BI476 - Biostatistics Case Study

Assignment 2: Observational Studies

2018 Spring

Learning Objectives

After successfully completing this section, you will be able to:

- Explain the differences among the metrics: ratio, proportion, and & rate.
- Define and calculate prevalence (and be able to distinguish between point prevalence and period prevalence). Be able to explain the use of prevalence in public health.
- Define and distinguish between cumulative incidence and incidence rate, and describe their strengths and limitations.
- Define and identify confounding.
- Identify three ways to control confounding in the design phase of a study, and identify the strengths and weaknesses of each approach
- Describe ways to control for confounding in the analysis phase of a study.
- Grasp the techniques to analyze a (paired/unpaired) case-control study, cohort study.
- Know how to compute the odds ratios, risk ratios, risk difference as well as their respective confidence intervals.

1 Background Knowledge

1.1 Observational studies

1.1.1 Descriptive study

The desriptive study has the advantages of being cheap, fast, easy to collect information. This kind of study is often used when little is known about the disease. The goal of these studies is to describe the occurrence of disease in population and generate hypotheses.

- Case report describe a single case.
- Case series describe a group of cases.
- Cross-sectional study (prevalence study) is a snapshot of health at any defined time point or very short time period, which can describe the attributes of a population based on a sampling from the population.
- Correlational (ecological) study is a aggregation-level study, which evaluate the correlation between exposure and disease at a group-level rather than at an individual-level.

1.1.2 Analytical studies

The analytical studies seek to identify and explain the causes of diseases and usually calculate a numerical metric to quantify the effect of a potential risk (exposure) factor. The aim of analytical studies is to determine the strength, importance and statistical significance of epidemiological associations.

- Case-control study is an observational, retrospective study.
- Cohort study is an observational study, which can be either retrospective or prospective.
- Randomized trials are non-observational retrospective studies.

1.2 How to evaluate the performance of screening/diagnosis

- **True Positive (TP, 真阳性)**
- **False Positive (FP, 假阳性)**
- **True Negative (TN, 真阴性)**
- **False Negative (FN, 假阴性)**
- **False Discovery Rate (FDR)**
- **Positive predictive value (PPV，阳性预测值)** is the probability that subjects with a positive screening test truly have the disease.
- **Negative predictive value (NPV, 阴性预测值)** is the probability that subjects with a negative screening test truly don't have the disease.
- **Sensitivity (敏感性)**: the accuracy of a screening test in identifying disease in people who truly have the disease.

$$\text{Sensitivity} = \text{True Positive Fraction} = P(\text{Screen Positive}|\text{Disease})$$

- **Specificity (特异性)**: the probability that non-diseased subjects will be classified as normal by the screening test.

$$\text{Specificity} = \text{True Negative Fraction} = P(\text{Screen Negative}|\text{Disease Free})$$

- **Receiver's Operating Characteristic Curve (ROC，受试者操作特征曲线)**

2 Confoundings

The term **extraneous variable** is a general term for variables that affect the DV and are linked to the IV. When extraneous are recognized during the design stage of the experiment, researchers use techniques to turn them into **controlled variables**. If extraneous variable go unrecognized, they become **confounded variables**.
The terms, **confounded, controlled**, and **extraneous** refer to variables that can influence the DV for one level of the IV differentially than they do for another level of the IV. However, **nuisance variables** affect the DV, but all levels of the IV are affected equally.

2.0.1 Confounding factors

There are three conditions that must be present for confounding to occur:

(1) The confounding factor must be associated with both the risk factor of interest and the outcome (i.e., an independent risk factor for outcome of interest, associated with exposure).

(2) The confounding factor must be distributed unequally among the groups being compared.

(3) A confounder cannot be an intermediary step in the causal pathway from the exposure of interest to the outcome of interest (i.e., not a result of the exposure of interest).

2.0.2 How to control for confounding in DESIGN stage?

- Randomization (controlled trials/experiments only)
- Matching (only when the confounding is known)
- Restriction (only to homogeneous subgroups)

2.0.3 How to control for confounding in ANAlYSIS stage?

- Stratification (Separate analysis by stratification of the confounding)
- Restriction (Restrict analysis only to some of the levels)
- Multivariate analysis (including potential confoundings as covariates)

3 Exercises

1. (10 points) Define the following terms in your own words.
   (1) Prevalence

   Solution: Prevalence is the proportion of subjects/individuals with some attribute or outcome (event) at a point of time or in a period of time.

   (2) Prevalence ratio

   Solution: Prevalence Ratio indicates how large is the prevalence of an event/outcome in one group of subjects/individuals (with characteristics/attribute) relative to another group (without the characteristics/attributes).

   (3) Cumulative incidence

   Solution: Cumulative incidence is the probability of developing disease over a stated period of time; it is an estimate of risk.

   (4) Incidence rate

Solution: The incidence rate is a measure of the number of new cases ("incidence") per unit of time ("rate").

(5) Risk

Solution: Risk is the probability of occurrence of a new event over a period of time among those who are at risk for event occurrence at the beginning of the follow up period.

(6) Risk ratio (RR)

Solution: Risk Ratio or Cumulative Incidence Ratio indicates how more or less likely one a group of individuals/subjects with attribute/characteristics (exposure) is to develop/acquire a health outcome or condition over the follow up period relative to the other group of unexposed.

(7) Odds

Solution:

(8) Odds ratio (OR)

Solution:

(9) Confounding factors

Solution: confounding is a type of bias that can lead to a distortion of the association between an exposure and an outcome that occurs when the study groups differ with respect to other factors that influence the outcome.

(10) Bias

Solution: A systematic error in the design, recruitment, data collection or analysis that leads to a mistaken estimation of the true effects of the exposure and the outcome.

(11) Effect modification

Solution: Effect modification occurs when the magnitude of the effect of the primary exposure on an outcome (i.e., the association) differs depending on the level of a third variable (factor, covariate).

2. (10 points) True or False.

(1) **T** Case reports are not considered evidence-based study because they involve only one or several patients and are thus not systematic research.

(2) **F** Case reports can demonstrate causality or argue for the adoption of a new treatment approach.

(3) **T** The patient should be described in detail, allowing others to identify patients with similar characteristics.

(4) **T** A highly sensitive test in diagnosis will reduce the likelihood of false positives.

(5) **F** Specificity is the likelihood of having a disease if you have a negative test.

(6) **F** Predictive value of a positive test is the likelihood that persons with the disease will have a positive test.

(7) **T** The interviewer bias is a kind of information bias since the interviewer may incorporate his own bias into the way that the question is asked and mislead the respondents.

3. (10 points) Make choices on the following question.

(1) One hundred healthy troops go on a 1-year mission to a malariandemic area of North Africa. During their stay 5 new cases of malaria are dentified, for a rate of 5%. This number is a:

A. prevalence rate
B. mortality rate
C. incidence-density rate
D. cumulative-incidence rate

(2) Among untreated children under age 3, 20% of malaria cases are fatal. 20% is a:

A. case-fatality rate
B. mortality rate
C. attack rate
D. prevalence

(3) What is the name of statistician who invented the exact test for 2-by-2 table?

A. Pearson
B. Fisher
C. Galton
D. Kolmorov

(4) For a same data, which produces the larger P-vallue?

A. Pearson's uncorrelated chi-square test
B. Yates' continuity-corrected chi-square test?

(5) During the past year, 7 new cases of multiple sclerosis were diagnosed in your community of 100,000 people. At any one time during the year, the prevalence of multiple sclerosis in your community was probably

A. substantially higher than 7/100,000
B. substantially lower than 7/100,000
C. about 7/100,000
D. equal to the cause-specific mortality rate
E. Unknown

(6) Three years ago there was a multistate outbreak of illnesses caused by a specific and unusual strain of Listeria monocytogenes. As part of the investigation of this outbreak, CDC workers checked the food histories of 20 patients infected with the outbreak strain and compared them with the food histories of 20 patients infected with other Listeria strains. This study design is best described as which one of the following:

A. Analytical, experimental
B. Analytical, observational, case-control
C. Analytical, observational, cohort
D. Descriptive

(7) The initial studies establishing maternal diethylstilbesterol (DES) intake as a cause of vaginal adenocarcinoma in female offspring were case-control studies. This was probably largely because:

A. A couple of decades ago cohort studies hadn't been invented.
B. A woman taking DES was always rare.
C. The disease outcome is rare.
D. The investigators had probably just happened to have a number of cases in their practices.

(8) In a case-control study of alcohol intake and bladder cancer, cases and matched controls are each interviewed by interviewers who are not blinded as to whether the subject is a case or a control. Many of the interviewers are in fact convinced that drinking alcohol is a cause of bladder cancer. Is this likely to represent a bias?

A. No, because the interviewers can't affect whether the subjects are considered cases or controls; that's already decided
B. Yes, but it's hard to predict the direction of the bias.
C. Yes, and would predispose to a rejection of the null hypothesis.
D. Yes, and would predispose to an acceptance of the null hypothesis.

(9) A published study follows a large group of women with untreated dysplasia of the uterine cervix, documenting the number who improve, stay unchanged, or progress into cervical cancer. This study design is best described as which one of the following:

A. Analytic, experimental
B. Analytic, observational, cohort
C. Analytic, observational, case/control
D. Descriptive, observational

(10) A community assesses a random sample of its residents by telephone questionnaire. Obesity is strongly associated with diagnosed diabetes. This study design is best described as which one of the following:

A. Case-control
B. Cohort
C. Cross-sectional
D. Experimental

(11) Data suggest that older age is a risk factor for fall. You decide to assess this by identifying the next 100 fall victims in the emergency room and comparing them for age with 100 emergency room patients who did not fall. What is this study?

A. Cross-sectional
B. Cohort
C. Case-control
D. Randomized double-blinded clinical trial

(12) The risk of a smoker developing lung cancer was found to be 10 times higher than that of a non-smoker in Shanxi. Since Shanxi has over 30 naturally occuring asbestos sites and several mines, the researchers suspected asbestos mining was a confounder. When the researchers adjusted for the confounding effect of working in the asbestos mines, they found that smokers were only 7 times more likely to develop lung cancer.

What was the magnitude of confounding in this study, and was the relationship between smoking and lung cancer truly confounded by asbestos mining exposure?

A. Asbestos mining made the the relationship between smoking and lung cancer appear 30% higher than it truly was. Because the percent difference between the crude and adjusted measures of risk was more than 10%, confounding was present.

B. Asbestos mining made the the relationship between smoking and lung cancer appear 30% lower than it truly was. Because the percent difference between the crude and adjusted measures of risk was more than 10%, confounding was present.

C. Asbestos mining made the the relationship between smoking and lung cancer appear 43% lower than it truly was. Because the percent difference between the crude and adjusted measures of risk was more than 10%, confounding was present.

D. Asbestos mining made the the relationship between smoking and lung cancer appear 43% higher than it truly was. Because the percent difference between the crude and adjusted measures of risk was more than 10%, confounding was present.

(13) Which of the following is NOT a selection bias?

A. Surveillance bias

B. Non-response bias

C. Inappropriate choice of control

D. Misclassification bias

4. (15 points) Each year cardiologists perform procedures to blocked coronary arteries only to have may of these repaired arteries re-clog (restenosis) afterwards. A study sponsored by the NIH Heart, Lung and Blood Institute was performed to determine whether prior infection with cytomegalovirus was predictive of arterial restenosis. In 21 of the 49 patients with serologic evidence of cytomegalovirus infection, re-growth of arterial plaque was noted. In contrast, only 2 of the 26 patients seronegative patient had restenosis.

   (1) Calculate the risk ratio of restenosis associated with CMV infection. Include a 95% confidence interval.

   (2) Try to interpret results.

   (3) Conduct a chi-square test of $H_0 : RR = 1$.

5. (5 points) How to correct for the effect of confounding factors on the association between the outcome and exposure of interest? Can you post some of the commonly-used methods.

Solution: At the design stage:

- Randomization
- Matching
- Restriction

At the stage of analysis:

- Stratified analysis
- Adjustment
- Restriction

6. (20 points) Read the three articles and answer the corresponding questions.

   - Chambers C., et al. Selective serotonin-reuptake inhibitors and risk of persistent pulmonary hypertension of the newborn. NEJM 2006; 354(6): 579-587.
   - Smedby K., et al. Autoimmune and chronic inflammatory disorders and risk of non-hodgkin lymphoma by subtype. JNCI 2006; 98(1): 51-60.
   - Teo, K. et al. Tobacco use and risk of myocardial infarction in 52 countries in the INTERHEART study: A case-control study. Lancet 2006; 368(9536): 647-658.

   (1) Is this article paired or not?

   (2) Print out the contingency tables.

   (3) Can you compute each effect size (OR) of the association, also the corresponding 95% confidence interval, and the p-value.

   (4) Did the authors use appropriate inclusion and exclusion criteria to avoid effects of confoundings? Why?

7. (30 points) A study planned to determine the prevalence of HIV-seropositivity in female prison inmates. And the association between HIV and intravenous drug use was studied. The individual records are stored in the data `prison.RData`.

   (1) Write down the cross-tabulated results by intravenous drug use (IVDU) and HIV.

   (2) Calculate the prevalence of HIV in each group.

   (3) Calculate the prevalence ratio associated with IVDU and then interpret your results in your own words.

   (4) Calculate a 95% confidence interval for the prevalence ratio. Interpret your results.

   (5) Use Yates' continuity corrected chis-square test to derive a P-value for the association. Show all hypothesis testing.

   (6) Do we need the Fisher's exact test here?

   (7) Replicate the analysis in R.

   (8) Suppose you were to plan a study in a prison population to see if ethnic group is an independent risk factor for HIV. You want to achieve 90% power with $\alpha = 0.05$ (two-sides). We will use a equal number of study subjects in each ethnic group. Determine the sample size you need to detect a **two-fold** difference in prevalence.

(9) Detect the sample size needed to detect a 50% increase in risk.

8. (10 points) A valid population-based case-control study has been done to explore the relationship between chili-pepper consumption and gastric cancer. The study has produced the following data:

	Gastric cancer		
Exposure	Cases	Controls	Total
Yes	204	552	
No	9	145	
Total			

(1) Estimate the case-exposure fraction (CaE) and the control-exposure fraction (CoE).

(2) Estimate the relative risk (odds ratio) and explain what this OR means in a sentence.

(3) Estimate the attributable AIE(%) $= \frac{\text{OR}-1}{\text{OR}} \times 100$ and explain what it means in your own words.

(4) Estimate the attributable AIT(%) $= \frac{\text{AIE}}{\text{CaE}} \times 100$ and explain what it means in a sentence.