediation while in figure 8.5 presented by a backdoor path $A \leftarrow W \rightarrow C$ and presented by both in figure 8.6

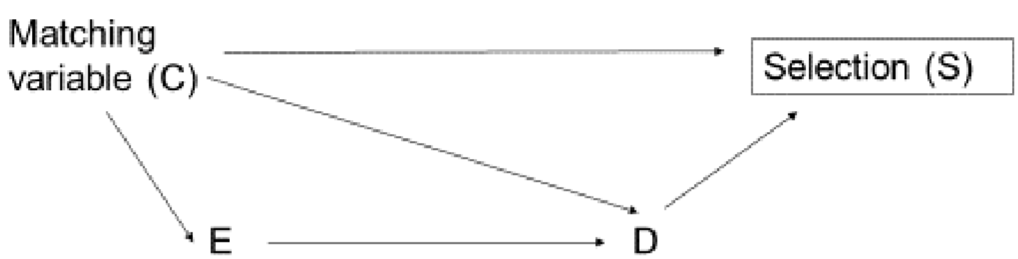
Selection bias in case-control studies

DAG



- E: Estrogen use
- D: CHD
- F: Hip fracture
- C: Selection into the study

Matched case-control designs are inherently biased



The structural definition of selection bias

> **Selection bias** to refer to all biases that arise from conditioning on a common effect of two variables, one of which is either the treatment or a cause of treatment, and the other is either the outcome or a cause of the outcome.

Selection bias, similar to confounding bias, is a violation of the exchangeability assumption.

Common examples of selection bias

- Differential loss to follow-up or informative censoring.
- Missing data bias, or non-response bias.
- Healthy worker bias.
- Self-selection bias or volunteer bias.
- Selection affected by treatment received before study entry aka **prevalent-user bias**
- Immortal-time bias is a mix of selection and misclassification bias.

Which designs are prone to selection bias?

- All of them, even randomized experiments.
- Randomization fixes confounding but not selection.
- Selection bias is more likely to occur with designs that are built on selection by default i.e. case-control design
  - Friendly advice, whenever you have the full cohort, please don't conduct a case-control study.

Which analyses are prone to selection bias?

- Conventional covariate adjustment in treatment-confounder feedback setting.
- Cox regression.

Selection without bias

- RCTs are conducted among volunteers willing to enter the experiment. So those volunteers select into the trial.
- However, this is not what we mean here by selection bias.
- Based on our definition, the selection variable should be a **common effect** of the treatment or a cause of the treatment and the outcome or cause of the outcome.
- Since volunteering participation happened **before** treatment assignment, there is no bias.
- The self-selection bias we mentioned earlier is about agreeing to continue in the trial after being treated.

The distinction between confounding and selection bias

Example

- A: Physical activity.
- Y: Heart disease
- C : Being a firefighter
- L: Parental socieconomic status
- U: Attraction towards physical activity

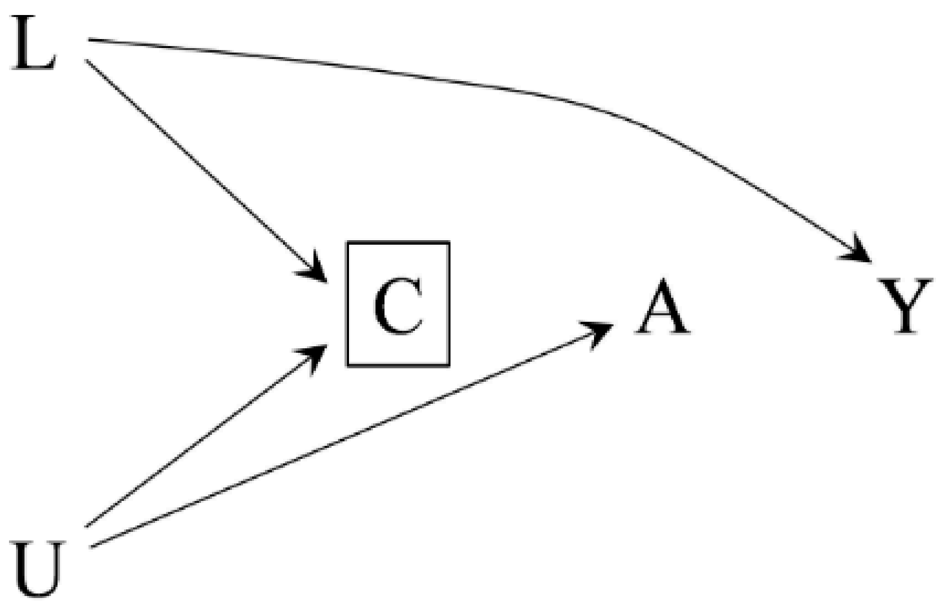
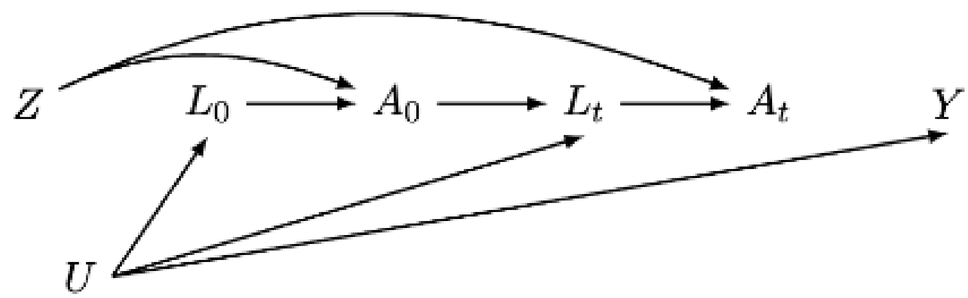


Figure 8.7

Advantages of using the structural approach

1

It can guide the choice of the analytic method



2

It can help is study design and data collection.

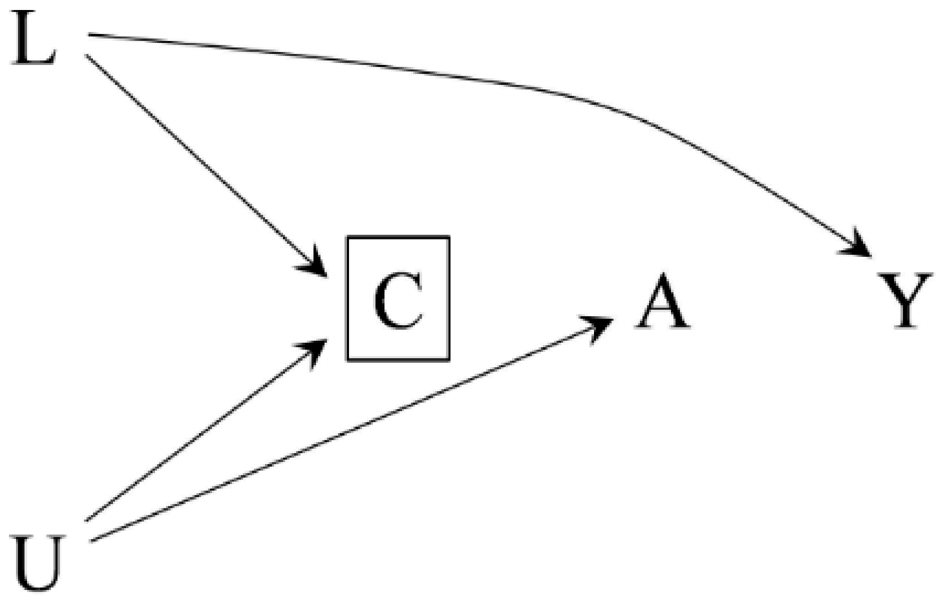


Figure 8.7

3

Selection bias resulting from conditioning on pre-treatment variables (e.g., being a firefighter) could explain why certain variables behave as "confounders" in some studies but not others.

4

Causal diagrams enhance communication among investigators and may decrease the occurrence of misunderstandings.

Important reminder

- DAGs ignore the magnitude or direction of selection bias and confounding.

- It is possible that some noncausal paths opened by conditioning on a collider are weak and thus induce little bias.
- It is not an "all or nothing" issue, in practice, it is important to consider the expected direction and magnitude of the bias

Selection bias in hazard ratios

- A: Treatment (protective)
- Y_1 and Y_2: Death at time 1 and time 2.
- U: Protective Haplotype

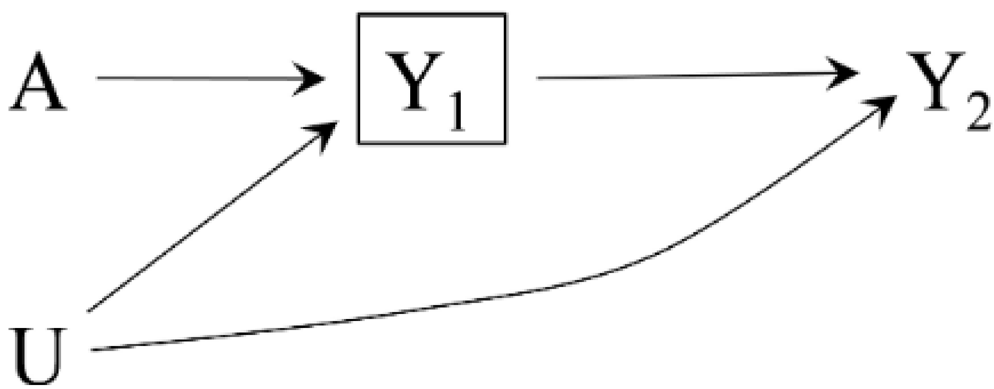


Figure 8.8

Measures of effects

Risk ratio

$$aRR_{AY_1} = \frac{Pr[Y_1 = 1|A = 1]}{Pr[Y_1 = 1|A = 0]}$$

$$aRR_{AY_2} = \frac{Pr[Y_2 = 1|A = 1]}{Pr[Y_2 = 1|A = 0]}$$

Hazard ratio

$$HR_{AY_1} = aRR_{AY_1} = \frac{Pr[Y_1 = 1|A = 1]}{Pr[Y_1 = 1|A = 0]}$$

$$HR_{AY_2} = aRR_{AY_2|Y_1=0} = \frac{Pr[Y_2 = 1|A = 1, Y_1 = 0]}{Pr[Y_2 = 1|A = 0, Y_1 = 0]}$$

In conclusion, we have two issues:

- The estimand changed.
- Selection bias

Avoiding selection bias

New estimand

- Similar to the interaction chapter, we will view selection or censoring as an intervention. If we are able to satisfiy the causal identification assumption with c, then this estimand can be estimated using observed data

$$\frac{Pr[Y^{a=1,c=0} = 1]}{Pr[Y^{a=0,c=0} = 1]}$$

- This reads as the effect of A on Y had everyone got A and remained uncensored vs everyone not getting A and remained uncensored.
- Weighting can be a good approach to achieve this (See example).

References

Hernán, Miguel A., Sonia Hernández-Díaz, and James M. Robins. 2004. "A Structural Approach to Selection Bias." *Epidemiology* 15 (5): 615–25. http://www.jstor.org/stable/20485961.