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.	Identify *associations* that may signal causation.
In Experiments, we have the power to…	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 singleton?

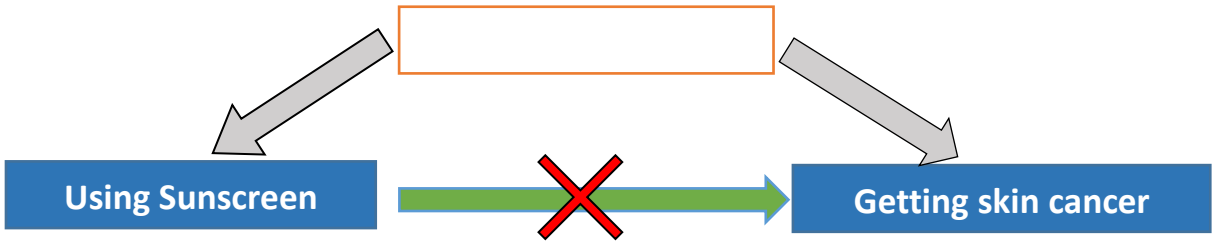
Reasons for completing an observational study.

- 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?

Reflection Questions

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

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

9.3. In science, what does it mean to model?

9.4. Consider this example: the number of drowning incidents on any particular day is highly correlated with national ice cream sales on those days. But one of these does not cause the other. What might be a confounder that we could add to this model?

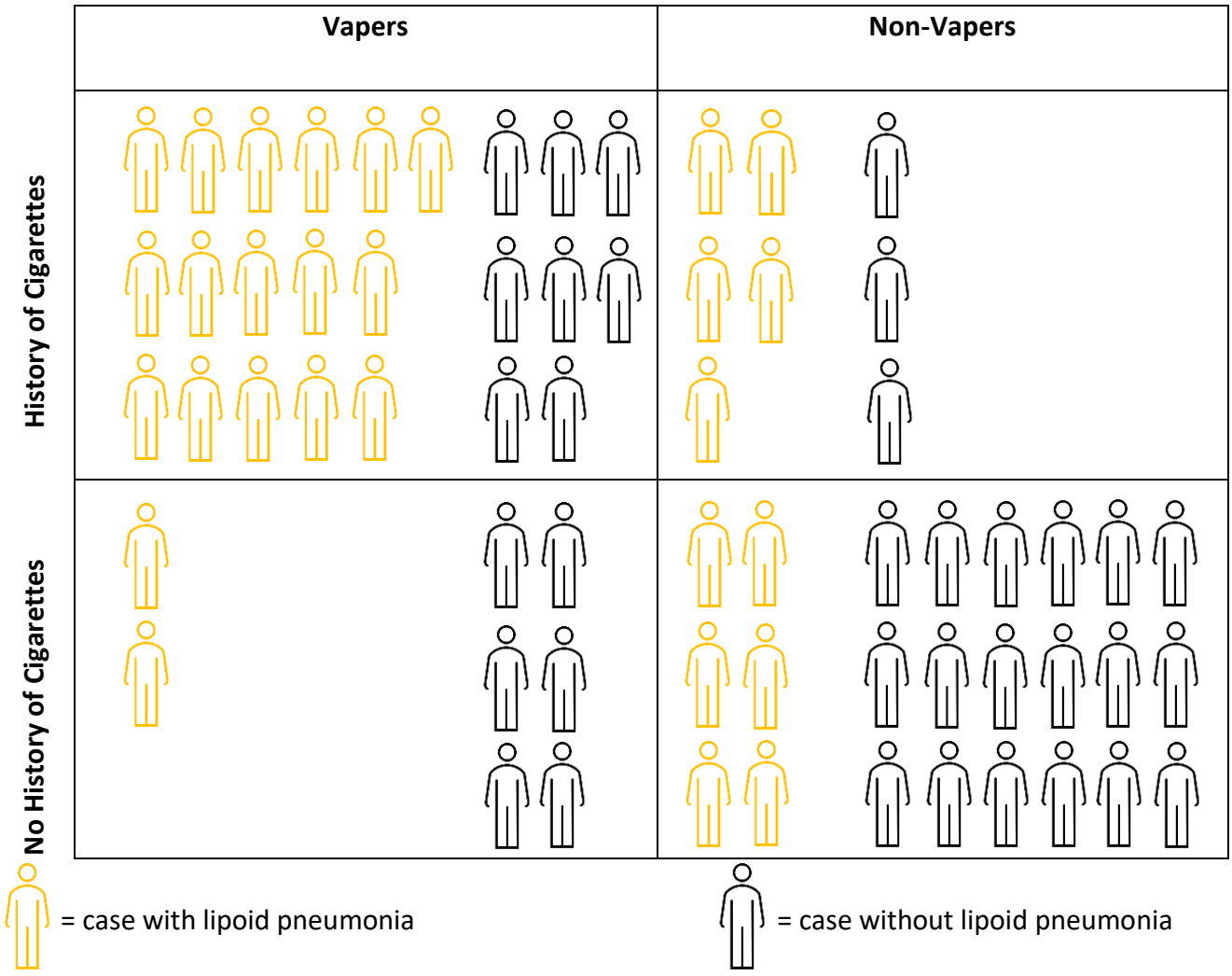
9.5. How is a mediator different from a confounder? Can you think of a mediator that might facilitate a causal chain between having an earlier bedtime and having higher grades?

Stratification - Controlling for Confounders in Observational Settings

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!

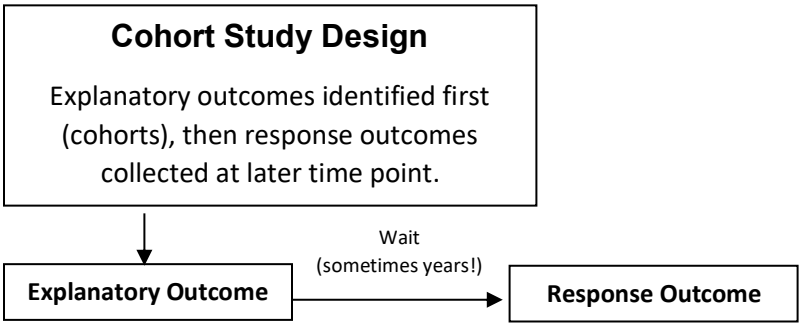
Stratification is the analytical process of breaking down our comparison groups (e.g., vapers and non-vapers) into smaller subgroups based on a potential confounder to see if differences still show up in the more targeted comparisons! **Start by comparing the risk of lipoid pneumonia between the vapers and non-vapers.**

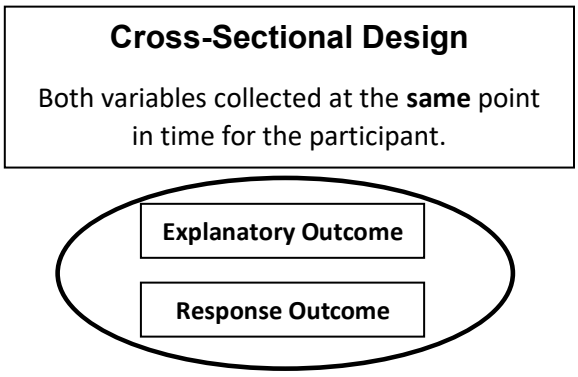
Table 2. Visualizing stratification

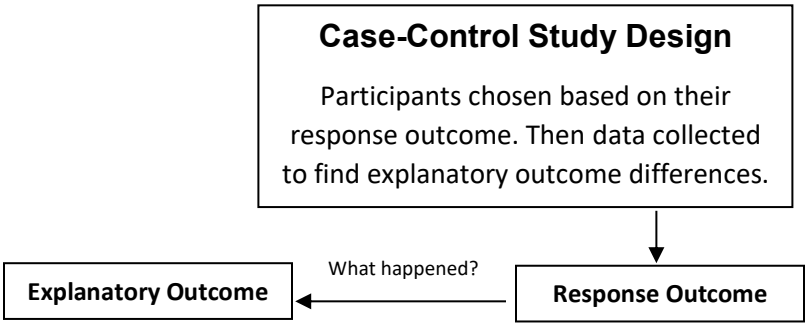


Different Observational Study Designs

- Different observational designs lend themselves to different advantages and analytical options! These design differences hinge on when or in what order we collect the response and explanatory data.
    - **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 differ from cross-sectional studies in having some type of longitudinal element. We have explanatory outcomes, and then at a later point in time, we see if there are differences in response outcomes.
        - 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 based on having or not having certain response outcomes. Then data for one or more explanatory variables are collected ***separately***, perhaps by asking the participants or through records.
        - 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. Then ask them if they have a history of vaping.







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 =

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 Design: We could collect data that preserves the **natural incidence** of lung cancer. For example, a prospective cohort study would identify people who smoke and don't smoke. Then over time, we could track who is diagnosed with lung cancer after some amount of time to see what differences emerge.

Let's say we identified 300 people regular and 300 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	41	259	**300**
Non Smoker	2	298	**300**
Totals	**43**	**357**	**600**

We calculate a sample statistic estimate for RR, then find a 95% confidence interval for the true RR.

$$\frac{41/300}{2/300} = 20.5\ (\_\_\_\_\_,\ \_\_\_\_\_)$$

This interval does include the true RR of 30, and this approach won't show bias in estimating RR in the long-run. However, the confidence interval is exceedingly wide.

Shouldn't 600 be a large enough sample size? Where does the uncertainty stem from?

And here lies an occasional weakness for experiments, cross-sectional studies, and cohort studies.

- When studying a _____ incidence response outcome, we may need a *very* large _____ _____ in order to estimate the true risk in both explanatory groups with reasonable precision.
- When that isn't feasible, we may turn to case-control studies where we directly find people with each response, rather than using a natural incidence sampling approach.

Table 4. Smoking and Lung Cancer (Case Control)

	Cancer	No Cancer	Totals
Smoker	213	28	**241**
Non Smoker	87	272	**359**
Totals	**300**	**300**	**600**

Case Control Design: Let's instead find 300 people with lung cancer and compare them to 300 people without lung cancer, and then ask them about their history of smoking.

Given that approximately 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:

We can again calculate a sample statistic estimate for RR, then find a 95% confidence interval for the true RR.

$$\frac{213/241}{87/359} = 3.65\ (3.02, 4.40)$$

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

By directly finding people with and without cancer, we no longer have _____.
The numerator and denominator no longer represent the conditional risk for cancer given smoking status.

Introducing "Odds"

- Risk is simply the probability of an adverse event occurring. It's the number with out of the total. *If throwing a 6-sided die, the probability of throwing a 1 is 1/6 (or approximately 0.167).*

$$\textbf{Risk} = \mathrm{P}(\text{outcome}) \approx \frac{\#\ \text{Cases with}}{\text{Total}\ \#\ \text{cases}}$$

- "Odds" also assesses the likelihood of an adverse event occurring, but not as a probability. It's the number with divided by the number without. *The odds of throwing a 1 would be 1 to 5 (or 0.2).*

$$\textbf{Odds} = \frac{\mathrm{P}(\text{outcome})}{\mathrm{P}(not\ outcome)} \approx \frac{\#\ \text{Cases with}}{\#\ \text{Cases without}}$$

- The true advantage of an odds calculation comes in the odds ratio measure. While a relative risk calculation is sensitive to the type of sampling done, an odds ratio is not.

$$\textbf{Relative Risk (RR)} = \frac{\text{Risk}_A}{\text{Risk}_B} \qquad \textbf{Odds Ratio (OR)} = \frac{\text{Odds}_A}{\text{Odds}_B}$$

- The construction of an odds ratios allows it to proportionally balance out the asymmetry in our explanatory group sizes. The numerator term is affected proportional to the denominator term!

$$\text{RR} = \frac{213/241}{87/359} = 3.65\ (3.02, 4.40) \qquad \text{OR} = \frac{213/28}{87/272} = 23.78\ (14.98, 37.75)$$

- As a result, when working with case-control data, one should use an odds ratio rather than relative risk, and it may often be used as a reasonable estimate for relative risk!

Understanding ***why*** *the odds ratio is insensitive to the bias in incidence sampling is beyond the scope of this course. It requires a deeper dive into some conditional probabilities and equivalencies!*

Advantages and Disadvantages of Cohort and Case-Control designs

- Cohort studies allow for _____ observation.
  - They can often provide more data in real time and sometimes help researchers better construct causality arguments.
  - However, they can take a long time and may need lots of participants in rare incidence cases.
- Case-control studies allow for _____ data collection.
  - We can directly identify people with this rare outcome rather than wait for it to happen.
  - However, we may have to rely on participants' memory of previous activity or exposure, or find records, and we typically don't have the same detailed accounts of timing as cohort studies.

Odds Ratios vs. Relative Risk (let's summarize the important parts!)

- ✓ 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!
- ✓ Be aware that an OR will ***exaggerate*** the effect in comparison to relative risk.
  - In other words, an OR tends to be farther from **1** than the true RR will be.
  - The **larger the sample size**, the closer OR tends to be in approximating RR.
- ✓ An OR is still valid in other designs and occasionally reported in other designs! But RR is often preferred when appropriate because it is the simpler, more straightforward measure to report.

Reflection Questions

9.6. 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) a causal link between two variables?

9.7. 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?

9.8. 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 (Videos available in the Ch 9 module on Canvas!)

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.