Case-control study

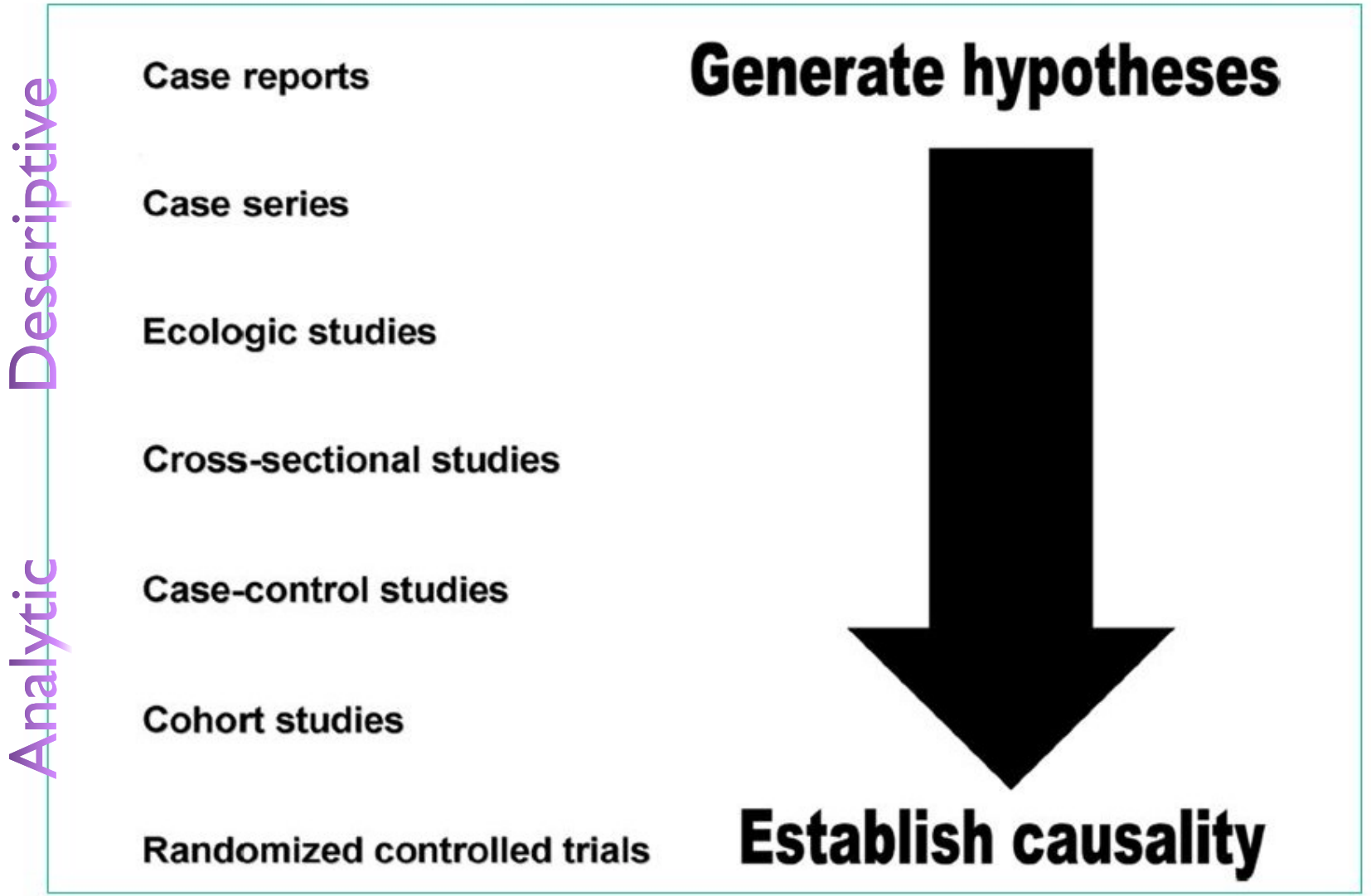
Joint Introductory Course on Epidemiology and Biostatistics 2019

Thanawadee Chantian, MD, MPH
FETP Thailand, Bureau of Epidemiology

Outline

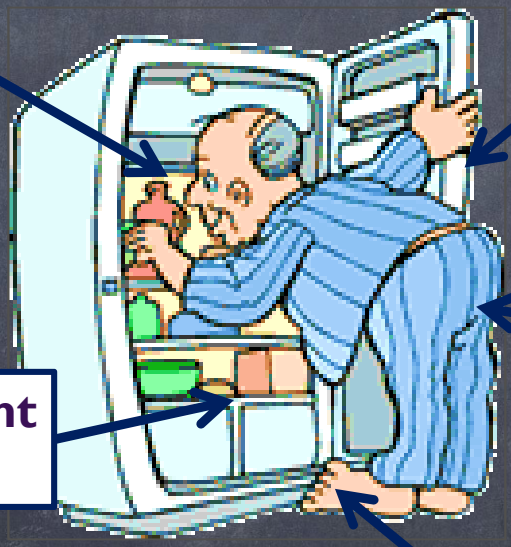
- Case-control design concept
  - When to use
  - Source of population
  - Identifying case
  - Identifying control
- Measurement
- Matching
- Potential bias in case-control studies

Hierarchy of Epidemiologic Study Design



Tower & Spector, 2007 (www)

When to use a case control approach



Rare disease

Shorter in duration

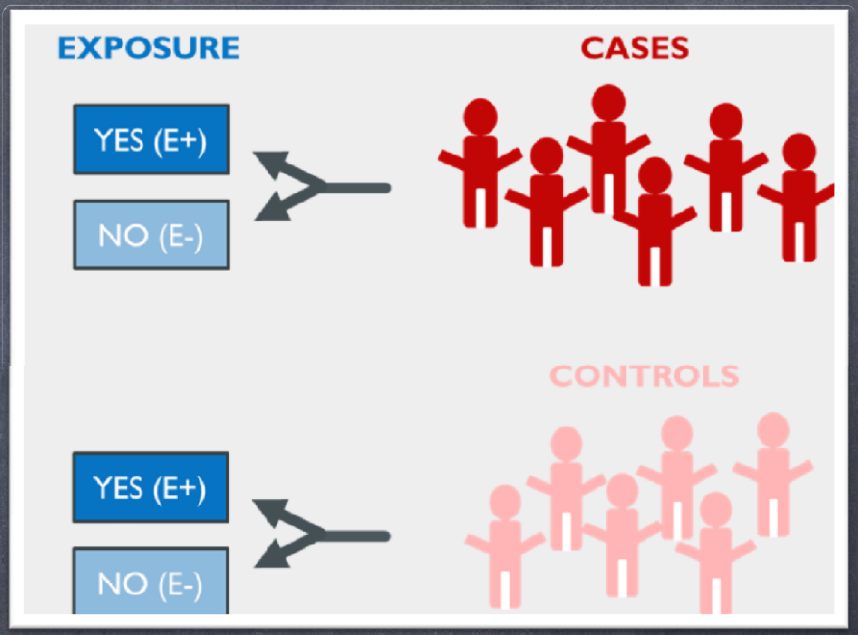
In an existing cohort study: new questions arise

Case ascertainment system in place

Limited resources

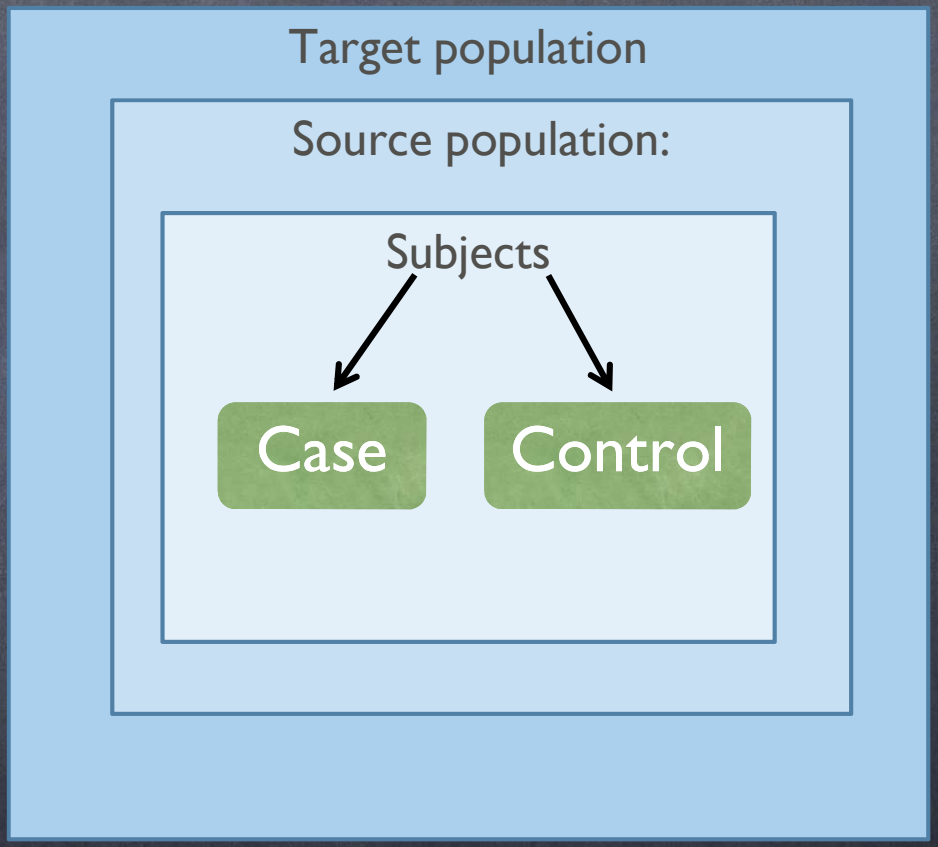
Case-control: study design concept

1. Identify people with the disease of interest (cases)
2. Identify people free of the disease (control)
3. Assess their prior exposure
4. Calculating the odds of exposure in each group
5. Compare the exposure between case and control (odds ratio)



Identifying the source population

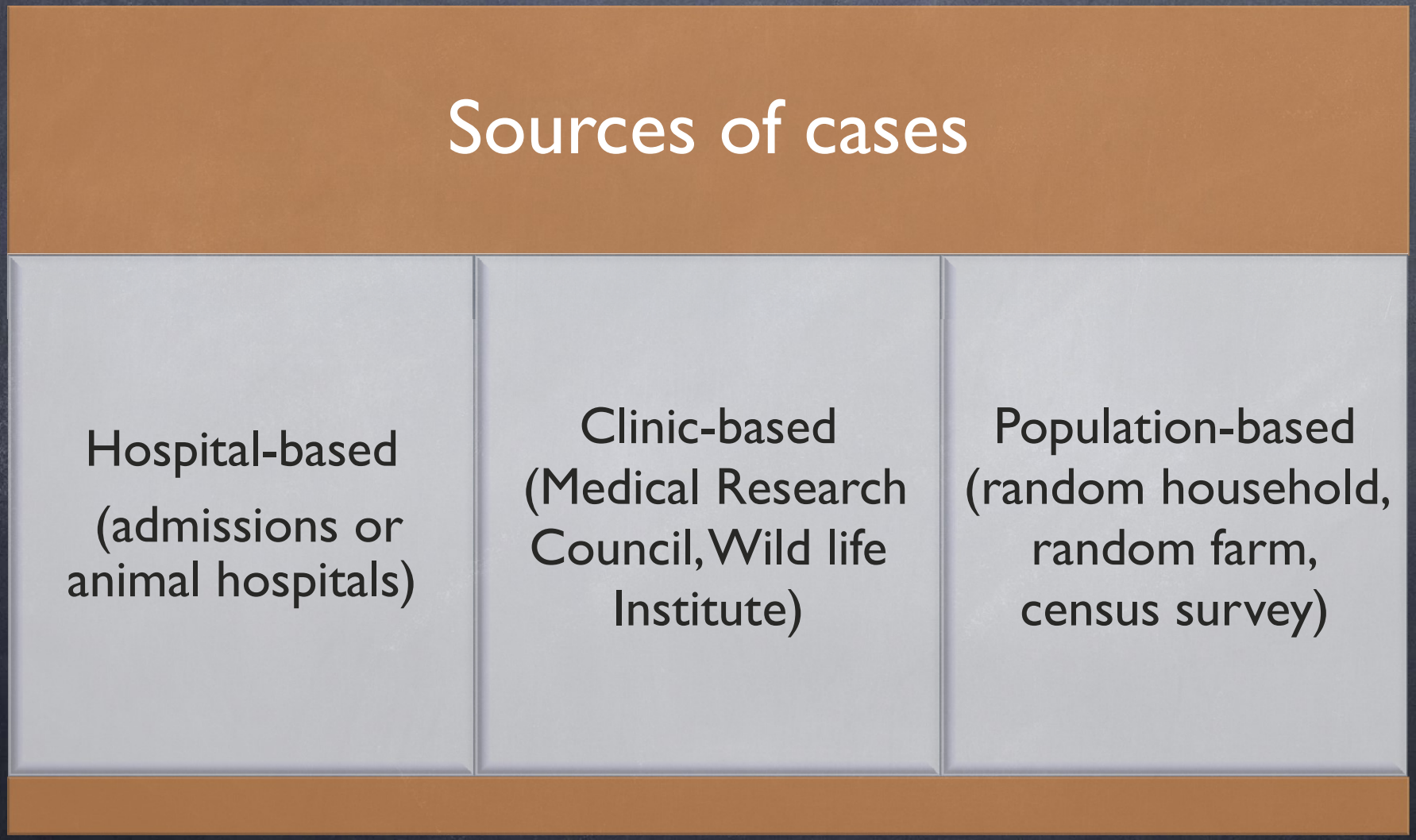
- Target population: population to which it might be possible to extrapolate results from a study.
- Source population: population from which the study subjects are drawn.
  - Ideally, it is the population from which case and control arise



Identifying cases

- Individual from the source population who meet case definition
- Selected cases should represent the cases developing in the source population
- Preferable to enroll incident cases

Identifying cases



Identifying controls

- Controls should be selected from the source population that gives rise to the cases
- Purpose: providing a good estimate of the level of exposure one would expect from the general population
  - The exposure distribution among the control should be the same exposure distribution among the source population
- Controls *must* be selected independently of their exposure status

Identifying controls

Hospital controls

- Easy to find and willing to participate
- The controls should
  - Have diseases that are unrelated to the exposure being studied
  - Have diseases with similar referral pattern as the cases
- Lead to bias if exposure related to hospitalization exposure

Identifying controls

Community control

- Exposure distribution is more similar to the source population

Methods

- Random-digit dialing: using a system that randomly selects telephone number from a directory, time consuming
- Random or systematic sampling from a list
- Neighborhood
- Patients' friends or relatives: willing to participate, similar preferences can lead to overmatching

Example : Hospital-based case-control study

- Cases are patients treated for severe psoriasis at the Mayo Clinic (Rochester, MN)
  - People come to the Mayo Clinic from all corners of the world
- Q: What is the source population?
- A: All people in the world who would go to the Mayo Clinic for severe psoriasis (were they to get severe psoriasis)
  - This population cannot be enumerated

*Rothman p 89

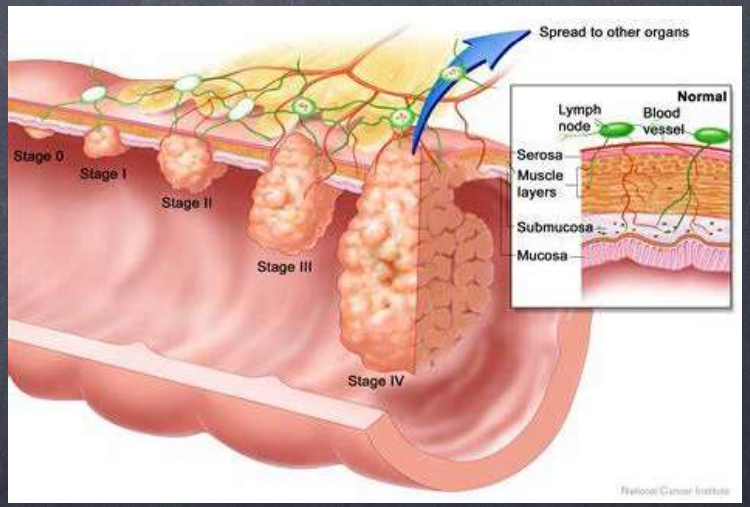
Example : Hospital-based case-control study

- Some options for the Mayo Clinic study
  - Limit to cases living in the United States, and then recruit controls using neighborhood controls or random digit dialing
  - Limit to cases living in Minnesota, and sample controls from Minnesota drivers licenses or a marketing list
  - Use the people who go to the Mayo Clinic for other diseases as the control group – note that this disease must not be caused by the exposure of interest
- Lots of issues and for option #3 (called a "hospital-based case-control study"); we generally just hope for the best
  - Hospital-based case-control studies are often considered "weaker" studies because the controls (sick people with other diseases) may not be representative of the non-diseased members of the source population regarding exposure

Example: Colon cancer and diet

- Case-control study to investigate the association between diet and colon cancer in 5-county Atlanta over a 5 year period
  - There is a SEER cancer registry that conducts active surveillance on incident cases of cancer among people residing in 5-county Atlanta at the time of diagnosis.
- How would you sample controls?
  - Choose controls living in Atlanta for

each case at the time that case is diagnosed



- This is called a "population-based" case-control study

Example: Colon cancer and diet

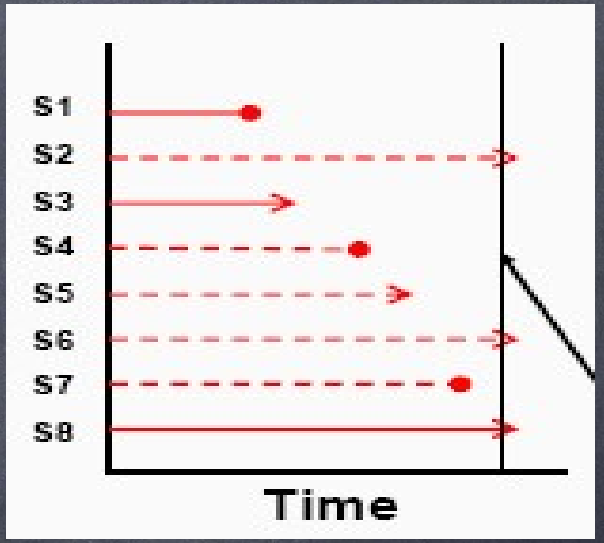
- Possibilities for control selection include:
  - Random-digit dialing (quite possibly obsolete)
  - Drivers licenses and state ID cards
  - Purchase a marketing list
  - Neighborhood controls
  - Friend controls
  - Others?

- No option is perfect, and thus some individuals who are at risk for disease (and who would be included in the study as cases if they developed disease) might not be identified as potential controls

Identifying control: cumulative incident sampling

- Controls selected from those who are free of diseases after all cases are identified
  - Not dynamic source population
  - Follow up time is short
- Can only be conducted within a closed cohort where exposed and non-exposed subjects are followed an equal amount of time

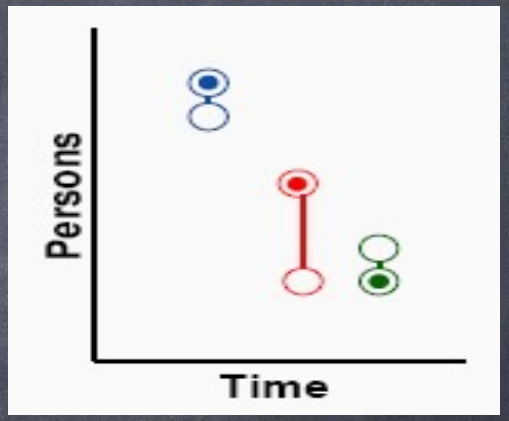
S2, S6 and S8 are selected as controls for cases S1, S4, and S7.



Sampling after all cases have occurred

Identifying control: incident density sampling

- Studying incident disease, where person-time is important
- Controls selected from those who are free of diseases at the time of each incident case
- Dynamic source population
- Follow up time is short
- Controls may be enrolled as case later
- Individuals may be selected as a control more than once



Ratio of controls-to-cases

- In a case-control study every time a case is identified you could choose to recruit:
  - 1 control from the source population
  - 2 controls from the source population
  - 3 controls from the source population…
- Choosing multiple controls per case is a good strategy when you are concerned that you might not be able to recruit enough cases to have adequate power to reject the null hypothesis (assuming that the alternative is true)
  - Note that *if you can find a sufficient number of cases*, then a 1:1 ratio of cases-to-controls is the best strategy

1:1 ratio

	Exposed	Unexposed	Total
Cases	55	35	90
Controls	45	45	90

OR = 1.57 (0.87, 2.85)

2:1 ratio

	Exposed	Unexposed	Total
Cases	55	35	90
Controls	90	90	180

OR = 1.57 (0.94, 2.64)

3:1 ratio

	Exposed	Unexposed	Total
Cases	55	35	90
Controls	135	135	270

OR = 1.57 (0.97, 2.57)

And so on…

1:1 ratio: OR = 1.57 (0.87, 2.85)

2:1 ratio: OR = 1.57 (0.94, 2.64)

3:1 ratio: OR = 1.57 (0.97, 2.57)

4:1 ratio: OR = 1.57 (0.98, 2.53)

5:1 ratio: OR = 1.57 (0.99, 2.51)

6:1 ratio: OR = 1.57 (1.00, 2.50)

7:1 ratio: OR = 1.57 (1.00, 2.49)

8:1 ratio: OR = 1.57 (1.01, 2.48)

9:1 ratio: OR = 1.57 (1.01, 2.47)

Advantage of case-control study

- Cheaper and quicker than prospective cohort studies
- Useful for studying rare diseases
- Sample sizes are smaller than for cohort studies
- Allows for study of several exposures

Disadvantage of case-control study

- Can only study one disease at a time
- Cannot directly estimate the risk of diseases
- Easily mislead to biased result due to control selection
- Not useful for studying rare exposures
- Temporality of exposure and disease may not be certain

Case-control study measure

Case-control study example: outbreak of acute diarrhea at resort A

The event occurred right after seafood night

Case-control study example: outbreak of acute diarrhea at resort A

EXPOSURE → HEALTH OUTCOME

Eating raw oyster → Diarrhea

Cases

Individuals with diarrhea

Controls

A small sample of disease-free guests

	Ate raw oyster	Did not eat raw oyster	
Cases	6	4	**10**
Controls	5	15	**20**
	11	19	30

Case-control study example: outbreak of acute diarrhea at resort A

- We cannot calculate risks or risk ratios
- Why not?
  - We don't know the total number of **exposed** and **unexposed** individuals
  - The number of individuals who do or **do not** get the disease is **fixed** by the investigator
- We need an alternative…

	Ate raw oyster	Did not eat raw oyster	
Cases	6	4	**10**
Controls	5	15	**20**
	11	19	30

Is there an association?

STEP 1 Measure Frequency → In a case-control study, we measure the frequency (odds) of exposure

STEP 2 Compare Frequencies

STEP 3 Interpret Comparison

What are Odds?

$$\text{odds} = \frac{\text{the probably of an event ocurring}}{\text{the probability of an event NOT occurring}} = \frac{p}{1-p}$$

This depends on the definition of the **event**.

In sports betting…..

$$\text{odds} = \frac{\text{the probably of NOT winning}}{\text{the probability of winning}}$$

Case-control study example: outbreak of acute diarrhea at resort A

Step 1: Measure Frequency

- What is the **probability** of exposure (eating raw oyster) among **cases**?
- What are the **odds** of exposure among **cases**?

	Ate raw oyster	Did not eat raw oyster	
Cases	6	4	**10**
Controls	5	15	**20**
	11	19	30

Case-control study example: outbreak of acute diarrhea at resort A

Step 1: Measure Frequency

- What is the **probability** of exposure (eating at the salad bar) among controls?
- What are the **odds** of exposure among controls?

	Ate raw oyster	Did not eat raw oyster	
Cases	6	4	**10**
Controls	5	15	**20**
	11	19	30

Case-control study example: outbreak of acute diarrhea at resort A

Step 2: Compare Frequencies

- What is the ratio that compares the odds of exposure among **cases** to the odds of exposure among **controls**?

	Ate raw oyster	Did not eat raw oyster	
Cases	6	4	**10**
Controls	5	15	**20**
	11	19	30

Case-control studies – Measure of association: Odds Ratio (OR)

$$\text{Odds Ratio} = \frac{\text{the odds of exposure among cases}}{\text{the odds of exposure among controls}}$$

$$= \frac{a/b}{c/d} = \frac{ad}{bc}$$

"Cross-Product"

	Exposed (E+)	Unexposed (E-)	
Cases (D+)	a	b	M_1
Controls (D-)	c	d	M_0
	N_1	N_0	N

Units	Unitless
Ranges	0 to + ∞
>1	Exposure may increase disease frequency
=1	Exposure may not affect disease frequency
<1	Exposure may decrease disease frequency

Case-control study example: outbreak of acute diarrhea at resort A

Step 3: Interpret Comparison

OR = 4.5

The **odds** of **eating raw oyster** among **individuals with diarrhea** were **4.5** times higher than the **odds** of **eating raw oyster** among **guests without diarrhea**

	Ate raw oyster	Did not eat raw oyster	
Cases	6	4	**10**
Controls	5	15	**20**
	11	19	30

Case-control studies – Measure of association: Interpreting the odds ratio

The **odds** of **exposure** among **cases** are **XX** times higher than the **odds** of **exposure** among **controls***

The **odds** of **eating raw oyster** among **individuals with diarrhea** were 4.5 times higher than the **odds** of **eating raw oyster** among **controls who did not have diarrhea**

Conditions for the OR to approximate the RR

1. The distribution of exposure in controls must be representative of the exposure distribution in the source population
2. Cases must be incident cases (i.e. representative of all cases in the source population)
3. Disease must be rare

NOTE: Caveats 1 and 3 go hand-in-hand….if disease is not rare, the controls cannot be representative of the source population

Cohort Study: Rare Disease

	E+	E-	
D+	200	100	300
D-	9,800	9,900	19,700
Total	10,000	10,000	20,000

What is the risk ratio?

What is the odds ratio?

Cohort Study: VERY Common Disease

	E+	E-	
D+	6,000	3,000	9,000
D-	4,000	7,000	11,000
Total	10,000	10,000	20,000

What is the risk ratio?

What is the odds ratio?

Matching

Matching in case-control study

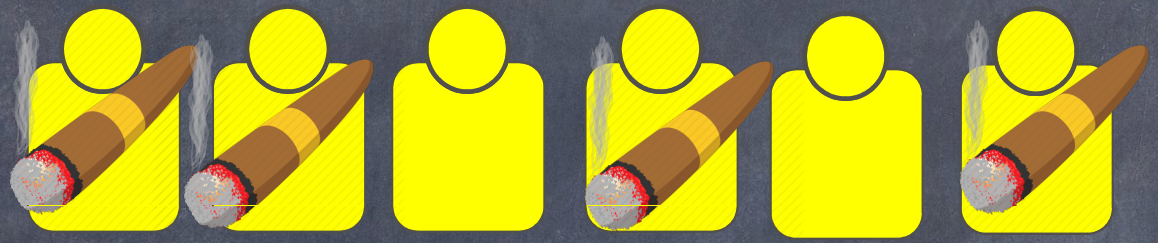
- In a case-control study, matching refers to the purposeful selection of controls so that they are identical (or nearly so) to the cases *with respect to the distribution of one or more potentially confounding factors*
- Matching on confounders is common in case-control studies
- If you decide to match, match only on strong confounders

Case-control study of lung cancer and uranium mining

- **Exposure? Uranium mining**
- **Outcome? Lung cancer**
- Is there something that could be associated with **both uranium mining** and **lung cancer** that might influence our results?
  Smoking!

Case-control study of lung cancer and uranium mining

CASES



6 Cases:
4 are smokers

CONTROLS

6 Controls:
4 are smokers

Matching continuous variables

Suppose we have a **case** who is a 42 year old male, and we want to match on age and sex…

- Category Matching
  - Divide controls by age group: 30-34, 35-39, 40-44, 45-49, etc…
  - Select a male **control** aged 40-44 years
- Caliper matching
  - Select a male **control** aged 42 ± 5 years

Individual matching: 4 possible combinations

Case	Control	Number of pairs
Exposed	Exposed	W
Exposed	**Not Exposed**	**X**
Not Exposed	**Exposed**	**Y**
Not Exposed	Not Exposed	Z

NOTE: These are PAIRS, not individuals!

Data Layout

CASE	CONTROL: Exposed	CONTROL: Unexposed
Exposed	W	X
Unexposed	Y	Z

Matched odds ratio

NOTE: Only considers **discordant pairs**

$$\text{Matched Odds Ratio (mOR)} = \frac{X}{Y}$$

- Concordant pairs (pairs that cases and controls have same exposure history)
- Odds ratio is the ratio of the discordant pairs

$$\text{mOR} = \frac{\text{\# pairs that case is exposed and control is not exposed}}{\text{\# of pairs that control is exposed and case is not exposed}}$$

Matching – pros and cons

Pros

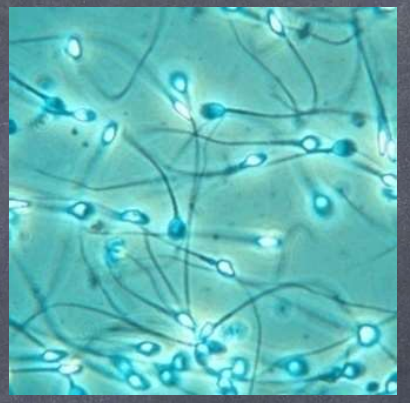
- Matching allows more precise estimation of odds of exposures of interest
- Avoids "wasting" power on factors that are already known to impact risk

Cons

- No estimates of risk available for matched-upon factors
- Increased cost/complexity
- Possibility of overmatching
- Loss of efficiency if matched on a factors associated with disease but not exposure

Matched case-control study example

- Matched case-control study to determine whether there's an association between working at a uranium mine and reduced sperm count
- **CASES**: 400 men with low sperm count diagnosed in a Utah clinic
- **CONTROLS**: 400 healthy men matched on race, age, area of residence, smoking and drinking habits



OUTCOME

EXPOSURE

Matched case-control study example: results

		CONTROL: Exposed	CONTROL: Unexposed
CASE	Exposed	W	X
	Unexposed	Y	Z

- Matched pairs in which both men worked in uranium mine: 8
- Matched pairs in which case had mine exposure but control did not: 18
- Matched pairs in which case had no mining background but control did: 4
- Matched pairs in which neither had worked in the mines: 370
- What is the matched odds ratio?

Potential biases in case-control study

- Selection bias
- Information bias
  - Memory bias
  - Recall bias
  - Interview bias

Selection bias

- Selection bias can occur in
  - The selection of cases if they are not representative of all cases within the population, or in
  - The selection of controls if they are not representative of the population that produced the cases

Fair Sampling	Diseased	Non-diseased
Exposed		
Non-exposed		

Selection Bias	Diseased	Non-diseased
Exposed		
Non-exposed		

Memory bias

Concerns:

- Can the cases and controls remember events, behaviors, physiologic status in the past? → ↑ Misclassification of exposure

Example:

- "How many times per week did you eat carrots when you were between the ages of 6 and 10?"

Recall bias

Concerns:

- Do cases recall prior events, etc. differently than controls?
- Mindset of someone with disease or caregiver: Is there something that I did that may have caused the disease?

Example:

- Mothers of babies born with congenital malformations more likely to recall (accurately or "over-recall") events during pregnancy such as illnesses, diet, etc.

Interview bias

Concerns:

- Differential interviewing of cases and controls, i.e., may probe or interpret responses differently

Example:

- Interviewer probes more fully of cases than of controls

Summary of Case-Control Study

- Cases and controls are enrolled based upon their **disease status** (not exposure)
- Efficient for rare diseases (Inefficient for rare exposures)
- Smaller sample size – Inexpensive
- Rapid conclusions
- Weakest Study design of observational studies (Higher potential for bias)
- **Uncertain temporal sequence of exposure and disease**

Summary of odds ratios

- Odds ratio (OR) = odds of exposure among cases / odds of exposure among controls = **ad/bc** (unmatched)
- In a matched case-control study: OR = X/Y (where X and Y represent discordant pairs)
- If disease is rare, controls are representative of the underlying population and incident cases are enrolled,

ORs ratio ≈ RR

Hat tip

- Dr. Saowapak Hinjoy. Case-control study.
- Dr.Matt Gribble. Introduction to Environmental Epidemiology. Environmental Epidemiology
- Prof. Jodie Guest and Dr. Kristine Wall. Case-control study. Principle of Epidemiology methods I. Rollins School of Public Health, Emory university