Chapter 9: Evaluating Causality with Observational Studies

Investigation: Since tobacco is a known carcinogen, vaping has been touted as a much safer alternative to smoking. But there is still a lot we don't know about the long-term effects of vaping. https://news.cancerresearchuk.org/2021/04/26/e-cigarettes-what-we-know-and-what-we-dont/

You are part of a medical research team exploring potential long-term dangers that might be caused by vaping. In particular, you would like to study whether e-cigarette vapor may directly increase the risk of lipoid pneumonia—a chronic condition that leads to asthmatic reactions and chronic coughing.

How might you realistically collect data that would help you determine if vaping actually ***causes*** an increased risk in lipoid pneumonia?

First: Design a study in which you have **no ethical constraints**. How might you best design this study in order to determine causality. *Jot down some ideas here!*

People might propose a randomized controlled design. Assign some people to vape and see what proportion get pneumonia. Compare to people who don't vape. Or some might propose a before and after. Likely though, there should be a theme of intervention.

Second: Design a study in which you **do** have ethical constraints. Nobody can be forced to complete anything they don't wish to. How might this change your design? *Jot down some ideas here!*

We can't force people to vape, so we may be limited to an experiment of people who were initially willing or ok with vaping (but then is it ethical to prevent them from vaping?).

Or, an observational design—we study people who choose to vape and people who don't. But these groups may be different in other ways. Folks may or may not suggest some idea of stratifying for other things

Save room for additional notes down here!

Depending on conversation, this space could be used to make a comparison between experimental design vs. observational study design. People may list possible confounders

Experiments	Observational Studies
Designed to identify *causal* relationships.	Identify *associations* that may signal causation
In Experiments, we have the power to… assign units to an intervention.	In Observational Studies, we can only… observe different variables and identify relationships.

Why aren't all studies experimental?

For each design below, consider whether you would address this investigation with an experiment or an observational study. If choosing observational study, why?

1. Do high levels of alcohol consumption during pregnancy increase the risk of premature birth?

Possible, but probably not ethical since we know alcohol consumption can be harmful

2. Does autism for teenagers affect their academic success and chances for college?

Not possible. We cannot assign teenagers to have or not have autism

3. Does a new therapy approach to improving mobility after surgery decrease time to full recovery as compared to standard therapy approaches?

Ethical if we are confident it's at least as good as what we have! We could assign people to new therapy or current therapy

4. Does eating more dairy increase the chance that a woman will conceive twins rather than a single fetus?

Possible, and not unethical (for most women), but it would be difficult to observe without a large sample

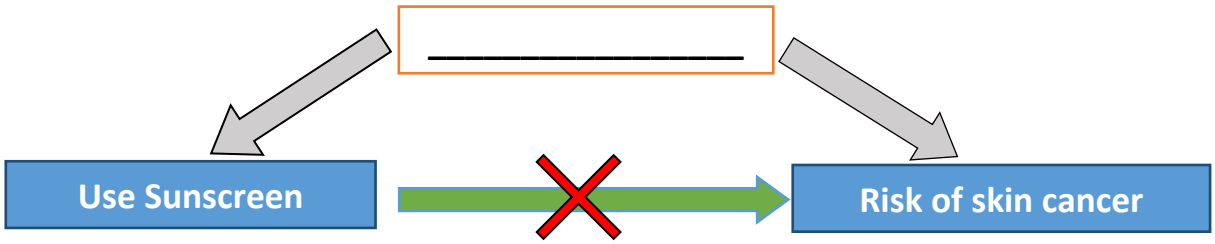
Reasons for completing an observational study

- It may be impossible or extremely difficult to assign participants to an intervention
- It may not be ethical to assign participants to an intervention if it increases risk of harm
- In special cases, the response being studied might be rare and difficult to reproduce without gathering a very large sample or waiting a very long time.
- Experiments are generally more expensive and may require time and extensive planning.

Evaluating Causality - Confounders and Mediators

- **Direct causation:** An explanatory variable directly causes changes in a response variable
- **Indirect causation:** An explanatory variable begins a causal chain, facilitated through one or more mediating variables, that will affect the response variable
- **Association without causation:** An explanatory variable is merely associated with a response variable. There is likely a confounding variable that explains why both outcomes occur together

Example of Confounding Variable. Consider a medical study to examine factors that might lead to melanomas (skin cancer). One researcher notes that people with melanomas were much more likely to have reported using sunscreen in the last year. Does that mean that sunscreen is causing skin cancer? What confounders explain this association?



Example: Somebody makes the observation that "**using a tanning bed**" may increase risk of skin cancer. Might that fit as a confounder to this relationship?

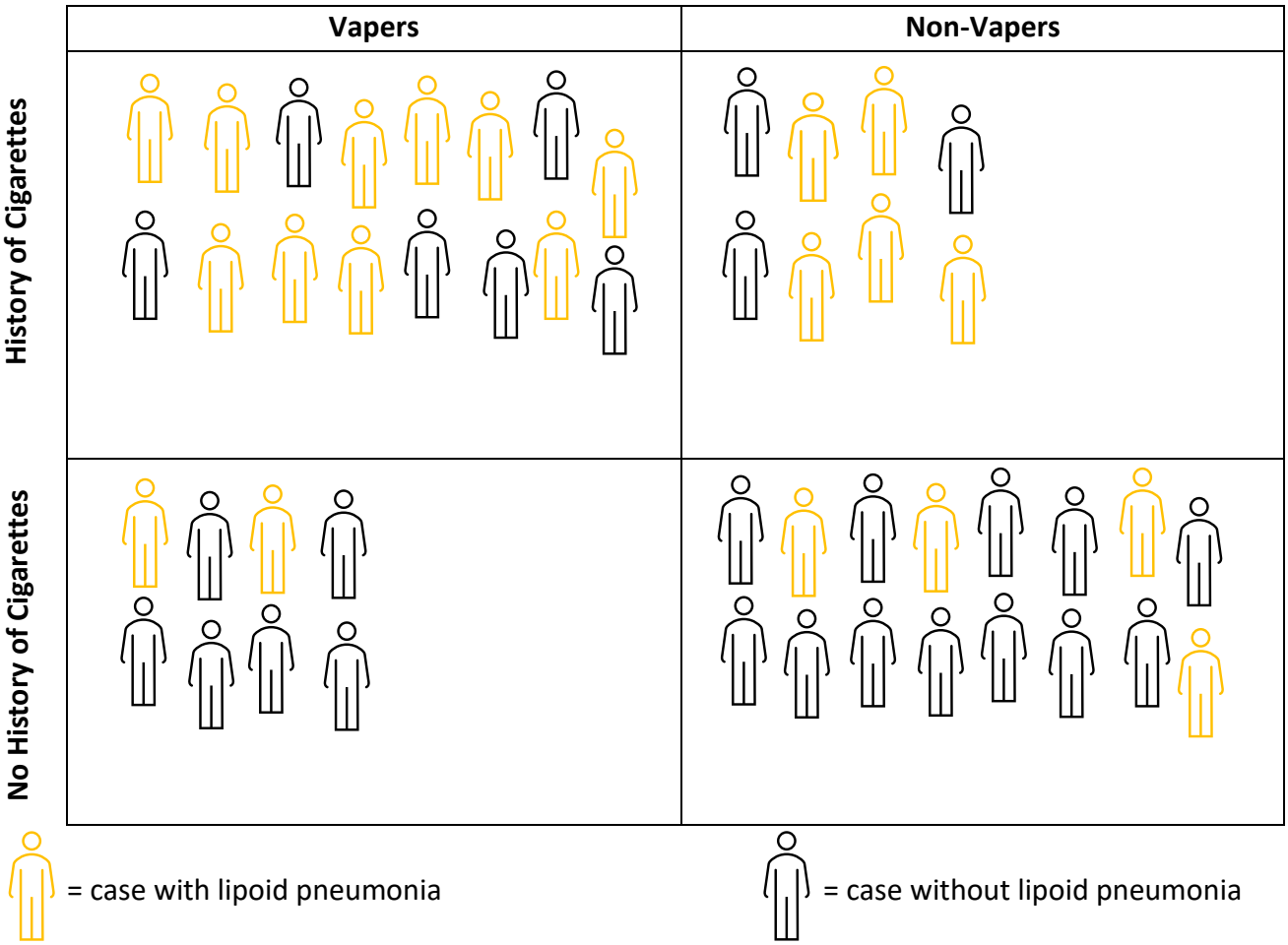
- In order for a variable to be a true confounder, it must be…
  - Truly causing (directly or indirectly) changes in the response variable
  - Be linked to the explanatory variable, but not necessarily in a known causal way

Example of Mediating Variable: People who earn more income tend to have longer lives. Does that mean that money itself is directly increasing lifespan? What mediates that relationship?

Stratification - Controlling for Confounders in Observational Settings

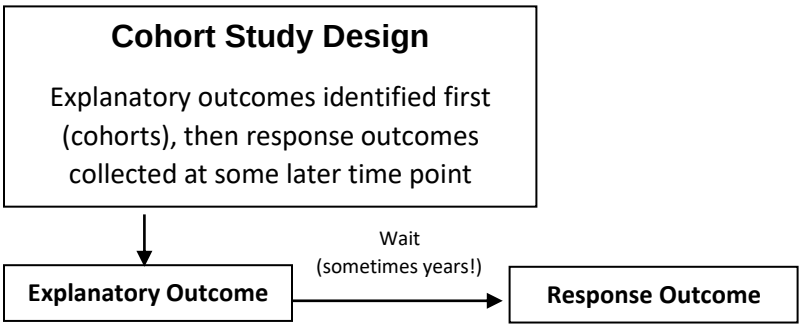
Investigation Revisited: Using an observational study design, we recruited vapers and non-vapers and observed whether vapers had a higher likelihood of a lipoid pneumonia diagnosis. One possible confounder to this relationship is a history of smoking. Let's draw a picture of our confounder diagram to represent that!

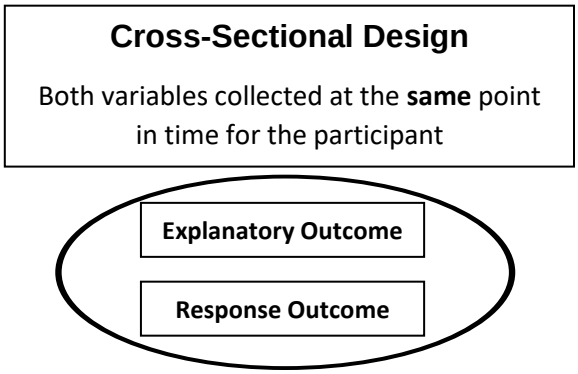
Stratification is the analytical process of breaking down our comparison groups (e.g., vapers and non-vapers) into smaller subgroups based on a confounder. Then we can see if their response outcomes are still different after "controlling for" differences based on the confounder.

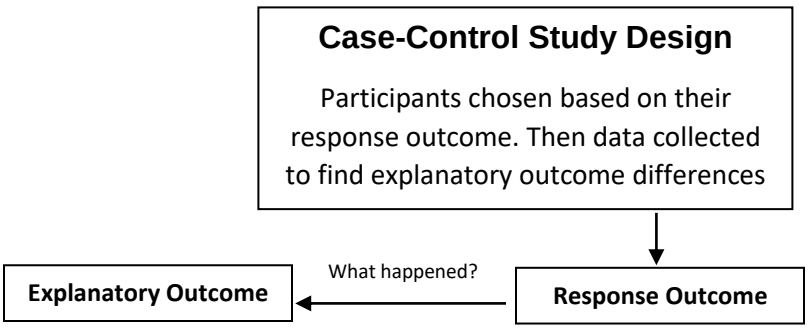


Different Observational Study Designs

- Different observational designs lend themselves to different advantages/disadvantages, and different analytical options! These design differences hinge on whether we collect the response variable or explanatory variable data at different time points. https://emj.bmj.com/content/20/1/54
    - **Cross-sectional Studies**
        - Cross-sectional studies collect both the explanatory and response outcome data for a single point in time. It's **data at a cross-section** of a participant's life.
        - We might use a survey to ask about one's vaping status and about current known health conditions.
    - **Cohort Studies**
        - Cohort studies involve identifying explanatory outcomes first, and then collect response outcomes at some later point in time—often because we need to wait and see!
        - Cohort studies are typically prospective in form, meaning that the response variable data is not available until a later time when we collect it.
        - We might identify vapers and non-vapers first, then wait several years to see if any differences emerge with their health.
    - **Case-Control Studies**
        - In case-control studies, researchers identify people who have had a certain response, and then look to see if there are any explanatory outcomes that differ.
        - Case-Control studies are typically retrospective in form, meaning that the explanatory variable data is not available until we collect it later.
        - We might identify people with lipoid pneumonia and compare them to similar people without lipoid pneumonia. Perhaps a history of vaping might explain the difference!







Optional Background Reading

Analytical Differences between Case Control and Cohort Designs

Example: Extensive research has found a link between smoking and lung cancer. It is estimated that…

- Approximately 15% of people who have smoked more regularly will develop lung cancer
- Approximately 0.5% of people who have smoked little to no cigarettes will develop lung cancer

Thus, the risk for lung cancer among smokers relative to non-smokers is… RR = $\frac{0.15}{0.005} = 30$

But let's say for now that we didn't actually know what the difference in risk was and we wanted to complete an investigation to more accurately estimate the risk for lung cancer in smokers relative to non-smokers.

Unit of observation: **one person**

Response variable: **Presence of Lung cancer**

Explanatory variable: **Status as regular or non-regular smoker**

Cohort or Cross-sectional Design: We could collect data that preserves the **natural incidence** of lung cancer. This might involve a "cross-sectional" survey, or a prospective "cohort study" where we sample people who have smoked or not, and then report the natural incidence rate of lung cancer in each group.

Let's say we identified 200 people regular and 200 non-regular smokers. We *might* get a sample like this:

We can find a 95% confidence interval.
https://istats.shinyapps.io/Association_Categorical/

$$\frac{26/200}{1/200} = 26\ (3.56, 189.76)$$

	Cancer	No Cancer	Totals
Smoker	26	174	**200**
Non Smoker	1	199	**200**
Totals	**27**	**373**	**400**

But this interval is very wide, reflecting our uncertainty about the true risk in the non smoking group. But it's not inaccurate—the true RR of 30 is not that far from our actual RR and is definitely in the interval!

Case Control Design: Since lung cancer is quite rare in our comparison group, maybe we could directly find 200 people with lung cancer and compare to 200 people without lung cancer, and then find out about their history of smoking. Given that 11.5% of U.S. residents are regular smokers, and given the known rate of lung cancer for each group, we *might* see a result like this:

Let's find the relative risk for lung cancer in this scenario and report the 95% confidence interval.
https://istats.shinyapps.io/Association_Categorical/

$$\frac{26/200}{1/200} = 4.52\ (3.46, 5.90)$$

	Cancer	No Cancer	Totals
Smoker	155	18	**173**
Non Smoker	45	182	**227**
Totals	**200**	**200**	**400**

This interval is much narrower, but…it's inaccurate! It's not even close to the true RR of 30.

The issue with the Case Control Design is that we no longer have "natural incidence sampling." That means that our response outcomes are not proportionally representative of the true risk to the population! But there is an analytical option we could try here:

- Introducing "Odds"
  - Risk is simply the probability of an adverse event occurring.
  - "Odds" also assesses the likelihood of an adverse event occurring, but it's constructed slightly differently than a simple probability.

Risk = $\mathrm{P(outcome)} \approx \frac{\#\ \text{Cases with}}{\text{Total}\ \#\ \text{cases}}$ **Odds =** $\frac{\mathrm{P(outcome)}}{\mathrm{P}(not\ outcome)} \approx \frac{\#\ \text{Cases with}}{\#\ \text{Cases without}}$

Relative Risk (RR) = $\frac{\mathrm{Risk_A}}{\mathrm{Risk_B}}$ **Odds Ratio (OR) =** $\frac{\mathrm{Odds_A}}{\mathrm{Odds_B}}$

The construction of an odds ratios allows it to proportionally balance out the incidence bias in our response outcomes. As a result, we should get an odds ratio that generally mirrors the true relative risk!

OR = $\frac{155/18}{45/182}$ = 34.83 (19.4, 62.5)

Odds Ratios vs. Relative Risk (this you should know!)

- ✓ In low incidence situations, you need very large samples to detect effects
- ✓ Case-control designs are an efficient option that doesn't require an *enormous* sample size, but in case-control designs, RR cannot be calculated accurately. But OR can be validly measured!
- ✓ An OR will ***exaggerate*** the effect in comparison to relative risk, but the **larger the sample size**, the closer OR will be in approximating RR.
  - RR will always be closer to 1
  - An OR is still valid in other designs, but RR is often preferred when appropriate.

Advantages and Disadvantages of Cohort and Case-Control designs

- Since Cohort studies allow for extended observation, researchers can monitor response outcomes when they happen and how they happen. As a result, they can help researchers better construct causality arguments.
- Case-control studies are advantageous in cases where time is short, or when the response outcome is a rare incidence situation. We can directly identify people with this rare outcome.
- To learn more about these design types and some of their specific advantages or disadvantages, check out this excellent article: https://emj.bmj.com/content/20/1/54

Chapter 9 Additional Practice

Practice: A study finds that people who carry lighters have a higher rate of lung cancer. Consider the following explanations and whether it is framed as a mediator, a confounder, or neither. Consider drawing a diagram of each to show what is affecting what.

Genetics—some people are more genetically prone to lung cancer than others.

1. Mediator
2. Confounder
3. Neither

Smoking cigarettes—people who smoke cigarettes have a higher rate of lung cancer and are also more likely to carry lighters

1. Mediator
2. Confounder
3. Neither

Lighter fluid—inhaling the fumes from lighters causes lung damage that leads to cancer

1. Mediator
2. Confounder
3. Neither

Radon—radon exposure raises one's risk for lung cancer

1. Mediator
2. Confounder
3. Neither

Identify whether each design below is an observational study or an experiment. If obs study, what type?

A survey conducted to college students asks whether they have a consistent bedtime on weeknights. This survey also asks how many hours of sleep they get a night. The team is curious if people who set a regular bedtime also get more sleep.

In another variation of this investigation, researchers took a group of students who did not set a regular bedtime and randomly chose some of them to choose a regular bedtime for 2 weeks. The others continued with life as normal. At the end of 2 weeks, the researchers compared the sleep amounts of those who stuck with the regular bedtime to those who continued without any change.

To determine how effective masks were in preventing the spread of COVID-19 in 2020, researchers identified cities that implemented a mask mandate and cities that did not. They then tracked the percentage of residents in each city who contracted COVID-19 over the following 4-month period.

A group of cardiologists identified patients with diagnosed heart disease. The researchers then looked back at medical records to determine which were prescribed a particular aspirin that the researchers suspected might have links to heart disease.

Chapter 9 Learning Goals

After completing this chapter, you should be able to…

- Distinguish an observational study design as one that observes variable relationships, but without the power to assign units to different interventions.
- Understand reasons why we might complete an observational study rather than an experiment.
- Recognize observational studies as having generally weaker claims to causality due to the difficulty of controlling for all possible confounders.
- Identify a confounder as a variable that explains why two variables are associated, but not causally linked.
- Identify a mediator as a variable that facilitates a causal chain between two variables.
- Recognize stratification as the analytical technique of controlling for a possible confounder to better infer causality in an observational study.
- Distinguish cross-sectional designs, case-control designs, and cohort designs as three different observational study options.
- Identify the efficiency advantage of a case-control design and the causality advantage of a cohort design.
- Relative risk vs. Odds ratios
    - Know that relative risk calculations are inaccurate in case-control data.
    - Recognize odds ratios as a replacement for relative risk slightly exaggerate the risk (farther from 1) compared to relative risk, but that approximate well in larger sample settings
- Determine if a study is observational or experimental.
- For an observational study described in context, distinguish what variables may be useful to control for and which may not be particularly useful to control for.