Case-control studies

Ph.D. course
University of Copenhagen
March 7, 2005

Aims

• Definition of case-control studies
• Know the difference between case-control and cohort studies
• Describe odds ratio
• Know the principles for selection of controls
• Be aware of bias and confounding in CC-studies
• Describe advantages and disadvantages of CC-studies

How are you going to test, if

• Amyl nitrite ('Poppers') is the cause of Kaposi’s sarcoma?
  - Case-control
• Hospitalisation with hip fracture is the cause of lung embolus?
  - Case-control
• Smoking is the cause of lung cancer?
  - Case-control
• Use of childcare centres is a cause of respiratory tract infections?
  - Cohort study
• Viagra acts against erectile dysfunction?
  - Randomised controlled study
The question (hypothesis) determines the method

- The method depends (among others) of
  - Type of disease
  - Frequency of disease
  - Characteristics of affected persons
  - Diagnostic methods

Analytical study types

- Determination of causes and effects

- Observational
  - Cohort studies
  - Case-control studies

- Interventional studies
  - Randomised, controlled studies

Cohort studies

- Cohort: Cohors (Latin): 1/10 of a legion
- Prospective (!)
- Starting point a population of healthy

![](TIME_diagram)
### Outcome in cohort study: Relative risk

<table>
<thead>
<tr>
<th></th>
<th>Sick</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>A</td>
<td>B</td>
<td>A + B</td>
</tr>
<tr>
<td>Non-exposed</td>
<td>C</td>
<td>D</td>
<td>C + D</td>
</tr>
<tr>
<td>Total</td>
<td>A + C</td>
<td>B + D</td>
<td>A + B + C + D</td>
</tr>
</tbody>
</table>

Relative risk = \[
\frac{A + B}{C + D}
\]

### With other words: Cohort studies measure
- Risk of disease among the exposed compared with the risk of disease among the non-exposed.
- The absolute risk may be calculated for both groups!

### Case-control studies
- If exact information of exposure and cases in a complete population is not available, a case-control study may be carried out.
- Starting point patients with a given disease.
- Compared with a control group (retrospective (!)).
- Frequency of the risk factor (‘cause’) among patients (cases) compared with frequency of the risk factor among controls.
- Rationale is that if there is an association between exposure and outcome, more cases than controls must have been exposed to the factor.
Case-control studies

- Became popular with the change from infectious disease epidemiology to chronic diseases. Why?
  - Western life-style diseases (cancer, heart diseases)
  - Diseases with long latent period
  - Most applicable when disease is rare
  - Study many possible risk factors / causes

- Today the most used analytical study type in epidemiology

Retrospective – prospective?

- CC-studies are often referred to as synonymous with retrospective studies, and cohort studies as prospective. Correct?

- No, retrospective refers to if all cases are identified at time of study start, which is the most usual in CC-studies

- However, do prospective CC-studies exist?
Prospective CC-study

Prospective/retrospective refers to time of registration of cases.

Outcome in case-control study: Odds ratio

- Odds: measure of frequency of exposure in group
- Measure of association, as you anticipate that if cases are more often exposed to a given exposure than controls, the exposure is probably a cause of the disease
- Odds have no unit
- Odds among cases = Number of cases exposed to risk factor / Number of cases not exposed
- Odds among controls = Number of controls exposed to risk factor / Number of controls not exposed
- Odds ratio: Odds for cases/odds for controls

Calculation of odds ratio

- 2x2 table
- OR = a/c / b/d = ad/bc
- P = Chi-square test/Fisher’s exact test

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>
|     | a+c   | b+d      | a+b+c+d
Odds ratio in practice.
Salmonella in Wales 1989

31 cases office workers
6 cases canteen staff
in total 37 cases
Hereof 3 attended
doctor, other identified
through interviews or
taecal tests

58 controls

Salmonella outbreak in Wales 1989

<table>
<thead>
<tr>
<th>Gastroenteritis</th>
<th>No gastroenteritis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eaten</td>
<td>Not eaten</td>
</tr>
<tr>
<td>Lunch 22/1 6</td>
<td>31</td>
</tr>
<tr>
<td>Lunch 23/1 18</td>
<td>19</td>
</tr>
<tr>
<td>Salad 12</td>
<td>24</td>
</tr>
<tr>
<td>Sandwiches 16</td>
<td>21</td>
</tr>
<tr>
<td>Chicken 4</td>
<td>33</td>
</tr>
</tbody>
</table>

Odds ratio

\[ \text{Odds} = \frac{\text{Number of persons exposed}}{\text{Number of persons not exposed}} \]

\[ \text{Odds for having eaten in the canteen January 22 for cases} \]
\[ = \frac{6}{31} = 0.193 \]

\[ \text{Odds for having eaten in the canteen January 22 for controls} \]
\[ = \frac{9}{48} = 0.188 \]

\[ \text{Odds ratio} = \text{Odds for cases/odds for controls} \]
\[ = 0.193/0.188 = 1.03 \]
All risk factors in Wales outbreak

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lunch 22/1</td>
<td>1.03</td>
</tr>
<tr>
<td>Lunch 23/1</td>
<td>2.93</td>
</tr>
<tr>
<td>Salad</td>
<td>5.20</td>
</tr>
<tr>
<td>Sandwiches</td>
<td>2.39</td>
</tr>
<tr>
<td>Chicken</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Odds ratio vs. relative risk

• Why is relative risk not used in case-control studies?

<table>
<thead>
<tr>
<th>Gastroenteritis</th>
<th>Cases</th>
<th>Controls</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lunch Jan. 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>6</td>
<td>31</td>
<td>37</td>
</tr>
<tr>
<td>No</td>
<td>9</td>
<td>48</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>79</td>
<td>94</td>
</tr>
</tbody>
</table>

RR = $4/15 / 31/79 = 1.02$ ???

• Because it is totally nonsense!

OR <> RR

• 1.400 in building, 37 cases
• How many got sick out of those having eaten in the canteen?
• Number of eating and number of sick unknown
• Sample of a population

• Therefore the rate (=risk) cannot be calculated in case-control study

But

• If disease is rare, then OR ~ RR
If disease is rare (a and c small), \( \text{OR} \sim \text{RR} \)

\[
\begin{array}{c|c|c}
\text{Disease} & \text{Cases} & \text{Controls} \\
\hline
\text{Exposure +} & a & b \\
\text{Exposure -} & c & d \\
\hline
a + c & b + d \\
\end{array}
\]

\[
\text{OR} = \frac{a}{c} \div \frac{b}{d} = \frac{bc}{ad}
\]

\[
\begin{array}{c|c|c|c}
\text{Disease} & \text{Cases} & \text{Controls} \\
\hline
\text{Exposure +} & 6 & 35,781 \\
\text{Exposure -} & 19 & 35,332 \\
\hline
25 & 35,781 & 35,322 \\
\end{array}
\]

\[
\text{RR} = \frac{6}{(6 + 35,781)} \div \frac{19}{(19 + 35,332)} = 0.56
\]

\[
\text{OR} = \frac{6 \times 35,322}{19 \times 35,781} = 0.56
\]

Generalisability – Confidence intervals

- As we have only got a random sample and not a whole population, we must calculate a measure of how certain our estimate (OR) is: the confidence interval
- The values between which the ‘true’ population estimate with 95% confidence is found

<table>
<thead>
<tr>
<th></th>
<th>OR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lunch Jan. 22</td>
<td>1.03</td>
<td>0.33 - 3.18</td>
</tr>
<tr>
<td>Lunch Jan. 23</td>
<td>2.93</td>
<td>1.21 - 7.09</td>
</tr>
<tr>
<td>Salad</td>
<td>5.20</td>
<td>1.65 - 16.4</td>
</tr>
<tr>
<td>Sandwiches</td>
<td>2.39</td>
<td>0.99 - 5.8</td>
</tr>
<tr>
<td>Chicken</td>
<td>1.64</td>
<td>0.38 - 7.61</td>
</tr>
</tbody>
</table>

With other words:
Case-control studies measure

- The extent of exposure among the sick compared with the extent of exposure among the healthy
- Odds ratio expresses this
- Odds ratio is not the same as risk, as the risk in the population is unknown (a sample of the population is drawn).
- But
- If the disease is rare (relatively), then OR \( \sim \) RR!
Case definition

- Demands precise definition
- Time, place, and person
- Colon cancer in DK 1973-88
- Myocardial infarction among 60-70 old males in DK 1973-88
- Salmonella in Wales
  - >3 loose stools/day between January 22 and 26.
  - Stayed in office building in Wales between January 22 and 26.
- Toxic shock syndrome
  - Fever, rash, scaled skin, hypotension, involvement of 3 or more organ systems (GI, muscles, mucous membranes, urinary tract, liver, blood, CNS), negative tests for various other microbes than staphylococci
- Working definition, may be refined during study work up (e.g., SARS)
- Likely, possible

Finding cases, examples

- Hospital source
  - Easy, but bias possible

- Certain localisation (restaurant outbreak)
  - Evident in actual case

- Population source (lung cancer, national register)
  - Often costly, but used in DK because of good registers

Prevalent and incident cases?

- Number of AIDS-cases in DK 1989
  - Prevalence, measure of disease burden
- Number of newly diagnosed AIDS-cases in DK 1989
  - Incidence, measure of risk
- In a CC study of risk of AIDS, what measure to use?
  - Incidence
- By including both incident and prevalent cases risk factors and factors determining disease course (cause/prognosis), and interpretation is difficult
- The hen and the egg: coffee may be a risk factor for gastric ulcer, but if you have a gastric ulcer, you drink less coffee because of stomach pains
- Important that exposure precedes outcome, therefore use incident cases
**Generalisability**

- Must cases reflect all persons with the disease?
- Myocardial infarction
  - All cases in Copenhagen County 1989, or
  - Males 45-74 år hospitalised 1989 on Herlev Amtssygehus?
- Big difference in biology (familiar hypercholesterolemia / calcification of blood vessels)
- Validity most important, not generalisability!

**Choice of controls**

- Crucial point - problematic!
- Must reflect the question whether the frequency of an exposure observed among cases is different than that among comparable individuals without the disease
- A representative sample of the population that the sick persons come from, must have the same risk of exposure as cases
- Must be selected at random from the population (randomised)
- A control is a case without the disease

**Example of control selection**

- Question: do certain genetic polymorphisms result in an increased risk of serious bacterial infections in childhood?
- Cases: children hospitalised at least once with a serious bacterial infection (meningitis, sepsicaemia, etc.) before age 2 years, born in DK by Danish parents and of normal birth weight and birth length, without concurrent diseases and having lived constantly in DK before age 2 years
- Controls?
  - The same, just without any hospitalisation for the same diagnoses
Types of controls

- Hospital controls
- Population controls
- Special groups
  - Friends, family members, spouses, neighbours
- Advantages and disadvantages by all types

Hospital controls

- Pro’s
  - Easy to find
  - Minimize recall bias (what is that and why?)
  - Subjected to the same specific and unspecific factors that made the cases attend this hospital
- Con’s
  - Sick by definition, do not represent the distribution of exposures in the background population (e.g. they smoke and drink more)
  - Yields a biased estimate (in what direction?)
  - In a study of smoking and bladder cancer, many smokers among controls resulting in dilution of estimate (weaker or negative effect)
  - In a study of coffee and bladder cancer, less coffee drinkers among controls resulting in enhancement of estimate (stronger or positive effect)
  - Controls from Frederikssund County Hospital vs. Steno Diabetes Center?

What categories of patients may be used as controls?

- Controls to lung cancer patients, patients with
  - Bronchitis?
  - Heart diseases?
  - Hip fractures?
  - Stomach ulcers?
  - Asthma?
- The diseases of the controls may be associated with the risk factors under study (positively/negatively), which is not desirable
Population controls

- Typically, when cases come from a particular population
- Examples
  - Households
  - Random digit dialing
  - Registers/voters lists
  - CPR
- Problems
  - Larger expenses
  - Hard to get hold of people (working, not at home - selection bias)
  - Recall bias
  - Less motivation
  - Problems with random digit dialing in the USA?

Special groups

- Neighbours, friends, family
- Advantages
  - Cooperative
  - Confounder control (how?)
- Disadvantages
  - More alike cases (result?)
  - Dilution of estimate

More control groups?

- Ideally one per case group
- But sometimes desirable with more groups. When?
- When no ideal control group can be selected (e.g. patient groups)
  - Breast cancer patients: Gynaecological cancers, non-cancer gynaecological patients, emergency operations
- When the distribution of exposure differs in a control group from the background population
- Pancreas cancer: Patient controls and non-hospitalised population controls
Number of controls per case?

- 1:1 best
- What is gained by more controls per case?
- If cases are hard to find, increased statistical strength

But

- 1:3 (=1 case + 3 controls) less statistical strength than 2 + 2 (2:2)
- Max. 1:4 (can be shown statistically), more controls waste of time and money

Finding cases and controls

- Avoid bias

  - Example: Study of PKU-cards
    - Controls card before and after case card in the box in the freezer
    - Advantages?
    - Can all cards be used as controls?

Information on exposure

- Numerous possibilities
  - Registers
  - Hospital files
  - Telephone interviews
  - Etc.

- CC-study of Hodgkin’s lymphoma and birth weight
  - Cases interviewed about birth weight in hospital
  - Controls information on birth weight from the Central Birth Register
  - OK?

- No, information must be obtained in the same way and from the same place/source from cases as well as controls, otherwise risk of bias
Problems in interpretation of CC-studies

• In all epidemiological studies
  – Chance (statistical strength)
  – Bias
  – Confounding

• Particular problems in CC-studies?
  – Bias!

Bias – systematic errors

• More than 30 types reported, most frequent are
• Selection bias
  Systematic differences between in study participants and non-participants
  – Smoking studies
  – Pregnant
  – Hospitalised salmonella patients
  – HIV-prevalence among blood donors
  – Formaldehyde in factory workers
  – ‘Healthy worker effect’
• Observational bias
  Systematic error in measurements
  – Different methods/laboratories
  – Recall bias
  – Non-differential misclassification

Selection bias

• Systematic skewness in selection of study participants
• Exposure/outcome differences between study participants and possible participants that refused participation or were not chosen
  – Hospitalised salmonella patients in a study of clinical symptoms and complications. Why?
    • Because hospitalised salmonella patients represent the severe cases, and are not representative of the majority of cases
  – Newspaper articles and worry for use of tampons and Toxic shock syndrome. Why?
    • Because tampon users for this reason more often went to doctor than non-users, and because doctors were more likely to diagnose toxic shock syndrome in tampon users
Observational bias

- Knowledge about disease status in case or observer (1) affects reporting or notification of exposure
  - Reporting/notification of headache/discomfort in vaccinated
  - Mefloquine/Lariam and neuropsychological side effects

- Recall bias

- Misclassification

Recall bias

- Cases better remember exposures than controls - Anders Koch remember perfectly what he had for lunch at Hotel Hans Egede September 9 in Nuuk, Greenland. Why?
  - He (and many others) got gastroenteritis

Misclassification

- Differential misclassification
  - Error in classification
    - Hypertension and cerebral hemorrhage: 80% of cases and controls with hypertension remembered this correct. 90% of cases and controls without hypertension remembered this correct
    - No direction, underestimates the "true" association

- Non-differential misclassification
  - Serious
    - Disease status affects exposure classification
      - Risk factors for Hepatitis C infection: Cases interviewed in clinic, controls were telephone interviewed blood donors without signs of Hep. C - what do they answer on drug and sexual practices?
      - Serious problem, increases or reduces the effect of the risk factor, depending on direction

Advantages case-control studies

- Quick (and dirty...) and cheap (relatively)
- Reproducible
- Less sick persons/cases necessary to show an effect
- Suitable in case of rare disease or diseases with long latency
- Possible to evaluate many risk factors concurrently
- Good at outbreak investigations (acute diseases)
- Most frequent analytical epidemiological method today

Disadvantages case-control studies

- Measurement methods back in time maybe not relevant today (HPV)
- Causality cannot with certainty be determined, rather associations
- Rates cannot be determined, only approximated in case of rare diseases
- Temporal associations between exposure and outcome difficult to evaluate
- Not well suited in case of rare exposures
- Control group may be difficult to select
- Big risk of confounding
- Big risk of bias
  - Selection bias
  - Recall bias because of memory problems and data back in time

Take home messages

- If exact information of exposure and cases in a complete population is not available, a case-control study may be carried out
- Well suited in case of rare diseases, long latency time and multiple risk factors
- Cheap and effective - the most frequent analytical study design today
- Association measured in Odds ratio (odds among cases divided by odds among controls) with confidence intervals to express the statistical uncertainty
- Selection of control group difficult - controls should ideally be cases who just haven't developed the disease
- Substantial risk of confounding and bias
- Causes hard to determine, only associations