Epidemiology (7) Chapter 5 Types of Epidemiologic Studies (2): Case control study

Minato NAKAZAWA, Ph.D.
<minato-nakazawa@umin.net>

CASE-CONTROL STUDIES

- Main drawback of cohort study
  - Necessity to obtain information on exposure and other variables from **large populations** to measure the risk or rate of disease
  - Usually only a tiny minority of those at risk develops the disease
- Case-control study aims at the same goal as a cohort study
  - More **efficient**, using **sampling**
  - **Properly carried out**, case-control studies provide information mirroring what could be learned from a cohort study
- Samples represents **a source population** (hypothetical study population in which a cohort study might have been conducted)
  - If a cohort study is done, the exposed and unexposed cohort are defined and the denominators are obtained from those populations, then the cases are identified for each cohort.
  - In a case-control study, the same cases are identified and classified according to whether they belong to the exposed and unexposed cohort. Instead of obtaining the denominators, **a control group is sampled** from the entire source population that gives rise to the cases. Individuals in the control group are then classified into exposed and unexposed categories.
- Control group is used to estimate the distribution of exposure in the source population. → Control has to be sampled independently of exposure status.

Nested Case-Control Studies

- (The right figure is slightly different from the textbook, thus the number below is also different from the textbook)
- In the source population, ¼ is exposed (48/192). Suppose that the cases arises during the 1 year follow-up.
- Assume all cases occurring at the end of the year.
  - In exposed cohort, 8 cases occurred within 48 person-years observation. IR(E) is 8/48=0.167
  - In unexposed cohort, 8 cases occurred within 144 person-years. IR(U)=8/144=0.056
  - IR(E)/IR(U) = 3
- Let's consider case-control study. Among the 48 control group, 12 are exposed. If the sample is taken independently of the exposure, the **same proportion** of **controls** will be exposed as the proportion of people (or person-time) exposed in the original **source population**, apart from sampling error. Cases are same as cohort study.
- Any case-control study can be considered as nested case-control study like this, while case-control study actually conducted within a well-defined cohort is referred as **nested case-control study** by epidemiologists. In occupational epidemiology, case-control study nested within an occupational cohort is common. Needed information is readily available.

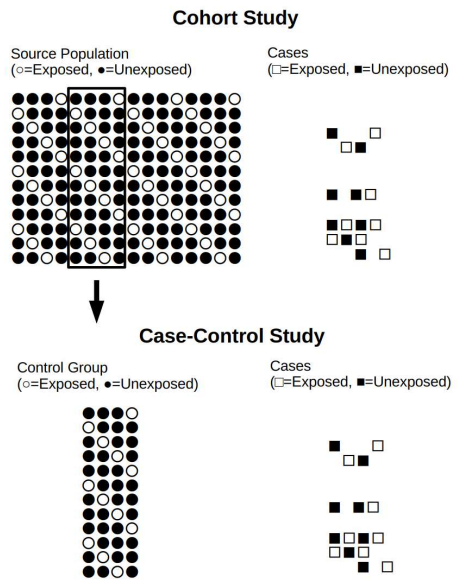

Figure 5-3. Schematic of a cohort study and a nested case-control study within the cohort shows how the control group is sampled from the source population.

An example of case-control study when the source population is difficult to identify

- The cases are patients treated for severe psoriasis at the Mayo Clinic.
- These patients come to the Mayo Clinic from all corners of the world.
- What's the specific source population?
  - We cannot identify it because we cannot know exactly who goes to the Mayo Clinic for severe psoriasis unless they develop severe psoriasis.
  - However, we can imagine a population around the world that constitutes the people who would go to the Mayo Clinic if they developed severe psoriasis.
    - This population is the source population in which the case-control study is nested and from which control-series would ideally be drawn.
    - In practice, the epidemiologists sample the controls from the patients with other disease in Mayo Clinic, because they might come to Mayo Clinic when they suffer from severe psoriasis.

Basic types of case-control studies

- The 3 basic types of case-control studies are defined by the 3 types of sampling controls
- The 3 types of **sampling controls** (if sampling is conducted independently from exposure, we can assume the sample reflects the distribution of exposure and unexposure in the source population)
  - Density-based sampling (Density sampling)
    - Controls are sampled to represent the distribution of person-time in the source population with respect to exposure
  - Cumulative sampling
    - Controls are sampled after the source population has gone through a period of risk, which is presumed to be over when the study is conducted (eg. A case-control study examining the effect of vaccination on the risk of influenza may be conducted at the end of influenza season, when the annual epidemic has ended. Control group is sampled from among those who didn't become cases during the period of risk)
  - Case-cohort sampling
    - Controls are sampled from the list of all people in the source population

Density Case-Control Studies

- Assume dichotomous exposure. Source population has 2 subcohorts, exposed (subscript 1) and unexposed (subscript 0).
- The a and b are the number of people who developed the disease. PT_1 and PT_0 are amounts of person-time at risk. The control series contains c exposed people and d unexposed people.
- The ratios c/PT_1 and d/PT_0 are called as **control sampling rates** for the exposed and unexposed components of the source population. ad/bc (cross product ratio or odds ratio) provides the estimate of IRR.

$$I_1 = \frac{a}{PT_1}, \quad I_0 = \frac{b}{PT_0}$$

$$\frac{c}{d} = \frac{PT_1}{PT_0}, \quad \frac{c}{PT_1} = \frac{d}{PT_0}$$

$$\frac{I_1}{I_0} = \frac{a/PT_1}{b/PT_0} = \frac{a}{b} \times \frac{PT_0}{PT_1} = \frac{a}{b} \times \frac{d}{c} = \mathrm{OR}, \quad \text{because} \frac{d}{c} = \frac{PT_0}{PT_1}$$

- Defining the source population
  - All patients are included as cases
  - Source population corresponds to the eligibility criteria for cases
  - If the cases are identified in a single clinic, the source population is **all people** who would attend that clinic and be recorded with the diagnosis of interest if they had the disease in question.
- Control selection
  - The probability of sampled as control is proportional to the person-time contribution to the denominators of incidence rates in the source population.
  - Until a person becomes a case, the person is included in the denominator of IRR
  - One way: Choose controls from the unique set of people in the source population who are at risk of becoming a case. This unique set changes from one case to another. It's referred as the **risk set**. (**risk-set sampling**)
  - During 3 years study, a person who selected as control in the 1st year develops disease in the 3rd year, the person becomes case. If so, the person has to be counted as both case and control.
  - Even one person can be counted twice or more as control (eg. hepatitis A and raw shellfish ingestion within the previous 6 weeks)
- Example (Table 5-5 and 5-6)
  - Table 4-7 and hypothetical control series
  - Instead of conducting cohort study, by density case-control study, 56 cases were identified, who are all cases in the 2 cohorts. Control series were 500 women.
  - Exposure distribution of controls mirrors the exposure distribution of the person-time in the source population.
  - Of the 47027 person-years of experience in the 2 cohorts, 28010 (59.6%) are related to radiation exposure. 500 multiplied by 0.596 becomes 298, the controls with radiation exposure.

Table 5-5 Hypothetical case-control data of breast cancer with/without radiation exposure

Radiation	Yes	No	Total
Breast cancer	41	15	56
(Person-years)	(28010)	(19017)	(47027)
Control series	298	202	500
Rate (/10000 yr)	14.6	7.9	11.9

$$\mathrm{IRR} = 14.6/7.9 = 1.86$$

Table 5-6 Case-control data alone from 5-5

Radiation	Yes	No	Total
Breast cancer cases	41	15	56
Controls	298	202	500

$$\mathrm{OR\ (Odds\ Ratio)} = (41/15)/(298/202) = \frac{41 \times 202}{15 \times 298} = 1.85$$

OR=IRR (with rounding error)

* Density case-control studies can estimate rate ratios!

Cumulative Case-Control Studies

- In cumulative case-control studies or case-cohort studies, each control represents a certain number of people, corresponding to cohort studies in closed population and measure risks. Effect measure is RR, not IRR.
- Sampling controls from the entire source population at the end of follow-up, which is **from the noncases that remain after the cases have been identified**. Often conducted at the end of epidemic or specific but time-limited risk period.
  - eg. The effect of specific drug exposure during early pregnancy on the occurrence of birth defects. Identify cases who are born with birth defects. Typically control series are sampled from babies born without birth defects. Such controls may not represent the experience of entire source population, because some babies who were at risk of birth defects may die before birth and cannot be included in controls. Thus this way of sampling controls leads to overestimate RR.
- RR can be estimated as OR (=ad/bc), where a and b are the number of exposed and unexposed cases, c and d are the number of exposed and unexposed controls. If the disease is rare (**rare disease assumption**), the experience of cases will be a small part of the overall experience of the source population and **OR is very close to RR**. If the risk for disease is **high**, OR obtained in cumulative case-control studies **overestimate** RR.

Table 5-7. Cumulative sampling vs case-cohort sampling

	Exposed	Unexposed	RR or OR
Cases	40	10	
Cohort denominator	100	100	RR=4.0 = (40/100) /(10/100)
Controls (cumulative)	20	30	OR = 6.0 = (40/10) /(20/30)
Controls (case-cohort)	25	25	OR = 4.0 = (40/10) /(25/25)

- Let's assume half of 200 people in closed cohort were exposed. All cases included and 50 controls by cumulative sampling. At the end, noncases were 150 (60 in exposed and 90 in unexposed).
- In cumulative sampling, exposure distribution of controls represents the exposure distribution of noncases at the end, thus the numbers of controls of exposed and unexposed are 50x(60/(60+90))=20 and 50x(90/(60+90))=30.
- OR=(40/10)/(20/30)=6.0
- If the risks are 4% in exposed and 1% in unexposed, RR is still 4, but the noncases at the end are 96 in exposed and 99 in unexposed, then exposure distribution in controls of exposed and unexposed are 50x(96/(96+99))=24.6~25 and 50x(99/(96+99))=25.4 ~ 25. OR=(40/10)/(25/25)=4.0 (4.1 if 24.6 and 25.4 are used instead).

Case-Cohort Studies

- Sampling controls from the entire source population (at the beginning of follow-up).
- It's used even if the subjects are followed for various amounts of time.
- Each control represents a fraction of the total **number** of people in source population, rather than a fraction of the total person-time. Thus the numbers of controls of exposed and unexposed are 50x(100/200) and 50x(100/200), respectively.
- Since sampling proportion is unknown, actual risks cannot be calculated. But OR is valid estimates of RR.
- No need of rare disease assumption.
- Case-cohort design is more convenient than density case-control design. Especially the same control group can be used to compare with various case series.
- A person selected as a control may also be a case (same as density case-control studies). Theoretically, no problem arises. The control series in a case-cohort study is a sample of the entire list of people who are in the exposed and unexposed cohorts. In cohort study, every person in numerator of risk is also included in the denominator. Similarly, if we sample controls at the start of the study, control sampling represents people who were free of disease. Only later, someone gets disease then becomes case. See "Modern Epidemiology" for case-cohort study in detail.

Table 5-8. Hypothetical case-cohort data for John Snow's natural experiment.

Water company	S&V	Lambeth
Cholera deaths	4093	461
Controls	6054	3946

- From the data in Table 5-1, assume that John Snow conducted case-cohort study instead natural experiment.
- Take 10000 controls to represent the distribution of 2 water companies.
  - 10000x(266516/(266516+173748))=6054
  - 10000x(173748/(266516+173748))=3946
- OR = (4093/461)/(6054/3946) = 5.79 = RR
- The result is essentially same as Snow's value. If Snow knew the case-cohort study and the only the numbers of each water-company users from business records, obtaining the information for each person was not necessary.

Sources for control series

- Ideal method = population-based study: sample controls directly from the source population of cases within a geographic area (***general population control***).
  - The at-risk subset of the population is the source population for cases, who met the study inclusion criteria for age, sex, other factors.
  - If a population registry is available, control sampling becomes easy through random sampling.
  - If no registry nor roster is available, **random-digit dialing** is useful but with a few challenges.
    - It assumes that every case can be reached by telephone
    - Every telephone has equal probability of being called, but households vary in the number of people, in the amount of time someone is at home.
    - Making contact with a household may require many calls at various times of day and various day of the week
    - Some telephone numbers are used for business, not for residential
    - The increase of telemarketing and the availability of caller identification has further compromised response rates to cold calling. Obtaining a control subject meeting specific eligibility characteristics can require dozens of calls
    - Answering machines, multiple phone numbers in one household, …
  - If a geographic roster of residences is unavailable, without enumerating them all, matching is convenient (after a case is identified, one or more controls in the same neighborhood are recruited)
- ***Hospital control***: not population-based, drawing a control series from patients treated at the same hospitals or clinics as the cases.
  - The source population does not correspond to the population of the geographic area, but only to those who would attend the hospital or clinic if they contracted the disease under the study.
  - Any nonrandom sampling of controls may not be independent from exposure. Hospitalized patients with other diseases may have higher possibility to be exposed (one exposure may cause several kinds of diseases)
    - One way to avoid it is exclude patients of diseases with the same causes from controls. Exclusion should be based on the cause of hospitalization used to identify the study subject (not on previous disease).
    - A variety of diagnosis has the advantage of diluting any bias that may result from including as the control series only a specific diagnostic group that turns out to be related to the exposure.
- Proxy sampling: If impossible to identify the actual source population for cases, it's still possible to sample control series with the same exposure distribution as the source population for cases. eg. Case-control study to examine the relationship between ABO blood type and female breast cancer. The brothers of the cases are not part of the source population, but the distribution of ABO blood type are same, and thus the brothers can be a control series.

Prospective and retrospective case-control studies

- Retrospective: Cases have already occurred when the study begins
- Prospective: Investigator must wait until cases will occur
- Usually cohort study is prospective and case-control study is retrospective, but there are retrospective cohort studies and prospective case-control studies
- Some textbook claim that the cases should represent all persons with the disease and that controls should represent the entire non-diseased population. It's misleading. Cases can be defined in any way that the investigator wishes and need not represent all cases. The case definition implicitly defines the source population of cases, from which the controls should be drawn. Cases and controls should represent this source population, not entire nondiseased population

Case-crossover studies

- Malcolm Maclure, The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events, *American Journal of Epidemiology*, Volume 133, Issue 2, 15 January 1991, Pages 144–153, https://doi.org/10.1093/oxfordjournals.aje.a115853
- A case-control version of the crossover study
- All the subjects are cases. The control series is represented by information on the exposure distribution drawn from the cases themselves, outside of the time window during which the exposure is hypothesized to cause the disease
- Only for an appropriate study hypothesis
  - The effect of the exposure must be brief
  - The disease event ideally will have an abrupt onset
- Maclure's example: Whether the sexual intercourse causes myocardial infarction. The period of increased risk after sexual intercourse was hypothesized to be 1 hour (in fact, 2 hours in the paper by Maclure).
  - The cases would be a series of people who had a myocardial infarction
  - Then each case would be classified as exposed if the person had sexual intercourse within the hour preceding the myocardial infarction. Otherwise, the case would be classified as unexposed.
  - There is no separate control series. The control information is obtained fro the cases themselves: The average frequency of sexual intercourse for each case during a period (eg. 1 year) before the myocardial infarction occurred.
  - Unchangeable characteristics (even unmeasured) are the same between cases and controls.
  - The comparison assumes that both exposure and confounding don't systematically change along with time, but the exposure must be something that varies from time to time for a person (Like blood type, unchangeable exposure cannot be examined by case-crossover study).
- It's impossible to escape from the confounding by trend, stratification by time-slice and calculation of pooled odds ratio is applied (Zhang Z. Case-crossover design and its implementation in R. Ann Transl Med. 2016;4(18):341. doi: 10.21037/atm.2016.05.42)

Cross-sectional vs longitudinal studies

- All cohort studies and most case-control studies rely on data in which exposure information refers to an earlier time than that of disease occurrence, making the study longitudinal (It assures the temporality in Hill's checklist of causation).
- Cross-sectional studies: All of the information refers to the same point of time. Snapshots of the population status for exposure and disease
- A cross-sectional study cannot measure disease incidence, because risk or rate calculations require information across a time period.
- Cross-sectional study can assess disease prevalence. It's possible to use cross-sectional data to conduct a case-control study if the study includes prevalent cases and uses concurrent information about exposure.
- Sometimes cross-sectional information is used because it's considerd a good proxy for longitudinal data.

RESPONSE RATES

- In a cohort study, if a substantial proportion of subjects cannot be traced to determine the disease outcome, the study validity can be compromised.
- In a case-control study, if exposure data is missing on a sizable proportion of subjects, it can likewise be a source of concern. The concern stems from the possibility of **bias** from selectively missing data, which is a form of **selection bias**.
- The more missing outcome in cohort study and the more missing exposure in case-control study, the greater the potential for selection bias.
- **Response rates**: the proportion with the disease outcome corresponding to the response in a cohort study and the proportion with exposure information corresponding to the response in a case-control study.
  - If the response rate is less than **70% to 75%**, the study is criticized as doubtful. Differential no-response may occur.
- In cohort studies, better strategy is to concentrate efforts more on follow-up than on recruitment. In case-control studies, if the participants know their exposure status, getting high levels of participation is important, if the participants don't know the exposure status, low recruitment into a case-control study is less of a concern.

COMPARISON OF COHORT AND CASE-CONTROL STUDIES

- Cohort study
  - Complete source population denominator experience tallied
  - Can calculate incidence rate or risks, and their differences and ratios
  - Usually very expensive
  - Convenient for studying many diseases
  - Can be prospective or retrospective
- Case-control study
  - Sampling from source population
  - Can calculate only the ratio of incidence rates or risks (unless the control sampling fraction is known)
  - Usually less expensive
  - Convenient for studying many exposures
  - Can be prospective or retrospective