BI476: Biostatistics - Case Studies

Lec02: Observational Studies

Maoying,Wu (ricket.woo@gmail.com)

Dept. of Bioinformatics & Biostatistics
Shanghai Jiao Tong University

Spring, 2018

Goals of the lecture

- To understand the cross-sectional study along its strength and weakness.

Goals of the lecture

- To understand the cross-sectional study along its strength and weakness.
- To distinguish from the concepts of prevalence and incidence.

Goals of the lecture

- To understand the cross-sectional study along its strength and weakness.
- To distinguish from the concepts of prevalence and incidence.
- To define the organization of a case-control study.

Goals of the lecture

- To understand the cross-sectional study along its strength and weakness.
- To distinguish from the concepts of prevalence and incidence.
- To define the organization of a case-control study.
- To calculate and interpret the odds ratio (OR) as well as the confidence interval.

Goals of the lecture

- To understand the cross-sectional study along its strength and weakness.
- To distinguish from the concepts of prevalence and incidence.
- To define the organization of a case-control study.
- To calculate and interpret the odds ratio (OR) as well as the confidence interval.
- To understand the strength and weakness of CCS.

Goals of the lecture

- To understand the cross-sectional study along its strength and weakness.
- To distinguish from the concepts of prevalence and incidence.
- To define the organization of a case-control study.
- To calculate and interpret the odds ratio (OR) as well as the confidence interval.
- To understand the strength and weakness of CCS.
- To understand the concept of bias and confounding.

Goals of the lecture

- To understand the cross-sectional study along its strength and weakness.
- To distinguish from the concepts of prevalence and incidence.
- To define the organization of a case-control study.
- To calculate and interpret the odds ratio (OR) as well as the confidence interval.
- To understand the strength and weakness of CCS.
- To understand the concept of bias and confounding.
- To define and organize a cohort study.

Goals of the lecture

- To understand the cross-sectional study along its strength and weakness.
- To distinguish from the concepts of prevalence and incidence.
- To define the organization of a case-control study.
- To calculate and interpret the odds ratio (OR) as well as the confidence interval.
- To understand the strength and weakness of CCS.
- To understand the concept of bias and confounding.
- To define and organize a cohort study.
- To calculate and interpret the relative risk or risk ratio (RR).

Goals of the lecture

- To understand the cross-sectional study along its strength and weakness.
- To distinguish from the concepts of prevalence and incidence.
- To define the organization of a case-control study.
- To calculate and interpret the odds ratio (OR) as well as the confidence interval.
- To understand the strength and weakness of CCS.
- To understand the concept of bias and confounding.
- To define and organize a cohort study.
- To calculate and interpret the relative risk or risk ratio (RR).
- To understand the different biases, and know how to detect them.

Goals of the lecture

- To understand the cross-sectional study along its strength and weakness.
- To distinguish from the concepts of prevalence and incidence.
- To define the organization of a case-control study.
- To calculate and interpret the odds ratio (OR) as well as the confidence interval.
- To understand the strength and weakness of CCS.
- To understand the concept of bias and confounding.
- To define and organize a cohort study.
- To calculate and interpret the relative risk or risk ratio (RR).
- To understand the different biases, and know how to detect them.
- To understand the difference between association and causation.

Outline

Next Section ...

Measures of Disease Frequency

A fundamental task in medical studies is to quantify or measure the occurrence of a disease in a population. Obtaining a measure of the disease occurrence is one of the first step in understand the disease of interest.

Ratios $\frac{a}{b}$, where a is NOT part of b

Measures of Disease Frequency

A fundamental task in medical studies is to quantify or measure the occurrence of a disease in a population. Obtaining a measure of the disease occurrence is one of the first step in understand the disease of interest.

Ratios $\frac{a}{b}$, where a is NOT part of b

Proportions $\frac{a}{b}$, where a is INCLUDED in b

Measures of Disease Frequency

A fundamental task in medical studies is to quantify or measure the occurrence of a disease in a population. Obtaining a measure of the disease occurrence is one of the first step in understand the disease of interest.

- Ratios $\frac{a}{b}$, where a is NOT part of b
- Proportions $\frac{a}{b}$, where a is INCLUDED in b
- Rates $\frac{a}{b}$, where a is the number of affected in a given time interval, while b is the population at risk over the same interval.

Prevalence

Proportion of cases in the population at a given time, indicating how widespread the disease is.

$$\text{Prevalence} = \frac{\text{\#cases observed at time } t}{\text{\#individuals at time} t}$$

- An outbreak of diarrhea on a cruise ship in a day.
- 8 persons out of 86 on the ship are exhibiting signs of diarrhea.
- The prevalence of diarrhea at this particular time is $8/86 = 0.092$

Incidence

- Cumulative incidence rate (CIR)

$$\mathrm{CIR} = \frac{\text{\#newly disease cases in a time interval}}{\text{\#individuals-at-risk}}$$

  - In the cruise ship example
  - 12 persons developed diarrhoeal disease after 5 days of the outbreak.
  - The indicidence rate was therefore $12/86 = 0.14$
- Incidence density rate (IDR)
  - **rate of occurrence of new cases**
  - conveys information about the risk of contracting the disease.

$$\mathrm{IDR} = \frac{\text{\#individuals of newly disease}}{\text{total time for all disease-free individuals-at-risk}}$$

Outcome and Exposure

A good research statement of a medical problem is often simple, considering one exposure and one outcome variable (e.g. organic solvents and brain tumor).

- Primary outcome
- Secondary outcome

Research questions ask about the associations between exposures and outcomes.

> *Is proximity of the residents to the coke-works site associated with hospital admission of children for repiratory problems?*

A hypothesis is then formulated to predict the result.

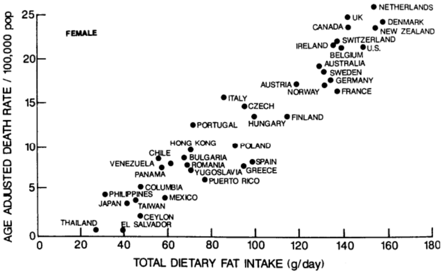
> *The risk of hospital admission for respiratory problems for children increases with proximity to the coke-works site.*

Epidemiological Studies: Categories

- Descriptive Studies
  - Correlational (Ecological) studies
  - Case reports; Case series
  - Cross-sectional studies
- Analytic Studies
  - Case-cohort study
  - Cohort study
  - Intervention study
- Time-dependent
  - Retrospective study
  - Prospective study

Ecological Fallacy

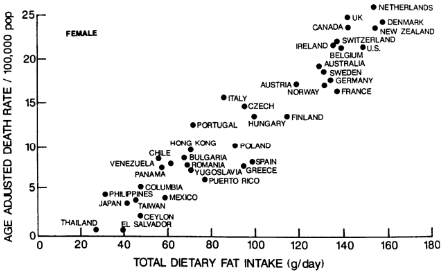
Higher fat intake causes breast cancer?



- Ecological fallacy is an error in reasoning, usually based on mistaken assumptions.

Ecological Fallacy

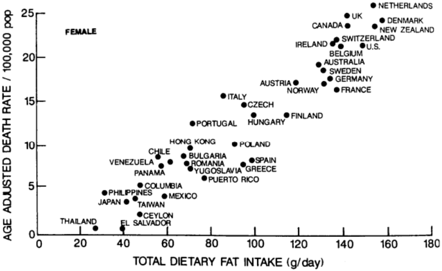
Higher fat intake causes breast cancer?



- Ecological fallacy is an error in reasoning, usually based on mistaken assumptions.
- The ecological fallacy occurs when you make conclusions about individuals based only on analyses of group data.

Ecological Fallacy

Higher fat intake causes breast cancer?



- Ecological fallacy is an error in reasoning, usually based on mistaken assumptions.
- The ecological fallacy occurs when you make conclusions about individuals based only on analyses of group data.
- In this fat intake example, we do not know individuals having breast cancer were actually consumed high amount of fat.

Next Section ...

Case Report: Example

Haemorrhagic-fever-like changes and normal chest radiograph in a doctor with SARS.

Wu E., et al. Lancet 2013; 361(9368): 1520-1521.

This case report is written by the patient, a physician who contracted SARS, and his colleague who treated him, during the 2003 outbreak of SARS in Hong Kong. They describe how the disease progressed in Dr. Wu and based on Dr. Wu’ s case, advised that a chest CT showed hidden pneumonic changes and facilitate a rapid diagnosis.

Kaposi’s sarcoma in homosexual men-a report of eight cases.

Hymes KB. et al. Lancet 1981; 2(8247): 598-600.

This case report was published by eight physicians in New York city who had unexpectedly seen eight male patients with Kaposi’ s sarcoma (KS). Prior to this, KS was very rare in the U.S. and occurred primarily in the lower extremities of older patients. These cases were decades younger, had generalized KS, and a much lower rate of survival. This was before the discovery of HIV or the use of the term AIDS and this case report was one of the first published items about AIDS patients.

Case Report: Definition

Case report is an article that describes and interprets an individual case, often written in the form of a detailed story.

Case reports often describe:

- Unique cases that cannot be explained by known diseases or syndromes
- Cases that show an important variation of a disease or condition
- Cases that show unexpected events that may yield new or useful information
- Cases in which one patient has two or more unexpected diseases or disorders
- the lowest-level of evidence, but also the first-line evidence.

A good case report will be clear about the importance of the observation being reported.

If multiple case reports show something similar, the next step might be a case-control study to determine if there is a relationship between the relevant variables.

Case Report: Pros and Cons

Advantages

- Help the identification of new trends or diseases;
- Help detect new drug side effects and potential uses (adverse or beneficial)
- Educational - a way of sharing lessons learned
- Identifies rare manifestations of a disease

Case Report: Pros and Cons

Advantages

- Help the identification of new trends or diseases;
- Help detect new drug side effects and potential uses (adverse or beneficial)
- Educational - a way of sharing lessons learned
- Identifies rare manifestations of a disease

Disadvantages

- Cases may not be generalizable
- Not based on systmatic studies
- Causes or associations may have other explanations (confounding)
- Can be seen as emphasizing the bizarre or focusing on misleading elements.

Case Series: Example

TNF inhibitors treatment vs. chronic cutaneous sarcoidosis

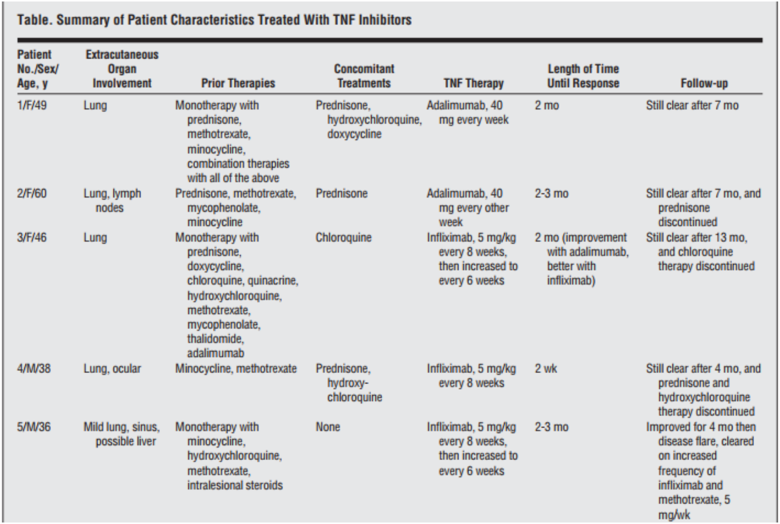


Table. Summary of Patient Characteristics Treated With TNF Inhibitors

Patient No./Sex/ Age, y	Extracutaneous Organ Involvement	Prior Therapies	Concomitant Treatments	TNF Therapy	Length of Time Until Response	Follow-up
1/F/49	Lung	Monotherapy with prednisone, methotrexate, minocycline, combination therapies with all of the above	Prednisone, hydroxychloroquine, doxycycline	Adalimumab, 40 mg every week	2 mo	Still clear after 7 mo
2/F/60	Lung, lymph nodes	Prednisone, methotrexate, mycophenolate, minocycline	Prednisone	Adalimumab, 40 mg every other week	2-3 mo	Still clear after 7 mo, and prednisone discontinued
3/F/46	Lung	Monotherapy with prednisone, doxycycline, chloroquine, quinacrine, hydroxychloroquine, methotrexate, mycophenolate, thalidomide, adalimumab	Chloroquine	Infliximab, 5 mg/kg every 8 weeks, then increased to every 6 weeks	2 mo (improvement with adalimumab, better with infliximab)	Still clear after 13 mo, and chloroquine therapy discontinued
4/M/38	Lung, ocular	Minocycline, methotrexate	Prednisone, hydroxy-chloroquine	Infliximab, 5 mg/kg every 8 weeks	2 wk	Still clear after 4 mo, and prednisone and hydroxychloroquine therapy discontinued
5/M/36	Mild lung, sinus, possible liver	Monotherapy with minocycline, hydroxychloroquine, methotrexate, intralesional steroids	None	Infliximab, 5 mg/kg every 8 weeks, then increased to every 6 weeks	2-3 mo	Improved for 4 mo then disease flare, cleared on increased frequency of infliximab and methotrexate, 5 mg/wk

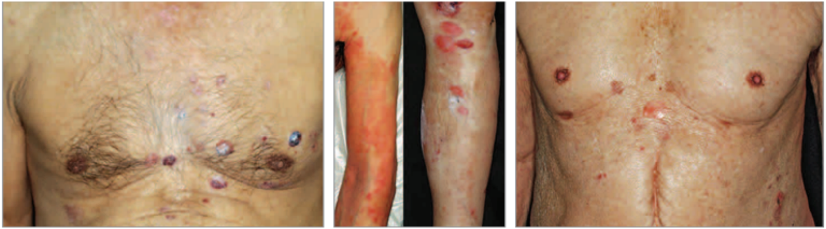
Case Series: Example

Bullous pemphigoid

Clinical and Immunological Profiles of 14 Patients With Bullous Pemphigoid Without IgG Autoantibodies to the BP180 NC16A Domain

Kenta N. et al. JAMA Dermatol. 2018;154(3):347-350.

This case series examines the association of nonreactivity of IgG to the noncollagenous 16A domain using the enzyme-linked immunosorbent assay and the chemiluminescent enzyme immunoassay and severity of disease course in 14 patients with bullous pemphigoid.



Next Section ...

Cross-sectional Studies: Definition

- A cross-sectional study explores the disease and risk factor patterns in a representive part of the population, in a narrowly defined time period.
- Primarily, this study provides information on prevalence of disease and risk factors.
- It also can seek associations, generate and test hypotheses. And, by repetition, cross-sectional study can be used to measure changes.
- Ideal cross-sectional study is of a geographically-defined, representative sample of the population within a slice of time and space.

Cross-Sectional Studies: Example

Association between sleep and hypertension

Insomnia with objective short sleep duration is associated with a high risk for hypertension.

Vgontzas AN et al. Sleep 2009. 32:491-497.

Table : Cross-sectional results of insomnia and hypertension

	Hypertension(+)	Hypertension(-)	
Insomnia(+)	121	78	199
Insomnia(-)	837	705	1542
	958	783	1741

Prevalence of hypertension: $958/1741 = 55\%$

Prevalence of insomnia: $199/1741 = 11\%$

Prevalence of hypertension among insomnia: $121/199 = 61\%$

Prevalence of hypertension among non-insomnia: $837/1542 = 54\%$

Prevalence ratio: $0.61/0.54 = 1.13$

Odds ratio: $\frac{121/78}{837/705} =$

Cross-Sectional Studies: Example

The effect of sunless tanning products on tanning behaviors

A Cross-sectional Study Examining the Correlation Between Sunless Tanning Product Use and Tanning Beliefs and Behaviors.

Rachel E. et al. Arch Dermatol. 2012;148(4):448-454.

Next Section ...

Case-Control Studies: Definition

An epidemiological study in which a group of persons with the disease of interest (case) and a group of persons with similar features to the case group but without the disease (control) are selected to compare the proportion of persons exposed to a risk factor of interest (exposure) in order to elucidate the causal relationship of the risk factor of interest and the disease.

From **case series**, **personal experience**, and **others stuffs**,

- Almost 99% of patients suffered from condition Y had a history/evidence of exposure to X.
- What is the problem in this case study?
  - It will never prove a causal relationship due to lack of control group.
- It will generate a hypothesis testable using epidemiological methods:
  - Persons with disease Y were more likely to have been exposed to factor X comparing to persons without the disease, but similar in other aspects.

That's the case-control study, a retrospective, observational study.

Case-control studies

Table : 2 × 2 contingency table

	Cases	Controls
Exposed(+)	a	b
Exposed(-)	c	d

Case-control studies

Table : 2 × 2 contingency table

	Cases	Controls
Exposed(+)	a	b
Exposed(-)	c	d

What we can answer

- What are the odds that a case was exposed? a/c
- What are the odds that a control was exposed? b/d
- What is the odds ratio? $\frac{a/c}{b/d} = \frac{ad}{bc}$

Note: OR is calculated in different fashion in pair-match case-control study.

Case-control studies

Table : 2 × 2 contingency table

	Cases	Controls
Exposed(+)	a	b
Exposed(-)	c	d

What we can answer

- What are the odds that a case was exposed? a/c
- What are the odds that a control was exposed? b/d
- What is the odds ratio? $\frac{a/c}{b/d} = \frac{ad}{bc}$

Note: OR is calculated in different fashion in pair-match case-control study.

What we CAN'T answer

- the prevalence of disease in exposed and not-exposed.
- the incidences of disease in exposed and not-exposed.
- the relative risk to determine if there is an association between the exposure and the disease.

Case-Control Studies: Computing Odds Ratio (OR)

Table : 2 × 2 contingency table

	Cases	Controls
Exposed(+)	*a*	*b*
Exposed(-)	*c*	*d*

Case-Control Studies: Computing Odds Ratio (OR)

Table : 2 × 2 contingency table

	Cases	Controls
Exposed(+)	a	b
Exposed(-)	c	d

Computing OR

- The OR: $\widehat{\mathrm{OR}} = \frac{a/c}{b/d} = \frac{ad}{bc}$
- The standard error: $\mathrm{SE}\log\mathrm{OR} = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$
- The 95% confidence interval of $\log\mathrm{OR}$: $\widehat{\log\mathrm{OR}} \pm Z^{-1}_{1-\alpha/2}\mathrm{SE}\log\mathrm{OR}$
- *p*-value can be computed using either `Chi-square test` or `Fisher's exact test`.
- In matched case-control study, McNemar's test is used to compute the *p*-value.

Case-Control Study Issues

Bias and confounding

In case-control studies, the inferred association can be introduced by the subject selection process, the problem of data collection, and/or the other factor(s) that are related to both exposure and the outcome under study.

Only when the two concerns - bias and confounding are addressed, the causality relationship can be derived.

Bias

Bias: Any systematic error in the design/subject recruitment, data collection, and/or data analysis that results in a mistaken estimation of the true exposure-outcome association.

Selection Bias: Error due to systematic differences in characteristics between those selected/non-selected into a study; or systematic differences in which cases and controls, exposed and non-exposed subjects are selected so that distorted association is observed.

Smoking-AMI Association

- If hospitalized cases of AMI are selected as cases, and heavy/long-term smokers are more likely to die outside of hospital (massive MI), the smoking-MI association would be underestimated.
- If hospitalized AMI and hostpitalized non-MI patients are selected as cases and controls, and smokers are more likely to be hospitalized for other condition (e.g. pulmonary), the smoking-AMI association would be underestimated.

Selection Bias

- Exposure has influence on the process of case assessment: the exposure prevalence in cases is biased.
- Exposure has influence on the process of control selection: the exposure prevalence in controls is biased (e.g. use chronic bronchitis patients as controls for study of smoking and CHD, smoking and LC.)
- Selective survival and selective migration: the exposure prevalence in prevalent cases may be biased compared to incident cases.

Selection Bias

- Exposure has influence on the process of case assessment: the exposure prevalence in cases is biased.
- Exposure has influence on the process of control selection: the exposure prevalence in controls is biased (e.g. use chronic bronchitis patients as controls for study of smoking and CHD, smoking and LC.)
- Selective survival and selective migration: the exposure prevalence in prevalent cases may be biased compared to incident cases.

Case-control association is valid only when

- Cases are selected as representative sample of all people with such disease.
- Controls are selected as a representive sample of all people without the disease in the same population as cases.

Information Bias

Misclassification bias

Information bias Systematic error due to inaccurate measurement and/or classification of study variables. It can occur in the classification of outcome, exposures, and covariates.

Information Bias

Misclassification bias

Information bias Systematic error due to inaccurate measurement and/or classification of study variables. It can occur in the classification of outcome, exposures, and covariates.

Example

recall bias Lung cancer cases are more likely to recall their smoking history (quantitatively or qualitatively), which will lead to over-estimation of smoking-cancer relationship.

missing bias Overweighed persons are more likely to have missing data in untrasound examination of carotid arteries.

report bias Females are more likely to under-report of their weight, and males over-report.

Confounders (Confounding)

Confounding is a situation in which a measure of the association between exposure and outcome is distorted because of the relationship between the exposure and the other factor (the confounder) that influences the outcome under study.

Confounding is the most important concept because it impacts the **validity** of an observational study.

Confounders (Confounding)

Confounding is a situation in which a measure of the association between exposure and outcome is distorted because of the relationship between the exposure and the other factor (the confounder) that influences the outcome under study.

Confounding is the most important concept because it impacts the **validity** of an observational study.

Factor X is a confounder only if it meets the 3 criteria

- It is an risk factor of the outcome under study, independent of the risk factor under study.
- It is associated with the exposure under study.
- It is not in the causal pathway between the exposure and the outcome.

How to avoid confounding association?

In the design and conduct of study

- Matching (individual or group) by potential confounders - in case-control studies.
- Collect information on potential confounders.

How to avoid confounding association?

In the design and conduct of study

- Matching (individual or group) by potential confounders - in case-control studies.
- Collect information on potential confounders.

In the analysis of study

- Conceptualize your potential confounders.
- Stratify by potential confounders to identify major confounders.
- Stratify to derive stratum-specific exposure-outcome association.
- Statistical adjustment of confounders.

Matched Case-Control Studies

- Select controls who are identical (similar) to cases on potential confounders.

Matched Case-Control Studies

- Select controls who are identical (similar) to cases on potential confounders.
  - Pair/Individual match or frequency match

Matched Case-Control Studies

- Select controls who are identical (similar) to cases on potential confounders.
  - Pair/Individual match or frequency match
  - Better control for confounding, especially when the distribution of a confounder do not have much overlap between the cases and source population (e.g., if cases of myocardial infaction tends to be older)

Matched Case-Control Studies

- Select controls who are identical (similar) to cases on potential confounders.
  - Pair/Individual match or frequency match
  - Better control for confounding, especially when the distribution of a confounder do not have much overlap between the cases and source population (e.g., if cases of myocardial infaction tends to be older)
  - Analytical methods: Matched paired analysis or conditional logistic regression

Matched Case-Control Studies

- Select controls who are identical (similar) to cases on potential confounders.
  - Pair/Individual match or frequency match
  - Better control for confounding, especially when the distribution of a confounder do not have much overlap between the cases and source population (e.g., if cases of myocardial infaction tends to be older)
  - Analytical methods: Matched paired analysis or conditional logistic regression

Table : Pair-matched data

Cases	Controls	
	Exposed	Non-exposed
Exposed	a	b
Non-exposed	c	d

How to compute the odds ratio (OR)?

Matched Case-Control Studies

- Select controls who are identical (similar) to cases on potential confounders.
  - Pair/Individual match or frequency match
  - Better control for confounding, especially when the distribution of a confounder do not have much overlap between the cases and source population (e.g., if cases of myocardial infaction tends to be older)
  - Analytical methods: Matched paired analysis or conditional logistic regression

Table : Pair-matched data

Cases	Controls	
	Exposed	Non-exposed
Exposed	a	b
Non-exposed	c	d

How to compute the odds ratio (OR)?

$$\mathrm{OR} = \frac{b}{c}$$

Matched Paired Data Analsysis

Table : Pair-matched data

Cases	Controls	
	Exposed	Non-exposed
Exposed	a	b
Non-exposed	c	d

Matched Paired Data Analsysis

Table : Pair-matched data

Cases	Controls	
	Exposed	Non-exposed
Exposed	a	b
Non-exposed	c	d

Computing OR

- a and d are concordant pairs; b and c are discordant pairs.
- The odds ratio: $\widehat{\mathrm{OR}} = \frac{b}{c}$
- The standard error: $\mathrm{SE}_{\log \hat{\mathrm{OR}}} = \sqrt{\frac{1}{b} + \frac{1}{c}}$
- The confidence interval of log OR:

$$\log \widehat{\mathrm{OR}} \pm Z_{1-\alpha/2} \mathrm{SE}_{\log \hat{\mathrm{OR}}}$$

- The regular McNemar's Chi-square test:

$$\chi^2_{\mathrm{McN}} = \frac{(b-c)^2}{b+c} \sim \chi^2_{df=1}$$

- The continuity-correct McNemar's Chi-square test:

$$\chi^2_{\mathrm{McN}} = \frac{(|b-c|-1)^2}{b+c} \sim \chi^2_{df=1}$$

A Paired Match Analysis

estrogen and endometrial carcinoma

Association Of Exogenous Estrogen And Endometrial Carcinoma

Donald C. Smith et al. NEJM 1975 Dec 4;293(23):1164-7.

- Retrospective case-control study;
- 317 patients with endometrium carcinoma
- 317 matched controls with other gynecologic neoplasms

Table : Paired study of association between estrogen and endometrial carcinoma

	Controls		
Cases	Exposed	Non-exposed	Total
Exposed	39	113	152
Non-exposed	15	150	165
Total	54	263	317

A Paired Match Analysis

Association Of Exogenous Estrogen And Endometrial Carcinoma

Donald C. Smith et al. NEJM 1975 Dec 4;293(23):1164-7.

- Retrospective case-control study;
- 317 patients with endometrium carcinoma
- 317 matched controls with other gynecologic neoplasms

Table : Paired study of association between estrogen and endometrial carcinoma

	Controls		
Cases	Exposed	Non-exposed	Total
Exposed	39	113	152
Non-exposed	15	150	165
Total	54	263	317

$$\mathrm{OR} = 113/15 = 7.5$$

Estrogen and Endometrial Carcinoma

Misuse of unmatched analysis

Association Of Exogenous Estrogen And Endometrial Carcinoma

Donald C. Smith et al. NEJM 1975 Dec 4;293(23):1164-7.

Table : Unmatched study of association between estrogen and endometrial carcinoma

	Cases	Controls	Total
Exposed	152	54	206
Non-exposed	165	263	428
Total	317	317	634

Estrogen and Endometrial Carcinoma

Misuse of unmatched analysis

Association Of Exogenous Estrogen And Endometrial Carcinoma

Donald C. Smith et al. NEJM 1975 Dec 4;293(23):1164-7.

Table : Unmatched study of association between estrogen and endometrial carcinoma

	Cases	Controls	Total
Exposed	152	54	206
Non-exposed	165	263	428
Total	317	317	634

$$\mathrm{OR} = \frac{152/54}{165/263} = 4.5$$

Estrogen and Endometrial Carcinoma

Misuse of unmatched analysis

Association Of Exogenous Estrogen And Endometrial Carcinoma

Donald C. Smith et al. NEJM 1975 Dec 4;293(23):1164-7.

Table : Unmatched study of association between estrogen and endometrial carcinoma

	Cases	Controls	Total
Exposed	152	54	206
Non-exposed	165	263	428
Total	317	317	634

$$\mathrm{OR} = \frac{152/54}{165/263} = 4.5$$

Conclusion

- The unmatched OR (4.5) is biased towards null compared to the matched one (7.5).
- The stronger the confounders, the biased the association.

Case-control Studies: Strengths and Limitations

Advantages

- Small sample size (rare disease)
- Less time (disease with long induction and latent period)
- Less expensive (small N, efficient for exposure that is expensive to measure)

Case-control Studies: Strengths and Limitations

Advantages

- Small sample size (rare disease)
- Less time (disease with long induction and latent period)
- Less expensive (small N, efficient for exposure that is expensive to measure)

Disadvantages

- Inefficient for rare exposure
- Selection bias (Are cases representative? Do controls represent the source population?)
- Challenge in measurement of exposure (measure exposure after the occurrence of disease)
- Difficulty in determine temporality
- Only one outcome is studied
- Incidence of disease cannot be studied, though OR is intended to estimate incidence ratio

Exercise: Case-control study

Tabacco use and acoustic neuroma

Role of tobacco use in the etiology of acoustic neuroma

Palmisano S et al. Am J Epidemiol. 2012 Jun 15;175(12):1243-51.

- What kind of study? case-control study
- How many cases and controls? 451 cases and 710 population-based controls.
- How did the author match the cases and controls?
- Can you write down the 2 × 2 contingency tables here?
- How did the author analyze the data?
- What was the effect size used in this article? and the confidence interval?
- Why did the author analyze the data in males and females, respectively?

Next Section ...

Cohort study

- Observational study, with selection into study on basis of exposure status at the beginning of the study
- Either prospective study or retrospective study

Flu Vaccine Study

An epidemiology case study

Influenza Vaccination and Reduction in Hospitalizations for Cardiac Disease and Stroke among the Elderly.

Kristin Nichol et al.: NEJM 2003;348:1322-32.

These investigators used the administrative data bases of three large managed care organizations to study the impact of vaccination in the elderly on hospitalization and death. Administrative records were used to whether subjects had received influenza vaccine and whether they were hospitalized or died during the year of study.

Flu Vaccine Study

Summarization table

The table below summarizes findings during the 1998-1999 flu season.

Table : flu vaccine study data in 1998-1999

	Vaccinated	Unvaccinated
	$N = 77,738$	$N = 62,217$
Hospitalized due to pneumonia or influenza	495	581
Hospitalized due to cardiac disease	888	1026
Deaths	943	1361

If the exposure is vacination and outcome of interest is death, how to assess the association?

Flu Vaccine Study

Contingency table

Table : flu vaccine study data in 1998-1999

	Vaccinated $N = 77,738$	Unvaccinated $N = 62,217$
Hospitalized due to pneumonia or influenza	495	581
Hospitalized due to cardiac disease	888	1026
Deaths	943	1361

If the exposure is vacination and outcome of interest is death, how to assess the association?

Table : flu vaccine and deaths in 1998-1999

	Dead	Alive	
Vaccinated	943	76,795	77,738
Non-vaccinated	1361	60,856	62,217

Solutions

Relative Risk (RR) and Risk Difference (RD)

Table : A general 2 × 2 contingency table

	Disease	Non-Disease	
Exposed	a	b	r_1
Non-exposed	c	d	r_2

RR approach

$$\mathrm{RR} = \frac{a/r_1}{c/r_2}$$

- measuring the strength of the association.
  - $\mathrm{RR} = 1$ suggests no association.
  - $\mathrm{RR} \to 1$ suggests weak association.
  - $|\mathrm{RR}| >> 1$ suggests a strong association.

RD approach

$$\mathrm{RD} = \frac{a}{r_1} - \frac{c}{r_2}$$

- a better measure of public health impact.
- How much impact would a prevention have?
- How many people would benefit?

Risk Ratio (RR)

measuring the association between vaccination and death

Table : flu vaccine and deaths in 1998-1999

	Dead	Alive	
Vaccinated	943	76,795	77,738
Non-vaccinated	1361	60,856	62,217

$$\mathrm{RR} = \frac{\mathrm{CI}_e}{\mathrm{CI}_u} = \frac{943/77738}{1361/62217} = 0.554$$

Conclusion

- There is a strong association between flu vaccination and death?
- Vaccination can protect ...?
- How about the hospitalization due to two diseases?

Risk Ratio: Confidence Interval

Table : A general 2 × 2 contingency table

	Disease	Non-Disease	
Exposed	a	b	r_1
Non-exposed	c	d	r_2

Risk Ratio: Confidence Interval

Table : A general 2 × 2 contingency table

	Disease	Non-Disease	
Exposed	a	b	r_1
Non-exposed	c	d	r_2

Confidence interval

- $\hat{p}_1 = \frac{a}{r_1}$; $\hat{p}_0 = \frac{c}{r_2}$
- $\widehat{\mathrm{RR}} = \frac{p_1}{p_0}$
- $\mathrm{SE}_{\log \mathrm{RR}} = \sqrt{\frac{1-\hat{p}_1}{r_1\hat{p}_1} + \frac{1-\hat{p}_0}{r_2\hat{p}_0}}$
- The confidence interval of $\log \mathrm{RR}$: $\log \widehat{\mathrm{RR}} \pm z_{1-\alpha/2}\mathrm{SE}_{\log \mathrm{RR}}$
- Hypothesis testing: Chi-square test or Fisher's exact test

Risk Difference (RD)

Attributable risk

Table : flu vaccine and deaths in 1998-1999

	Dead	Alive	
Vaccinated	943	76,795	77,738
Non-vaccinated	1361	60,856	62,217

$$\mathrm{RD} = \mathrm{CI}_e - \mathrm{CI}_u = \frac{943}{77738} - \frac{1361}{62217} = -0.0097 = -97/100000$$

per year.

Conclusion

- Flu vacinnation can reduce the death rate by 97 per 100,000 population per year.
- Flu vaccination has a strong protective effect ....?
- How about the hospitalization due to two diseases?

Different conclusions for RR and RD

Sometimes ...

Table : Annual Mortality Per 100,000 (CI)

	LC	CHD
Smokers	140	669
Non-smokers	10	413
RR	14	1.6
RD	130	256

Note: LC - lung cancer; CHD - coronary heart disease

Conclusion

- Smoking is a stronger risk factor for ...?
- Smoking is a bigger public health problem for ...?

Benefits and Risks for Different Diseases

Protective and risk effects of aspirin

Table : RRs and RDs of aspirin on some heart diseases

	Aspirin	Placebo	Risk	Risk
	(/10,000)	(/10,000)	Ratio	Diff
MI	125.9	216.6	0.59	-100
Stroke	107.8	88.8	1.2	19
- Ischemic	82.4	74.3	1.1	8
- Hemorrhagic	20.8	10.9	1.9	10
Upper GI ulcer	153.1	125.1	1.2	28
- with hemorrhage	34.4	19.9	1.7	15
Bleeding	2699.1	2037.3	1.3	690
- Transfusion need	43.5	25.4	1.7	18

The risk difference is in /10,000 persons per year.

Rare Outcomes - RR or RD?

It depends ...

If we are going to discuss rare, but severe possible complications of influenza vaccine, would it be better to look at the RR or the RD?

Observed frequencies

- Exposed people: 2/100,000
- Non-exposed people: 1/100,000

Choices

- RR=2: those exposed had two times the risk! (OMG!)
- RD=1/100,000: the exposed group had an excess risk of 1 case per 100,000 subjects (NO THAT IMPORTANT!).

Attributable Risk %

Attributable proportion

What % of risks in the exposed group can be attributed to having had the exposure?

The proportion (%) of disease in the exposed group that can be attributed to the exposure, i.e., the proportion of disease in the exposed group that could be prevented by eliminating the exposure:

$$\mathrm{AR\%} = \frac{\mathrm{RD}}{I_e} = \frac{0.053 - 0.013}{0.053} \times 100 = 75\%$$

Interpretation

75% of risks occurring in patients who had the exposure could be attributable to the exposure.

Cohort study：Pros and cons

Strengths

- Accurate exposure
- Subjects in cohorts can be matched, limiting the influence of confounders.
- Temporal relationship: short-term or long-term outcome
- Multiple outcomes: cerebral dysfunction, myocardial infarction, etc.
- Very rich study design, but it is diffcult and often act as the second step.
- Easier and chpeaper than an RCTs.

Limitations

- Cohorts can be difficult to identify due to confounders.
- No randomization leads to imbalance in patient features.
- Blinding/masking is difficult.
- More difficulties related to both costs and feasibility because of the longitudinal follow-ups
- Different lost-to-follow-up (LTFU) rates for healthier and diseased groups

Limitations of Observational Studies

- Patients who receive on-pump or off-pump CABG may differ in many ways besides their type of surgery that could be related to the outcome under study (confounding).
- In observational studies, on- vs. off-pump CABG was related to:
  - extent and location of heart disease
  - obesity
  - chronic obstructive pulmonary disease (COPD)
  - experience of practitioner with off-pump
  - etc.
- These differences could explain differences in rates of endpoints such as mortality or complications seen in observational studies between on- vs. off-pump patients.
- We can control for known confounders through univariate stratification, or multivariate analysis (MVA), but we can't control for unknown, unmeasured, or unmeasurable differences.
- The amount of uncontrolled bias and confounding could be as large as the effect under assessment.

Propensity score

A method to minimize the confounding in observational studies

- In RCTs, the proper randomization provides a balanced distribution of both observed and unobserved factors between different study arms, which minimize confounding.
- Observational studies often suffer because such balance cannot be met.
- Multivariate regression can provide adjusted estimate for the main exposure by including the confounders in a model.
- However, when more than the allowable number of covariates need to be adjusted, we may need to consider dimensionality reduction.
- Propensity score adjustment is a method of control for potential confounders for the association between a pre-specified exposure and outcome, which provides a quasi-RCT setting for observational data analysis.

The propensity score analysis works by creating a single score estimate on the probability to exposure, which is used to control for many potential confounders at once without a loss of analytical power, is considered the most aggressive adjustment for measured confounders.

Aspirin Use and All-Cause Mortality

A propensity score matching and adjustment

Aspirin Use and All-Cause Mortality Among Patients Being Evaluated for Known or Suspected Coronary Artery Disease: A Propensity Analysis

Gum PA et al. JAMA, 2001 Sep 12; 286(10):1187-94.

- Prospective cohort study
- Cleveland Clinic, 1990-1998
- Mean follow-up 3.1 years
- 6174 patients undergoing stress echocardiography for evaluation of known or suspected coronary disease

Exercise: Propensity Score Matching

Association of statin use and cataracts

Association of statin use with cataracts: a propensity score-matched analysis.

Leuschen J, et al. JAMA Ophthalmol. 2013 Nov;131(11):1427-34.

- What was the primary outcome?
- What was the exposure in this study?
- Which study design did this study adopt?
- Why do we say it is a cohort study other than a case-control study?
- What were the variables used to compute the propensity scores? Why and how?
- How did the authors compare the baseline characteristics between statin-users and non-statin users?
- Why did the authors conduct sensitivity analysis?
- What is the difference between primary analysis and secondary analysis?
- Read the comments and response supplementing this article, and give your own comments.

Observational Studies: Summary

- Cross-sectional Study

Observational Studies: Summary

- Cross-sectional Study
- Case-control study
  - Matched case-control study

Observational Studies: Summary

- Cross-sectional Study
- Case-control study
  - Matched case-control study
- Cohort study

Observational Studies: Summary

- Cross-sectional Study
- Case-control study
  - Matched case-control study
- Cohort study

Important issues

Observational Studies: Summary

- Cross-sectional Study
- Case-control study
  - Matched case-control study
- Cohort study

Important issues

- Low response rate in cross-sectional study

Observational Studies: Summary

- Cross-sectional Study
- Case-control study
  - Matched case-control study
- Cohort study

Important issues

- Low response rate in cross-sectional study
- Low follow-up rate in cohort study

Observational Studies: Summary

- Cross-sectional Study
- Case-control study
  - Matched case-control study
- Cohort study

Important issues

- Low response rate in cross-sectional study
- Low follow-up rate in cohort study
- Selection bias in case-control study

Observational Studies: Summary

- Cross-sectional Study
- Case-control study
  - Matched case-control study
- Cohort study

Important issues

- Low response rate in cross-sectional study
- Low follow-up rate in cohort study
- Selection bias in case-control study
- Information (misclassification) bias in all studies

Observational Studies: Summary

- Cross-sectional Study
- Case-control study
  - Matched case-control study
- Cohort study

Important issues

- Low response rate in cross-sectional study
- Low follow-up rate in cohort study
- Selection bias in case-control study
- Information (misclassification) bias in all studies
- Confounding in all studies

Observational Studies: Summary

- Cross-sectional Study
- Case-control study
  - Matched case-control study
- Cohort study

Important issues

- Low response rate in cross-sectional study
- Low follow-up rate in cohort study
- Selection bias in case-control study
- Information (misclassification) bias in all studies
- Confounding in all studies
- Use the most appropriate analytical methods

Design a Study Protocol

(A) What is your Research Question and Hypothesis

(B) Target population

(C) Explanatory variable (Exposure)

(D) Response variable (Outcome)

(E) Extraneous variable (Potential confounders)

(F) Anatomy of the design

(G) Planned analyses

(H) Sample size requirements

(I) Study limitations

Statistical Testing Choices

Continuous outcome (mean)

- Two independent groups
  - *t*-test (parametric)
  - Wilcoxon Mann-Whitney rank-sum test (nonparametric)
- 2+ independent groups
  - ANOVA (parametric)
  - Kruskal-Wallis test (nonparametric)
- Paired data
  - Paired *t*-test (parametric)
  - Wilcoxon signed-rank test (nonparametric)
- 2+ dependent groups
  - ANOVA
  - Friedman's test

Statistical Testing Choices

Continuous outcome (mean)

- Two independent groups
  - *t*-test (parametric)
  - Wilcoxon Mann-Whitney rank-sum test (nonparametric)
- 2+ independent groups
  - ANOVA (parametric)
  - Kruskal-Wallis test (nonparametric)
- Paired data
  - Paired *t*-test (parametric)
  - Wilcoxon signed-rank test (nonparametric)
- 2+ dependent groups
  - ANOVA
  - Friedman's test

Categorical outcome (Percentage)

- Independent groups (2+)
  - χ^2-test
  - Fisher's exact test

Questions?