Epidemiology

Fall semester 2007

Descriptive & analytic epidemiology II
Case-control studies

---

Descriptive and analytic study types

<table>
<thead>
<tr>
<th>Descriptive studies</th>
<th>Analytic studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case reports/series</td>
<td>Case-control studies</td>
</tr>
<tr>
<td>Correlational studies</td>
<td>Cohort studies</td>
</tr>
<tr>
<td>Cross sectional surveys</td>
<td>Randomised/Intervention trials</td>
</tr>
</tbody>
</table>

---

A patient series

Carcinoma of the penis and cervix

“… Case 3. – Presented with 5-year history in November, 1969, aged 47. He had massive penile condylomata with squamous carcinomatous change and invaded ingual nodes. Died in 1977. His wife presented with carcinoma of the cervix in 1971 at the age of 43. She had a squamous cell carcinoma and stage III disease. Died 27 months later.”

Cartwright and Sinson, 1980; Lancet: 1: 97

---

Correlational studies

---

Ecological studies

In correlational studies, measures that represent characteristics of entire populations are used to describe disease in relation to some factors of interest such as age, calendar time, utilization of health services or consumption of a food, medication or other product. (H&B p 102)

---

Occurrence of gonorrhoea

England and Wales

[Graphic showing gonorrhoea incidence per 100,000 women over calendar years]

Beral, Lancet, 25 May 1974
Exercise

Measure of mortality from cervical cancer in England/Wales & Scotland per birth cohort
Adapted from Beral, Lancet, 1974

- Insert data for cervical cancer mortality in hand-outs
- Discuss what is shown in the figure
- Can this method be used to test hypotheses?

Kaposi’s sarcoma

Fig. 1. Kaposi’s sarcoma and non-Hodgkin’s lymphoma incidence among men, per 100,000 people per year, age-standardized to the 1970 U.S. population, shown on a linear and log scale to illustrate both the absolute and relative changes in nine Surveillance, Epidemiology and End Results (SEER) registries and in the San Francisco area registry only, from 1973 through 1998. Years with no cases were set arbitrarily at 0.12 cases in the log scale.

Correlational analysis

“The presence of a correlation does not necessarily imply the presence of a valid statistical association. Conversely, lack of a correlation in such studies does not necessarily imply the absence of a valid statistical association.”

Cross sectional surveys

“A third type of study is the cross-sectional or prevalence survey, in which exposure and disease status are assessed simultaneously among individuals in a well-defined population.”

Helicobacter pylori antibody prevalence Greenland 1998

Figure 1. Seroprevalence by age group among 427 persons born in Greenland and living in Denmark. Greenland (167 persons with inconclusive tests excluded).

Koch et al. 2004
HP seroprevalence pattern childhood

- At what age are Greenlanders infected?
- At what age are children in other countries infected?
- What are possible causes of the differences?
- Differences in infection patterns
- Cohort effect

Risk factors for HP seropositivity (adjusted), Greenland

<table>
<thead>
<tr>
<th>No. of older siblings</th>
<th>% (95% CI)</th>
<th>OR (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>74 (35)</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>1-2</td>
<td>97 (32)</td>
<td>1.44 (0.37 – 5.00)</td>
<td>0.01</td>
</tr>
<tr>
<td>2+</td>
<td>22 (64)</td>
<td>0.59 (1.45 – 10.7)</td>
<td>0.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distance to nearest older sibling</th>
<th>% (95% CI)</th>
<th>OR (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1 years</td>
<td>32 (53)</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>2 years</td>
<td>22 (36)</td>
<td>0.57 (0.1 – 3.37)</td>
<td>0.03</td>
</tr>
<tr>
<td>3-4 years</td>
<td>31 (42)</td>
<td>1.31 (0.3 – 5.00)</td>
<td>0.03</td>
</tr>
<tr>
<td>5+ years</td>
<td>34 (21)</td>
<td>0.17 (0.03 – 0.9)</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Risk factor assessment in correlational studies

- Measure association at time of study
- Information on prior exposures may be biased
- Time of outcome (seropositivity) unknown

Cross sectional surveys

<table>
<thead>
<tr>
<th>Job</th>
<th>Job</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>80</td>
</tr>
<tr>
<td>Diseases</td>
<td>20</td>
</tr>
<tr>
<td>Prevalence</td>
<td>20%</td>
</tr>
</tbody>
</table>

Cross sectional surveys

<table>
<thead>
<tr>
<th>Job</th>
<th>Job</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>80</td>
</tr>
<tr>
<td>Diseased</td>
<td>10</td>
</tr>
<tr>
<td>Prevalence</td>
<td>11%</td>
</tr>
</tbody>
</table>

Cross sectional studies

- Studies, in which outcome and exposure are determined simultaneously
- The observed outcomes are prevalent
- Data on risk factor associations will accordingly represent both survival and etiology
- It cannot be ruled out that the exposure under observation has changed after and maybe because of the outcome
- It cannot always be determined which came first, the exposure or the outcome
- Cannot test causation, but suggest associations to be tested in analytical studies
Descriptive & analytic epidemiology

Descriptive epidemiology
Advantages
• Cheap & quick
• May provide important overview
Disadvantages
• No information on the individual
• No control for confounding
• May involve bias
• Results may be ambiguous
• Can not test (causal) hypotheses

Analytic epidemiology
Advantages
• Information on the individual
• Control for confounding
• Results less ambiguous
• Can test (causal) hypotheses
Disadvantages
• Expensive
• Laboreous
• May involve bias

Analytical studies

Case-control studies

Aims
• Definition of case-control studies
• Know the difference between case-control and cohort studies
• Describe odds ratio
• Know the principles for selection of cases and controls
• 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

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
- Rare disease
- Long disease latency
- Fewer persons necessary to show effect

Why 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

All incident cases (why incident?)

Study start

End of 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: if the exposure is a cause of disease, then sick persons (cases) should be exposed more often than controls!
- 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>Disease</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>a + c</td>
<td>b + d</td>
</tr>
<tr>
<td></td>
<td>a + b + c + d</td>
<td></td>
</tr>
</tbody>
</table>

Calculate odds ratio:

<table>
<thead>
<tr>
<th>Disease</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>10</td>
<td>80</td>
</tr>
<tr>
<td>No</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>a + c</td>
<td>b + d</td>
</tr>
<tr>
<td></td>
<td>a + b + c + d</td>
<td></td>
</tr>
</tbody>
</table>

- OR = a/c / b/d = ad/bc = ???
- Odds for cases = 10/40 = 0.25
- Odds for controls = 80/20 = 4
- OR = 0.25/4 = 0.0625
- Interpretation in words?

Observation by chance?

- Could an OR of 0.0625 occur by chance (‘is it significant’)?
- A X² test tests if the observed numbers differ (significantly) from the expected numbers
- What are the expected numbers in each cell?

<table>
<thead>
<tr>
<th>Disease</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>10</td>
<td>80</td>
</tr>
<tr>
<td>No</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>60</td>
</tr>
</tbody>
</table>
Observed/expected numbers

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Disease</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>10</td>
<td>30</td>
<td>80</td>
</tr>
<tr>
<td>No</td>
<td>40</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>150</td>
</tr>
</tbody>
</table>

- Does the observed deviate from the expected (‘is this significant’)?

X² test

\[
X^2 = \frac{(O_a - E_a)^2}{E_a} + \frac{(O_b - E_b)^2}{E_b} + \frac{(O_c - E_c)^2}{E_c} + \frac{(O_d - E_d)^2}{E_d}
\]

If X²>3.84, then p<0.05!

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Disease</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>10</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>40</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

What is X²? 50 – p < 0.0001

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 faecal tests

1400 employees

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>Eaten</td>
<td>Not eaten</td>
</tr>
<tr>
<td>Lunch 22/1</td>
<td>6</td>
</tr>
<tr>
<td>Lunch 23/1</td>
<td>18</td>
</tr>
<tr>
<td>Salad</td>
<td>12</td>
</tr>
<tr>
<td>Sandwiches</td>
<td>16</td>
</tr>
<tr>
<td>Chicken</td>
<td>4</td>
</tr>
</tbody>
</table>

Odds ratio

- Odds = \( \frac{\text{Number of persons exposed}}{\text{Number of persons not exposed}} \)

- Odds for having eaten in the canteen January 22 for cases = \( \frac{6}{31} = 0.193 \)

- Odds for having eaten in the canteen January 22 for controls = \( \frac{9}{48} = 0.188 \)

- Odds ratio = Odds for cases/odds for controls

- Odds ratio for having eaten in the canteen at January 22 = \( \frac{0.193}{0.188} = 1.03 \)

All risk factors in Wales outbreak

<table>
<thead>
<tr>
<th></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>2.91</td>
</tr>
<tr>
<td>Salad</td>
<td>5.21</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>
</tr>
</thead>
<tbody>
<tr>
<td>Lunch Jan. 22</td>
<td>Yes</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>15</th>
<th>79</th>
</tr>
</thead>
</table>

RR = 6/15 / 31/79 = 1.02

- Because the calculation is nonsense!

### OR <> RR

- 1,400 in building, 37 cases
- How many got sick out of those having eaten in the canteen?

<table>
<thead>
<tr>
<th></th>
<th>Number of eating and number of sick</th>
<th>Sample of a population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unknown</td>
<td></td>
</tr>
</tbody>
</table>

- 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), then OR ~ RR

If disease is rare (a and c small), then OR ~ RR

<table>
<thead>
<tr>
<th>Disease</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure+</td>
<td>19,781</td>
<td>19,787</td>
</tr>
<tr>
<td>Exposure-</td>
<td>35,313</td>
<td>35,332</td>
</tr>
<tr>
<td></td>
<td>54,912</td>
<td>110,096</td>
</tr>
</tbody>
</table>

RR = (6 + 19,781) / (19,781 x 19) = 0.56

OR = 6 x 35,313 / 19,781 x 19 = 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

<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 – 7.01</td>
</tr>
</tbody>
</table>

### Time problem

- Temporal associations between exposure and outcome difficult to evaluate

- Measurement methods back in time maybe not relevant today (HPV)

### 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 ~ 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.
• Toxic shock syndrome
  – Fever, rash, scalded 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 (ex. SARS)
• Likely, possible

Finding cases, examples

• Hospital source
  – Easy, but bias possible
• Certain localisation (restaurant outbreak)
  – Evident in actual case
• Population source (lung cancer, register)
  – Often costly, but used in DK because of good registers

Prevalent or 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, septicaemia, 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
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

PARTICULAR STUDY TYPES: NESTED CASE-CONTROL

- A case-control study ‘nested’ into a cohort
- Eg. A study of genetic factors and myocardial infarction:
  - Cohort: Østerbro-undersøgelsen (thousands of persons)
  - Case-control: 200 cases and 200 controls from the cohort having genetic analyses done
- Why? Expensive or difficult to carry out genetic analyses on all subjects in the cohort or difficult to obtain detailed information for all

DENSITY CASE-CONTROL STUDY

- Controls selected to represent proportion of person-time for exposed an non-exposed controls
- Knowledge of person-time of cases necessary
- The chance of being selected as control is proportional to person-time experience in the source population
- If sampled properly, the Odds ratio from a density case control study design gives an estimate of incidence rate in study population

CASE-CROSSOVER STUDY

- Determines effect of two or more interventions
- Each case receives all interventions and acts as his/hers own control
- Only relevant when exposure varies from time to time within a person
- Evaluates exposures that trigger short-term effects (mobile phones and traffic accidents, sexual intercourse and myocardial infarction)
ALTERNATIVE: CASE-COHORT STUDY

- Controls have the same chance of being selected irrespectively of person-time spent
- Fraction of total number of people in study population rather than person-time
- Gives an estimate of risk ratio instead of incidence rate as in density CC study
- Control may also be a case!

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
- Causes hard to determine, only associations

Next lesson: pitfalls and challenges in case-control studies…