Author: Conf. Dr. Cosmina-Ioana Bondor

Lecture 3 – Case-control study

A ALWAYS

S SEEK

K KNOWLEDGE

Aim of the study

- a more general aspect

Examples:

- Evaluating an association between
  - a risk factor and a disease
  - a treatment and an outcome
  - between two factors
- Comparing
  - two diagnostic techniques
  - two treatments

Objective – what we want to demostrate

Main objective

- Study hypothesis – what will we study?
- Ex. Assess the association between alcohol consumption during pregnancy and the occurrence of a malformation in the child

Secondary objectives (what else can we study?)

- Ex. Quantify the importance of this link

Study objectives

= precise, practical steps

Major objectives:

- existence of a link between the risk/prognostic factor and the disease
- quantification of the importance of the link between the risk/prognostic factor and the disease
- difference between treatments
- difference between diagnostic tests
- quantification of the difference

Secondary objectives of the study.

- Other biological phenomena studied in the same study

3. Protocol – plan of the research

Study protocol

- The protocol is drawn up before the study begins
- Contain
  - the motivation of the purpose
  - the purpose of the study
  - the objectives
  - the response to all these questions:

Study protocol

- The protocol answer to all this questions:
  - Where will it take place?
  - How will it take place?
  - Who will be included/excluded in the study?
  - What will we measure/investigate/observe?
  - What measurement/investigation methods will be used?
  - What devices will we investigate/measure/observe with?
  - What will be the data analysis methods?
  - Who will be the investigators?
  - How will they be trained?
  - etc.

Study protocol

- It is written according to certain **writing standards**
- experimental studies:
  - templates provided by the agency:
    - European Medicines Agency (EMA)
- The entire team will read and review the protocol
- The protocol is submitted to an ethics committee!!!
- After validation, it cannot be changed
  - experimental studies (all)
    - their registration in the clinical trials register:
      - EU Clinical Trials Register

A. Study types

Research domain

Description of a new health phenomenon
Evaluation of a diagnostic procedure
Evaluation of a therapeutic procedure
Evaluation of a risk/ protective or prognostic factors

Description of a new health phenomenon

studies that describe

- a case
- or
- a series of cases

Evaluation of a diagnostic procedure

- Some disease need a new diagnostic procedure
  - compare the new diagnostic procedure with the older one

	Disease+	Disease-	Total
New test+	a	b	
New test -	c	d	
Total			

Sensibility, specificity, positive predictive value, negative predictive value

Evaluation of a therapeutic procedure

- therapy versus placebo
- or
- new theraphy versus old theraphy

Predictive factor

if present = high probability of having a positive response or lack of response to a particular therapy.

Ex. radiotherapy is associated with remission of prostate cancer
Ex. erythromycin is associated with improvement of oxygen saturation in case of pneumonia diagnosis

Evaluation of a risk/ protective or prognostic factors

Risk factor

if present = high probability of disease

Ex. pollution influences the presence of asthma in children,
Ex. the presence of nitrogen in drinking water influences fertility, etc.

Protective factor

if present = low probability of disease

Ex. physical activity prevents obesity,
Ex. sun exposure reduces the frequency of fractures, etc.

Prognostic factor

if present = high probability of recovery or disease

Ex. old age influences tumor recurrence,
Ex. physical activity influences maintenance of normal weight after gastric sleeve surgery/stomach reduction

Descriptive

Discribing a case or a series of cases

- Ex. a boy who ingest soy sauce in very high quatity
- Ex. Covid-19 first series of 1015 cases

Assessment of disease-related indices

- Ex. incidence (new case in population)
- Ex. prevalence (all cases in a population) exposure statistics to factors
- Ex. prevalence of periodontitis

Analytical

- Comparisons are made
- Connections are traced

Characteristics:

- uses statistical tests
- inferential statistical methods
- e.g.

Ex. evaluating a link between butter consumption and heart attack

Ex. comparing aspirin and placebo to reduce heart attack

Ex. comparing X-rays to CT scans in cancer diagnosis

according to the researcher's attitude towards the study subjects

observational versus experimental

experimental if

there is direct intervention by the researcher with implications on the event of interest

Who make the study?

Researcher is

- the Dentist - intervenes → Experimental study
- the Student - observe → Observational study
- the Client → Observational study

Study

Observational

researchers do not intervene

- neither on the subjects,
- nor on the evolution of the disease

Ex. evaluate the relationship between obesity and hypertension

Ex. compare subjects with or without heart attack to see the relationship with butter consumption

Ex. evaluating the consumption of different foods or physical effort until the onset of the infarction

Experimental

researchers intervene

- on subjects
- on the evolution of the disease

by

- administering treatments (aspirin vs. placebo),
- surgical interventions (appendectomy),
- various procedures, etc.

Characteristics:

- rigorously controlled
- suitable for inferring causality

in humans,

- mainly clinical studies
- randomized controlled trial

Study

Observational

Advantages

- easy to do,
- low cost

Disadvantages

- we cannot prove causality

Experimental

Advantages

- we can demonstrate causality
- the strongest study in the hierarchy of studies

Disadvantages

- difficult to organize
- ethical implications
- high costs

Study duration

Cross-sectional

Features:

- We observe subjects only once
- Quick access to information

Ex. evaluate the relationship between obesity and hypertension at a given time

Ex. evaluate the frequency of obesity in school children

Longitudinal

Definition:

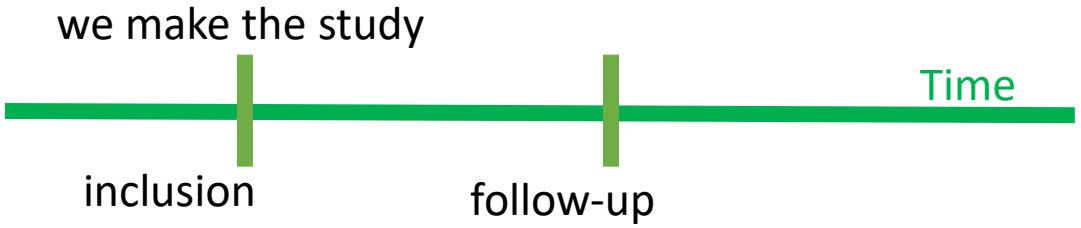
- collects information about subjects at multiple points in time

Characteristics:

- organization is more difficult
- access to information is longer
- cost can be significant

Ex: evaluates the relationship between excessive coffee consumption (decades) and the onset of osteoporosis

Ex. Evaluates the relationship between physical exertion and weight



Study duration

Cross-sectional

Advantages

- prevalence can be determined
- We can study more diseases
- easier to organize
- lower cost
- shorter duration

Disadvantages

- we cannot observe whether a factor precedes an outcome
- we cannot calculate incidence or relative risk (RIE, RIN, RR, RA)
- those who die are lost from the study

Longitudinal

Advantages

- better information compared to cross-sectional studies

Disadvantages

- more difficult to organize
- more expensive

Direction over time

Longitudinal retrospective

Definition:

observations/measurements of characteristics were made in the past (before the study began)

information collected from:

observation sheets

databases

Ex. coffee consumption, smoking, food consumed

Longitudinal prospective

Definition:

observations/measurements of characteristics are made after the study begins (we have no past information)

the researcher observes/measures the characteristics of interest directly

Ex. the relationship between physical effort and anxiety reduction

Longitudinal retrospective

Advantages

- applicable to rare diseases and long incubation
- easy to organize
- low cost
- short duration

Disadvantages

- risk of observation error
- risk of recall error
- we cannot calculate incidence or relative risk (RIE, RIN, RR, RA)

Longitudinal prospective

Advantages

- more precise information
- we can calculate incidence or relative risk (RIE, RIN, RR, RA)

Disadvantages

- requires a lot of personnel
- people lost from the study (risk of attrition)
- long duration
- possible change over time
- of diagnostic criteria
- influence of exposure factor
- difficult to organize,
- expensive

Exhaustive versus sampling

Exhaustive (all the population)

Definition:

- the entire population is studied.

Advantages

- perfect representativeness
- accurate results

Disadvantages

- takes a long time
- difficult to organize
- errors
  - multiple investigators,
  - high ammount of data
- expensive

Sampling

Definition:

- a sample (several) is studied
- subjects are drawn from the target population

Advantages

- easy to organize
- low cost
- short duration

Disadvantages

- the result is an estimate
- there is a risk of error
  - we may not be able to generalize the results well

Primary

Definition:

- involve collecting data directly from the source
  - experiments
  - surveys

Secondary

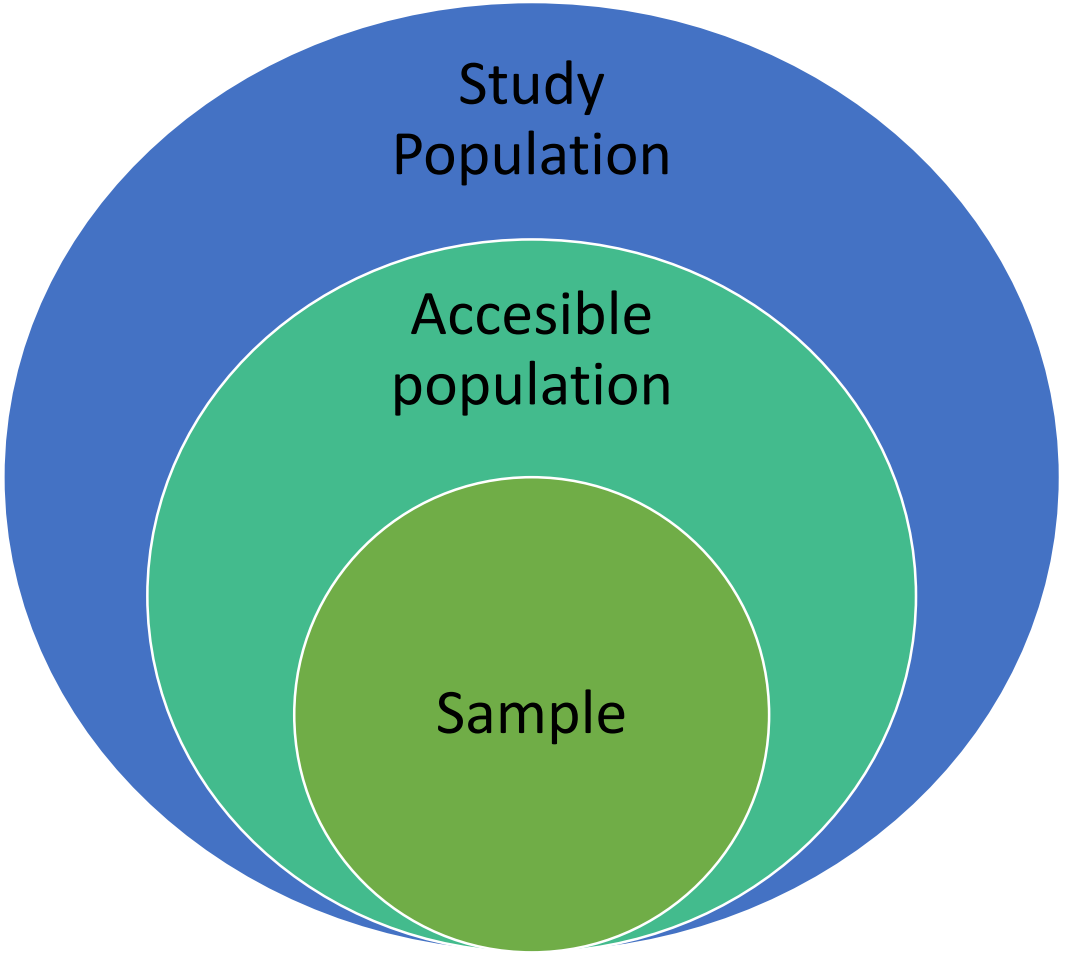
- summarising
- critiquing

past studies

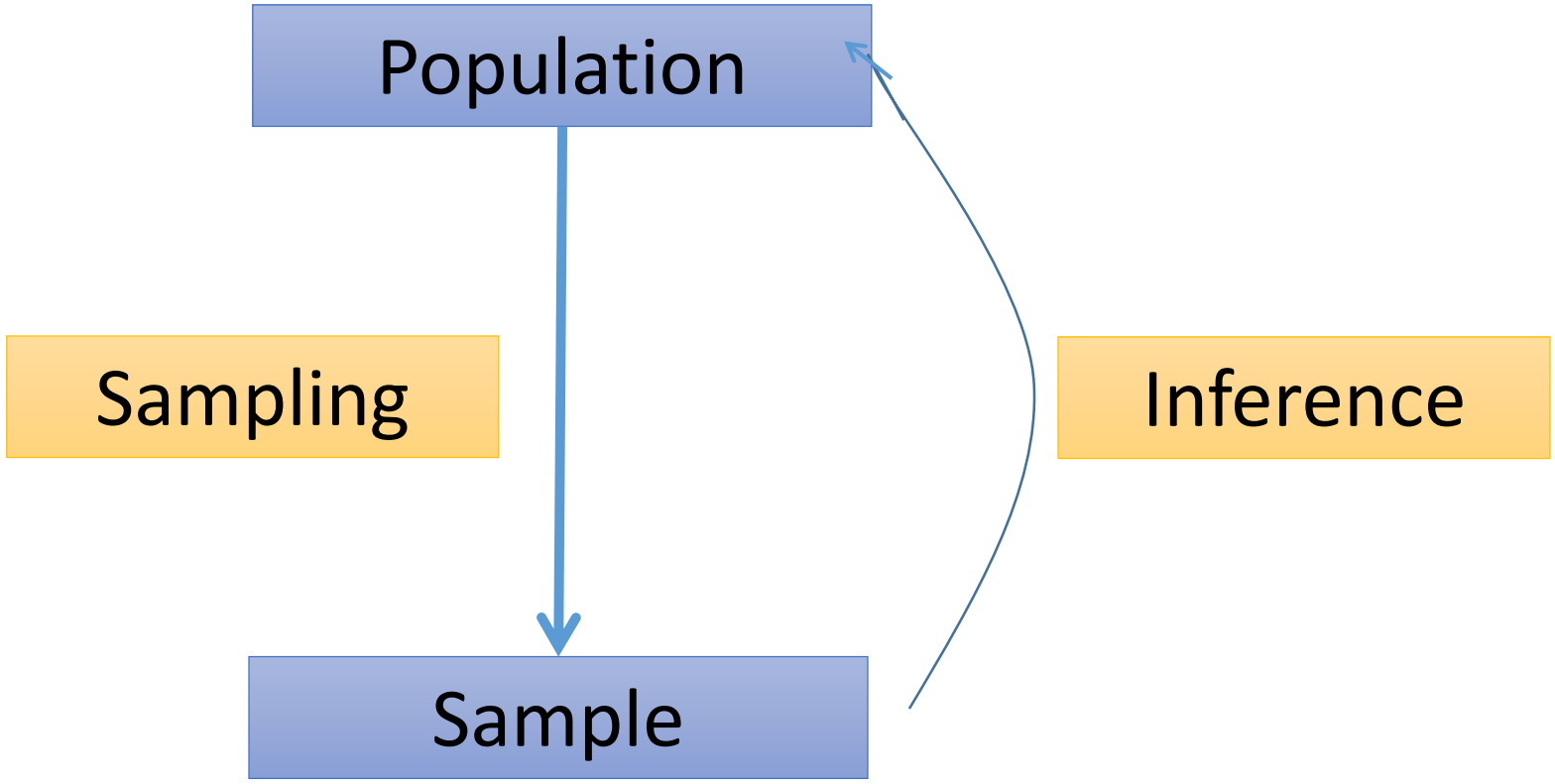
old published data

generate conclusions

Sampling



Research



Condition for a good estimation (inference)

Random sampling

Why random selection?

- Reduces the selection bias
- The representative sample for the population
  - should have the same distribution of main characteristics as the population
- where:
  - Important characteristics = in association with those studied

Sample size (number of individuals)

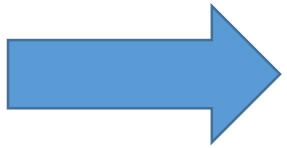
!!! the number of individuals is determined before the study

- depends on
  - what we want to demonstrate (means, frequencies)
  - how many samples? (groups of patients with the same characteristic)
  - how precise we want the study to be
    - the size of the confidence interval
      - large (not very precise)
      - narrow (precise)
  - how big we expect the difference between the groups to be
    - big
    - small

Example

we want to see

- if there is a difference in exposure to obesity between those who have osteoporosis and those who do not
- from the **literature** we know that
  - those with osteoporosis are 50% obese
  - the general population is 30% obese
    - 20% difference to be demonstrated



Determining the required sample size for comparing two proportions, one of 50% and the other of 20%.http://statpages.org/proppowr.html

Variables of interest

- qualitative?
- quantitative?

difference to be demonstrated=20%

Significance Level (alpha):	0.05	(Usually 0.05)
Power (% chance of detecting):	80	(Usually 80)
Group 1 Population Proportion:	.30	(Between 0.0 and 1.0)
Group 2 Population Proportion:	.50	(Between 0.0 and 1.0)
Relative Sample Sizes Required (Group 2 / Group 1):	1.0	(For equal samples, use 1.0)

level of error =5%

study power =80%

equal number of individuals in each group

Compute

Results:

Sample Size Required

	Group 1	Group 2	Total
"Classical" Calculation:	93	93	186
With Continuity Correction:	103	103	206

Example

we want to see

- if a change in diet decrease the total cholesterol?
  - from the literature we know
    - that those with hypercholesterolemia have an average of 230 cholesterol
- we want it to decrease by at least 20 to decide that the diet is effective
  - 20 difference to demonstrate
- standard deviations: 26, respectively 33

Determining the necessary sample size to compare two means, one of 230 and the other of 210 http://sampsize.sourceforge.net/iface/s2.html#nm

Assumptions:

level of error =5%	alpha =	5	(two-sided)
study power =90%	power =	90	
means difference = 20	m1 =	230	
	m2 =	210	
standard deviation 1	sd1 =	26	
standard deviation 2	sd2 =	33	
equal number of individuals in each group	n2/n1 =	1	

Results:

Estimated sample size:

n1 =	47
n2 =	47

E. Standardization of methods

- Defining measurement / observation / recording methods
  - Understandable, clear
  - Feasible (can be done)
  - Accurate
  - Reproducible – anyone can reproduce the study
- As for the equipment / laboratory
  - it is preferable to be one throughout the study

F. Establishing the data analysis plan

- Database creation
- Database validation
- Transfer method
- Persons who have access to the database
  - levels of accesibility
- Statistical methods that will be applied
- What indicators will be calculated

G. Other aspects

- Staff – training
- Financial side: will the participant be paid?
- Will the staff be paid?
- Where will the funds come from?
- Ethical considerations
  - Medical ethics rules
- Data protection
- Establishing the duration of each stage

Summary of the previous

- Researcher's attitude
  - observational
  - experimental
- Objective
  - descriptive
  - analytic
- Selection method
  - sample
  - exhaustive
- Selection method
  - sample
  - exhaustive
- Data collection duration
  - cross-sectional study
  - longitudinal study
    - prospective
    - retrospective

Case-control study

Objectives

- Define a case–control study and describe its structure.
- Explain how cases and controls are selected.
- Construct a 2×2 table and calculate the odds ratio.
- Interpret the association between exposure and disease.
- Identify common biases and confounding in case–control studies.
- Critically appraise a published case–control study.

Case control studies

- compare
  - Patients **with a disease (cases)**
  - Patients **without the disease (controls)**
- to determine whether they were previously exposed to a **risk factor**.

- Were cases more frequently exposed than controls?

Scenario

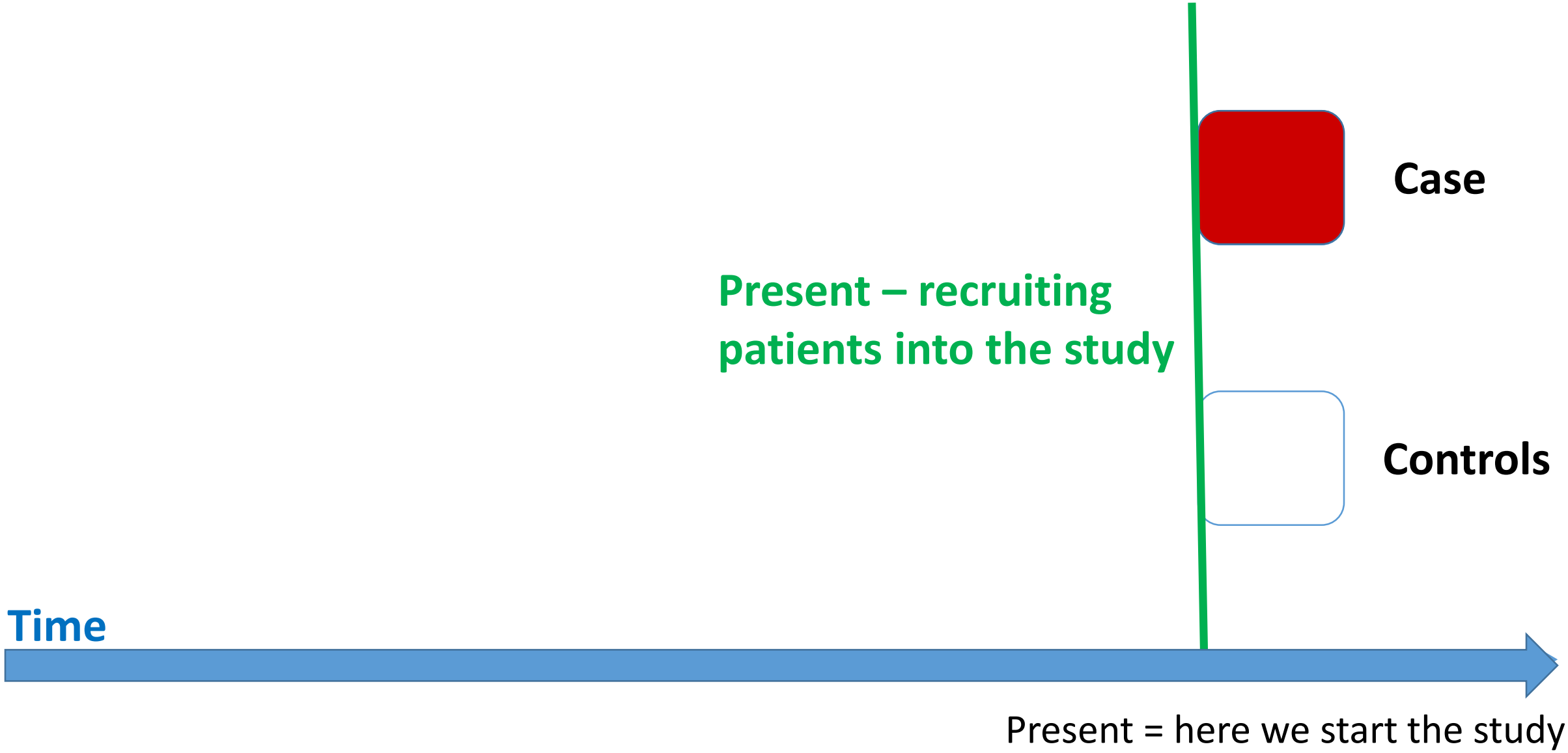
- Is oral cleft in new borns linked to maternal alcohol intake?

- Cases:
  - children **with** nonsyndromic **oral clefts**
- Controls:
  - children **without** nonsyndromic **oral clefts**

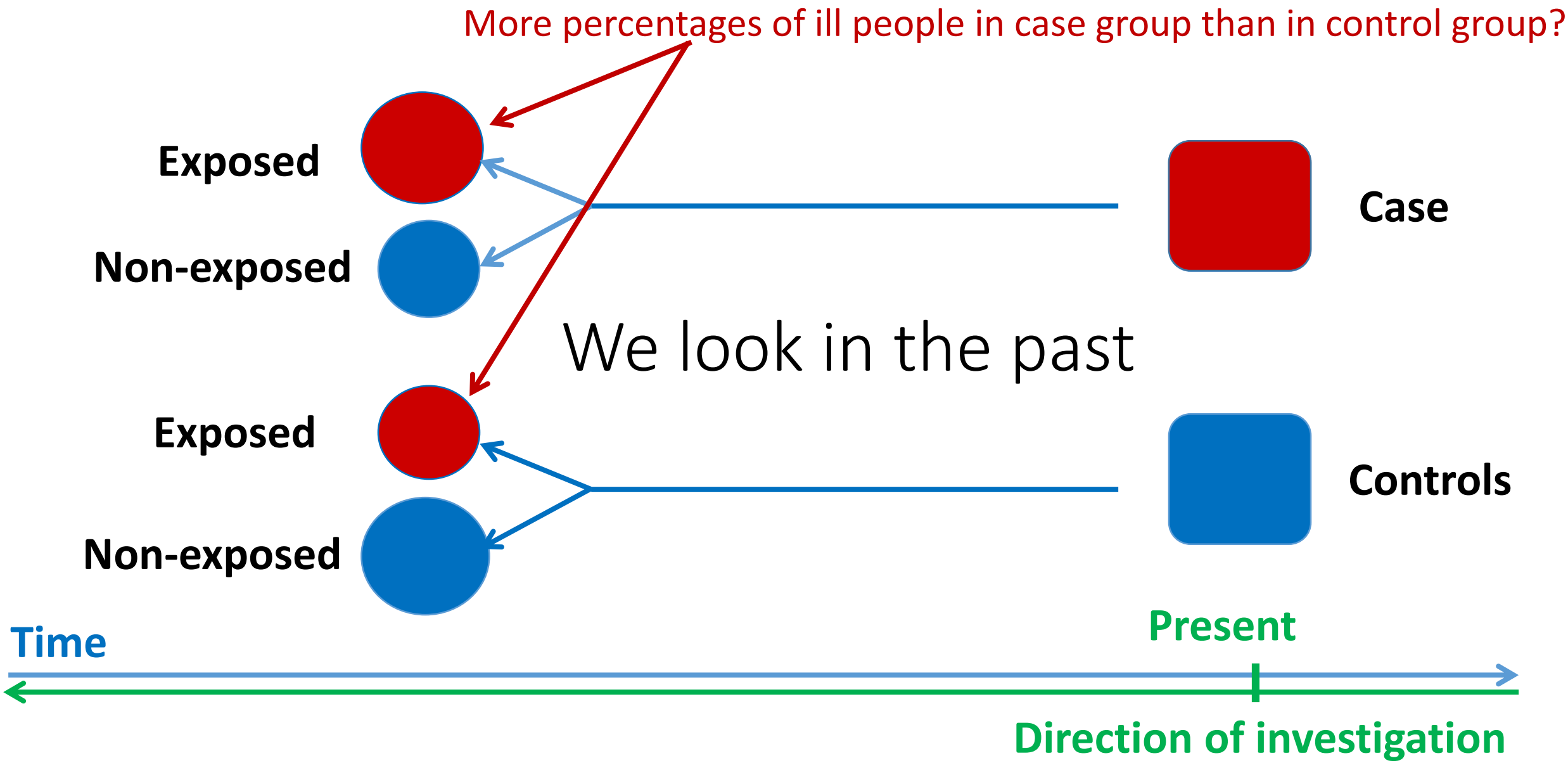
- **Exposure**
  - **maternal alcohol intake** above standard limits

- start with disease

- look backward in the past for exposure


Case
Present – recruiting
patients into the study
Controls
Time
Present = here we start the study

Greenberg RS. Retrospective studies. In: Encyclopedia of Statistical Science, Vol.8. Kotz S, Johnson NL editors. Wiley, 1988.



Greenberg RS. Retrospective studies. In: Encyclopedia of Statistical Science, Vol.8. Kotz S, Johnson NL editors. Wiley, 1988.

Case – control studies

1
- The doctor diagnoses the disease

2
- Patients with disease are recruited (case group)
- Individuals without the disease selected from the same population (control group)

3
- Investigate exposure to the factor in their past
- Exposure to the risk factor = presumed cause of the disease, precedes the disease

4
- if the percentage of those exposed to the factor in the case group is greater than the percentage of those exposed in the control group
  - we decide that there is a link between the risk factor and the disease
- if the percentage of those exposed to the factor in the case group is less than the percentage of those exposed in the control group
  - we decide that there is a link between the protective factor and the disease
- if the percentage of those exposed to the factor in the case group is equal to the percentage of those exposed in the control group
  - we decide that there is NO link between the protective factor and the disease

Case – control studies

Objectives

demonstrate the existence of a link between the disease and a possible prognostic factor

possible results

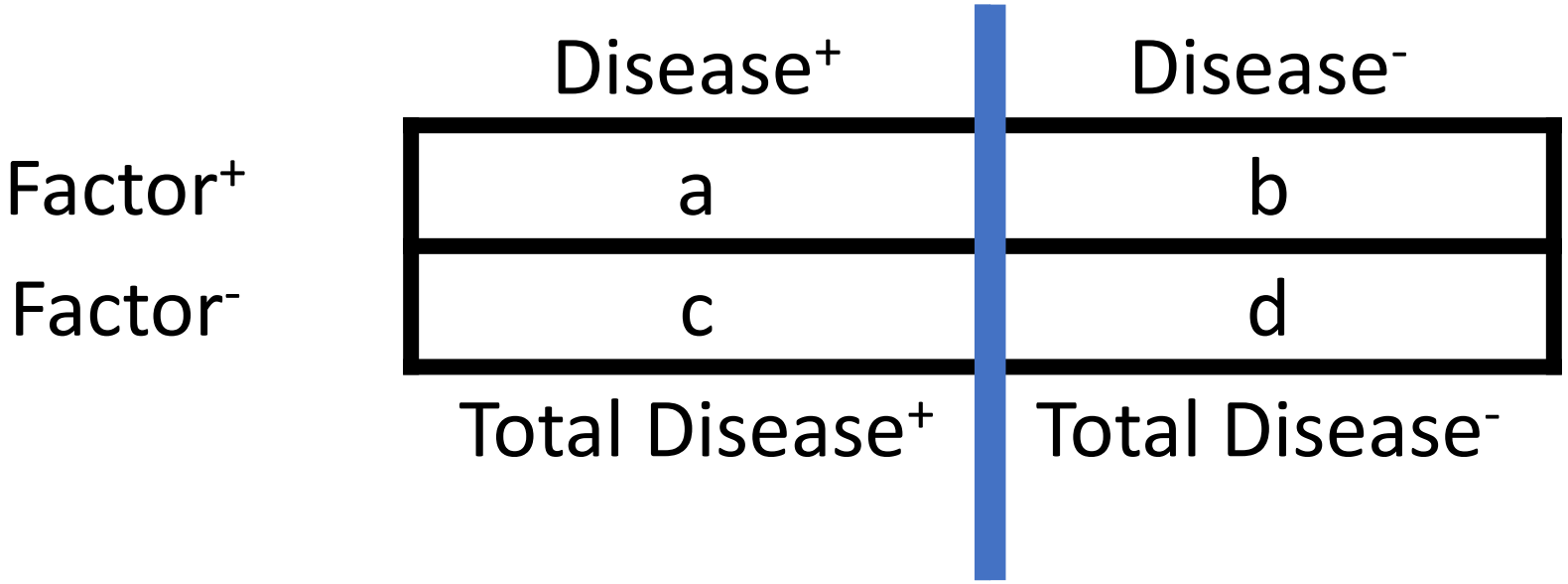
- the risk factor was present in a statistically significant **higher** percentage of patients with the disease compared to those without the disease
- the risk factor was present in a statistically significant **smaller** percentage of patients with the disease compared to those without the disease
- the risk factor was **not** present in a statistically significant **different** percentage of patients with the disease compared to those without the disease

quantify the link

how strong is the association?

exposure – the factor is present in the patient's past. e.g. the patient smokes, so she was exposed to cigarette smoke, or the patient's husband smokes, so the exposure was passive, the patient was exposed to cigarette smoke

Contingency table – Case – control studies

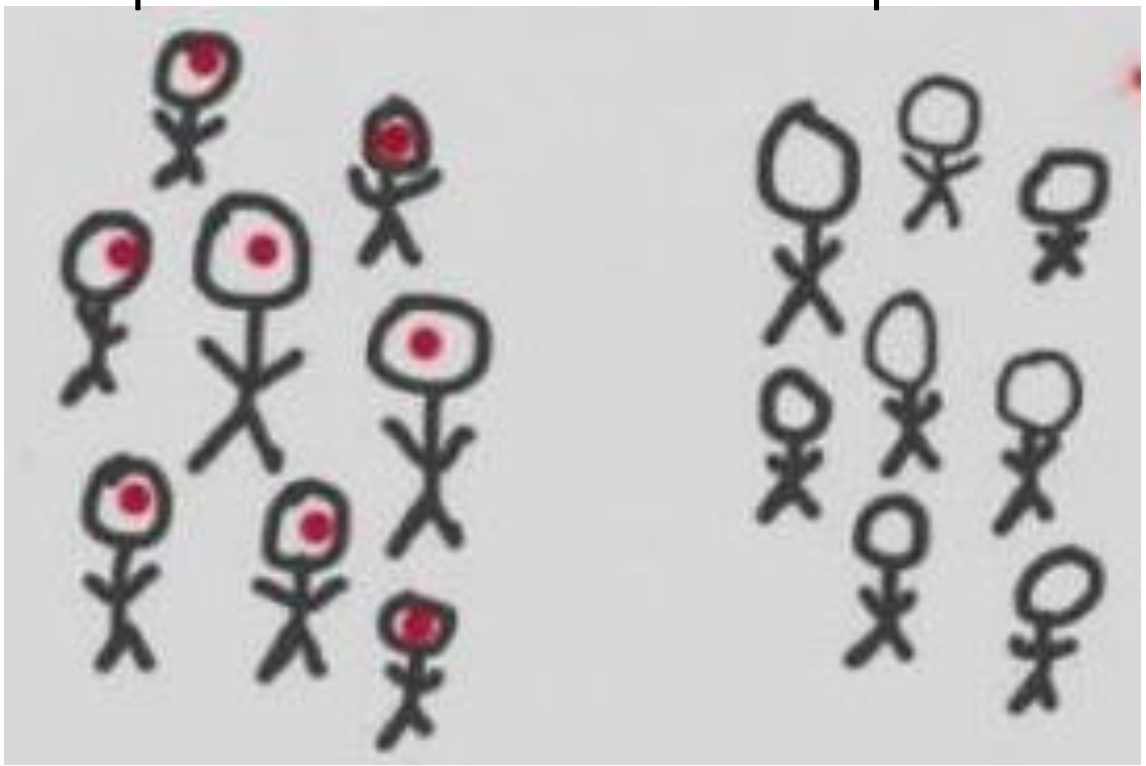


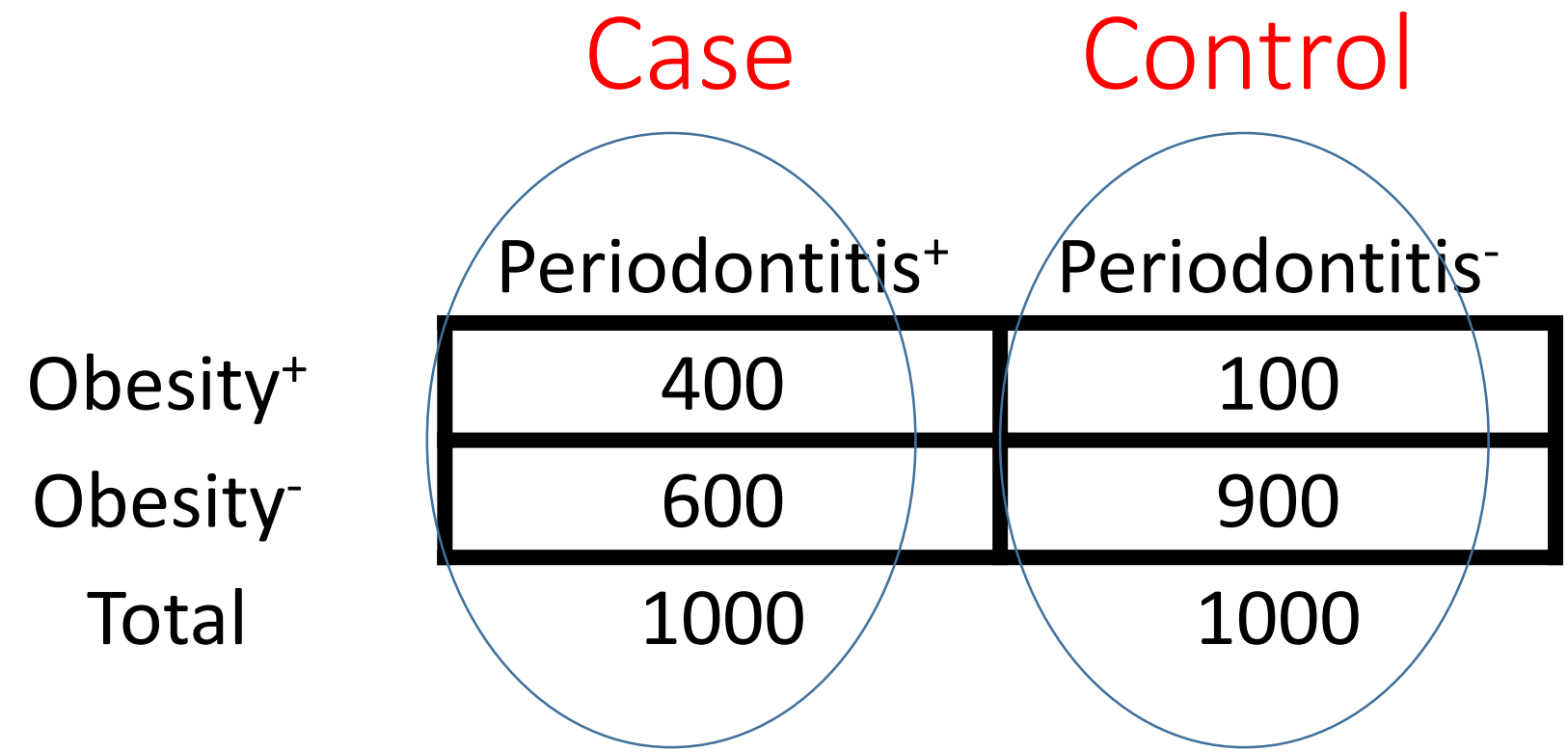
Disease^{+} - Disease present, Disease^{-} - Disease absent, Factor^{+} - factor present, Factor^{-} - factor absent

! Calculations only in the component on the right of the line or on the left of the line

Case-control study: Is there an association between smoking and periodontitis?

- 2 groups of patients
  - cases: with periodontitis
  - controls: without periodontitis
- obesity was noted retrospective
  - from the patients file



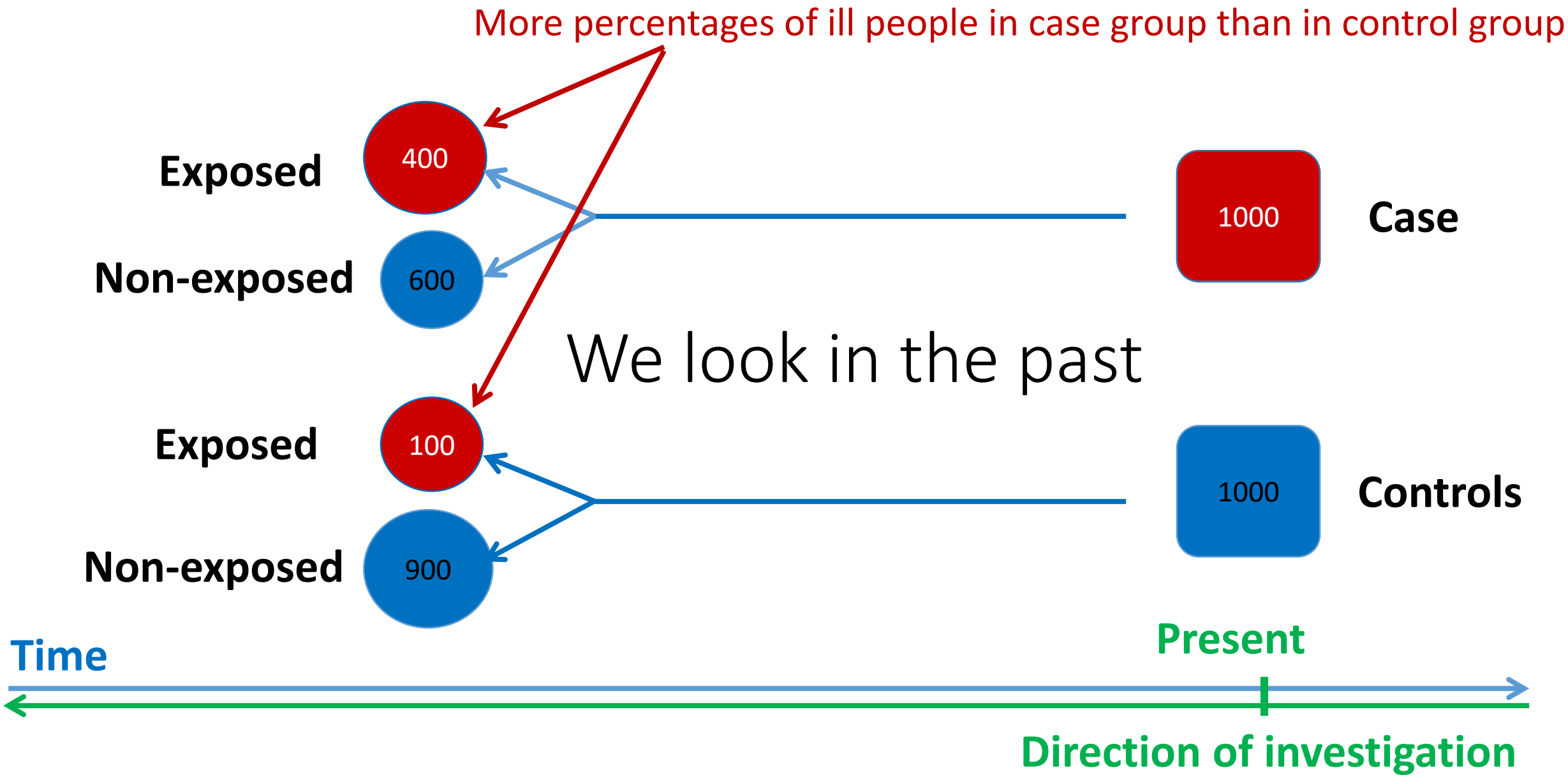


2000 subjects aged 60 years.

Group 1: with Periodontitis

Group 2: without Periodontitis

data on obesity are collected retrospectively from the record



Greenberg RS. Retrospective studies. In: Encyclopedia of Statistical Science, Vol.8. Kotz S, Johnson NL editors. Wiley, 1988.

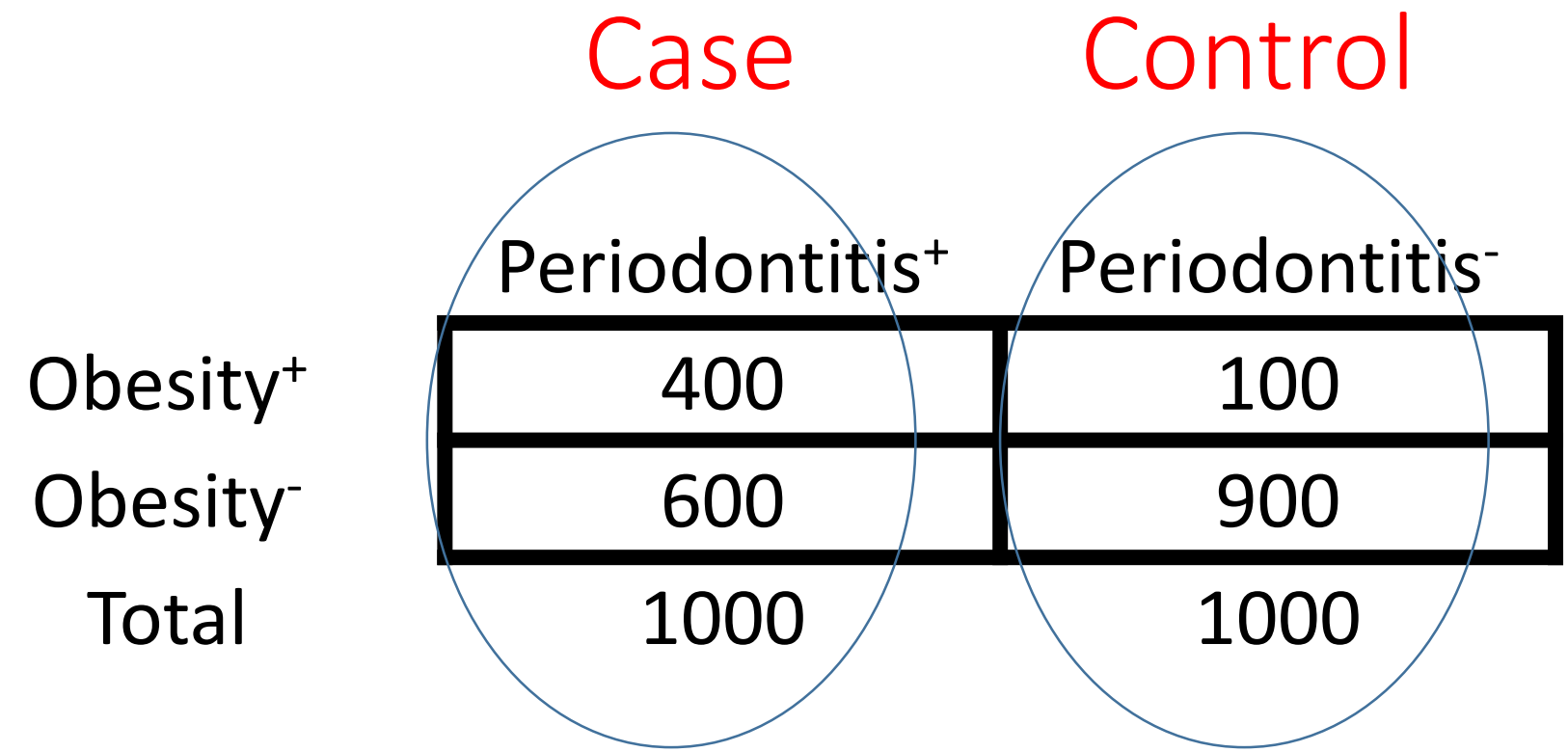
Case-control study characteristics

- cases must have the disease
  - clearly diagnosed
  - ex. Cases = patients with moderate/severe periodontitis
- controls should come from the same population
  - ex. attending the same dental clinic
- exposure information may come from
  - medical records
  - questionnaires

Study characteristics

- By the attitude of the researcher:
  - observational - the study does not involve an intervention, but anamnesis
- By the population included in the study:
  - sampling (select groups not whole population)
- By duration:
  - Longitudinal (patients are follow over time)
    - retrospective (we look in the past)

Data analysis



2000 subjects aged 60 years.

Group 1: Cases - with Periodontitis

Group 2: Control - without Periodontitis

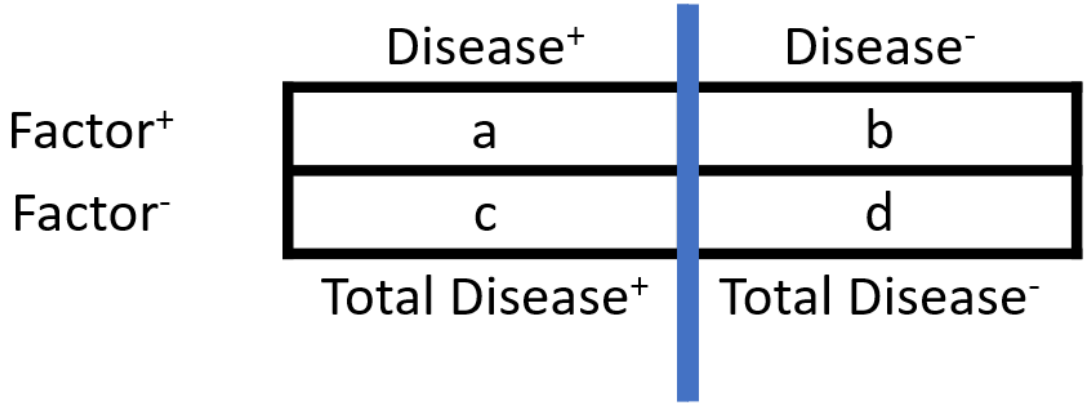
data on obesity are collected retrospectively from the record

- Odds among cases:

$$a/c$$

- Odds among controls:

$$b/d$$



Statistic for quantifying the factor-disease link

- Odds ratio:

$$\text{OR} = \frac{a/c}{b/d}$$

$$\text{OR} = \frac{a*d}{b*c}$$

B – disease, F - factor

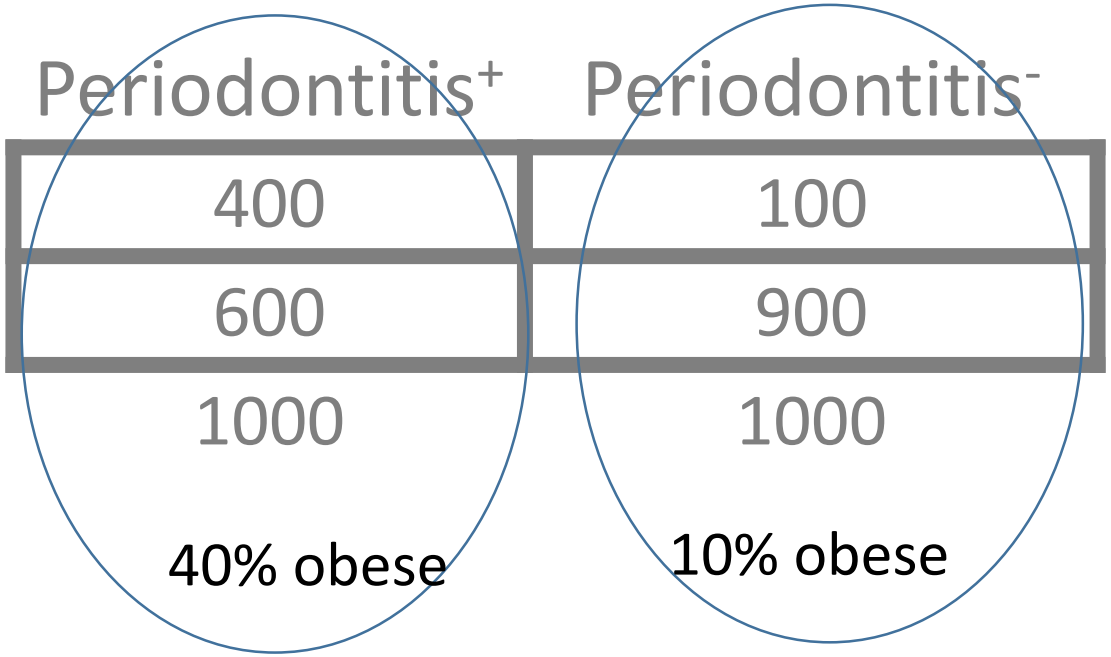
	B^+	B^-	
F^+	a	b	a+b
F^-	c	d	c+d
	Total B^+	Total B^-	Total=n

Interpretation - OR

- Clinical point of view
- OR<1 – protective factor – exposure is protective
- OR = 1 – no relationship
- OR >1 – risk factor – exposure increases disease risk

Study objective: obesity and periodontitis are dependent

	Case Periodontitis^{+}	Controls Periodontitis^{-}
Obese^{+}	400	100
Obese^{-}	600	900
Total	1000	1000



Interpretation

- statistical point of view:
  - OR = $\frac{400*900}{600*100}$=6
    - 6 times higher chance of periodontitis for those who are obese compared to those who are not

Statistical tests for the contingency table

Chi-square test

- H0: there is no association between the disease and the risk factor
- Ha: there is an association between the disease and the risk fa
- If <20% of the cells in the theoretical (expected) table have values <5, the **Fisher exact test** is used

- $p<0.05$ reject H0, accept Ha: there is an association between the risk factor and the disease
- $p\geq0.05$ fail to reject H0: there is NO significant association between the risk factor and the disease

! it does not result from the study which is the cause and which is the effect

Statistical test for the contingency table

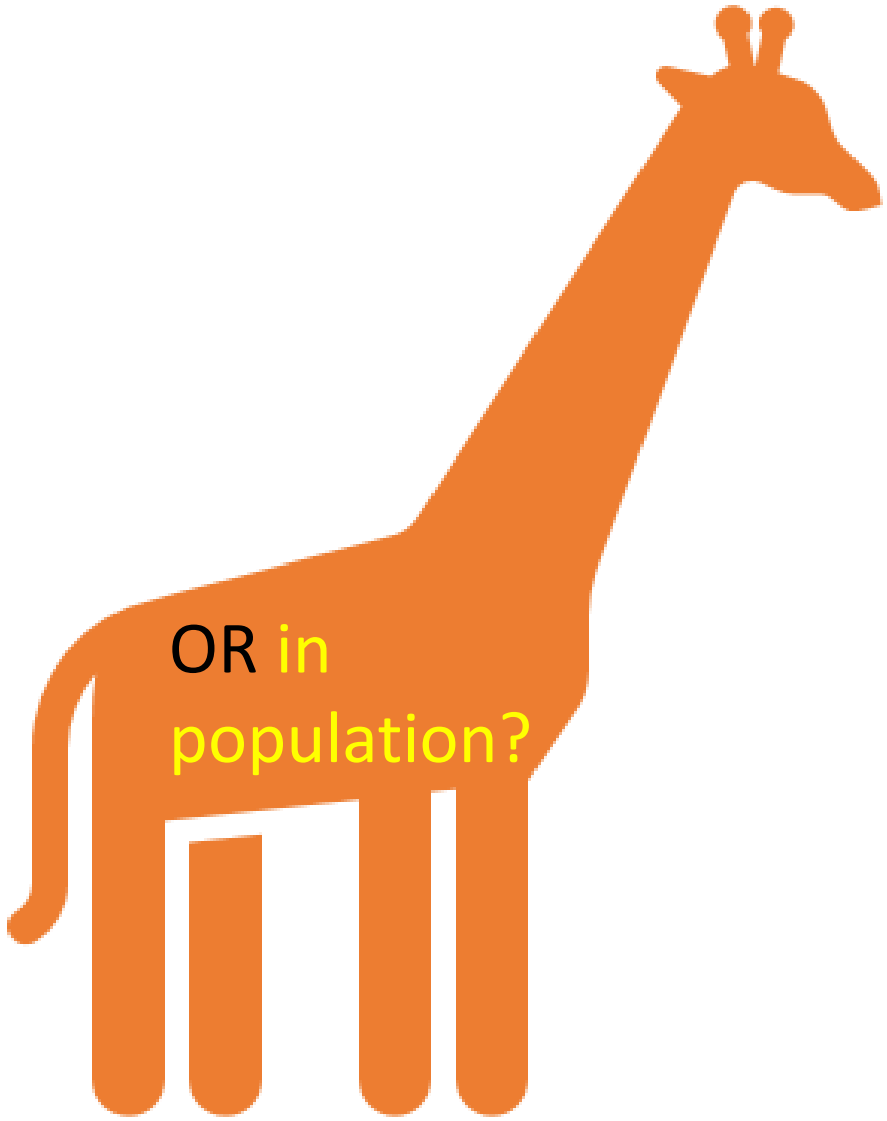
Chi-square test

- H0: there is no association between the periodontitis and obesity
- Ha: there is an association between the periodontitis and obesity

- $p<0.001 \rightarrow p<0.05$ reject H0, accept Ha: there is an association between the risk factor and the disease

- Odds ratio on the sample:

$$\text{OR} = \frac{a * d}{b * c}$$

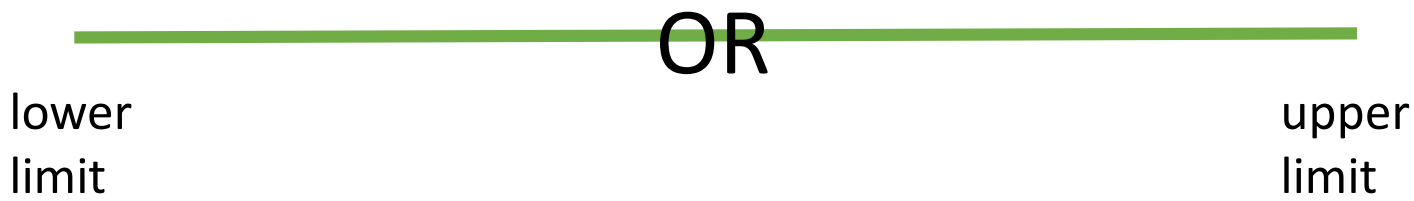


Objective: what would be the OR if we included all individuals (the target population) in the study?

Generalization of OR to the entire population

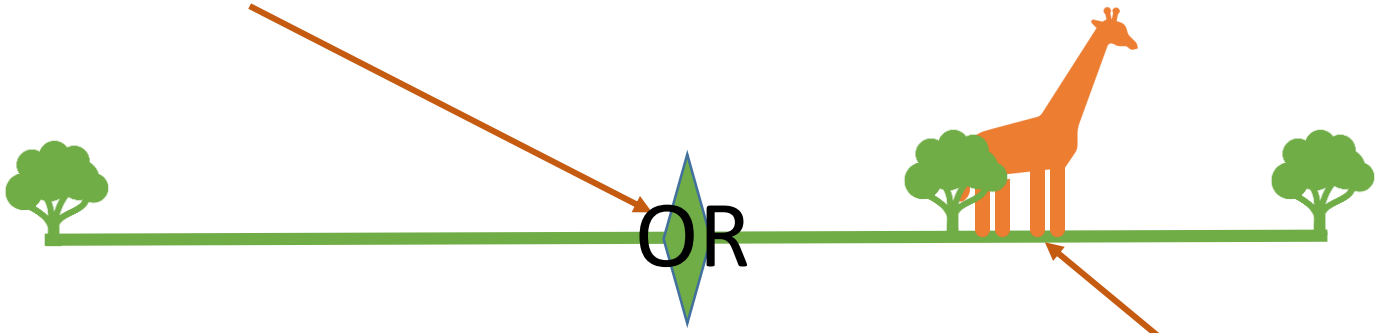
- calculating 95% **confidence interval**

- Ex. OR=6; 95% CI (4.7 – 7.6)
  - 4.7 lower limit of the interval
  - 7.6 upper limit of the interval
  - CI – confidence interval



95% confidence interval for OR

Estimating a fix value = OR calculated from the sample



Confidence interval

OR in population

- estimated with an interval
- population OR is in the estimated interval with 95% probability
- can be anywhere in the estimated interval, we cannot tell where is in the interval,
  - if we want to be more precise we will have to repeat the study in a bigger sample → estimate a narrow 95% CI
- there is 5% level of error = probability of 5% that the population OR is not in the interval

A possible interpretation of the confidence interval

OR="value"; 95% CI ("lower limit"; "upper limit")

1. Interpretation

- in the population OR
  - can be say with a 95% probability
    - it is between the "lower limit" and the "upper limit"
  - What it mean:
    - don't know exactly what the OR is in the population
    - would know if we did the study on the entire population,

2. Interpretation

- in the population OR
  - can be say with a 5% error
    - it is somewhere in the confidence interval

A possible interpretation of the confidence interval

We have also the associated p-value?

3. Interpretation

- If p is statistically insignificant ($p \geq 0.05$), then OR=1 is in the CI
  - possible in the population OR=1 → the factor and the disease are not associated
  - e.g. p=0.08; OR=1.3; 95% CI (0.9; 1.8) $1 \in (0.9; 1.8)$
- If p is statistically significant ($p < 0.05$) then OR=1 is not in the CI
  - in the population OR≠1 → the factor and the disease are associated
  - e.g. p=0.03; OR=1.3; 95% CI (1.1; 1.8) $1 \in (1.1; 1.8)$

- !!! OR = 1 – there is no relationship

A possible interpretation of the confidence interval:

Clinical interpretation

- wide interval
  - e.g. (1.1 – 27)
  - imprecise study
- narrow interval
  - e.g. (2.2 – 2.6)
  - precise study

- you have to appreciate
  - the range is wide, so the study is imprecise
  - the range is narrow, so the study is precise

A possible interpretation of the confidence interval:

Clinical interpretation

- lower margin much greater than 1
  - e.g. (4.2 – 11)
  - important risk factor
- lower margin close to 1 and upper margin close to 1
  - e.g. (1.2 – 1.8)
  - unimportant risk factor
- lower margin close to 1 and upper margin far from 1
  - e.g. (1.2 – 11)
  - risk factor of unclear importance

- you appreciate
  - just argue well

A possible interpretation of the confidence interval:

Clinical interpretation

- upper margin much less than 1
  - e.g. (0.2 – 0.25)
  - significant protective factor
- upper margin close to 1 and lower margin close to 1
  - e.g. (0.8 – 0.95)
  - unimportant risk factor
- upper margin close to 1 and lower margin far from 1
  - e.g. (0.2 – 0.95)
  - risk factor of unclear importance

- you appreciate
  - just argue well

Bias in Case–Control Studies

- are vulnerable to **bias (errors)**
- Bias can distort associations.

Selection Bias

- Occurs when:
- cases and controls are not from the same population
- lost of the deaths from the cases group
- Example:
  - Cases from hospital
  - Controls from general population

Bias in Case–Control Studies

- are vulnerable to **bias (errors)**
- Bias can distort associations.

Recall Bias

- Cases may remember exposures differently
- Example:
- Patients with oral cancer may recall tobacco use more accurately

Bias in Case–Control Studies

- are vulnerable to **bias (errors)**
- Bias can distort associations.

Confounding

- A **confounder** is a variable associated with both:
  - exposure
  - disease
- Example:
- Age may influence:
  - obesity
  - periodontitis

Case-control study advantages

- Low costs
- Relatively short duration
- Several exposure can be studied at the same time
- Useful
  - in the study of rare pathologies
  - in the case of a long time between exposure and the onset of the disease

Disadvantages of Case-control studies

- Cannot determine cause of disease
  - Temporal relationship sometimes unclear
    - we do not know which one was first: exposure or the disease
- Cannot calculate:
  - prevalence of disease in population
  - relative risk of disease in case of exposure
- can study only one objective (outcome, disease)
- vulnerable to bias

Critical Appraisal

— Questions to Ask

- When reading a case–control study:
  - Were cases clearly defined?
  - Were controls appropriate?
  - Was exposure measured reliably?
  - Were confounders controlled?

The problem of assessing causality (is the risk factor the cause of the disease?)

- A single observational (case-control) study is not enough to establish that the factor studied is the cause of the disease!

Take-Home Messages

- Case–control studies start with **disease status**
- They compare **previous exposure**
- Association is measured using **odds ratio**
- Careful control of **bias and confounding** is essential

Thank you!