Epidemiology Basics

Max Salvatore

2022-05-09

Contents

Chapter 1

Introduction

Hello and welcome! This book will contain notes on a variety of epidemiologic topics including study design and biases.

At present, this is a **not a functional book** and the content is **a work in progress** .

The current version of the book has not been reviewed for correctness and references.

Chapter 2

Confounding bias

2.1 Identifying confounding

2.1.1 Data-driven approaches

- There are many data-driven approaches that are used to identify confounding that are based on statistical approaches to examine associations in study data
  - Stepwise regression
    - Throw all suspected confounders into a model and remove those not associated with the outcome (e.g., p > 0.05) in a stepwise fashion
  - Change-in-estimate approach (10% rule)
    - Throw all suspected confounders into a model, retain those whose removal changes the exposure -> outcome effect estimate by > 10%

But confounding is about **causal** relationships, thus it is best to identifying confounding by using causal relationships

- The observed data structure and *a prior* theory or knowledge about the suspected data structure are used to identify confounding
- This is better than stepwise regression or the change-in-estimate approach, which use arbitrary rules based on statistical significance

2.1.2 Structural approach

2.1.2.1 Three criteria

A confounding factor must:

1. Be a cause of the outcome under study
2. Be associated with the exposure under study in the source population
3. Must not be caused by the exposure or disease

2.1.2.2 Criteria 1: Confounding factor must cause the outcome (either directly or indirectly)

- A confounding factor must be a cause of the outcome
  - May be an actual cause of the disease
  - May be a surrogate/proxy or indirect cause of the disease
    - Household income as a surrogate for a milieu of social factors correlated with income
    - Education as a proxy for literacy
  - Prior theory or knowledge (not the data itself) is used to determine the relation of the suspected confounding factor to the outcome

2.1.2.3 Criteria 2: Confounding factor must be associated with the exposure in the source population

- We can generally identify this directly from our data, however varies a bit by design
- Cohort Study
  - Cohort is source population. Therefore, this relationship can be determined from the observed study data.
- Case-control Study
  - In a C-C study, the controls are selected from the source population, however, the control group needs to be very large and have no selection bias or measurement error in order to accurately reflect the association in the source population.
  - External information can be used when available, or prior knowledge.

2.1.2.4 Criteria 3: The factor cannot be caused by the exposure or the disease

- A confounding factor must not be affected by the exposure or the disease.

- Two scenarios can occur here:
  1. The factor that we think is a confounder is actually an intermediate on the causal pathway between exposure and outcome (a **mediator**)
  2. The factor that we think is a confounder is a common outcome of the exposure and outcome of interest (a **collider**)

2.1.2.5 Intermediate on the causal pathway (Mediator)

- Here, low birth weight is on the causal pathway from maternal smoking and perinatal mortality

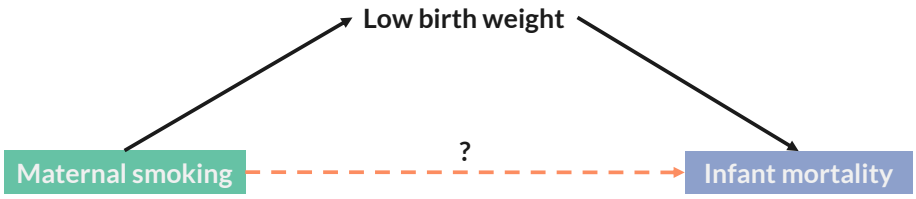

Figure 2.1: Example of mediator

- Low birth weight is a **mediator**. Adjusting for a mediator is referred to as "over-adjustment"

Another example: Fluoridation, Diet sugar, and Tooth decay

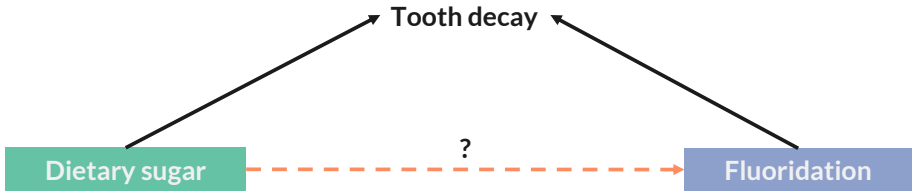

Figure 2.2: Example of a collider

2.1.2.6 Confounding is structural

- Confounding arises because of how, structurally, the variables are related to each other
  - Either how they are naturally related to each other, or, how they are related to each other after an investigator has changed the relationships by adjustmnet, matching, conditioning, etc.

How do we know what the proper structural relationships are?

- Directed Acyclic Graphs (DAGs)
  - NOT a method of data analysis
  - They are used to IDENTIFY confounders based on the assumptions we are willing to make
- DAGs help us to depict the assumed temporal structure of the relationships between our factors
  - Both in the state of nature
  - And after investigators have intervened on the natural structure by conditioning on a set of factors

So

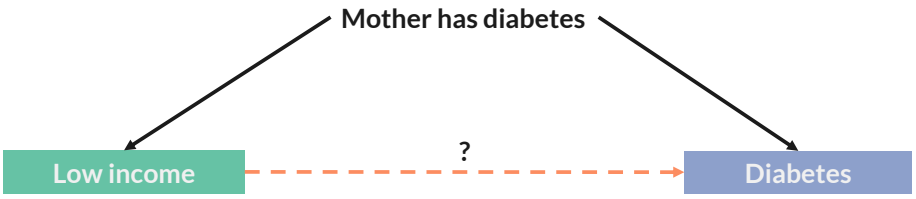

Figure 2.3: Example of confounding

- We want exchangeability of those with low vs. non-low income, condition on having a mother with diabetes (conditional exchangeability)
- The goal is to block all backdoor paths from the exposure to the outcome on the DAG

2.1.2.7 DAG rules for identifying confounding

1. To get from A to Y through a backdoor path, you can move along any path, regardless of the arrow's directionality

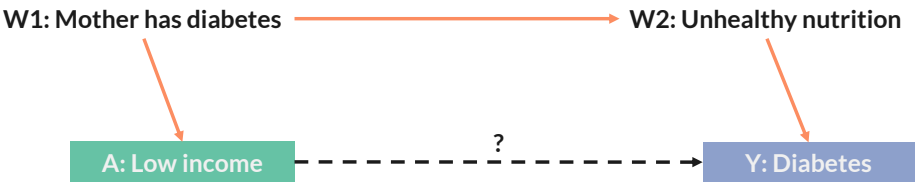

Figure 2.4: An open backdoor path

In the above example (Fig. 2.4), a backdoor path is open through the orange arrows. The effect of the exposure on the outcome is **not** identified.

2. Conditioning on a *common cause of the exposure and outcome* (**confounder**) closes the backdoor path

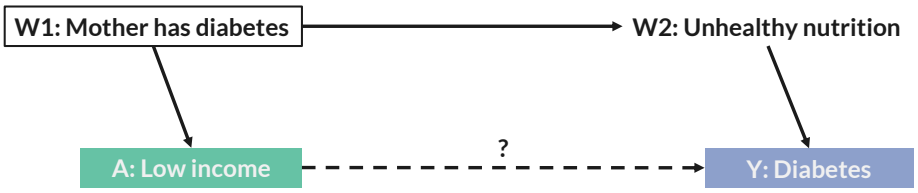

Figure 2.5: Conditioning on confounder W1 closes backdoor path

In the above example (Fig. 2.5), a backdoor path starting from A is blocked at W1. The effect of A (the exposure) on Y (the outcome) **is** identified.

3. **Unmeasured** factors (U) may still lead to confounding, even if you closed the backdoor path through measured factors

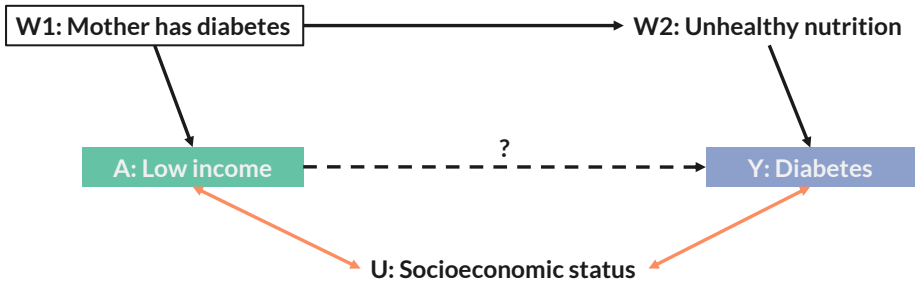

Figure 2.6: Unmeasured factors (U) may still lead to confounding

In the above example (Fig 2.6), a backdoor path is blocked at W1, but it is open through U. The effect of A on Y is **not** identified.

4. The existence of a collider will block the backdoor path

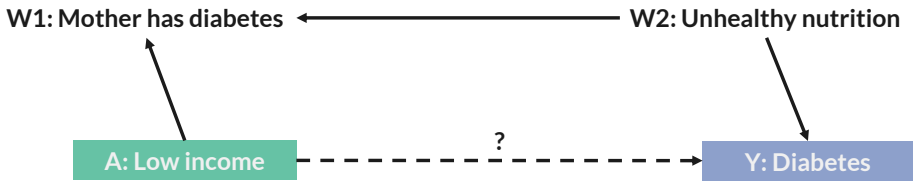

Figure 2.7: Colliders (W1) block backdoor paths

In the above example (Fig 2.7), a backdoor path starting from A is blocked at W1. The effect of A on Y **is** identified.

5. Condition on a collider will open the backdoor path

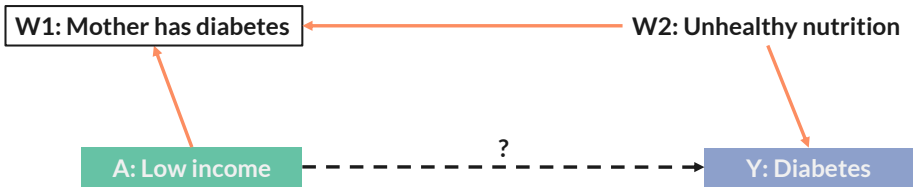

Figure 2.8: Colliders (W1) block backdoor paths

In the above example (Fig 2.8), a backdoor path starting from A is opened by condition on W1. The effect of A on Y is **not** identified.

6. Conditioning on a descendant (outcome) of a collider will open the backdoor path

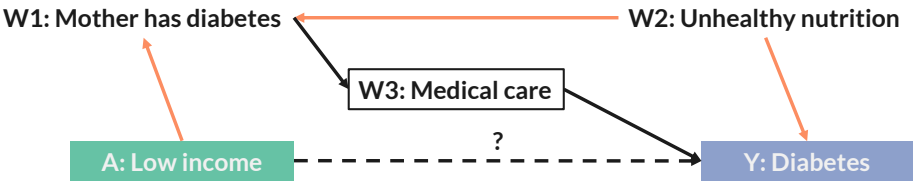

Figure 2.9: Conditioning on descendant of collider opens backdoor path

In the above example (Fig 2.9), a backdoor path starting from A is opened by conditioning on W3. The effect of A on Y is **not** identified.

2.1.2.8 Quantifying confounding

- Non-collapsibility of strata
  - When the association of exposure and outcome is different across the strata of a third variable identified as a confounder and the crude (non-stratified) association, then the data are not collapsible and confounding is present
  - To be continued….

Chapter 3

Selection bias

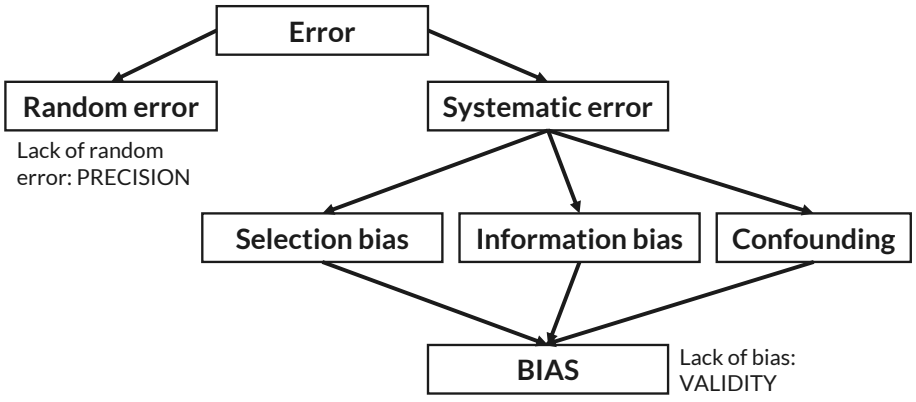

Figure 3.1: Types of error

There are three broad types of systematic error in epidemiologic studies: selection bias, information bias, and confounding.

Random error differs from systematic error in that its error gets smaller as the sample size (`n`) gets larger. Systematic error does not get better with larger `n`. Selection bias is a type of systematic error.

3.1 Selection bias definitions

Selection bias is:

- Distinct from confounding and information bias because of its mechanisms

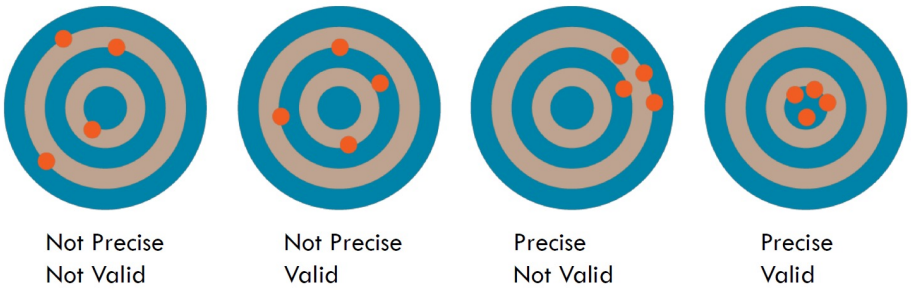

Figure 3.2: Visualization of precision and validity

- Similar to confounding because **exchangeability** is violated
  - $Pr[A = a]$ is not independent of Y^a

3.1.1 Definitions

- Traditional: "Selection bias is present when individuals have different probabilities of being included (or retained) in a study sample according to relevant characteristics, namely the exposure and outcome of interest." (Szklo & Nieto, 2019)
- Structural: "Bias resulting from conditioning on a common effect (a collider) of two variables, one of which is the exposure or associated with the exposure and the other is either the outcome or associated with the outcome." (adapted from Hernán, Hernández-Díaz, & Robbins 2004)

3.1.2 Selection bias in DAGs

3.1.2.1 Brief review of paths

- $A \rightarrow Y$ (A causes Y)
- $A \leftarrow C \rightarrow Y$ (A and Y share a common cause; aka confounding)
- $A \rightarrow \boxed{\text{S}} \leftarrow Y$ (A and Y share a common effect; S is a collider)
  - We must condition on the collider S (adjustment or restriction) or a on a descendant of a collider for us to detect a statistical association between A and Y in this scenario

3.1.2.2 Selection bias DAG examples

There are some more complex ways that selection bias can be captured in a DAG.

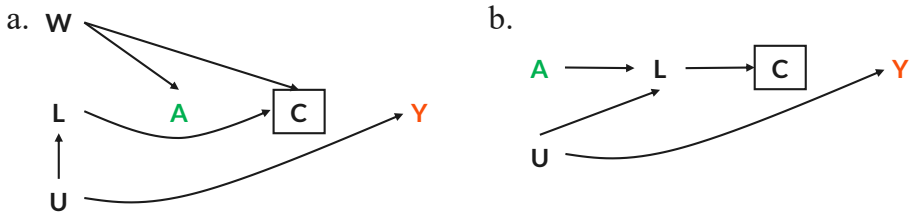

Figure 3.3: Examples of selection bias in DAGs

In Fig. 3.3.a., conditioning on the collider, C, opens a backdoor pathway between the exposure, A, and the outcome, Y. This path is $A \leftarrow W \rightarrow \boxed{C} \leftarrow L \leftarrow U \rightarrow Y$.

In Fig. 3.3.b., conditioning on C, the descendant of the collider, L, opens a backdoor path.

Here is a motivational example that displays the intuition behind a collider.

Rain $\rightarrow$ Wet sidewalk $\leftarrow$ Neighbor's sprinkler

Take, for example, the question: did it rain last night? Suppose we only observe when the sidewalk is wet. The sidewalk could be wet for two reasons: (1) it rained or (2) your neighbor ran the sprinkler. If we know the sidewalk is wet and that the neighbor's sprinkler is broken, then it probably rained. Knowing information about one of the causes of a collider gives us information about the other.

3.1.2.3 Traditional vs. structural definition

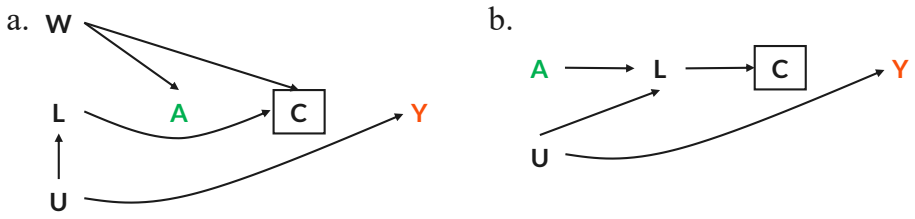

Figure 3.4: Examples of selection bias in DAGs

The above figure graphically depicts the traditional (Fig 3.4.a) and structural (Fig 3.4.b) definitions of selection bias.

3.1.3 Examples of selection bias

- There are many ways a study can be subject to selection bias

- Learning about well accepted types of selection bias can help with finding it in our own studies
- Differs somewhat by study design: case-control, cohort, and RCT

3.1.3.1 Selection bias in case-control studies

- "Berkson's bias"
  - Particularly relevant for hospital-based case-control studies
  - Occurs when **controls are not selected independent of exposure**
  - Case-control studies are thought to be particularly susceptible to selection bias because at the very least $Y \rightarrow \boxed{\text{S}}$

Take the DAG in Fig. 3.4.b to represent a hospital-based case-control study of the malnutrition (A) on depression (Y). Because of the study design, those with depression (Y) are more likely to be included in the study (C). Malnutrition by itself is likely to cause somebody to be admitted to the hospital (C). This is called "Berkson's Bias."

3.1.3.2 Selection bias in cohort studies

- Selection bias in cohort studies typically arises because of **loss to follow-up** or mortality related to both the outcome and the exposure.

Take the DAG in Fig. 3.3.a to represent a cohort study of occupational exposure (A) on risk of stroke (Y). The terrible work conditions of the job (W) is a common cause of exposure (A) and likelihood that a person will quit (C). Underlying health status (U) is a cause of stroke (Y) and causes a person to quit (C) through deteriorating physical health (L). This is also called the "healthy worker effect."

3.1.3.3 Selection bias in cross-sectional studies

- Cross-sectional studies are also susceptible to selection bias.
- Sometimes this is referred to as "incidence-prevalence bias."
- Prevalent cases with better prognosis or underlying health are more likely to show up in your study.

Take the DAG in Fig. 3.4.b to represent a study of the effect of folic acid (A) at conception on the prevalence of birth defects (Y). Only those babies born were included in the study (C).

3.1.3.3.1 Selection bias in randomized control trials (RCT)

- RCTs are often thought to be the "gold standard" of causal inference
- While they are less likely to be subject to confounding, they are equally likely to be subject to selection bias due to loss to follow-up.

Take the DAG in Fig. 3.3.b to represent a study of AZT (A) on development on AIDS (Y). Treatment A and illness severity (U) both cause side effects (L) which leads to dropout (C).

3.2 Key points

- Selection bias is a type of systematic error related to recruitment or retention of participants.
- Recruitment or retention must be related to exposure and outcome to cause bias.
- We can visualize selection bias on DAGs.
- All types of studies are subject to selection bias but it might look different depending on the study.

Chapter 4

Literature

Here is a review of existing methods.