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.

You are part of a medical research team exploring potential long-term dangers that might be caused by vaping. You specifically want 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 will 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!*

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!*

Save room for additional notes down here!

Table 1. Experiments vs. Observational Studies

Experiments	**Observational Studies**
Designed to identify *causal* relationships. In Experiments, we have the power to…	Identify *associations* that may signal causation. In Observational Studies, we can only…

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?

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

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

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

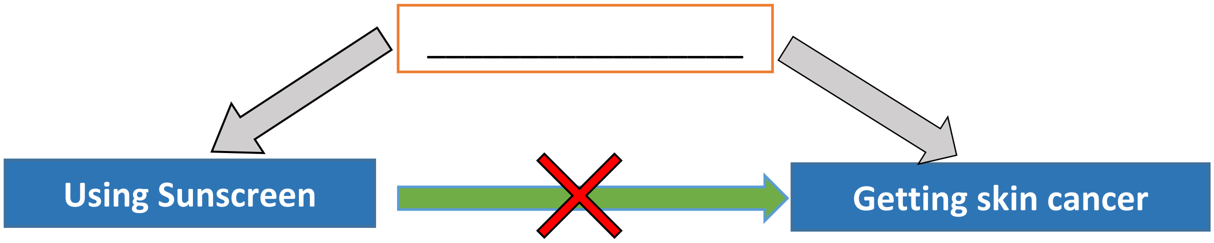
Reasons for completing an observational study.

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

Modeling Variables

- Theoretical modeling is a key part of science—it is the process of thinking about how _____________ ____________ to one another in some kind of process or system.
- Typically, when we propose a theoretical model, we are trying to identify ___________ mechanisms.
- Gathering data to examine those variables can increase our confidence as to whether our model may be correct! But we need to be cautious—especially with observational study data.
  - Two outcomes may be correlated, but not causally linked. This might be because a _________________ explains why the two are likely to occur together!
  - Two outcomes may be causally connected through a ______________ variable. But if the mediating variable is disrupted, the causality chain breaks.

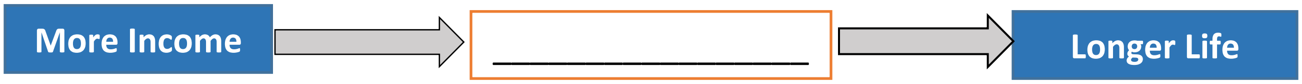
Example of Confounding Variable. Consider a medical study to examine factors that might lead to melanomas (skin cancer). One researcher notes that people with melanomas were much more likely to have reported using sunscreen in the last year. Does that mean that sunscreen is causing skin cancer? What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



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

- For a variable to be a true confounder, it must be…
  - Truly causing (directly or indirectly) changes in the ________________________
  - Be ____________ 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 mediator could we fit into this theoretical model?

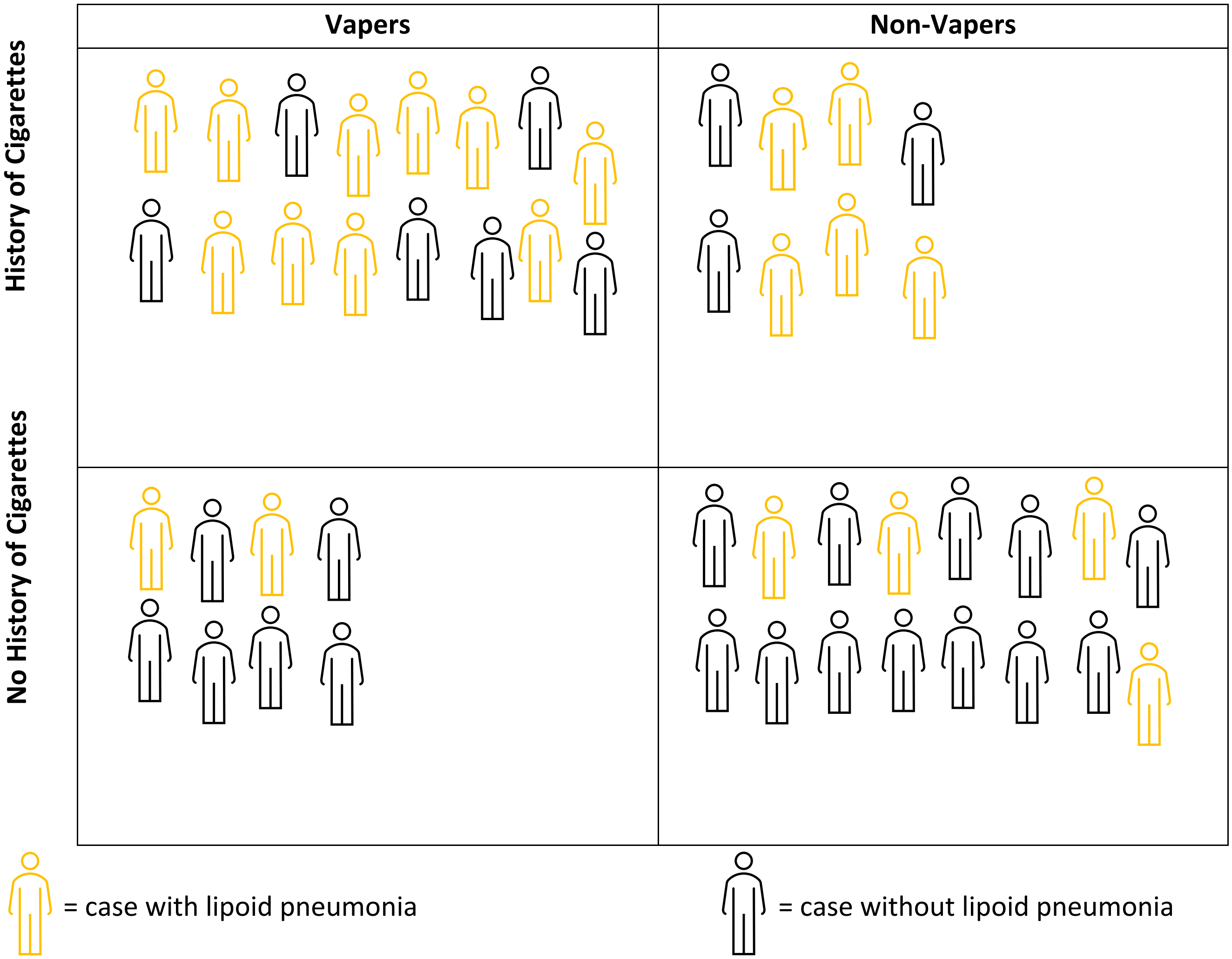


Stratification - Controlling for Confounders in Observational Settings

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 "history of smoking." Let's draw a picture of our confounder diagram to represent that!

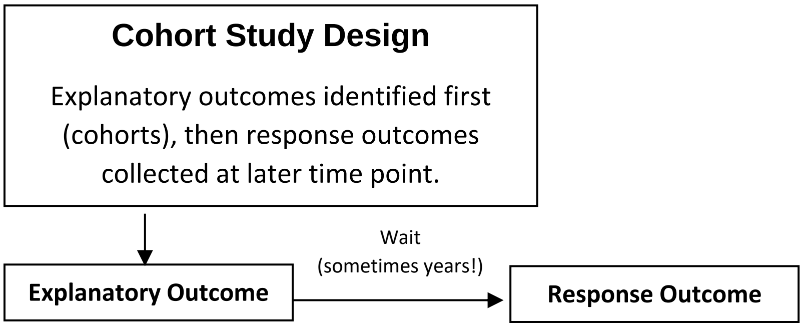
Stratification is the analytical process of breaking down our comparison groups (e.g., vapers and non-vapers) into smaller subgroups based on a potential confounder. Then we can see if their response outcomes are still different when making these subgroup comparisons!

Table 2. Visualizing stratification



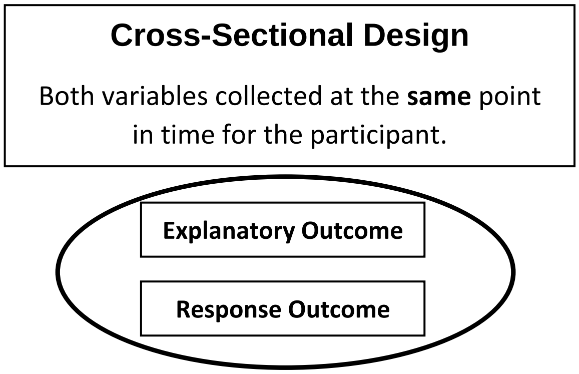
Different Observational Study Designs

- 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.
  - **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** ___________________ 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 collecting response outcomes at some later point in time—often because we need to wait and see!
    - Cohort studies are typically __________________ 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 ____________________ 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!



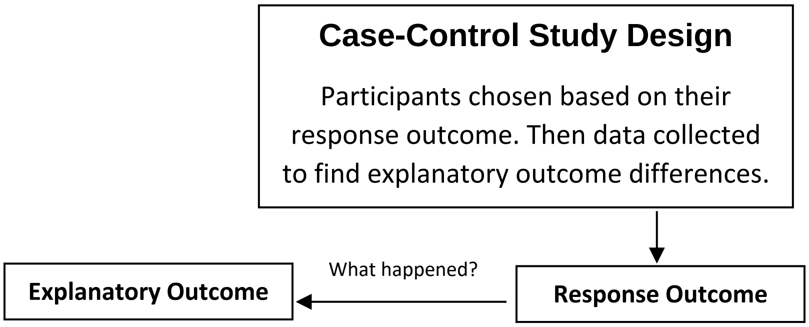
Cohort Study Design

Explanatory outcomes identified first (cohorts), then response outcomes collected at later time point.



Cross-Sectional Design

Both variables collected at the **same** point in time for the participant.



Case-Control Study Design

Participants chosen based on their response outcome. Then data collected to find explanatory outcome differences.

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 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:

Table 3. Smoking and Lung Cancer (cross sectional or cohort)

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

We can find a 95% confidence interval.

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

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!

Table 4. Smoking and Lung Cancer (Case Control)

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

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 them 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.

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

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

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!

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

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 ___.
- 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 ____________________________.
- Case-control studies are advantageous in cases where time is short, or when the response outcome is a _________ _________________ 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, here is an in-depth article on observational design.

Chapter 9 Reflection Questions

What distinguishes an experimental design from an observational design? Which one is better designed to identify causal relationships, and why?

What are common reasons why researchers may choose (or need) to use an observational study design?

In science, what does it mean to model? What are we trying to do?

Consider this example: drowning incidents and ice cream sales are highly correlated, but one of these does not cause the other. What might be a confounder that we could add to this model?

How is a mediator different from a confounder? Can you think of a mediator that might facilitate a causal chain between having an earlier bedtime and having higher grades?

In observational study contexts, what does it mean to stratify the data? How can that help us build an argument to either support (or question) causality?

Can you distinguish a cross-sectional design, cohort design, and case-control design? Which design is particularly helpful in situations where the response outcome of interest has a very low incidence?

For which observational study design(s) are we **not** able to accurately calculate relative risk? What ratio measure could we use instead?

Chapter 9 Additional Practice (if you need it!)

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.