Case control study

Soawapak Hinjoy, DVM, MSc, MPH, DrPH

Bureau of Epidemiology, Ministry of Public Health

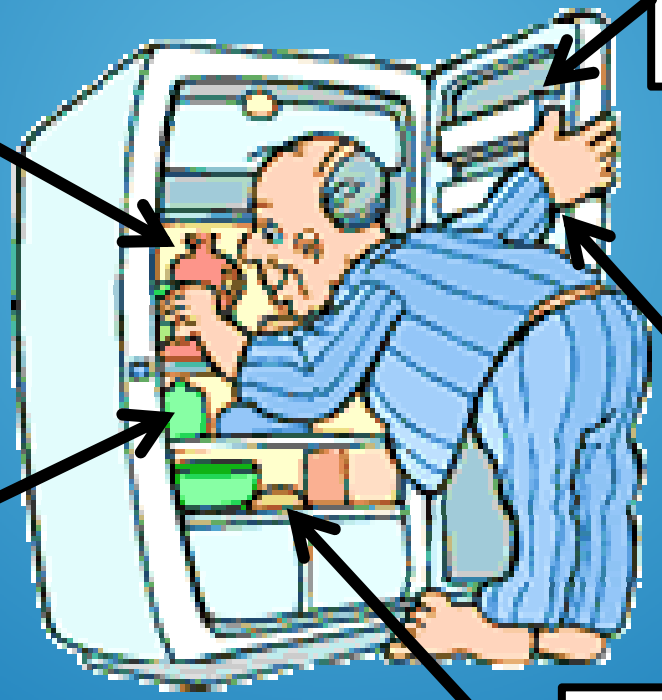
Outline

- Content
- Measurement
- Exercise

When to use a case control approach

Rare disease

Case ascertainment system in place



Shorter in duration

In an existing cohort study: new questions arise

Limited resources

Conduct of case-control study

Cohort study	Case control study
• Define research question and generate a hypothesis	• Define research question and generate a hypothesis
• Identify target population – generalize inference	• Identify target, and source populations
• Select source population base for selecting sample	• Recruit cases and controls
• Recruit study population enrolled in study	

Selection based on disease status in case control study

Conduct of case-control study

- Define research question and generate a hypothesis
- Identify target, and source populations
  - Cases and controls should have same target and source population
  - Sources may differ between cases and controls
- Recruit cases and controls
  - Incidence density sampling – controls selected at the time of each incident case
  - Cumulative incidence sampling – controls selected after all cases are identified
- Ascertain exposure status and characteristics that may influence the determination of the association between exposure and outcome status
- And, do it all without BIAS!!

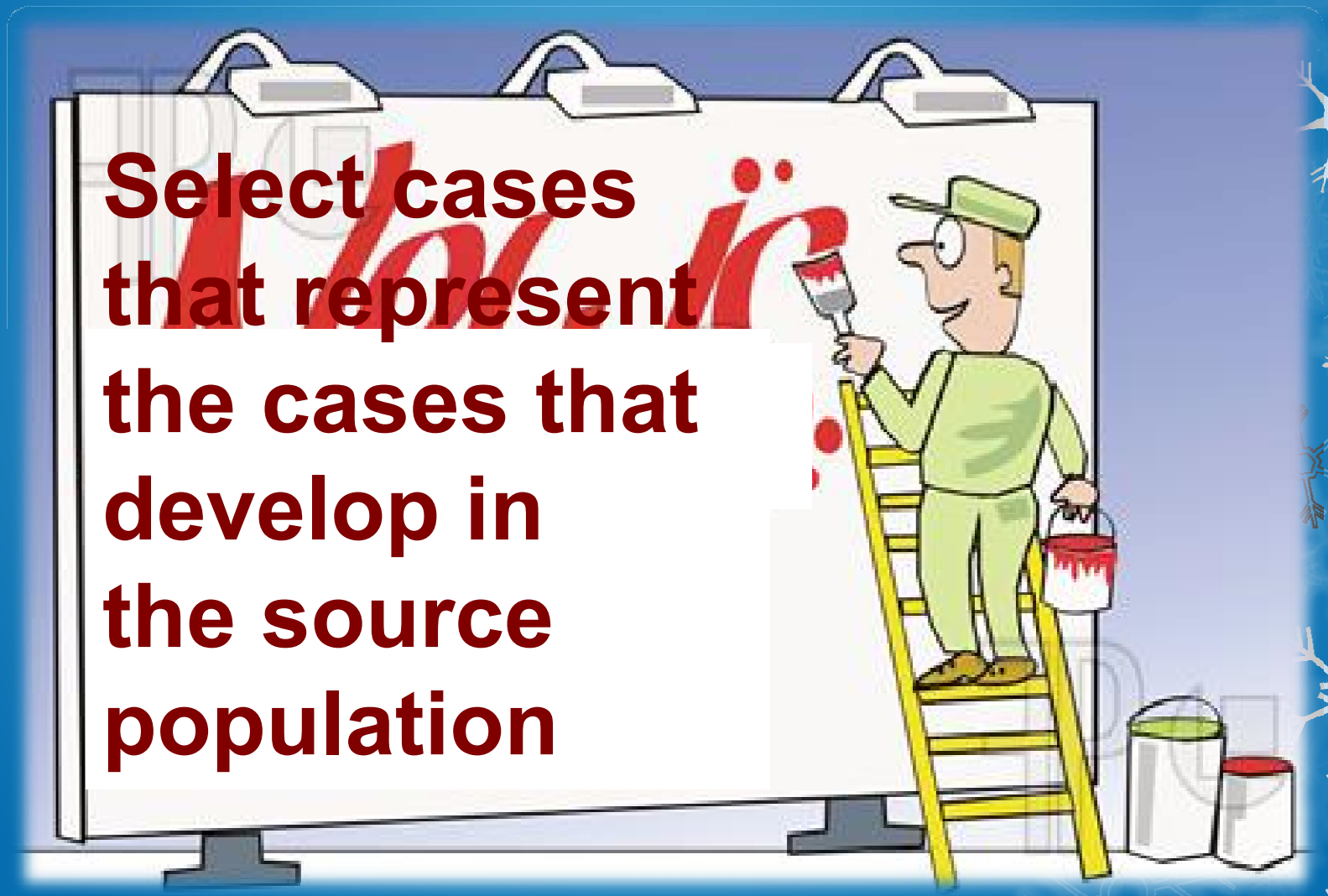
Target and source populations

Identify target and source populations

Target population: special/ general

- Special study populations are often defined by exposure opportunity and in whom outcome occurs with relatively high frequency, e.g.:
  - Dog sprayed under 1 year, exposed to breast cancer
  - Farm workers, exposed to pesticides
- General study populations are all those that are not special study populations
- Source population:
  - Community/ geographic
  - Clinic or hospital
  - Occupational
  - Sheep or pig populations

Ascertainment of Cases



Select cases that represent the cases that develop in the source population

Ascertainment of Cases

Sources of cases

Hospital-based	Clinic-based	Population-based
(admissions or animal hospitals)	(Medical Research Council, Wild life Institute)	(random household, random farm, census survey)

Do the cases represent all cases in the source population?

Ascertainment of Cases

INCIDENT CASE = all new cases occurring in a population within a certain period

PREVALENT CASE = all existing (new and old) cases who are present in a population regardless of when the case was diagnosed

Ascertainment of Controls

Select controls such that exposure distribution is the same as the distribution of person-time in the underlying source population

Ascertainment of Controls

Sources of controls

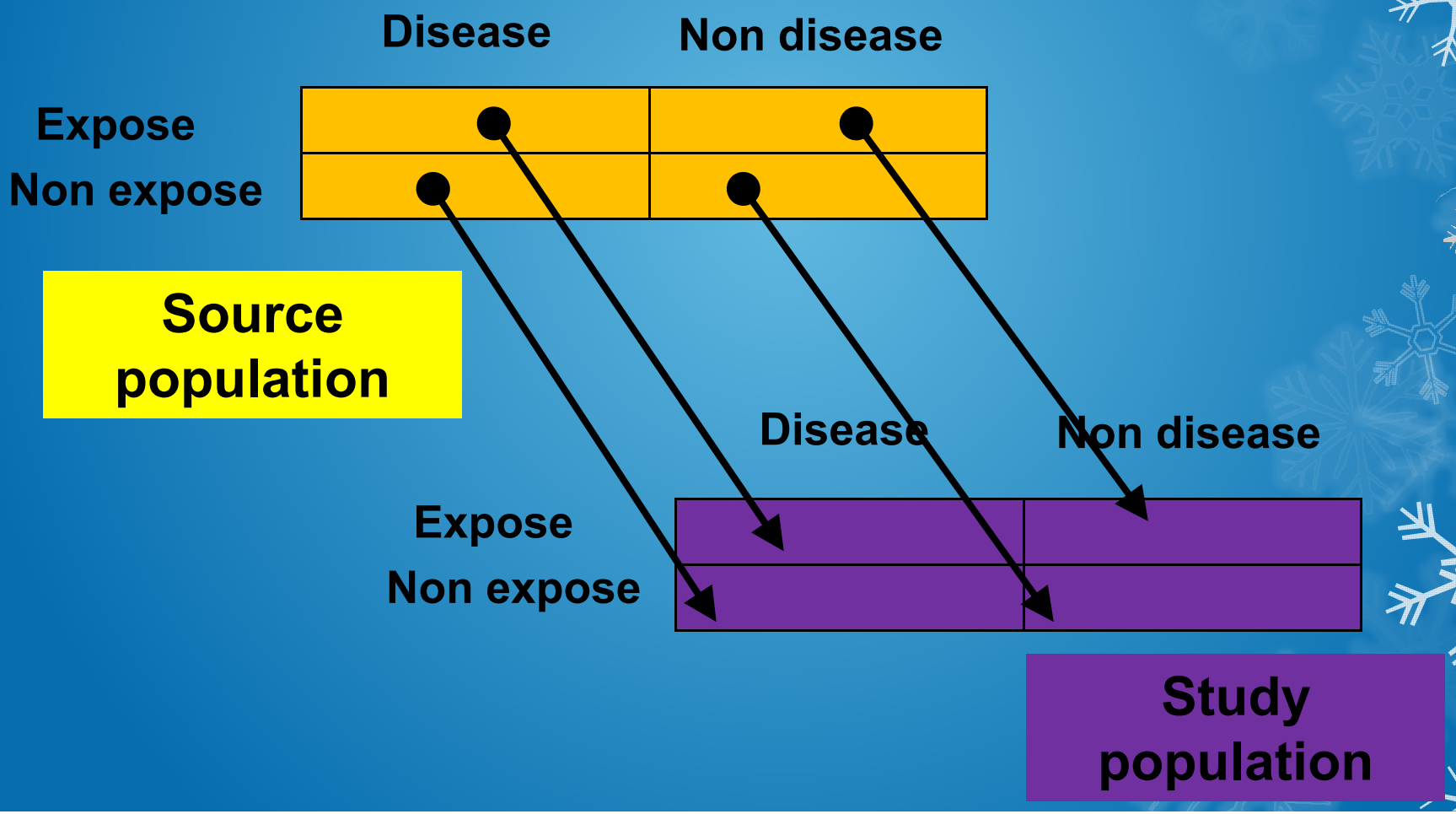
Hospital patients/ animal in clinic	Population of defined area	Probability sample of total population	Neighbors	Friends, siblings, spouses, relatives or associates of cases

Cases and controls should have same target and source population

Sources may differ between cases and controls

Should answer yes to: If developed disease of interest during study period, could they have been included as a case?

Ascertainment of cases and controls



Selection of controls

- Should have the same opportunity for being a case, if the outcome should have occurred
- Selection needs to be independent of exposure status
- Represents the exposure distribution of the source population
- Confirmation of lack of outcome/disease
- Avoid selection bias

Selection of controls

Sources of controls

Neighborhood controls
- Random sample of the neighborhood from which cases reside
- Census
- Convenience sampling

Neighbors

Concerns:
- If exposure is related to environment, are cases and controls too similar (SES, cultural factors)?

Selection of controls

Sources of controls

Generate listing of possible telephone numbers using area code (by random sampling)

Need to call at varying times before classifying household as non-response

Random digit dialing

Concerns:

If exposure is related to environment, are cases and controls too similar?

Ex. Low SES households may not have phones

Selection of controls

Sources of controls

Ask each case for list of possible friends who meet eligibility criteria
Randomly select among list

Friends or family members

Concerns:
May inadvertently select on exposure status, that is, friends because of engaging in similar activities or having similar characteristics/culture/tastes ("over-matching")
Since selection is dependent on case, may bias the selection by exposures

Selection of controls

Sources of controls

Sample of group of persons seen/treated in same medical care source as cases

Ease and accessibility

Hospital or clinic-based

Concerns:

May lead to bias if exposure related to the hospitalization

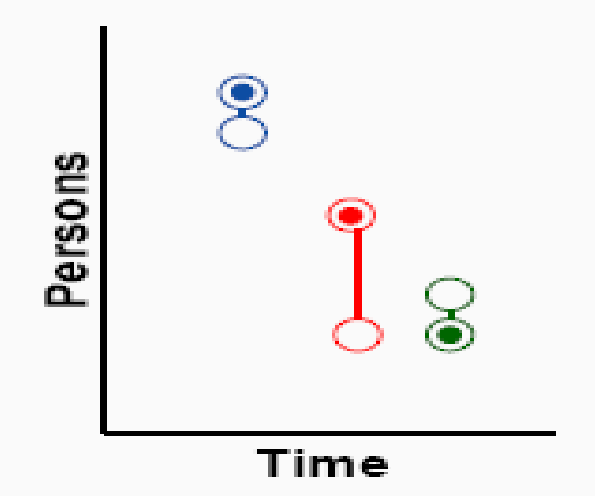
May still not be same underlying source population if case or control diagnosis represents referrals to the hospital but other group does not

Selection of controls

– INCIDENCE DENSITY SAMPLING =
Controls selected at the time of each incident case

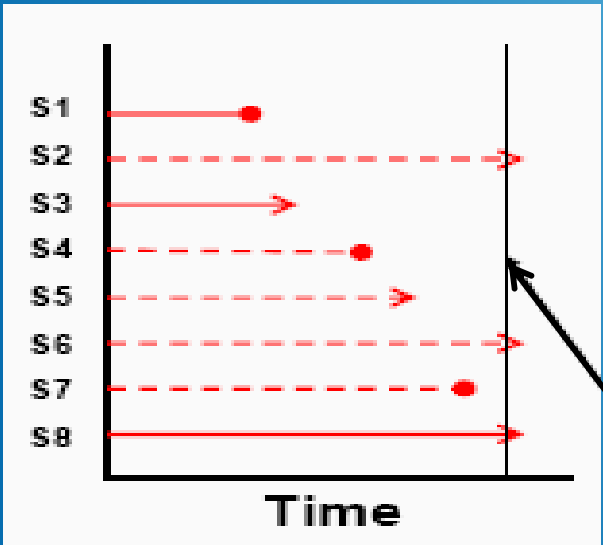
– CUMULATIVE INCIDENCE SAMPLING
= Controls selected after all cases are identified

Selection of controls



INCIDENCE DENSITY SAMPLING

Identifying cases and controls in short time period



CUMULATIVE INCIDENCE SAMPLING

Select all incident cases, but only those without outcome at end of study period as the controls

Sampling after all cases have occurred

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

Recruit cases and controls

Sources of cases

Sources of Controls

Develop case definition

Develop definition

Defining the cases

If continuum:

mild →moderate →severe

Which stage is under investigation?

Diagnostic criteria – level of certainty

- Subjective (self-report, symptoms)
  - ↑ capture of all persons with disease but also include those without true disease
  - ↑ misclassification
  - ↓ time to find cases, larger selection pool
  - → may dilute the effect
- Objective (clinic/pathologic/radiologic)
  - ↓ misclassification →certain that your cases truly have the disease
  - ↓ selection pool
  - → effect estimate is more accurate

Develop definition

Defining the controls

Controls must fullfil all the eligibility criteria defined for the cases apart from those relating to diagnosis of the disease

For example, if the cases are women with breast cancer aged 45 years and over, the controls must be selected from women in the same age group without the disease.

Ascertain exposure

Selection is on outcome status (cases, controls) → Exposures are then determined or ascertained

Exposure ascertainment	*Issues*
Active methods Questionnaire (self- or interviewer administered) Biomarkers **Passive methods** Medical records Insurance records Employment records School records	- Establish biologically relevant period - Capture is after outcome has developed (except for explicitly nested case-control studies) - Measurement occurs at one time - Information is historical

Bias

Observed sample

E (Exposure) → D (Disease) ✓

Selection Bias

Selection bias can result when the selection of subjects into a study or their likelihood of being retained in the study leads to a result that is different from what you would have gotten if you had enrolled the entire target population

Selection bias

Fair Sampling	Diseased	Non-diseased
Exposed		
Non-exposed		

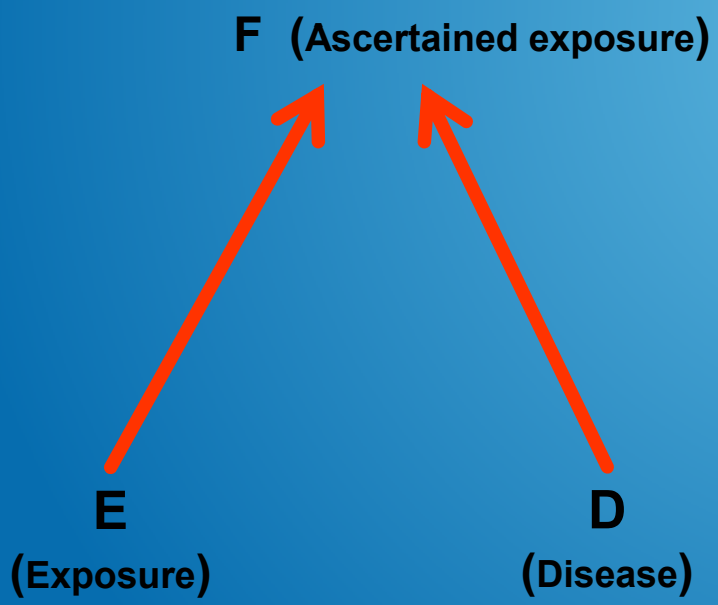
Selection Bias	Diseased	Non-diseased
Exposed		
Non-exposed		

We will find the average height of Americans based on a sample of NBA players.

Selection Bias?

Information bias (Memory bias)

All information is historic, so if relying on reporting by participants, accuracy depends on recall

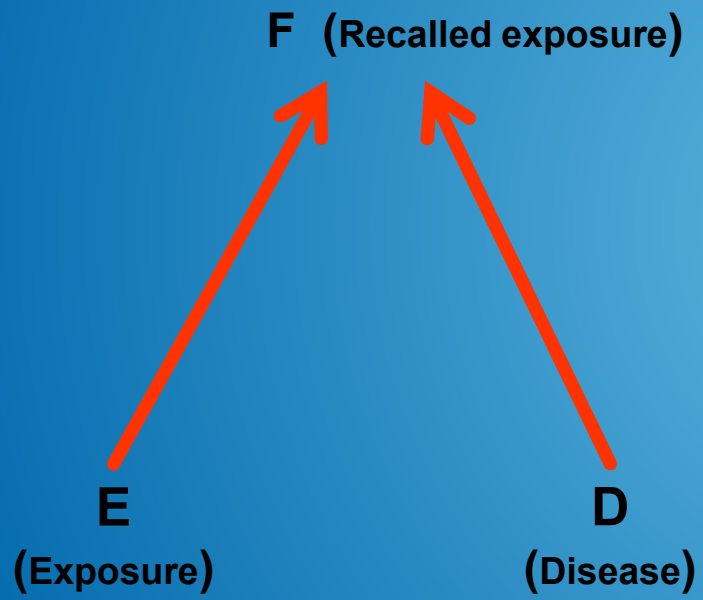


Concerns:

- **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?"

Information bias (Recall bias)



Concerns:

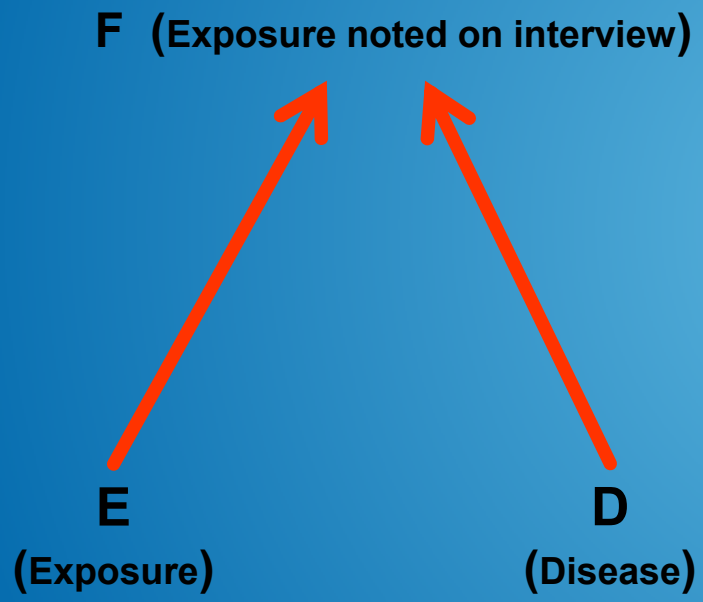
Do cases recall prior events, etc. differently than controls?

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.

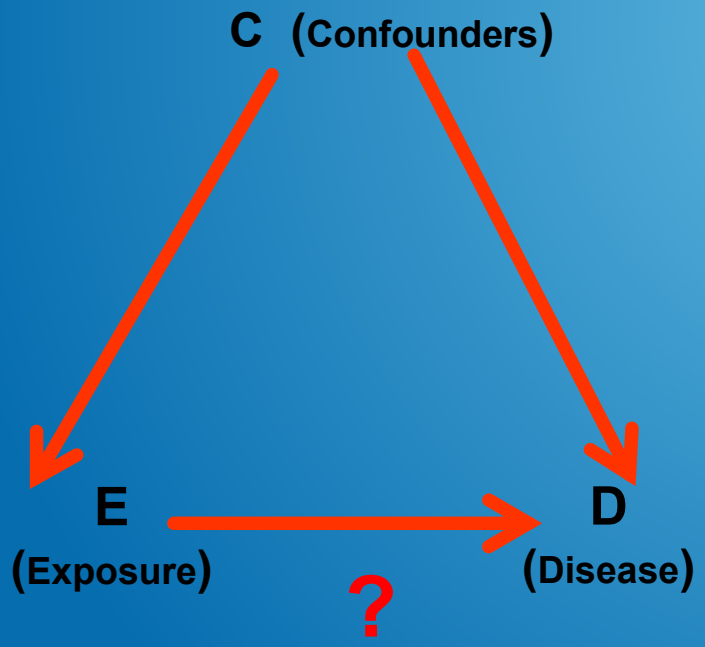
Information bias (Interview bias)



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

Confounder



Potential confounder

1. Determinant of disease
2. Associated with exposure
3. But not in causal pathway of exposure to disease

Outline

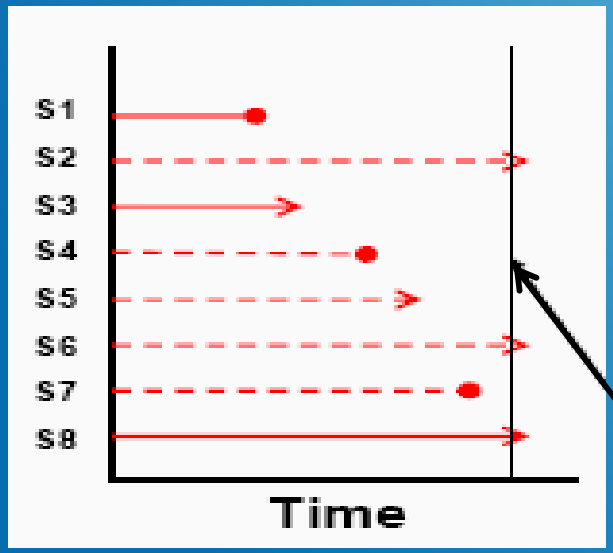
- Content
- **Measurement**
- Exercise

Analysis of case – control studies

– **INCIDENCE DENSITY SAMPLING** = Controls selected at the time of each incident case **(NOT TODAY)**

– **CUMULATIVE INCIDENCE SAMPLING** = Controls selected after all cases are identified

Start with



CUMULATIVE INCIDENCE SAMPLING

Select all incident cases, but only those without outcome at end of study period as the controls

Sampling after all cases have occurred

Why do we use Odds ration (OR) instead of Relative risk (RR) in case-control studies?

- RR: It is calculated by dividing the incidence rate of those exposed to the factor by the incidence rate of those not exposed to the factor
- $RR = \frac{\text{Incidence in the exposed}}{\text{Incidence in the nonexposed}}$
- RR cannot generally be calculated in a case-control study because the entire population has not been studied, so incidences are unknown

Odds are ???

Odds are simply a ratio of the probability that an event will occur **versus** the probability that the event will not occur
= *probability / (1-probability)*

If you go fishing and you catch 3 redfish and 1 trout

then the odds of catching a trout
= [(1/4)/(3/4)] = 1/3 = 0.33

≠

This differs from risk (or probability) the risk of catching a trout is equal to

(# of trout caught) / (total # of fish caught) = 1/4 = 0.25

Reference: http://intmedweb.wfubmc.edu/ebmreviews/odds.html

Odds ratio are ???

If you compare your luck with fishing with no bait versus fishing with bait and cast your line 100 times using each method

	# of times caught	# of times not caught	Total # of casts
With bait	50	50	100
No bait	2	98	100

The *odds* of catching a fish with the bait is [50/100] / [50/100] = 50/50 or 1.0

The *odds* of catching a fish with no bait is [2/100] / [98/100] = 2/98 or 0.02

Therefore, the *odds ratio* for catching a fish with the bait vs. no bait is 1.0/0.02 = 50

The *probability* of catching a fish with the bait is 50/100 or 0.50

The *probability* of catching a fish with no bait is 2/100 or 0.02

Therefore, the *relative risk* for catching a fish with the bait vs. no bait is 0.50/0.02 = 25

Reference: http://intmedweb.wfubmc.edu/ebmreviews/odds.html

An example: Odds ratio (OR)

- If we suspect that bait is associated with catching more fish,
- Then we could take 100 successful fishermen and compare them with 100 fishermen who were unable to catch any fish
- We can compare the *odds* of the use of bait in those who caught fish vs. those who were unable to catch fish by calculating the *odds ratio*

	Bait use	No bait	Total
Caught fish	40	60	100
Caught nothing	20	80	100

- The *odds* of bait use in those who caught fish is [40/100] / [60/100] = 40/60 or 0.67
- The *odds* of bait use in those who caught nothing is [20/100] / [80/100] = 20/80 or 0.25
- Therefore, the *odds ratio* bait use in successful vs. unsuccessful fishermen is 0.67/0.25 = 2.7

The odds of use of bait were 2.7 times greater in successful fishermen vs. unsuccessful fishermen in this study

Or

Successful fishermen were 2.7 times more likely to have used bait than unsuccessful fisherman in this study

Reference: http://intmedweb.wfubmc.edu/ebmreviews/odds.html

Use of odds in epidemiological study

	Disease	No disease	Total
Exposed	a	b	a+b
Unexposed	c	d	c+d
Total	a+c	b+d	

In cohort study: $OR = \frac{\text{odds of disease among exposure}}{\text{odds of disease among unexposed}}$

$OR = \frac{[a/a+b] / [b/a+b]}{[c/c+d] / [d/c+d]}$

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

In case-control study: $OR = \frac{\text{odds of exposure among diseased}}{\text{odds of exposure among non-diseased}}$

$OR = \frac{[a/a+c] / [c/a+c]}{[b/b+d] / [d/b+d]}$

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

Odds ratio (OR) and Relative risk (RR)

	Disease	No disease	Total
Exposed	a	b	N1
Unexposed	c	d	N0

$$RR = \frac{R_1}{R_0} \quad ; R_1 = a/N_1 \quad ; R_0 = c/N_0$$

$$a = R_1N_1 \quad ; c = R_0N_0 \quad ; b = (1-R_1)N_1 \quad ; d = (1-R_0)N_0$$

OR in cohort study

$$OR = \frac{a/b}{c/d} = \frac{R_1N_1/(1-R_1)N_1}{R_0N_0/(1-R_0)N_0}$$

$$= \frac{R_1N_1(1-R_0)N_0}{R_0N_0(1-R_1)N_1} = \frac{R_1}{R_0}\frac{(1-R_0)}{(1-R_1)}$$

Bias

$OR \approx RR$ If disease is rare (low incidence: less than 10%), the higher the risk, the bigger the bias

$$= \frac{R_1}{R_0} \times 1 \text{ (when } R \rightarrow 0)$$

Interpretation of OR

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

Odds Ratio (OR) in multiple levels of exposure

- Factors with multiple levels
- If collapse to a binary variable (2 levels), may lose information, i.e. Is there a dose response?
- Don't want too many levels, or insufficient numbers in each cell to make meaningful assessment
- Use same reference level
  - Test for trend
  - ORs are all relative

Odds Ratio (OR) in multiple levels of exposure

	Cases	Controls	OR
No exposure	A_0	B_0	1.0 (indicates reference)
Exp level 1	A_1	B_1	$\frac{A_1 B_0}{A_0 B_1}$
Exp level 2	A_2	B_2	$\frac{A_2 B_0}{A_0 B_2}$
.	.	.	.
.	.	.	.
.	.	.	.
Exp level k	A_k	B_k	$\frac{A_k B_0}{A_0 B_k}$

Summary

- Fundamental elements of the case-control study design and its relationship to cohorts
  - Selection of cases and controls
- Issues when selecting cases and controls
- Know your target population:
  - Generalizability
  - Help determine appropriate sources for controls

Summary

- Source populations
- Avoid misclassification
- Avoid selection bias: select cases and controls independent of exposure status!

Summary

- **Odds Ratio**

  - With outcomes of very low incidence in the underlying cohort and sampling of controls independent of exposure: OR ≈ RR

Strengths of case-control study

- **Efficient – typically:**

  shorter period of time

  not as many individuals needed (total cohort, thus non-diseased controls)

  cases are selected, thus particularly good for rare diseases

- **Informative – may assess multiple exposures and thus hypothesized causal mechanisms**

Acknowledgements

Most of the slides (in parts of contents) in this lecture come from sequence of courses in epidemiologic methods (751-753) in 2008-2009, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health

Kappa statistics-reliability

$$\frac{\text{Observed agreement (\%)} - \text{Agreement expected by chance alone (\%)}}{\text{100 \%} - \text{Agreement expected by chance alone (\%)}}$$

Kappa statistics

$$\frac{73 - 50}{100\ \% - 50}$$

Interpreting values of kappa

Value of kappa	Strength of Agreement
• 0.0	• No agreement better than chance alone
• < 0.20	• Poor
• 0.21 - 0.40	• Fair
• 0.41 - 0.60	• Moderate
• 0.61 - 0.80	• Good
• 0.81 - 1.00	• Very Good