

Epidemiological methods

Prof Hynek Pikhart

Department of Epidemiology and Public Health

University College London

Objectives:

At the end of the session students should be able to:

- Differentiate between different types of data
- Describe the structure of an epidemiological dataset
- Define and calculate measures of disease occurrence and measures of association
- Describe the basic features of the main types of epidemiological studies
- Explain the main features of bias, confounding, chance

Introduction

- Until 1950s, the term “epidemiology” was mainly used for studies of communicable diseases
- Later, it was suggested that a new field of study should be created to look at non-epidemic diseases
- The meaning of “epidemiology” was broadened to cover also non-communicable diseases

Definition of epidemiology

- A modern definition of epidemiology is thus very general:
- **Epidemiology is the study of the distribution and determinants of disease in population**

Much of epidemiological research is taken up trying...

- to establish associations between exposures and disease rates
- to measure the extent to which risk changes as the level of exposure changes
- to establish whether the associations observed may be truly causal (rather than being just consequence of bias or chance)

- Epidemiology has a major role in developing appropriate strategies to improve public health through prevention
 - public health has wider meaning in this sense; it is about the health of the whole population.
 - it does not cover only classic areas, such as immunization or monitoring of diseases, it also covers factors such as poverty, smoking, nutrition
- In this sense, epidemiology has a crucial role in trying to put into perspective the effects on population health of different risk factors.

Epidemiology

- The study of the **distribution** and **determinants** of the **frequency** of health-related outcomes in specified populations
- Quantitative discipline
- Measurement of disease / condition / risk factor frequency is central to epidemiology

Variables (outcomes/risk factors)

- Binary
 - Deaths (y/n)
 - Sex (m/f)
- Categorical (ordinal or nominal)
 - Frequency of drinking (never, 1-3 times a month, 1-3 times a week, 4 times a week or more often)
 - Severity of pain (none, some, a lot)
 - Marital status (single, married/in partnership, divorced, separated, widowed)
 - Country of birth (Czech R, Slovakia, Poland, Austria, Germany, Ukraine, Hungary)
- Continuous
 - BMI, blood pressure, etc.

What type of variable is...

- Self-rated health (Very poor, poor, average, good, very good)
- Total cholesterol concentration
- Economic activity (employed, unemployed, housewife, pensioner)
- Risk of CVD death in the next 10 years (score)
- Having lung cancer or not
- Quartile of income
- Sex
- Social class (upper, upper-middle, middle, working, lower)

What type of variable is...

- Self-rated health = Categorical (ordinal)
- Total cholesterol concentration = Continuous
- Economic activity = Categorical (nominal)
- Risk of CVD death in the next 10 years (score) = Continuous
- Having lung cancer or not = Binary
- Quartile of income = Categorical (ordinal)
- Sex = Binary
- Social class = Categorical (ordinal)

Binary outcomes: “cases” vs. “non-cases”

- Persons with disease = “cases”
- **Definition of case is crucial**
- E.g.
 - Obesity: $\text{BMI} \geq 30$
 - Hypertension: $\text{SBP} \geq 140$ mm Hg or $\text{DBP} \geq 90$ mm Hg or treatment
 - High cholesterol: ≥ 6.2 mmol/L
- Can be complex in clinical settings
(e.g. metabolic syndrome, depression, etc.)
- **But must always be clearly specified**

Measures of disease frequency

- Used for binary outcomes
- Require a numerator and denominator
 - = **number of persons with disease**
number of persons examined
- expressed as X per 1,000 persons (or per 100,000 etc.)

Numerators and denominators

Example:

- The number of cancer cases in the UK is 247,667 whereas in Belgium it is 47,948
- The UK has a bigger problem in numerical terms
- But do Belgians have lower risk of getting cancer?
 - Numerators alone are meaningless
 - We need both **numerators AND denominators**

Numerators and denominators

- The number of cancer cases in the UK is 247,667 whereas in Belgium it is 47,948
- **UK:** $247\,667 / 65\,000\,000 = 0.00381 = 381$ per 100 000
- **Belgium:** $47\,948 / 11\,000\,000 = 0.00436 = 436$ per 100 000

What we measure.....

a) Prevalence

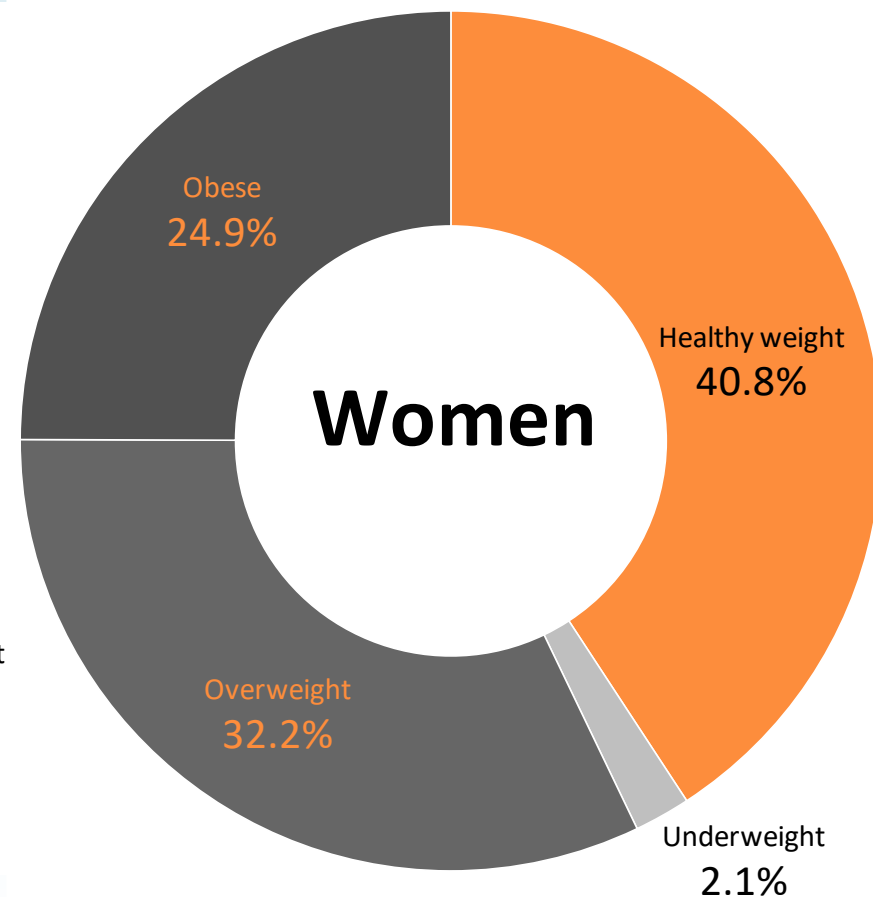
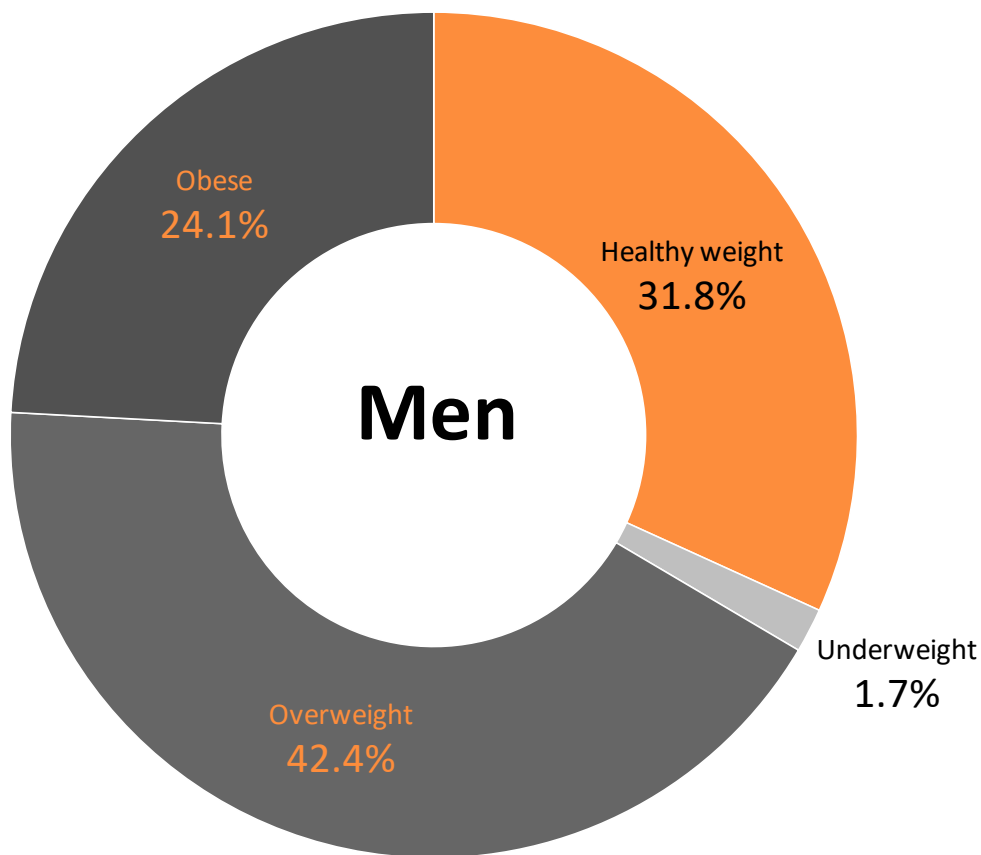
$$= \frac{\text{number of **existing** cases}}{\text{population of interest at a defined time}}$$

- Unable to work now for health reasons
- Occupational injury ever
- Ever wheezing or whistling in the chest
- Hangover in the last 12 months
- Headache today

NOTE a **denominator** is needed for prevalence

Adult prevalence by BMI status

Health Survey for England (2008-2010 average)



Adult (aged 16+) BMI thresholds

Underweight: $<18.5\text{kg/m}^2$

Healthy weight: 18.5 to $<25\text{kg/m}^2$

Overweight: 25 to $<30\text{kg/m}^2$

Obese: $\geq 30\text{kg/m}^2$

b) Incidence

$$= \frac{\text{number of **new** cases in a given time period}}{\text{total population at risk}}$$

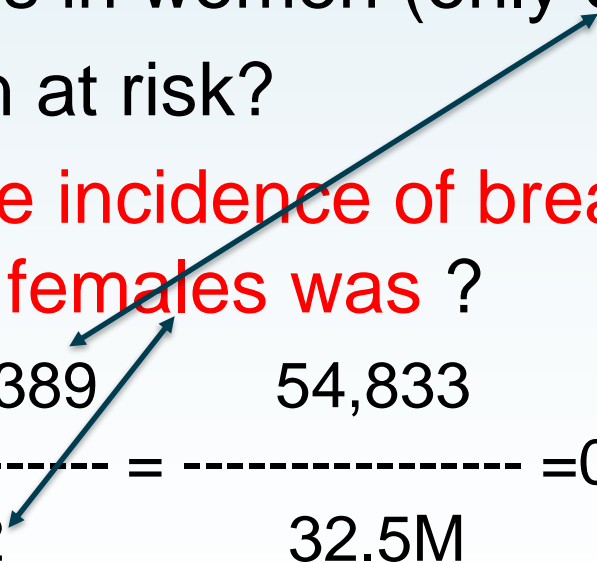
Exercise

- In 2014, **55,222** new cases of breast cancer were diagnosed in the UK.
- Approximately **65M** people in the UK
- Most cases in women (only **389** cases in men)
- Population at risk?
- Cumulative incidence of breast cancer in the UK in 2014 in females was ?

???

???

- In 2014, **55,222** new cases of breast cancer were diagnosed in the UK.
- Approximately **65M** people in the UK
- Most cases in women (only **389** cases in men)
- Population at risk?
- **Cumulative incidence of breast cancer in the UK in 2014 in females was ?**

$$\frac{55,222 - 389}{65\text{M} / 2} = \frac{54,833}{32.5\text{M}} = 0.001687 = 168.7 / 100,000$$


Example

3-year study with a sample size of 100, outcome of interest was fatal heart disease.

	<i>year 1</i>	<i>year 2</i>	<i>Study ends</i>
Developed outcome	6	5	4
Dropped out	4	10	-
Sample at risk	90	75	-

- 10 participants were followed for 1 year
- 15 participants were followed for 2 years
- 75 participants were followed for 3 years

Total person-years:

Incidence Rate:

3-year study with a sample size of 100, outcome of interest was fatal heart disease.

	<i>year 1</i>	<i>year 2</i>	<i>Study ends</i>
Developed outcome	6	5	4
Dropped out	4	10	-
Sample at risk	90	75	-

- 10 participants were followed for 1 year
- 15 participants were followed for 2 years
- 75 participants were followed for 3 years

Total person-years of follow up = $(10 \times 1) + (15 \times 2) + (75 \times 3) = 265$ person-years at risk

Incidence rate = $15 / 265 = 0.057 = 5.7$ cases per 100 person-years

Relationship between prevalence and incidence

- The prevalence of a health-related outcome depends both on the incidence rate and the time between onset and recovery or death
- **Prevalence = Incidence x Average disease duration**

Exercise

- Population of 10,000 people
- 10 new cases of cancer a year
- 20 registered cases at any time
- Average duration of (survival from) the cancer is...

Exercise

- Population of 10,000 people
- 10 new cases of cancer a year
- 20 registered cases at any time
- Average duration of (survival from) the cancer is...
 $20/10$ (prevalence/incidence) = 2 years

c) Mortality

- = **number of deaths / total population**
- Rate (or risk)
- = the number of deaths in a specified population
the number of that population **/per unit time**
- If the mortality rate is to be calculated in a given year, the mid-year population is usually used as the denominator
- Mortality rate is always expressed as deaths per X (e.g. 1,000 persons per year)

Example

- A city has a population of 900,000;
- 30,000 deaths occur in a 3-year period
- Mortality rate for the period = $\frac{30\ 000}{900\ 000}$
 = 0.0033 or 33 deaths per 1,000
 = 11 deaths per 1,000 per year

All mortality rates MUST include

- Number of deaths = *numerator*
- Population size (in which these deaths were counted) = *denominator*
- Time period (during which these deaths happened)

Exercise

Which piece of information were necessary to calculate following result:

1.5 deaths/10,000 population per day

1. All deaths occurred in hospital
2. The population is 29,661
3. 53% of deaths were males
4. The deaths occurred over 3 months
5. 404 deaths happened
6. Tuberculosis caused 17% of the deaths

cases/ population per number of days
....recalculated to 10,000

- The number of deaths = 5.
- The population size = 2.
- The time period in which the deaths occurred = 4.

Which piece of information were necessary to calculate following result:

1.5 deaths/10,000 population per day

cases/ population per number of days

....recalculated to 10,000

1. All deaths occurred in hospital
- ② The population is 29,661
3. 53% of deaths were males
- ④ The deaths occurred over 3 months
- ⑤ 404 deaths happened
6. Tuberculosis caused 17% of the deaths

Mortality rates

- **All-cause mortality rates**
- **Cause-specific mortality rate**
- **Crude mortality rates**
- **Standardized mortality rate**

- **All-cause mortality (rates):** refers to the total number of deaths per 1,000 people per year

- **Cause-specific mortality rate**

= total number of deaths due to a specific cause
(population at risk x period of time)

- **Crude mortality rates:** no care has been taken for age structure of the population
 - Counts all deaths
 - All cases
 - All ages and sexes
 - Denominator includes entire population
 - All ages and sexes

- **Standardized mortality rate** refers to a mortality rate which is age-standardized in order to permit comparisons between different countries, regions etc.

=also **age-specific mortality rate**

- Infant mortality rate
- Maternal mortality rate
- Under-5 mortality rate

Other commonly used measures

- **Perinatal mortality rate** is the number of neonatal and fetal deaths (stillbirths) per 1000 births
- **Case fatality rate** is the rate of death among people who already have a condition, usually in a defined period of time. usually measured as a decimal or as a percent.
- **Survival rate** is the proportion of people who remain alive for a given period of time after diagnosis of disease. E.g. breast cancer has 5-year survival rate around 70%.

Recently often measured and mentioned...

- **Number of excess deaths**

Hypothetical number of deaths caused by the emergency itself

Or

Number of deaths that would not have occurred had the emergency not happened

=Difference between pre-emergency mortality rate and mortality rate found during emergency \times population size

Most recent excess deaths

- Covid

Exercise & a bit of history

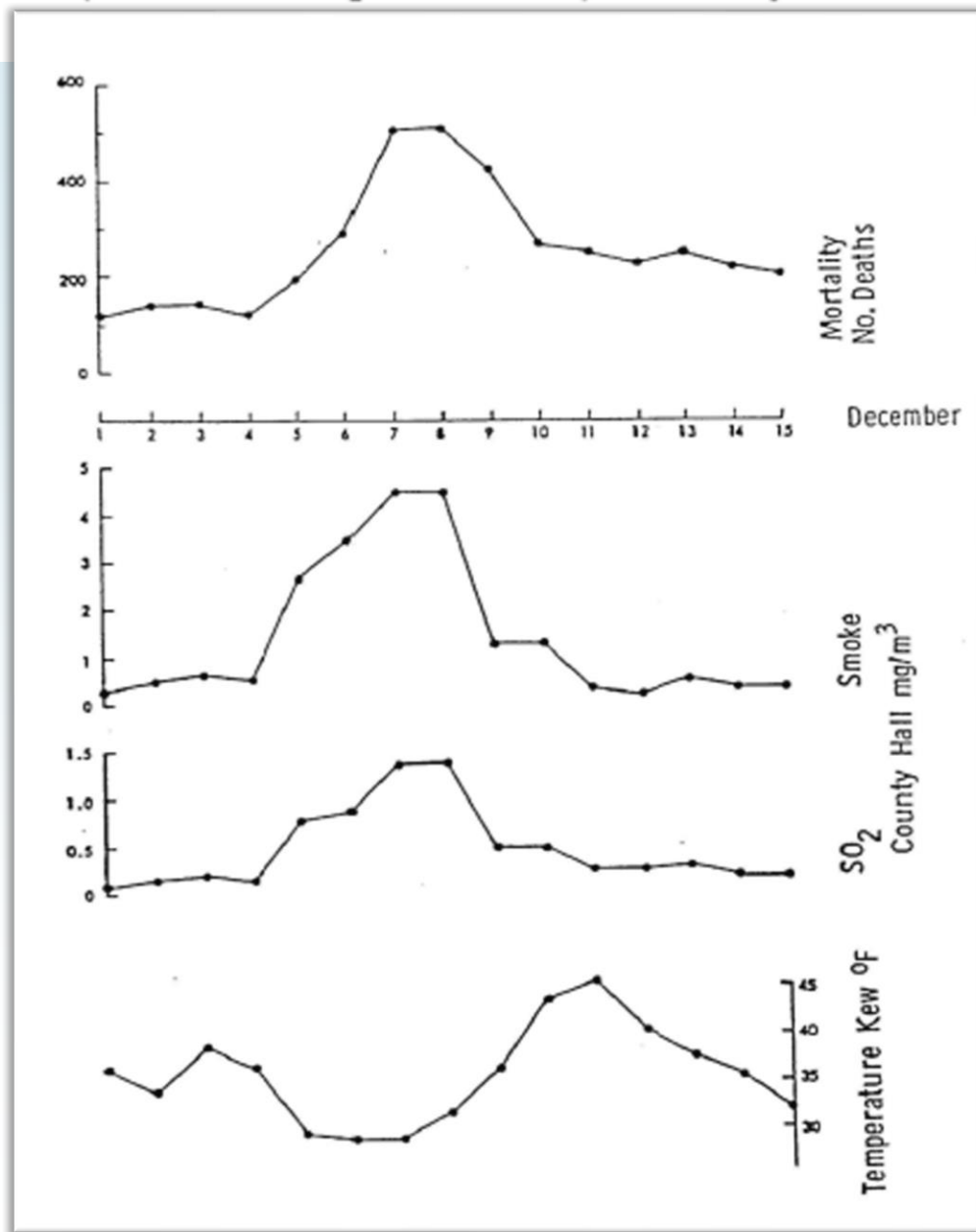
Smog in London



During the first half of **December of 1952**, the London area experienced periods of fog culminating in one of the most intense in memory lasting from the morning of Friday 5 December to early in the morning of Tuesday 9 December and then dispersed quickly when the weather changed.

Air pollution and meteorological factors, particularly low temperatures, were suggested as possible causative or contributory agents. A period of cold weather, combined with an anticyclone and windless conditions, collected airborne pollutants - mostly arising from the use of coal - to form a thick layer of smog over the city.

Figure 1 Numbers of deaths in London AC, mean daily temperature at Kew and mean atmospheric pollution (smoke and sulphur dioxide) at County Hall between 1st and 15th December 1952.



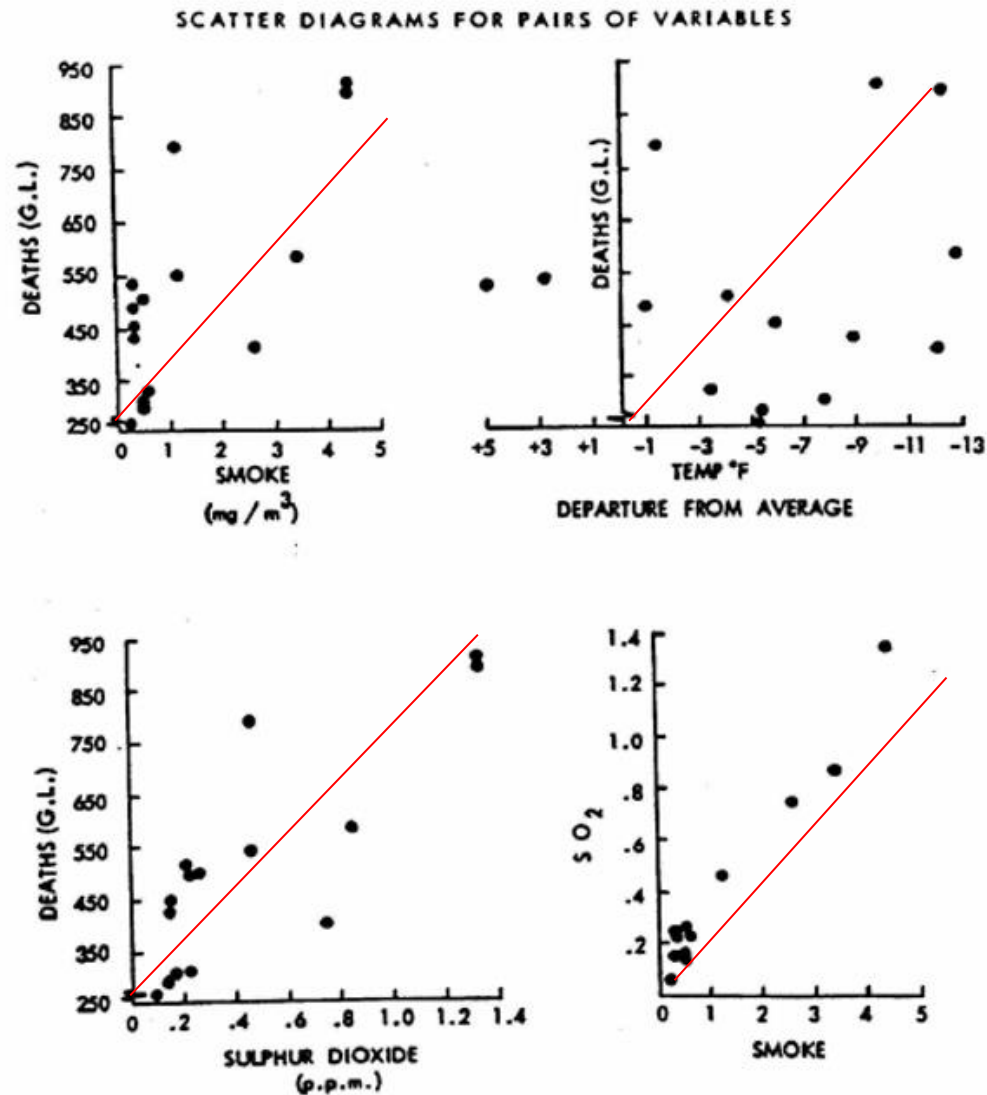
Q1. Comment on findings in Table 1

- a)** This is 'numerator data' referring to cases and not to population-based measurements such as incidence or prevalence. However, over the short period involved, the assumption of a constant population at risk is not unreasonable. You should have thought to plot the data (roughly).
- b)** Data presented by date of registration, not date of death —possible delays over the Christmas period.
- c)** November forms the 'baseline' against which the 'epidemic' can be assessed.
- d)** Comparison with other great towns suggests an epidemic specific to London, and most acute in Central London (London AC).

Q2. Define the period of excess mortality in London and its relationship to the prevailing weather conditions

- a) Compared to Dec 1-4, excess mortality peak at Dec 7-8 is 4-5 times baseline rate, and has not entirely resolved by 15th.
- b) Time course of smoke and SO₂ very similar to each other and closely followed by rise and fall in numbers of deaths. Latent period of 24-48 hours suggests an acute toxic mechanism.
- c) Temperature changes less well matched by changes in numbers of deaths.

Figure 2 Scatter diagrams showing the relationship between number of deaths in Greater London between 1st and 15th December 1952 with air pollution and meteorological factors measured during that time, and a scatter diagram showing the association between smoke and sulphur dioxide in that period.



Q3. How might you proceed to further investigate the influence of fog on mortality?

- a) Clarify the nature of the epidemic. Which causes of death were most affected? Which age groups? Were deaths confined to already sick?
- b) Obtain more detail on the hazardous 'exposure'. Examine the time course and geographical distribution of fog, smoke and sulphur-dioxide. Compare the effect on outdoor and indoor workers?
- c) Look for similar epidemics associated with fog elsewhere and in London in the past. (This was the first time that daily mortality returns had been examined — monthly figures would have shown a much less marked effect).
- d) Follow time course of mortality after the fog to investigate delayed consequences (cause-specific data by age + sex most useful).

Now, we have data and we know basic measures

- Let's consider **2 groups** of individuals
- An **exposed** group (group with risk factor of interest) and **unexposed** group (without such factor of interest)
- We are interested in comparing the amount of disease (mortality or other health outcome) in the exposed group to that in the unexposed group

Measures of association

- Risk of disease, rate of disease in different groups of population
- Comparison of risks/rates

(Absolute) Risk

- Risk is the probability of new occurrence of disease among individuals in an initially disease-free population during a defined time period
- To calculate a risk (r), we divide the number of **new cases** (d) in the defined period by the **population at risk** at the beginning of the **period** (N);
(d and N are referred to as the numerator and denominator, respectively)

$$r = d / N \text{ over a defined period}$$

- Risk is probability but is often multiplied by a suitable number (eg 100,000)

Example

- *In 1980, an annual risk of death was
14 per 1,000 in Kenya,
10 per 1,000 in France
26 per 1,000 in Malawi*

(United Nations, Demographic Yearbook)

Risk measures

- Risk in exposed (r_1)
- Risk in unexposed (r_0)

Risk ratio

$$RR = \frac{r1}{r0} = \frac{\text{incidence in the group with attribute/exposure}}{\text{incidence in a group without attribute/exposure}} = \frac{a / (a+b)}{c / (c+d)}$$

		DISEASE status		Total
		yes	no	
EXPOSURE status	yes	a	b	a+b
	no	c	d	c+d
Total		a+c	b+d	a+b+c+d

Risk difference

- the absolute difference between two risks (or rates)

$$RD = r_1 - r_0$$

$$[a / (a+b)] - [c / (c+d)]$$

		DISEASE status		Total
		yes	no	
EXPOSURE status	yes	a	b	a+b
	no	c	d	c+d
Total		a+c	b+d	a+b+c+d

- We can also have different **strata** of exposure
- We may calculate ratio measures for each strata = we compare measure of frequency in each level with measure of frequency in the baseline (unexposed) level

Example

Death rates from CHD in smokers and non-smokers by age

Age	Smokers rate	Non-smokers rate	Rate ratio
35-44	0.61	0.11	5.5
45-54	2.40	1.12	2.1
55-64	7.20	4.90	1.5
65-74	14.69	10.83	1.4
75-84	19.18	21.20	0.9
85+	35.93	32.66	1.1
ALL AGES	4.29	3.30	1.3

What can you say about this table?

Age	Smokers rate	Non-smokers rate	Rate ratio
35-44	0.61	0.11	5.5
45-54	2.40	1.12	2.1
55-64	7.20	4.90	1.5
65-74	14.69	10.83	1.4
75-84	19.18	21.20	0.9
85+	35.93	32.66	1.1
ALL AGES	4.29	3.30	1.3

The rate ratio decreases with increasing age.

It may suggest that the effect of smoking on the rate of CHD is higher in younger ages.

Odds of disease

- We can calculate risks, risk ratio, risk difference however the analysis is often based on **ODDS RATIOS**

Odds of disease/survival

- related measure of disease occurrence
- for a defined population and time period

$$\text{Odds} = \frac{\text{Cases}}{\text{Non cases}}$$

=by the time of observation

- In many situations, it may be easier to calculate odds ratio (**OR**) which is defined as

$$\frac{\text{odds of disease among exposed}}{\text{odds of disease among unexposed}} = \frac{a/b}{c/d} \quad \left(\frac{\text{odds}_1}{\text{odds}_0} \right)$$

$$\text{OR} = \text{odds}_1 / \text{odds}_0$$

		DISEASE status		Total
		yes	no	
EXPOSURE status	yes	a	b	a+b
	no	c	d	c+d
Total		a+c	b+d	a+b+c+d

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

Odds ratio as an approximation to the risk ratio

- For a rare disease, odds ratio is approximately equal to the risk ratio (because denominators are very similar)

If disease common:

Disease	Exposed	Unexposed	Total
Yes	50	25	75
No	50	75	125
Total	100	100	200

$$\frac{a}{a+b}$$

$$\frac{c}{c+d}$$

$$R_1 = 50/100 = 0.5 \quad R_0 = 25/100 = 0.25$$

$$RR = 2.0$$

$$\frac{a}{b}$$

$$\frac{c}{d}$$

$$Od_1 = 50/50 = 1.0 \quad Od_0 = 25/75 = 0.33$$

$$OR = 3.0$$

Measures of population impact

- **Population attributable risk (PAR)** is the absolute difference between the risk (or rate) in the whole population and the risk or rate in the unexposed group

$$PAR = r - r_0$$

Population attributable risk fraction (PAF)

- It is a measure of the proportion of all cases in the study population (exposed and unexposed) that may be attributed to the exposure, on the assumption of a causal association
- Also called the aetiologic fraction
 - the percentage population attributable risk
 - the attributable fraction

- If r is rate in the total population

$$\text{PAR} = r - r_0$$

$$\text{PAF} = \text{PAR}/r$$

$$(\text{PAF} = (r-r_0)/r)$$

Example

- 50 persons attended a garden party
- 25 of them developed diarrhoea in the next 3 days
- What was the risk of diarrhoea among the participants of the party?

- 25/50

Example II.

- 30 party visitors had a BBQ (minced meat)
- 24 of them developed diarrhoea

- 20 people did not eat BBQ
- 1 of them developed diarrhoea

- How would you calculate RR related to eating BBQ?

30 party visitors had a BBQ (minced meat)
24 of them developed diarrhoea

20 people did not eat BBQ
1 of them developed diarrhoea

- Risk among unexposed R_0 :
- $1/20$

- Risk among exposed R_1 :
- $24/30$

- Relative risk $RR=R_1/R_0=(24/30)/(1/20)=16$

Introduction to epidemiological study design

What is the purpose of Epidemiology?

“the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the prevention and control of health problems”

= Last's Dictionary of Epidemiology

It enables us to:

- Describe patterns of disease in populations
- Study the determinants or risk factors of disease
- Compare disease between groups
- Assess the effectiveness of interventions

Basic tool in epidemiology is the study

- foundations of a good research proposal is the **study design**
- defines how data or evidence is collected and can be used to compare disease between groups

Study

*“An **epidemiological study** is a statistical study on human populations, which attempts to link human health effects to a specified cause”* (wikipedia.org).

- Epidemiology studies *populations*, not individuals
- *Statistical study*: requires large number of people
- *Effects*: often means associations but here it means consequences
(i.e. disease, health condition)
- *Cause*: often means *risk factor*, because *cause* implies *causal association* which is very difficult to demonstrate in epidemiology

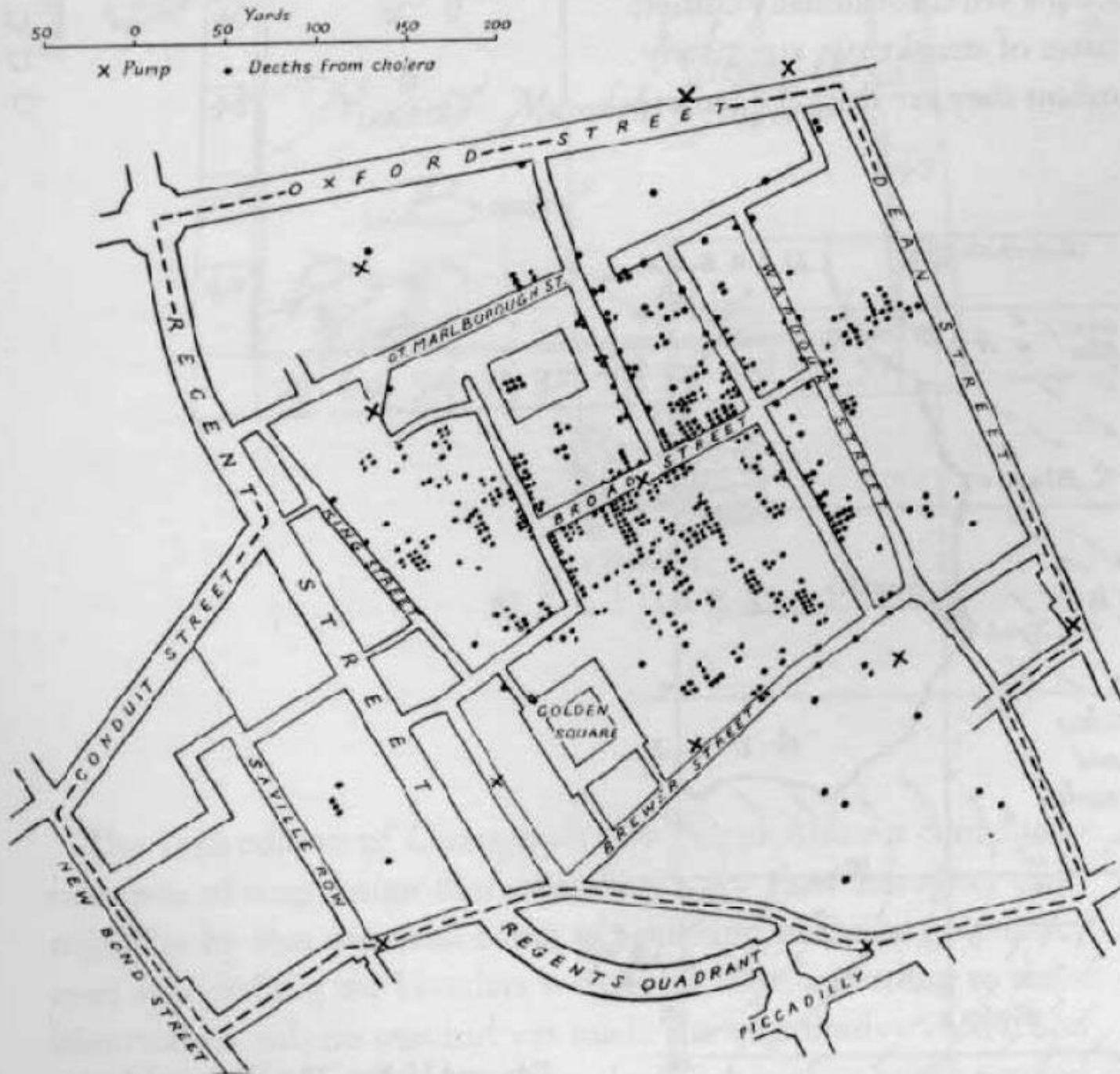
Exercise

John Snow and cholera

- **Introduction**

John Snow (1813-1858) was a physician in London who was distinguished for, among other things, administering chloroform to Queen Victoria at the birth of two of her children. He is best known for his studies of cholera, in particular of two outbreaks which occurred in London in 1848-49 and 1853-54.

- **Background: the 1848/49 cholera epidemic in London**
- Cholera periodically swept across Europe during the nineteenth century
- After a severe epidemic in 1832, the disease next appeared in London in 1848.
- The severity of this epidemic (approximately **15,000** recorded deaths from cholera) led to considerable discussion in the medical press. Mortality was particularly severe in the **low-lying areas along the banks of the Thames River**; hypotheses about what caused cholera included living in lower regions and the existence (contested at the time!) of microbes.



- **Classic exercise about a classic incident**
- *London in 1850:*
 - *no electricity*
 - *no tarred roads*
 - *full of horses and cows*
 - *low standard of hygiene*
 - *squalor*
 - *sewage drained into the Thames*

Q1. Relevance of data to Snow's hypothesis:

Note the thoroughness of the study - data on 330/334 = 99% deaths. A huge effort. Appear consistent with Snow's hypothesis, but need rates. It is possible that the ratio of the number of houses or persons supplied with water by Southwark + Vauxhall (S+V) compared to number supplied by Lambeth (L) is 286/14, thus the numbers of deaths are as expected.

- **Q2. Deaths per houses data**

- He inferred that these data supported his hypothesis that cholera was transmitted.

Risk S+V = $1,263/40,046 = 31.54$ per 1000 houses

L = $98/26,107 = 3.75$ per 1000 houses

Other = $1,422/256,423 = 5.55$ per 1000 houses

- But we should also consider:

a) perhaps S+V houses were bigger, divided into flats?

b) perhaps S+V houses in poorer, lower, denser area? Note that S+V area is lower lying, along the river.

Better data because he obtained denominators. Thus, he could look at deaths in relation to the number of houses receiving water from each company. Although he did not know the number of *people* actually living (and drinking) in each house, it was the best approximation he could get of the number of people “at risk”.

- **Q3.** The spatial clustering of cases suggests that the pump in Broad Street is the source – the density of cases decreases in all directions from this pump. However, note that (again) no denominators are given – how many people lived near to the Broad Street pump compared to the other pumps?

- **Q4.** Pump closed on the 8th - epidemic was almost over by then. So it was not the removal of the pump handle that caused the epidemic to stop (although this had great publicity value).
- Epidemic may have stopped as a result of:
 - exhaustion of susceptibles (local people had already become ill or had fled)
 - dilution of contamination

Epidemiology = comparison

- 550 cases of stomach cancer

- 550 cases of stomach cancer in Hertfordshire

- 550 cases of stomach cancer in Hertfordshire in 2005

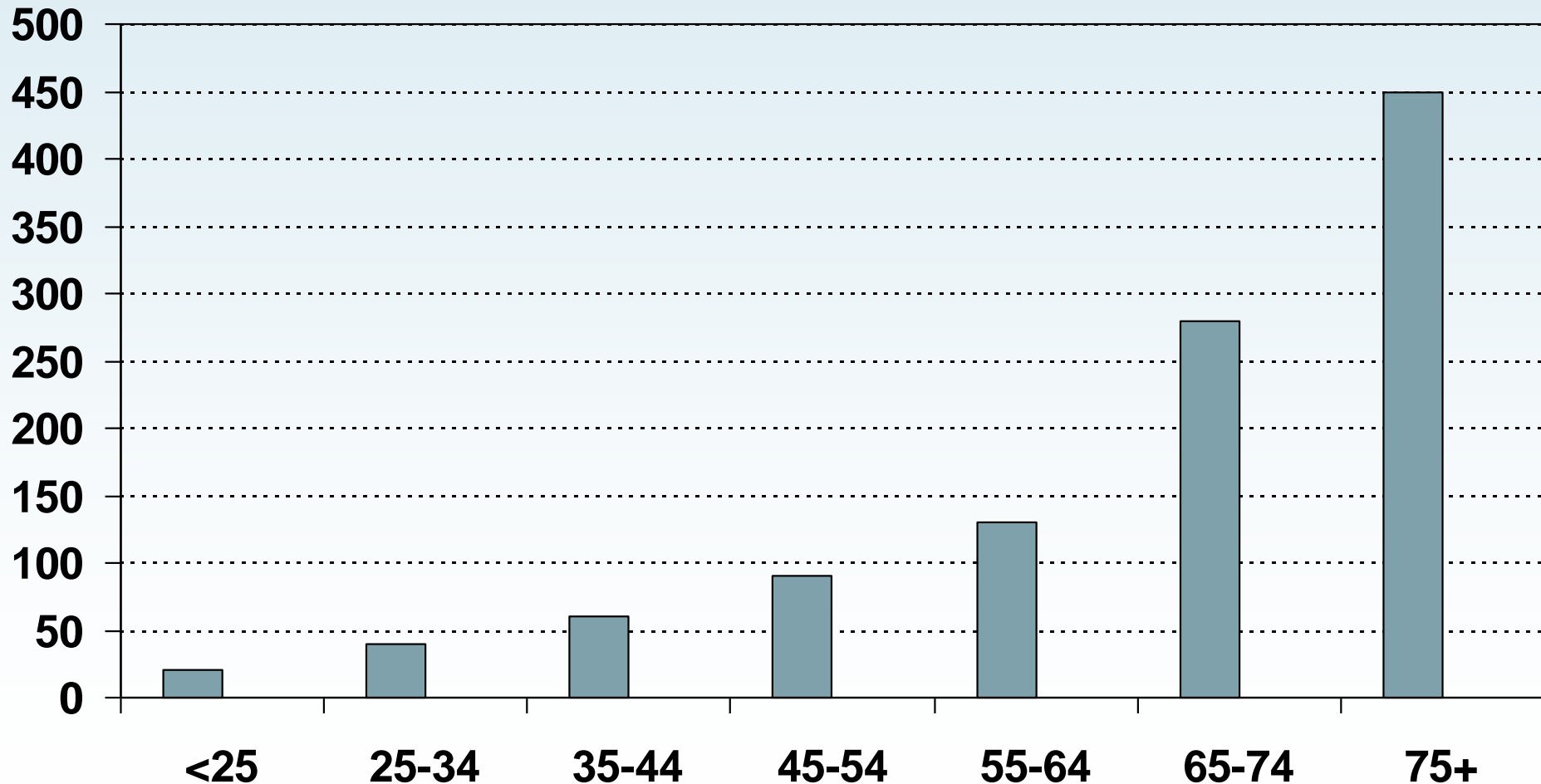
- 550 cases of stomach cancer in Hertfordshire in 2005
- Population 550,000

=> Rate

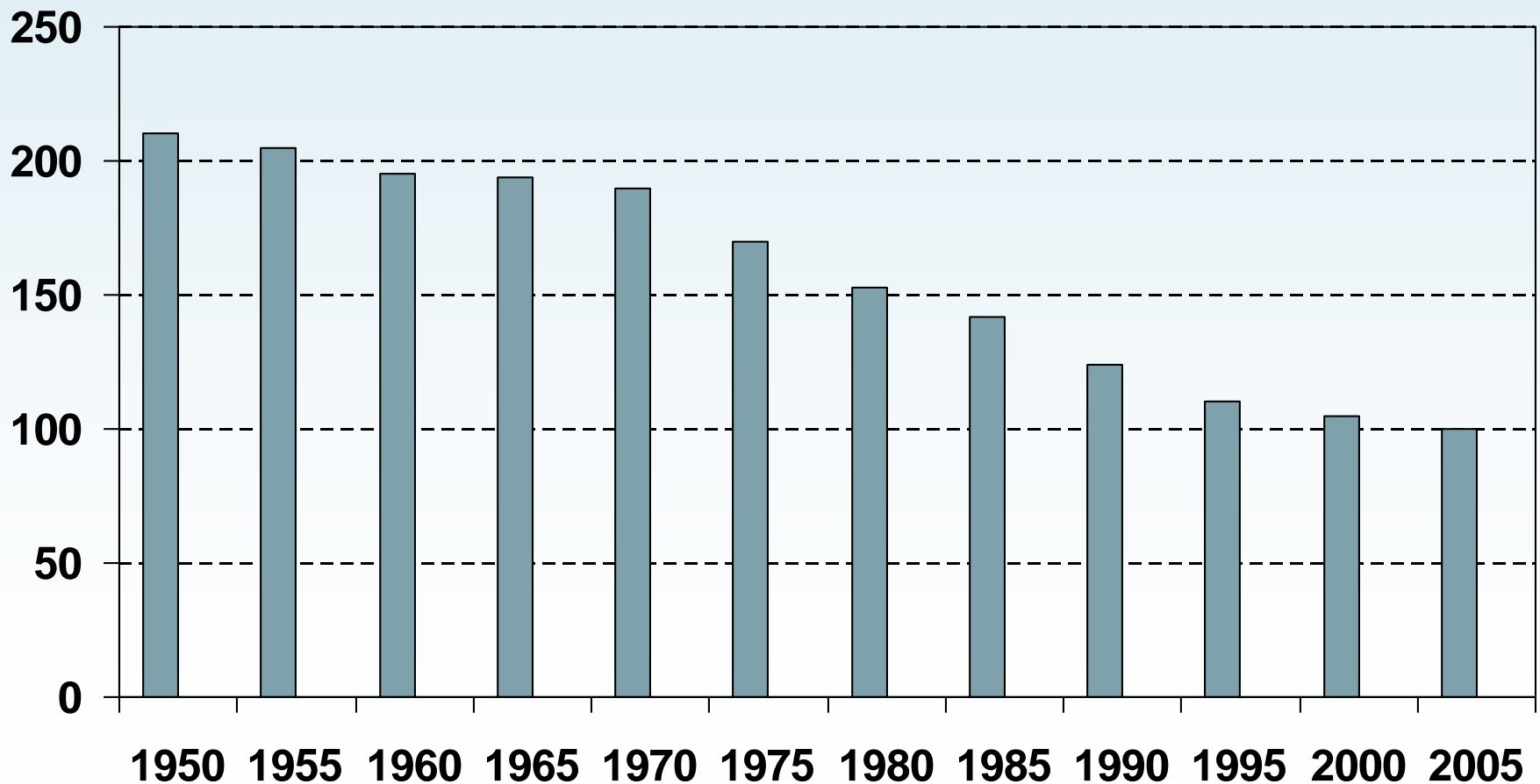
- 550 cases of stomach cancer in Hertfordshire in 2005
 - Population 550,000
- => Rate 100/100,000

What else would be our interest?

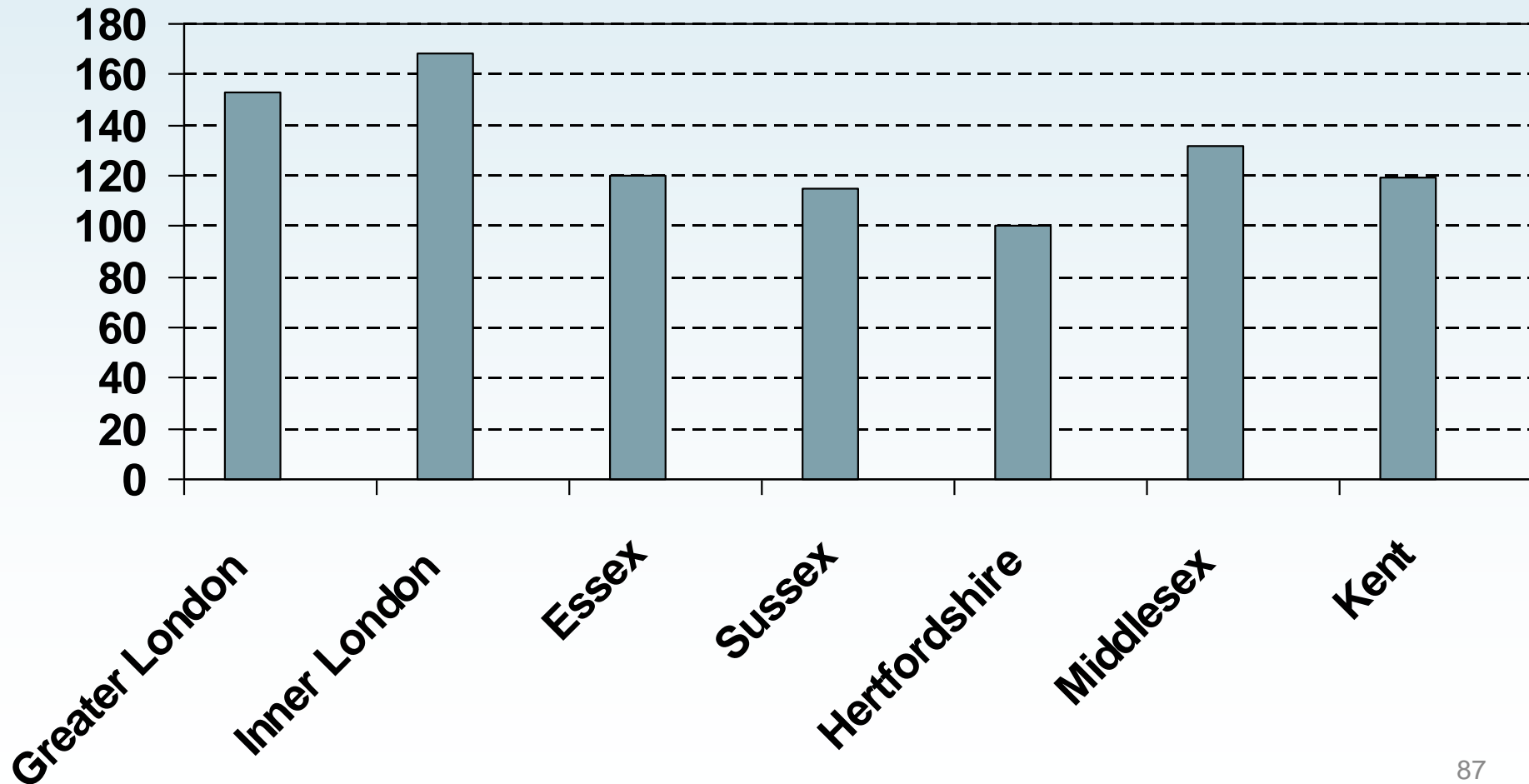
Stomach cancer by age group, 2005, per 100,000



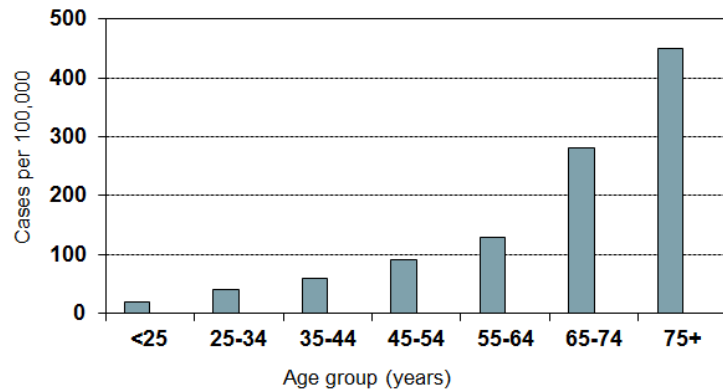
Stomach cancer in Hertfordshire, 1950-2005, per 100,000



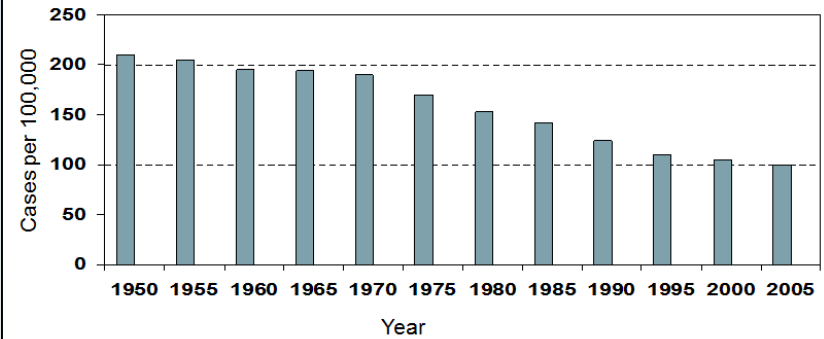
Stomach cancer in SE England in 2005, per 100,000



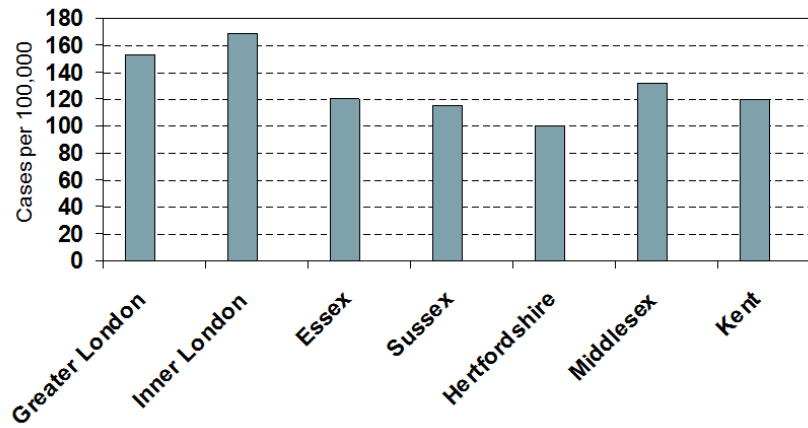
Stomach cancer in Hertfordshire in 2005 by age group per 100,000



Stomach cancer in Hertfordshire, 1950-2005, per 100,000



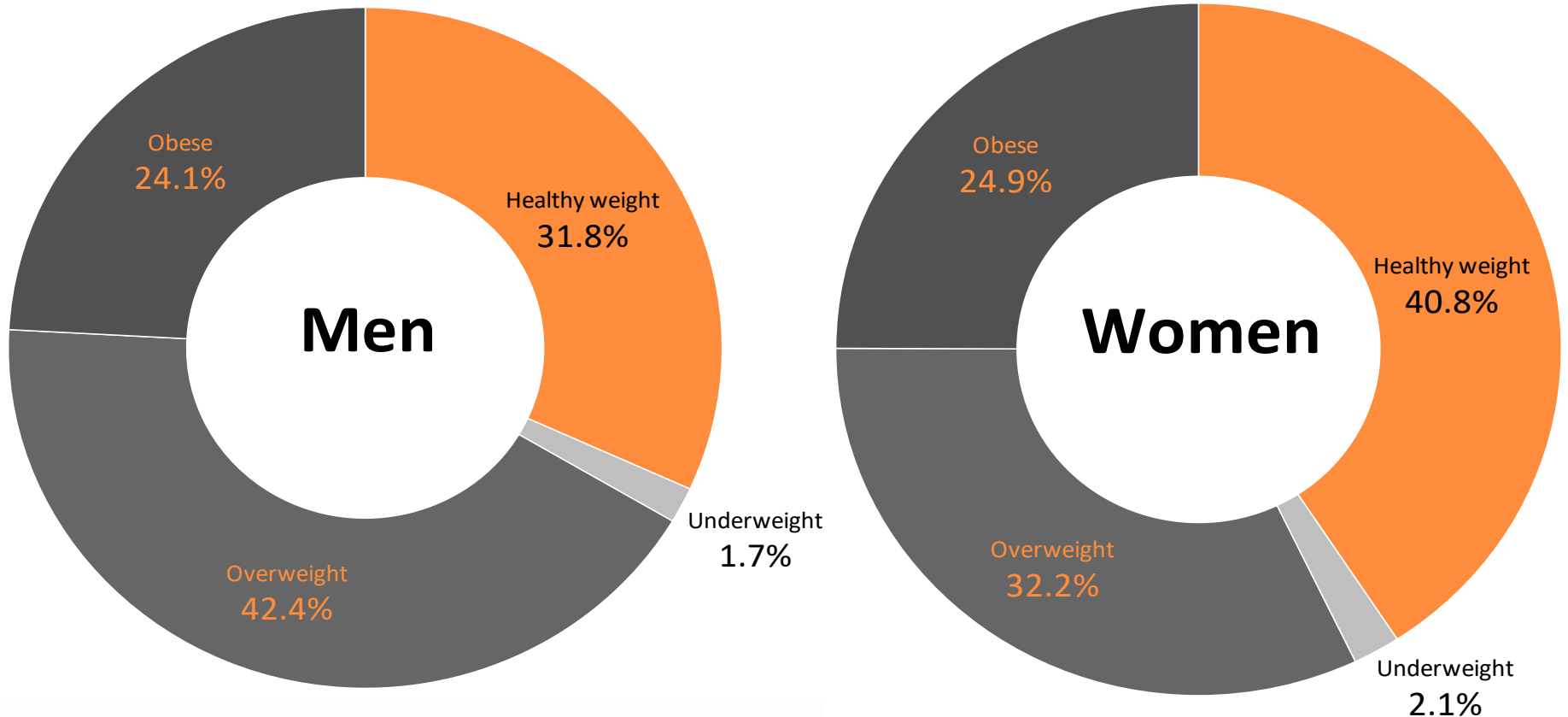
Stomach cancer in SE England in 2005, per 100,000



We compare

Adult prevalence by BMI status

Health Survey for England (2008-2010 average)

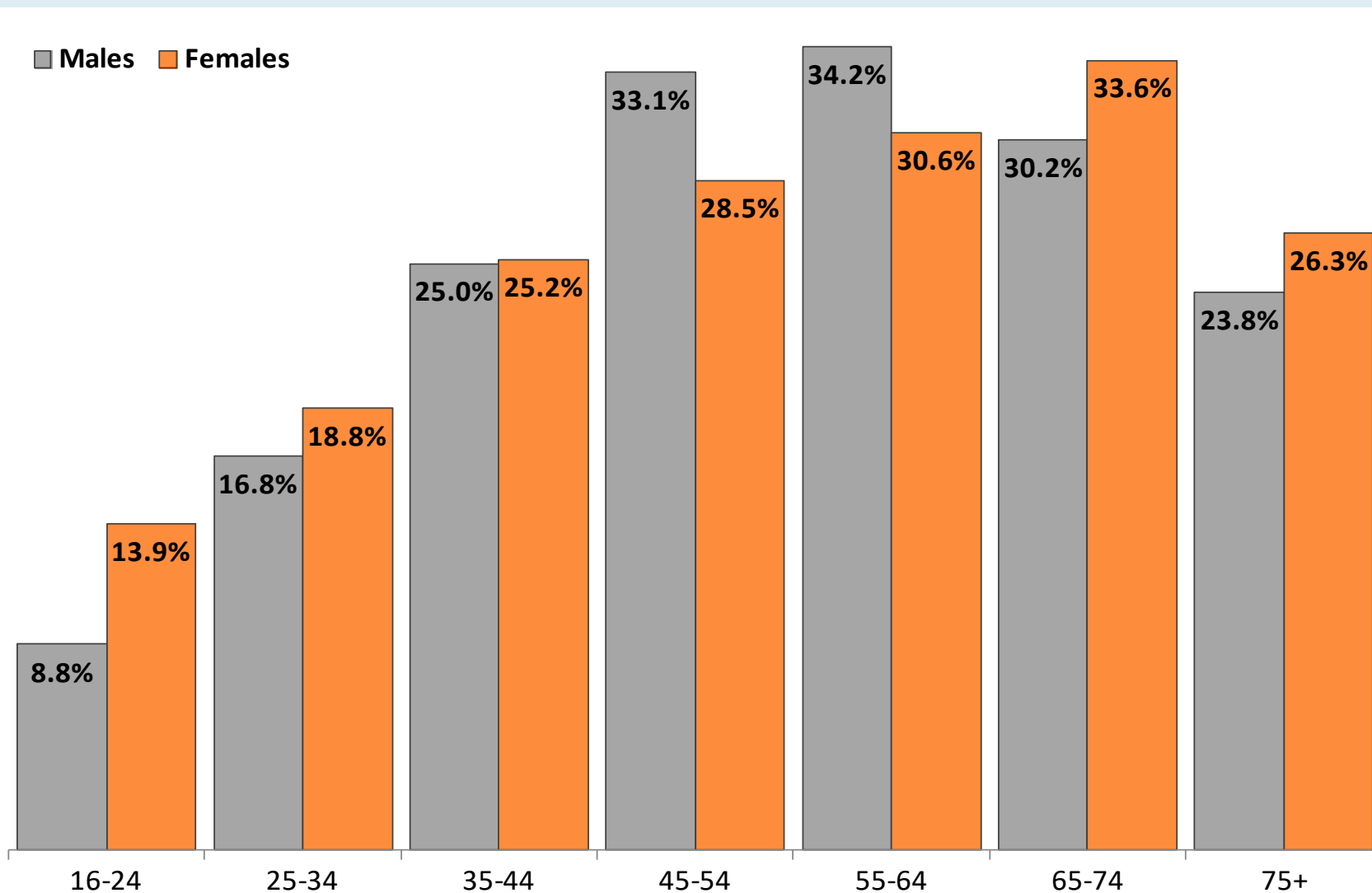


Adult (aged 16+) BMI thresholds

- Underweight: $<18.5\text{kg/m}^2$
- Healthy weight: 18.5 to $<25\text{kg/m}^2$
- Overweight: 25 to $<30\text{kg/m}^2$
- Obese: $\geq 30\text{kg/m}^2$

Adult obesity prevalence by age and sex

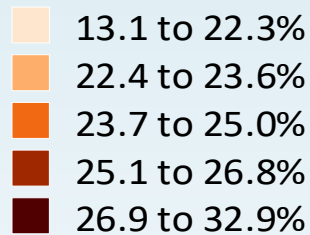
Health Survey for England 2008-2010



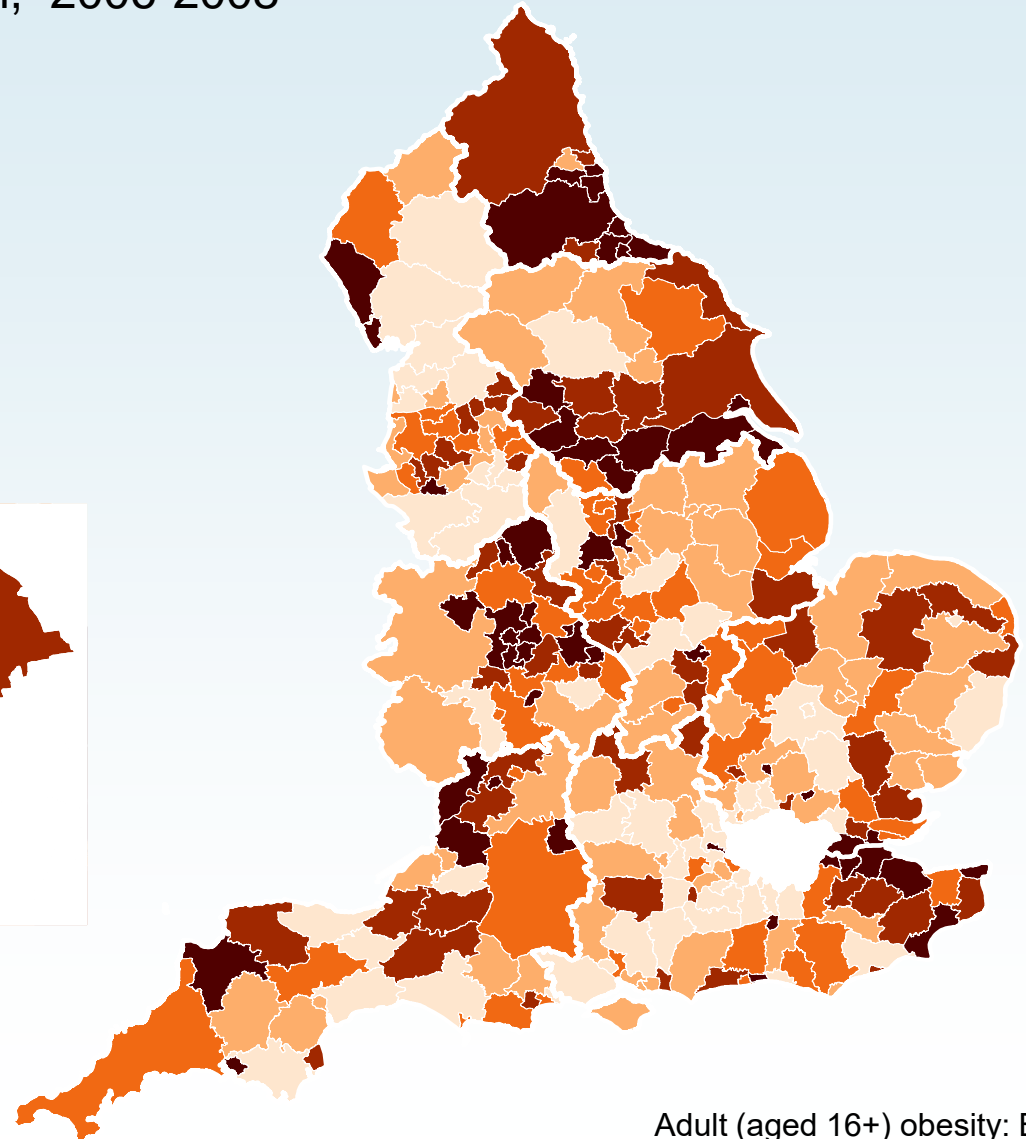
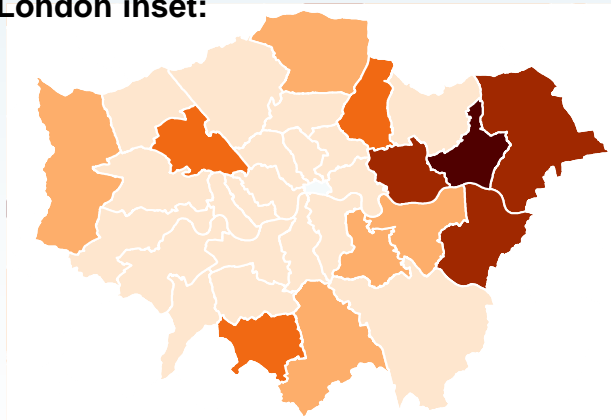
Adult (aged 16+) obesity: BMI \geq 30kg/m²

Adult obesity prevalence modelled estimates

National Centre for Social Research, 2006-2008



London inset:

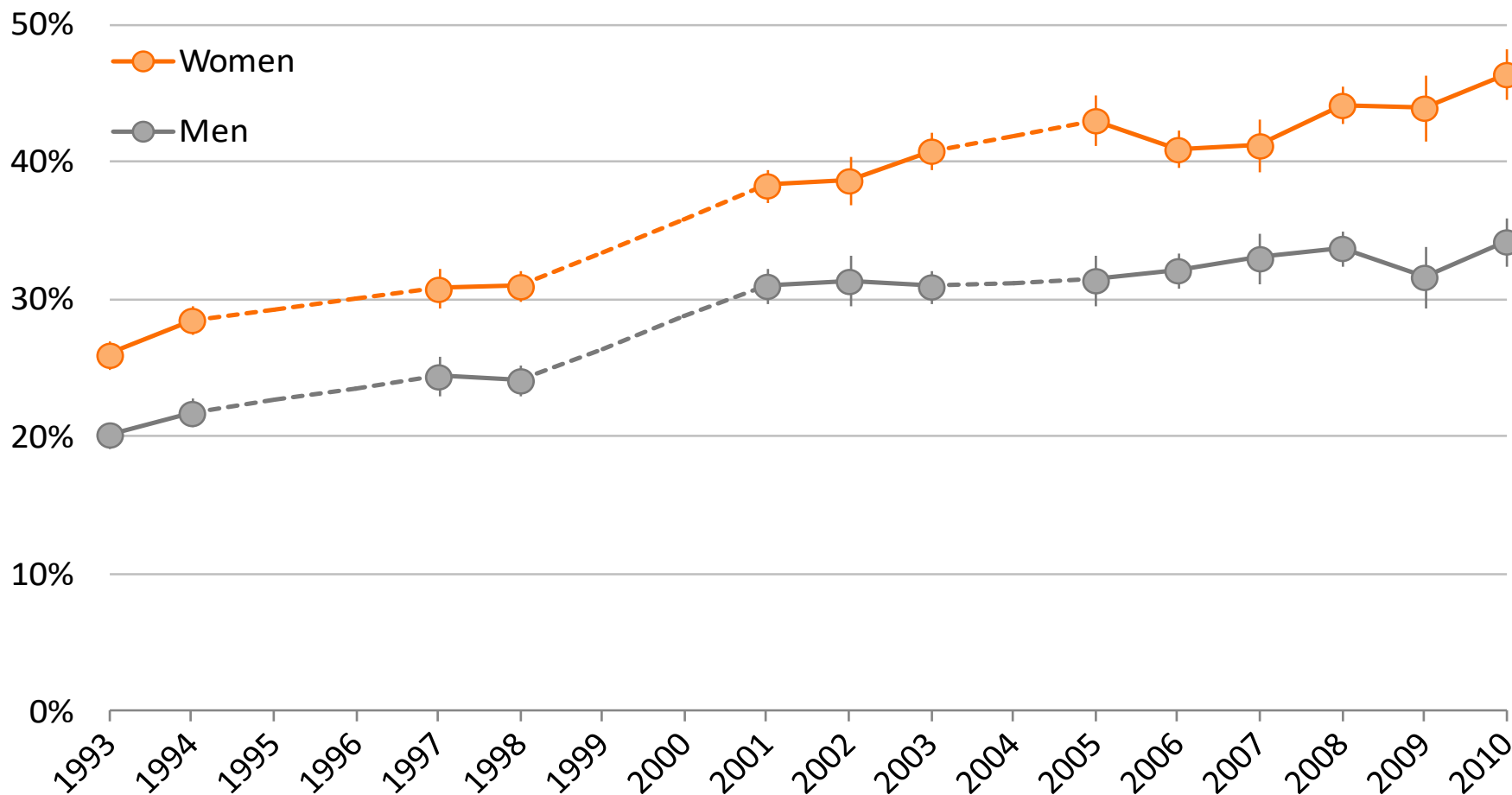


Obesity prevalence (%)
 by Local Authority

Adult (aged 16+) obesity: BMI \geq
 30kg/m²

Trend in raised waist circumference among adults

Health Survey for England, 1993 - 2010



The chart shows 95% confidence limits
Adults aged 16+ years

Raised waist circumference defined as >102cm for men and >88cm for women

Epidemiology = comparison

- Type of comparison (= type of study) depends on purpose.
- E.g.
 - *Describe* the disease / condition
 - Study (*analyse*) its determinants / causes
 - Study (*analyse*) prevention / treatment

Two primary criteria

- Descriptive vs. analytical
- Observational vs. interventional

Descriptive vs. analytical studies

- describe a pattern of occurrence of a disease: ***descriptive studies*** (always observational)
- to analyse the relationship between a disease and an exposure of interest: ***analytical studies*** (can be both observational and interventional)

Descriptive studies

- Describe patterns of disease occurrence
- **Fast and cheap BUT often do not allow proper comparisons**
- Useful for:
 - health services planning
 - hypothesis formulation in research
- Usually based on existing data:
 - mortality
 - reporting of diseases (infections, STDs, cancers...)
 - hospital and medical records
 - Census
 - employment statistics etc.

Four basic questions :

WHAT ? Who? Where? When?

Person (Who?)

age, sex, marital status, social class etc.

Place (Where?)

Geography within countries (cancer atlases etc.)
or internationally

(Japanese more stomach ca than in USA)

! *Special case - migrant studies*

Time (When?)

Changes over time:

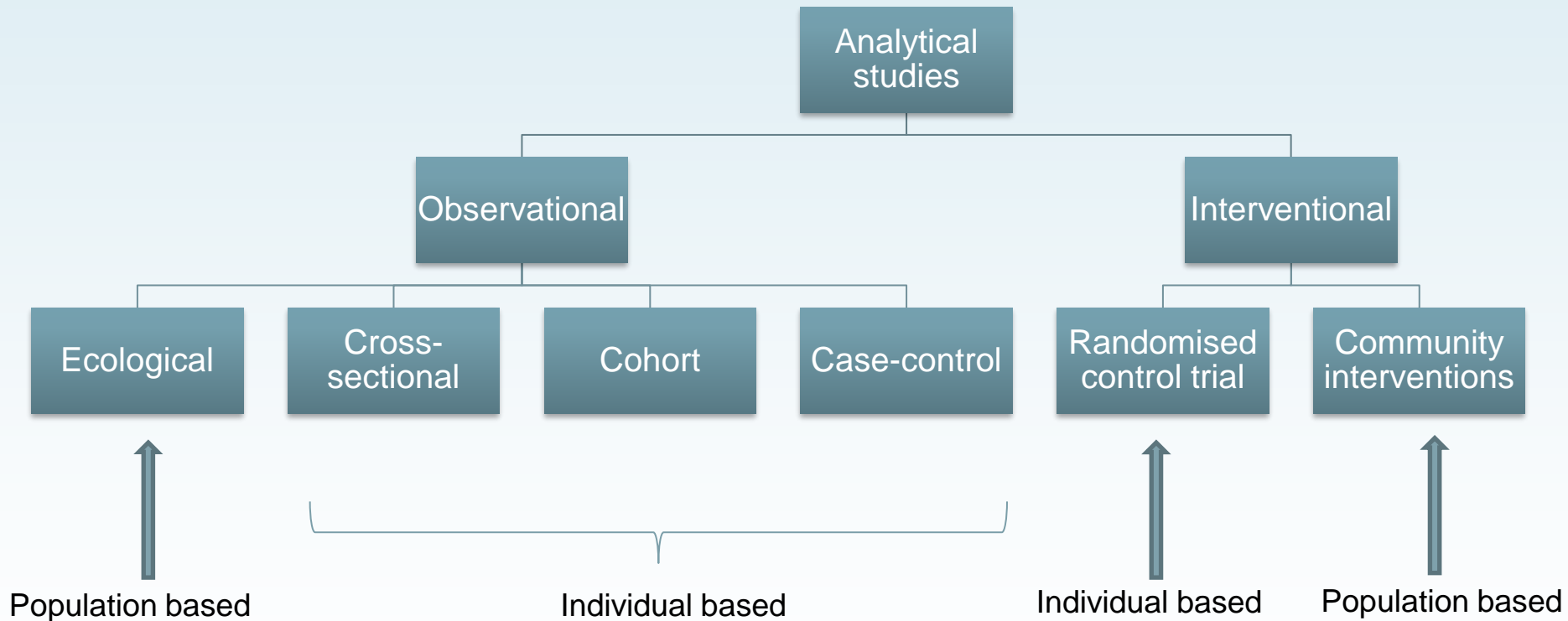
- sudden onset of diseases (thalidomide, toxic shock sy)
- seasonal pattern (births, deaths, infections, etc.)
- secular trends

All in
relation
to the
“**What**”

Analytical studies

- Analysed relationship between exposure and disease
- Often used in aetiological research
- Include
 - ecological studies
 - cross-sectional studies
 - cohort studies
 - case-control studies
 - interventional studies (RCT, prevention trials etc)

Analytical studies



Observational Studies

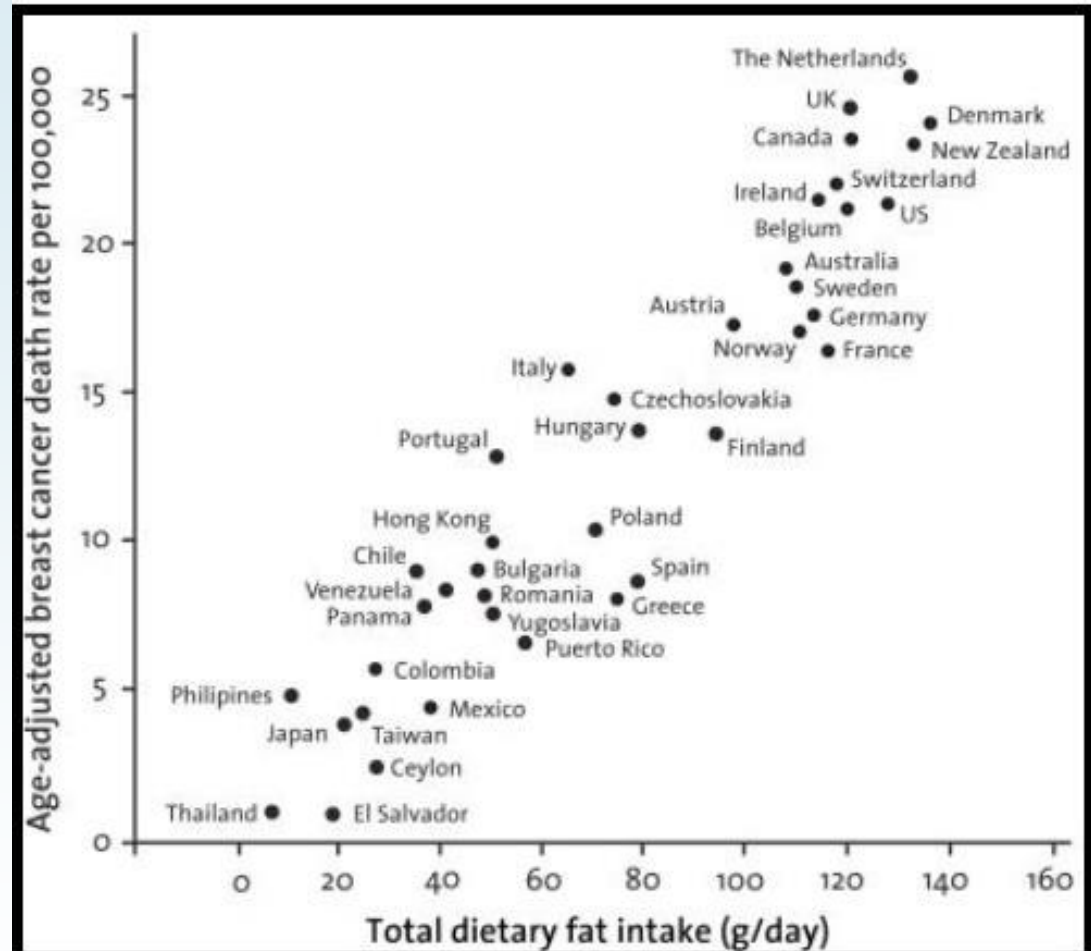
- Sampling
 - **Determined by** outcome and/or exposure
 - Examples of exposure: smoke, physically active, SES
 - Examples of outcome: disease or state of ill health
- Timing
 - Single point in time
 - Retrospective (CAVE how questions worded)
 - Prospective (from now → future)

Observational vs. interventional studies

- ***Observational studies*** -observe the populations or individuals under study
 - descriptive studies
 - ecological studies
 - cross-sectional studies
 - cohort studies
 - case-control studies
- ***Interventional studies*** -where the investigators intervene, e.g. they assign exposure or a health measure to a particular individuals or groups
 - prevention studies
 - randomised clinical trials
 - community interventions

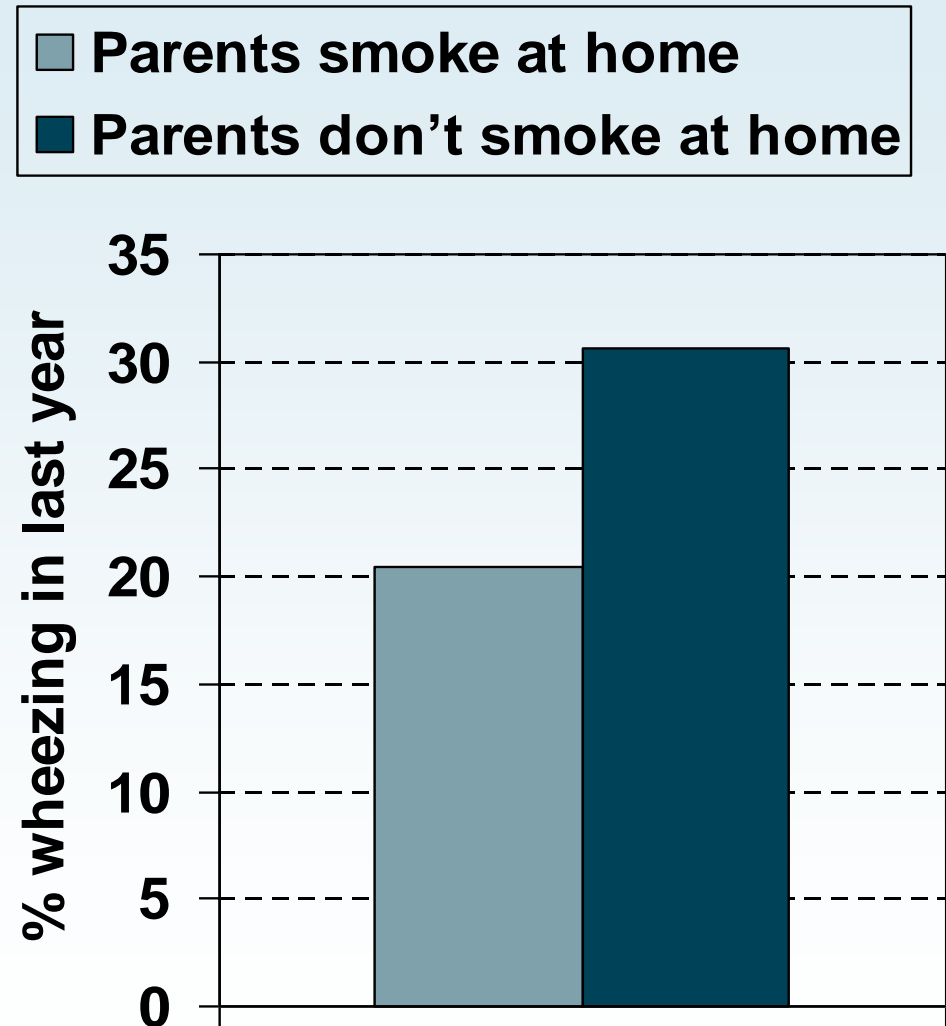
Ecological studies

- Grouped data
- Geographical or time-series
- Cheap & quick
- Useful to generate hypotheses
- **Ecological fallacy**
= it is wrong to extrapolate from groups to individuals



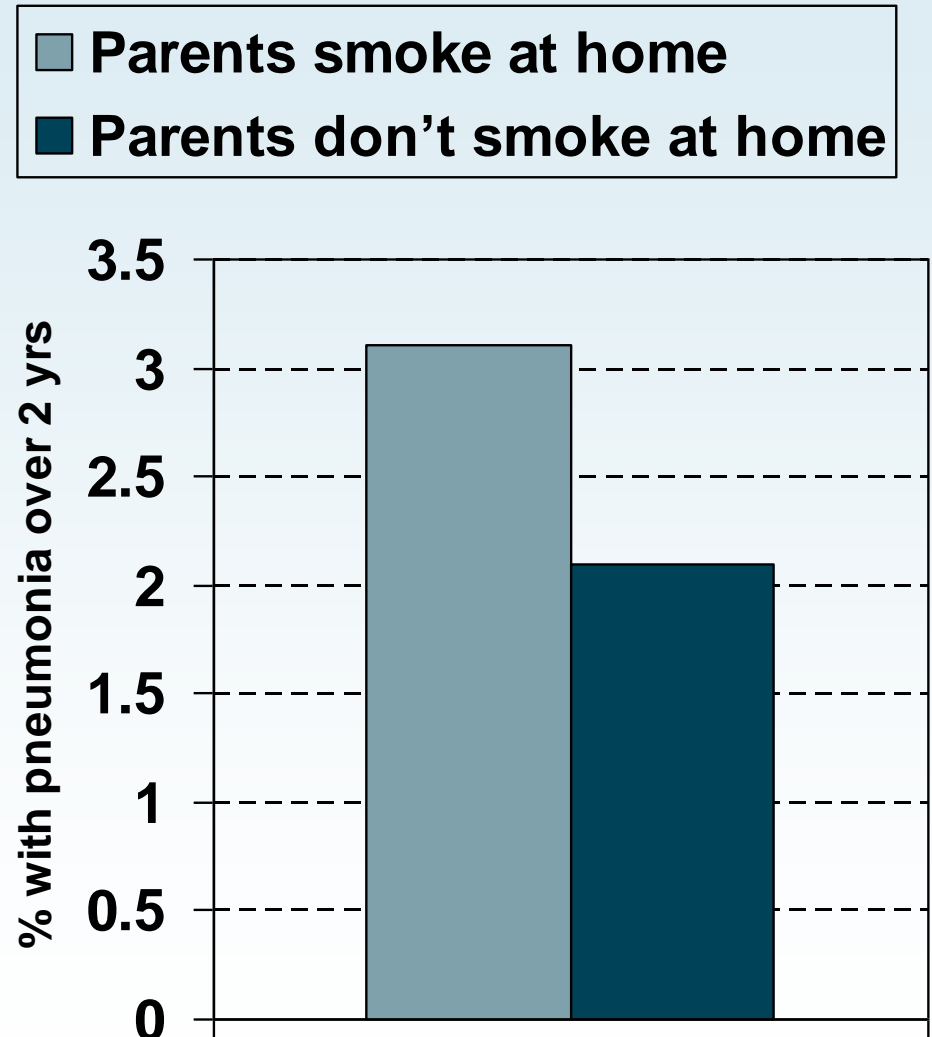
Cross-sectional studies

- All data collected at one point in time
- Prevalence
- Relatively cheap & quick
- Useful to estimate burden of disease
- Difficult to make causal inference



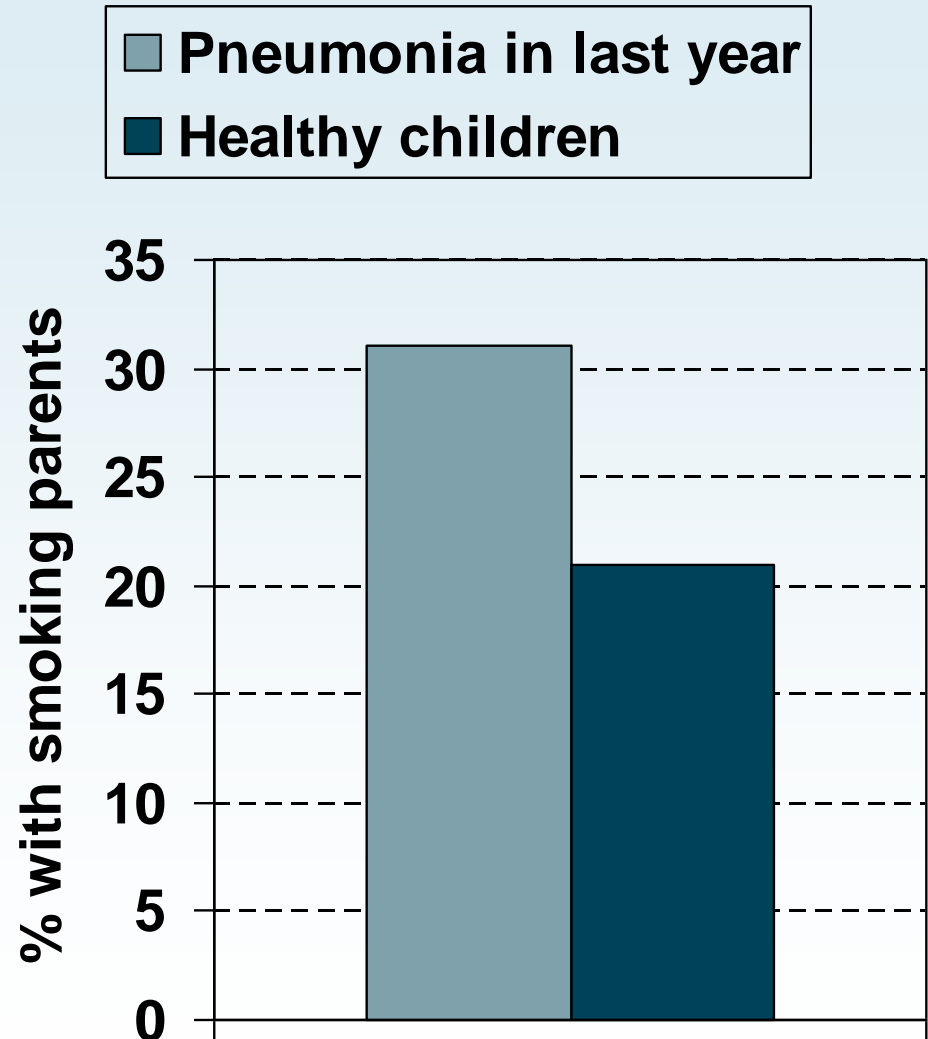
Cohort studies

- Exposure measured in healthy individuals
- Follow up
- Incidence
- Time consuming & expensive
- Temporality clear
- Possibly the “best” observational design



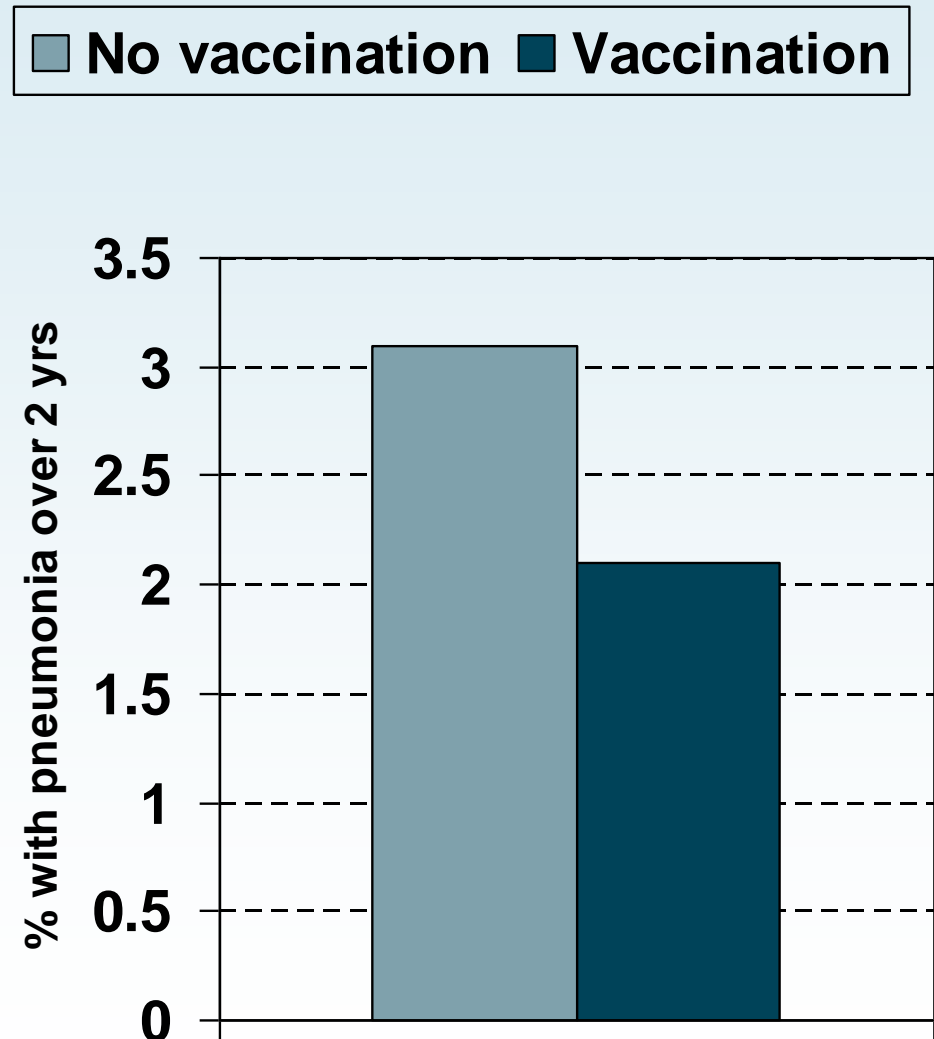
Case-control studies

- Cases vs. controls (current status)
- No follow up
- Asking about exposure in past
- No incidence or prevalence
- No need to wait for cases → quick
- **Temporality may be a problem**
- Good for exposures stable over time



Interventional studies

- Exposure allocated by researchers (often randomly)
- Follow up
- Incidence
- Time consuming & expensive
- Temporality clear, no confounding / bias
- Gold standard



Observational studies

Interventional studies

Data from groups

Data from individuals

Data from groups

Data from individuals

Descriptive

Analytic

Descriptive

Analytic

Community trial

Clinical trial, Individual trial

Ecological study

Cross-sectional

Cross-sectional

Cohort

Case-control

Types of comparisons in different types of studies

Study design	Type of comparison
Ecological studies	Comparing disease frequency between populations
Cross-sectional studies	Comparing disease frequency between persons with and without characteristic of interest IN ONE TIME
Cohort studies	Comparing disease incidence between exposed and unexposed persons IN MORE TIME POINTS
Case-control studies	Comparing frequency of (PAST) exposure between cases and healthy controls
Interventional studies	Comparing incidence of events in persons exposed to the intervention of interest and in control group

Applications of different observational and analytical study designs

	Ecological	Cross sectional	Case control	Cohort
Investigation of rare disease	++++	-	+++++	-
Investigation of rare exposures	++	-	-	+++++
Examining multiple outcomes	+	++	-	+++++
Studying multiple exposures	++	++	++++	+++
Measurement of time relationships between expo and outcome	+	-	+	+++++
Direct measurement of incidence	-	-	+ ¹	+++++
Investigation of long latent period	-	-	+++	+++ ²

¹ incidence only if the sampling fraction known for both cases and controls

² if historical cohort

Summarization

Rates

- What can you name and define?

What types of studies do you know

What means retrospective and prospective?

- What study uses which approach?

Exercise

Study design overview

15 mins

Exposure= independent variable (e.g., smoking)

Outcome= dependent (e.g., lung cancer)

Exercise

1,000 of retired police workers was followed for 25 years. Half of them were regular alcohol drinkers, and there were 20 cases of liver cancer in this group. In the rest of the group, there was found 10 cases of the cancer.

- i) Build up the table and calculate absolute risk for each group
- ii) What is the relative risk/risk ratio among regular drinkers in comparison with others?
- iii) Calculate odds ratio for the same association. What do you think about results ii) and iii) ?

i)

	Cancer	No cancer	Total	Absolute risk
Regular drinkers	20	480	500	0.04
Non regular/non drinkers	10	490	500	0.02
Total	30	970	1000	0.03

ii) **What is the relative risk among regular drinkers in comparison with others?**

$$RR = 0.04 / 0.02 = 2.00$$

Dates of study entry, diagnosis and end of follow up (dropout or death) would be needed to calculate person-years for the denominator.

iii) **Calculate odds ratio for the same association. What do you think about results ii) and iii) ?**

$$OR = \frac{a \times d}{b \times c} = \frac{20 \times 490}{10 \times 480} = \frac{9800}{4800} = 2.04$$

Results in b) and c) are very similar. We have very rare outcome in this calculation and therefore OR and RR are similar and we can say that OR is good approximation of RR

Puerperal fever

Ignaz Semmelweis (1818-1865) began his medical career in 1844 in obstetrics and midwifery at the Vienna General Hospital (Allgemeines Krankenhaus). There were two obstetric divisions in the hospital: patients in the first division were examined by doctors and medical students, while midwives attended to the patients in the second division. Semmelweis noticed that there were more maternal deaths in the first division than the second division.

In this exercise you will follow Semmelweis' steps investigating the problem.

a. Calculate the total and year specific mortality rate for the 6-year period (1841-6) in the first and second divisions (fill the empty cells in the table above).

Year	First division			Second division		
	Births	Deaths	Mortality rate	Births	Deaths	Mortality rate
1841	3036	237	0.08	2442	86	0.04
1842	3287	518	0.16	2659	202	0.08
1843	3060	274	0.09	2739	169	0.06
1844	3157	260	0.08	2956	68	0.02
1845	3492	241	0.07	3241	66	0.02
1846	4010	459	0.11	3754	105	0.03
TOTAL	20042	1989	0.1	17791	696	0.04

b. Do you agree with Semmelweis' claim that there were more deaths in the first division?

c. Is it necessary to calculate the mortality rates for each year in order to compare the two divisions?

Year	Births	Deaths	Mortality rate
Jan-April 1846	1193	194	0.16
May-Aug 1846	1039	140	0.13
Sep- Dec 1846	1120	125	0.11
Jan-Apr 1847	1240	84	0.07
TOTAL	4592	543	0.12
INTERVENTION			
May-Aug 1847	1076	50	0.05
Sep-Dec 1847	1059	42	0.04
Jan-Apr 1848	1155	14	0.01
May-Aug 1848	1107	7	0.006
TOTAL	4397	113	0.03

Was Semmelweis' intervention successful?

Briefly comment on the importance and implications of this finding in terms of epidemiology and clinical practice.

We have study, we have basic results from analysis...

.....we must know how to interpret findings

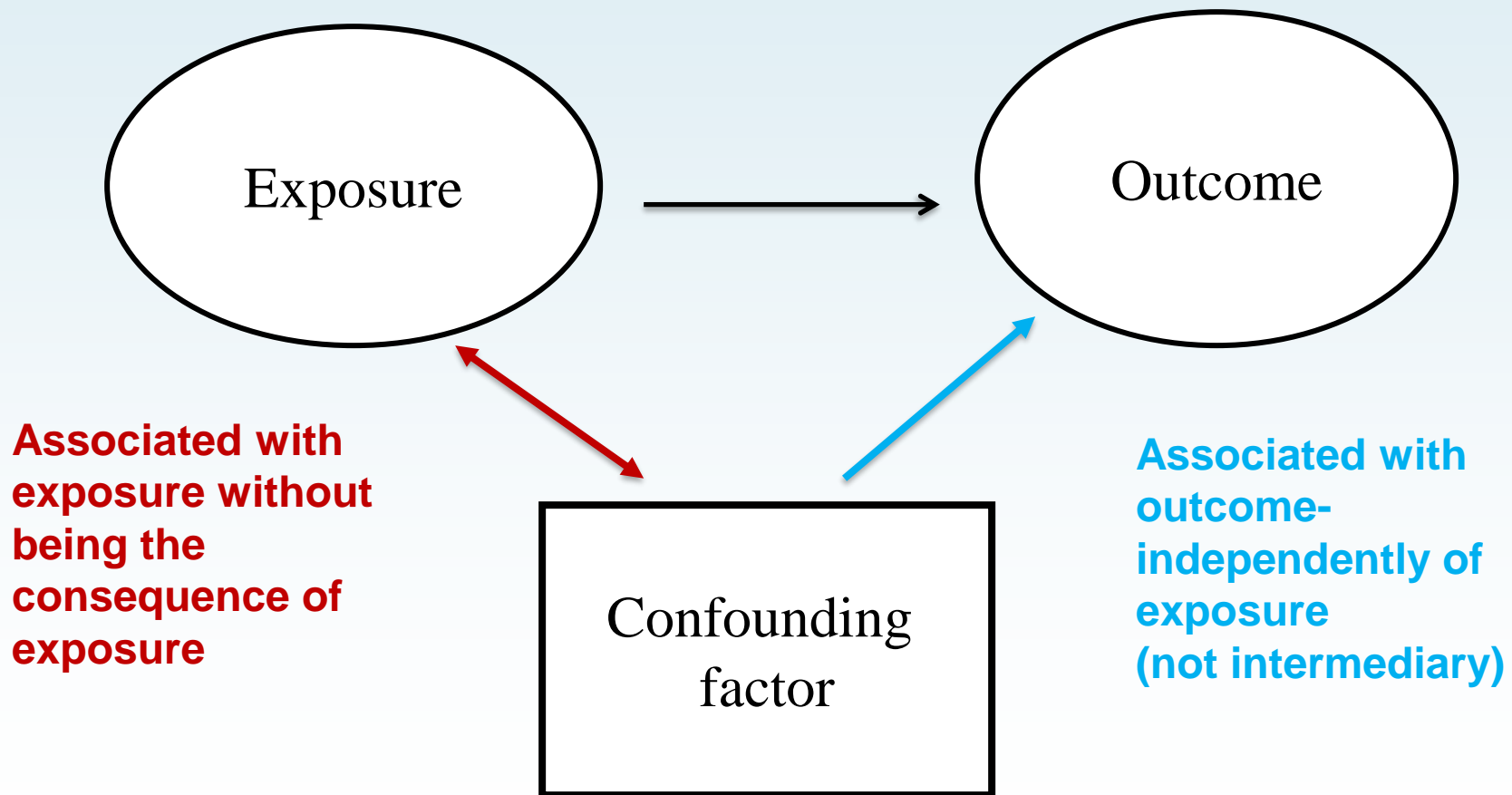
Three major issues in interpretation of any epidemiological study

- Chance (random variation) – statistics
 - Bias (i.e. systematic error)
 - Confounding
-
- Only if all of these have been excluded, you may start thinking of a causal association

Confounding

- Situation when a third factor is associated with both exposure and disease
- Association between “exposure” and disease may not be causal; instead, it is due to a third factor which is associated with both exposure and disease.

Confounding



EXAMPLE

Case-control study of alcohol and lung cancer

	Alcohol	No alcohol
Cancer	450	300
No cancer	200	250

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

Estimated odds ratio = 1.9

The same data stratified by smoking:

	Non-smokers		Smokers	
	<u>Alcohol</u>	<u>No alcohol</u>	<u>Alcohol</u>	<u>No alcohol</u>
Cancer	50	100	400	200
No cancer	100	200	100	50
Estimated odds ratio	1.0		1.0	

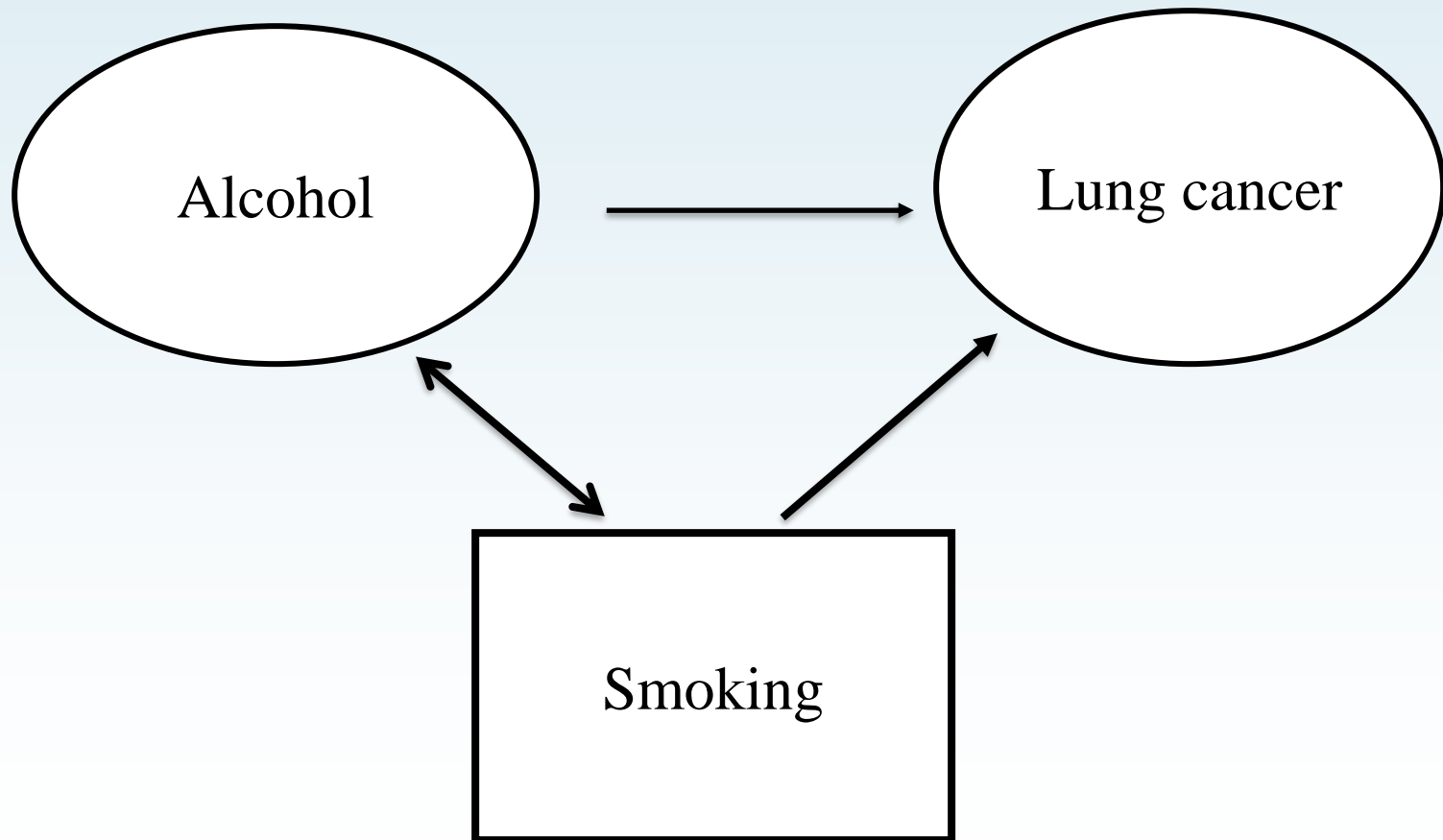
Alcohol and smoking in controls (=no cancer)

	Alcohol	No alcohol
Smokers	100	50
Non-smokers	100	200

Non-drinkers: 1 in 5 were smokers.....50 from 250

Drinkers: 1 in 2 were smokers.....100 from 200

Confounding



Most common confounders:

- Gender (men have higher mortality and more risk factors; women higher morbidity)
- Age (risk of most diseases increases with age)
- Socioeconomic status (risk of most diseases higher in lower SE groups)
- Ethnic group
- Smoking
- Alcohol
- etc...

Control of confounding

Design

- Randomisation
- Restriction
- Matching

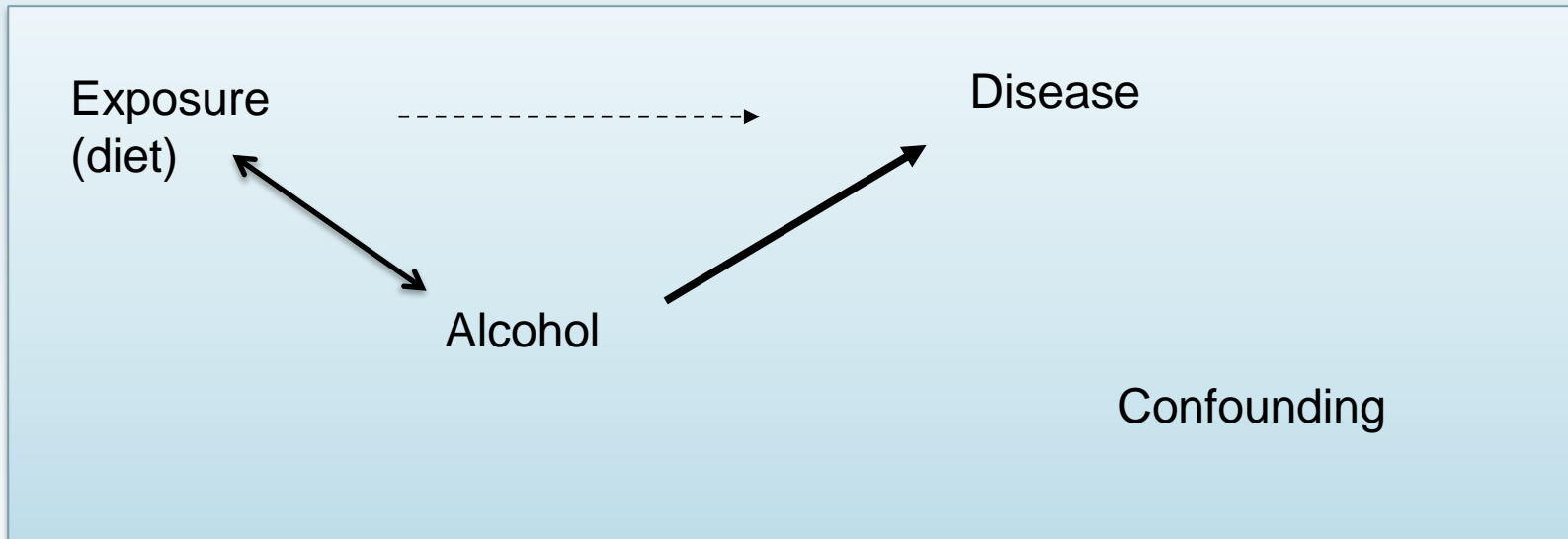
Analysis (if data collected)

- Stratification
- Regression modelling

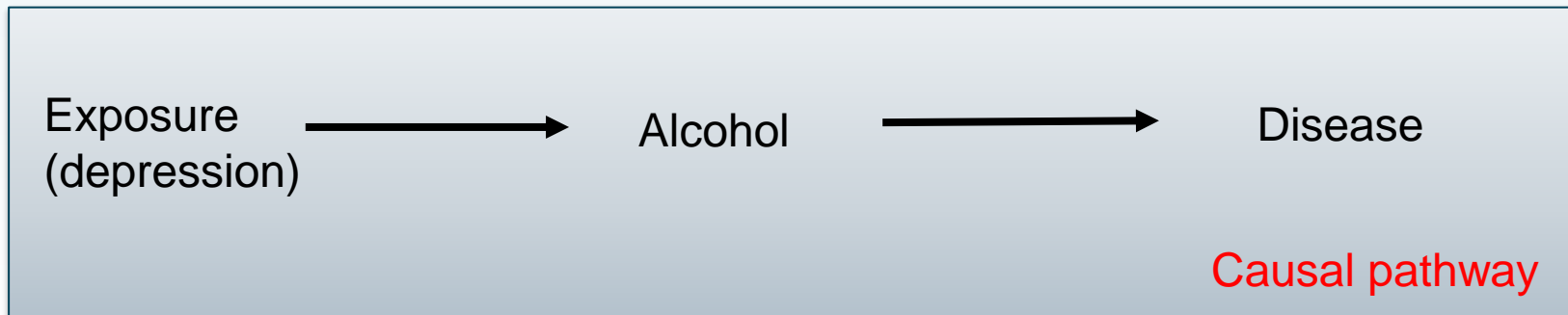
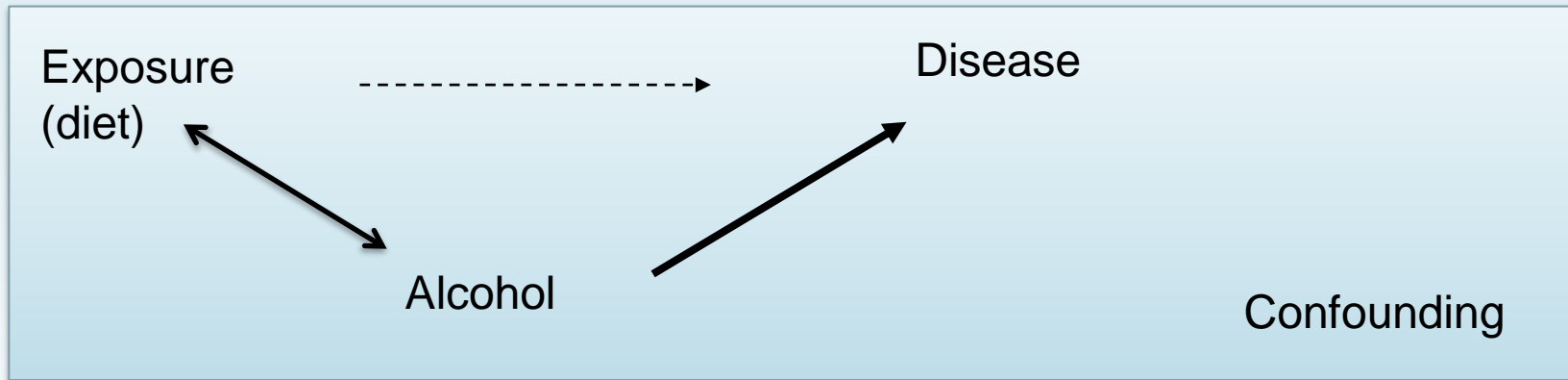
Residual confounding

- Unmeasured confounding factors or measurement error in confounding factors may lead to residual confounding.
- The possibility of residual confounding cannot be completely eliminated in observational studies

Confounding vs. causal pathway



Confounding vs. causal pathway



Statistically, confounding is the same as causal pathway
The difference is conceptual – i.e. in your head!

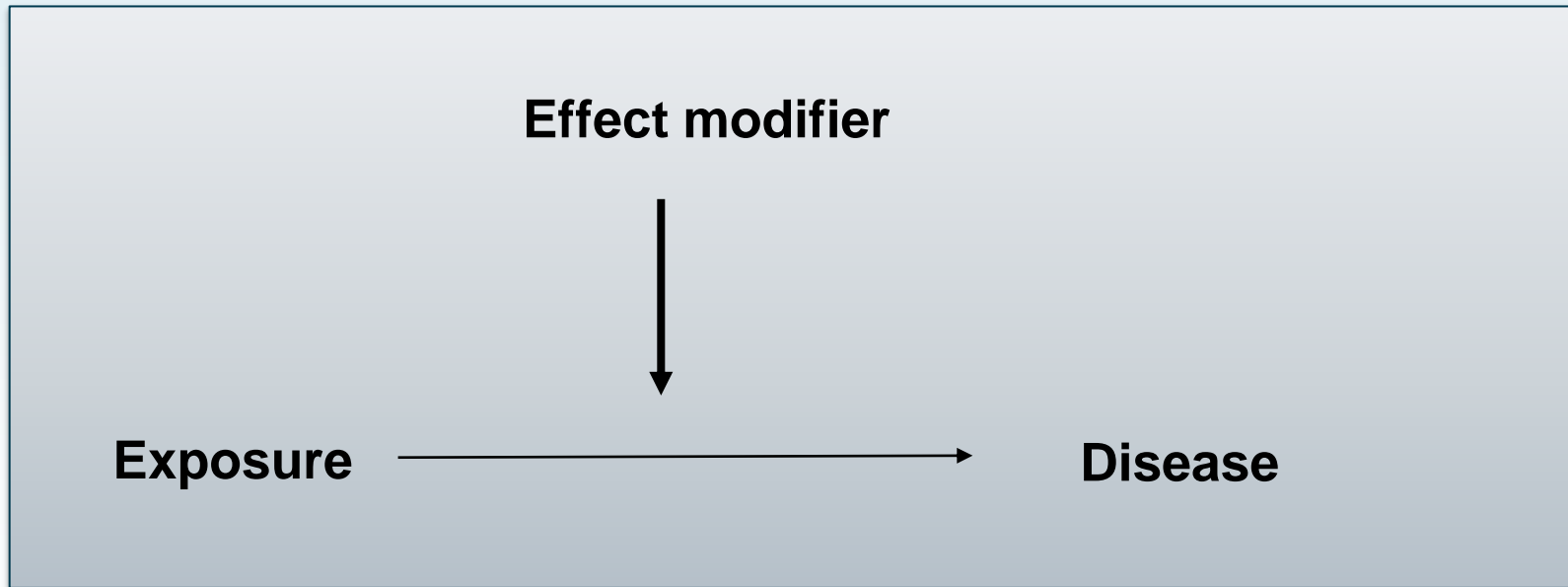
Effect modification (interaction)

- the effect of exposure on disease is dependent on the level of a third factor

or

- a moderator specifies on whom or under what conditions another variable (exposure) will operate to produce the disease.

Effect modification



Finding out the different influence in different strata

=exploring association between exposure (independent variable) and outcome (dependent variable) within different strata of the 3rd characteristic

age groups

sex

achieved education level

geographical area

Example from yesterday

Death rates from CHD in smokers and non-smokers by age

Age	Smokers rate	Non-smokers rate	Rate ratio
35-44	0.61	0.11	5.5
45-54	2.40	1.12	2.1
55-64	7.20	4.90	1.5
65-74	14.69	10.83	1.4
75-84	19.18	21.20	0.9
85+	35.93	32.66	1.1
ALL AGES	4.29	3.30	1.3

The rate ratio decreases with increasing age.

It may suggest that the effect of smoking on the rate of CHD is higher in younger ages.

EXAMPLE

CHD, smoking and age in British doctors study (rates per 100,000)

	Non-smokers		Heavy smokers	
	Rate	RR	Rate	RR
<45	7	1.0	104	14.9
45-54	118	1.0	393	3.3
55-64	531	1.0	1025	1.9

Positive and negative effect modification

- **Positive:**
 - “susceptibility factor” or “vulnerability factor”,
 - its presence (or higher values) strengthens the association between exposure and disease.
- **Negative:**
 - “resiliency factor” or “buffering factor”
 - its presence (or higher values) weakens the association between exposure and disease

Identification of effect modification

- **Stratified analysis**
- Compare effect estimates in strata
- Assess differences in effects by significance tests (p-value for heterogeneity)
- Pooled estimates (e.g. standardised) **not appropriate** when there is an interaction
- *Please note that genuine & meaningful interactions are rare*

Confounding vs. interaction

Confounding

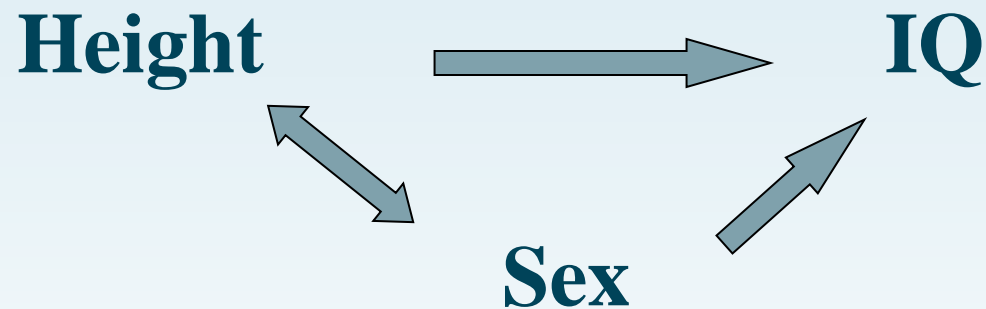
- Alternative explanation
- Distorts the “truth”
- Efforts to remove it to get nearer to the “truth”
- When present, stratum specific effects are similar to each other but different from the overall crude effect.

Effect modification

- One factor modifies effect of another factor
- It is genuine, not artefact
- Property of the relationship between factors
- We should detect and describe it but not remove it.

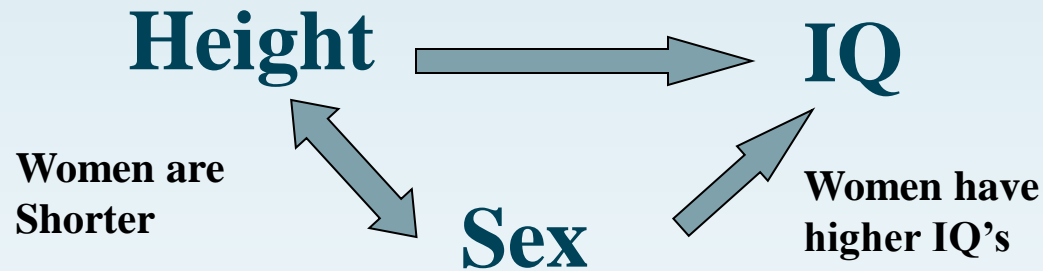
Example:

Height and IQ – real association or not?



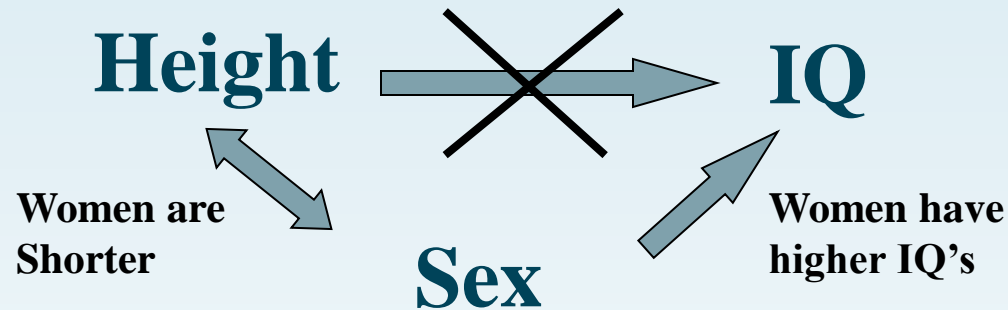
=High negative association between height and IQ

Height and IQ



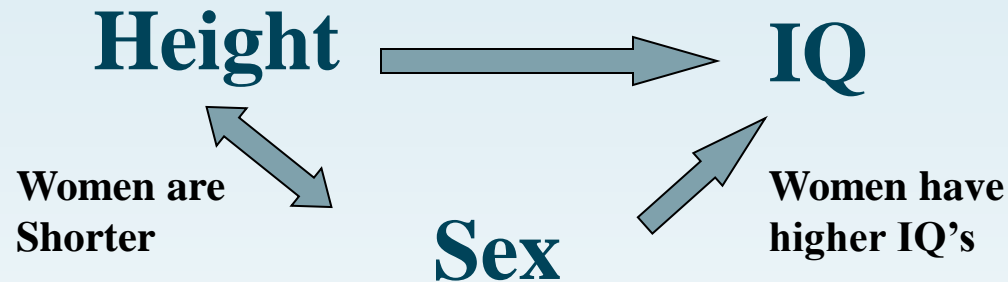
- Find out that Sex is related to Height and that Sex is related to IQ
- Therefore, Sex is a ***potential*** confounder

Height and IQ



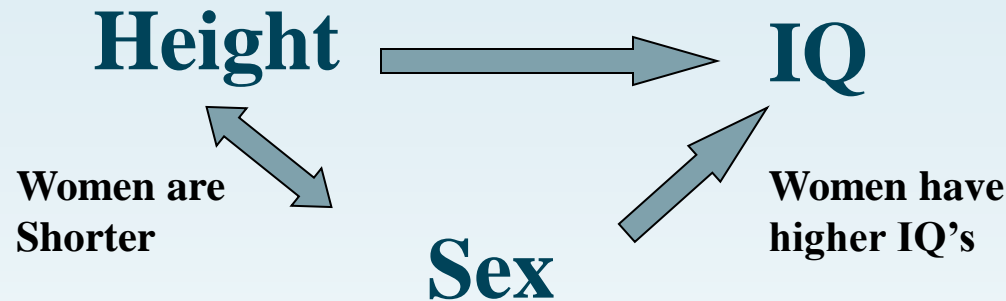
If after adjustment for Sex there is
NO association between height and IQ,
then Sex was a confounder

Height and IQ



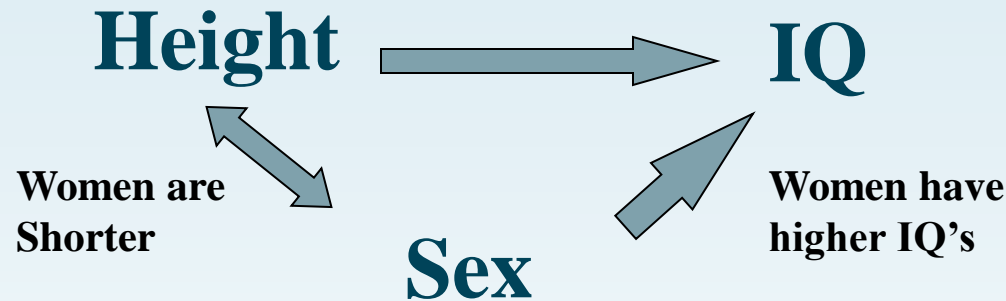
If after adjustment for Sex there is still a strong negative association between Height and IQ, then Sex is not a confounder

Height and IQ



If, after adjustment for Sex, there is still an association between Height and IQ, but the nature and/or strength of the association changes with Sex (=different for M and F), then Sex is an Effect Modifier.

Height and IQ



If there is no association between Sex and IQ, then

Sex cannot be a confounder

Likewise, if there is no association between Sex and height, then **Sex cannot be a confounder**

The confounder must be related to both Exposure and Outcome

Three main categories of alternative explanation

- Chance - random error
- **Bias** - systematic error
- Confounding – third factor explaining an association

Bias

- is a systematic error in the design of an epidemiological study which leads to a distortion or error in the study results
- an association will allow to be distorted if error is differential

Validity

- A study's results and conclusions are **valid** when they reflect the true relationship in the study population
- To assess the validity of findings we need to consider **alternative explanations** for the observed associations

Bias can affect

- Estimate of one variable
- Estimate of association between variables

Errors (biases) may be

a)

- Non-differential vs. differential
 - error in one variable not related to / dependent on the value of other variables
 - error in one variable is related to value of other variable

Example: sex differences in HDL-cholesterol

non-differential – badly calibrated measurement of HDL-cholesterol does not bias estimate of mean sex difference (the error cancels out)

differential - measurement of HDL-cholesterol in different single sex studies using different labs: biases estimate of mean sex difference – unless labs carefully calibrated against an external standard.

E.g. cases and controls analysed in different labs!

Errors (biases) may be

b)

- Selective vs. informative
 - Related to selecting subjects into study
 - Related to collecting information

Selection bias

- due to errors in the way sample is recruited
- a distortion that results from procedures used to select subjects or their participation
- resulting in a difference in the characteristics between those who are included in the study and those in study population but not included in the study sample

- The study sample
 - representative or random sample better than volunteers
 - high response rate (>70%)
- Follow-up participation in longitudinal study
- Item non-response

If non-response is related to the exposure and/or outcome, then the study may produce invalid findings

e.g. sick smokers may refuse to participate more often than healthy smokers

Particular concern in case-control studies because **exposure and disease are both present** at time of recruitment

Hospital-based studies are problematic because cases are filtered: not all cases go to hospital, not all cases get the correct diagnosis

e.g. a hospital-based study of depression will involve severe cases only

Information bias

- due to errors in way in which information collected from the sample
- errors in the way information about exposure or disease collected

=> misclassification - putting subjects in wrong category
inaccurate estimates of occurrence of effect size, or even direction of association

e.g., exposed as unexposed, case as control

Important types of information bias include

- **Reporting/recall bias**: by study participants
- **Observer bias**: in measurements by research personnel
- **Diagnostic bias**: probability of detection or correct identification of disease across study groups or over time

Misclassification may be

- **Random** - above / below
- **Systematic** – all in one direction
- **Non-differential** (error in one variable not related to / dependent on the value of other variables)
- **Differential** (error in one variable is related to value of other variable)

Non-differential misclassification:

- Tend to bias estimates towards null
- Cholesterol machine giving random readings
- Underestimated effect traditionally seen as less of problem than overestimate

Differential misclassification

- Can distort associations, and can produce spurious associations

i) Reporting bias

- May underestimate some behaviours eg alcohol, smoking
- In CS or CC studies when exposure & disease assessed at same time – bigger problem
- eg depression and poor physical health
- Often not conscious – placebo effect

ii) Recall bias

- Particular problem in case control studies or as part of retrospective part of longitudinal study
- Case may have better recall of exposure
- Eg., mothers of babies with congenital abnormality
- Diarrhoeal illness and food consumption

iii) Observer bias

- investigator classifies exposure differently in cases / control

or

- the investigator diagnoses disease differently in exposed / unexposed participants

=> the results are distorted

iv) Interviewer bias

- Interviewer may probe cases more closely for exposure
- May look for endpoint more carefully in those exposed to disease

=> **Study must be blinded**

v) Detection bias

- Differences may occur in accessing medical care
- Differences in diagnostic criteria
- These differences may be associated with exposure eg social class / country
- Hence detail paid to ascertainment and validation of endpoints

What can we do to prevent / reduce bias?

Selection bias

- random sampling from study population
- strategies to reduce non –response eg repeat mailings, offering different times at clinic
- proper choice of control group in case-control studies

Recall / reporting bias

- recall bias : try to obtain objective information on past exposures wherever possible or use proxy informants
- reporting bias – include lots of different questions so that subjects are hypothesis blind
- trials should be controlled and blinded

Observer bias

- investigators blind to case / exposure status wherever possible
- use standardised instruments and protocols, back translations
- ideally use centralised measurement or calibrate instrument
- periodic check on staff to check for differences in procedures

Detection / diagnostic bias

- aim for population - based ascertainment of cases
- follow 'Standardised diagnostic criteria'

Assessment of bias

- Non-responders questionnaire
- Baseline characteristics of those lost to follow can be analysed and compared to those remaining in study
- Objective validation of self-reported information
- Sensitivity analyses to estimate effect of bias

Bias: the silent menace

- Cannot be assessed numerically
- No software to identify bias
- If there is flaw in the design of the study increasing numbers will not get rid of it !
- Can only be assessed by careful evaluation of the design

Publication bias

High-impact journals prefer clear, positive results! 😞

Bias in systematic reviews

Form of selection bias arising if null studies are not published
If not included the overall estimate is biased upwards.

Minimised by searching grey literature, trial registers and
conference proceedings to include null/negative results

e.g. the 'drug effectiveness cycle' (β -blocker-mortality
example), selective serotonin reuptake inhibitors in treating
depression

Publication bias

Failure to publish

- a negative or inconclusive trial result
- a small trial may be abandoned

Duplicate publication

- a large treatment effect
- need for research output

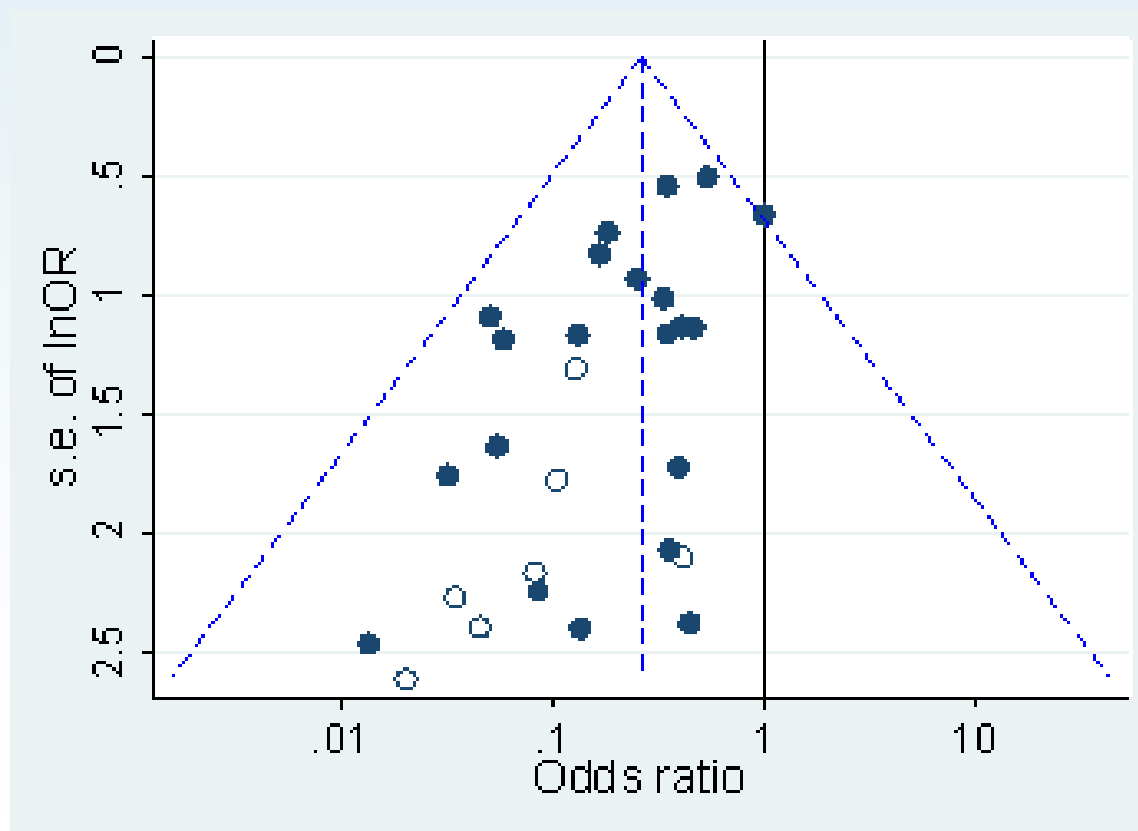
E.g. nine trials of ondansetron (antiemetic) in 23 (!) publications
(Tramer et al BMJ 1997)

How to avoid publication bias

- To make sure studies are not double counted
- To search for unpublished studies (e.g. contact researchers directly)
- To use non-English language publications
- Statistical checking (funnel plots: smaller studies report more extreme results)
- Registration of studies and to make sure all results are in public domain (not yet fully achieved)
- Trial registration: assigns unique trial identification numbers, and to record other basic information about the trial so that essential details are made publicly available
- From 2004 International Committee of Medical Journal Editors (ICMJE) would consider trials for publication only if they had been registered before the enrolment of the first participant.

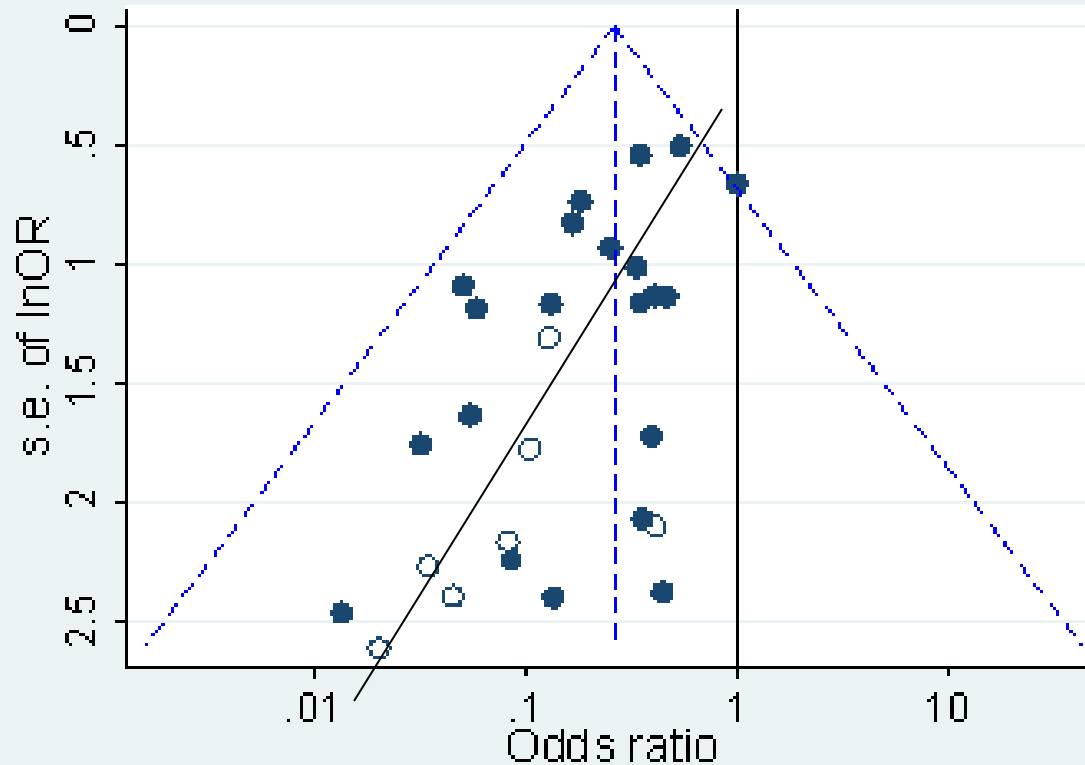
Funnel plot:

asymmetrical plot in the presence of bias: some smaller studies (open circles) are of lower methodological quality and therefore produce exaggerated effect estimates

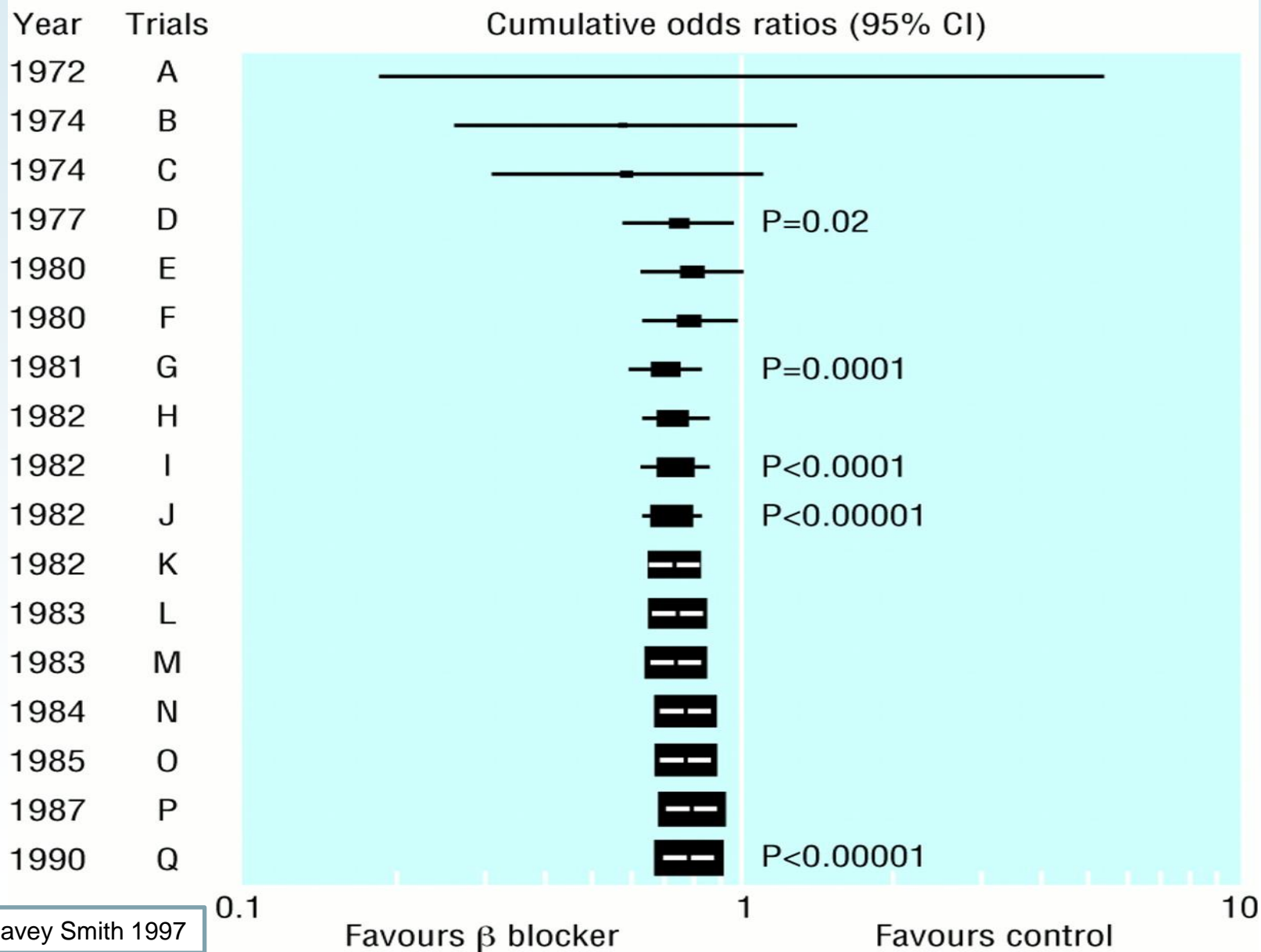


Funnel plot:

asymmetrical plot in the presence of bias: some smaller studies (open circles) are of lower methodological quality and therefore produce exaggerated effect estimates



Beta-blockers and total mortality after MI: meta-analysis



Conclusions:

- All studies are imperfect
- Most studies are subject to measurement error and various biases
- The question is: **are the results valid enough for my purpose?**

Three major issues in interpretation of any epidemiological study

- Chance (random variation) – statistics
 - Bias (i.e. systematic error)
 - Confounding
-
- Only if all of these have been excluded, you may start thinking of a causal association

Causality

1/ we find an association between exposure and outcome

2/ we need to ask whether the association is causal
= does the exposure cause the outcome?

What is a cause?

Rothman (1986):

An event, condition, or characteristic that plays an essential role in producing an occurrence of the disease. Source - Modern Epidemiology.

- Something that has an effect
- Alters disease frequency or health status

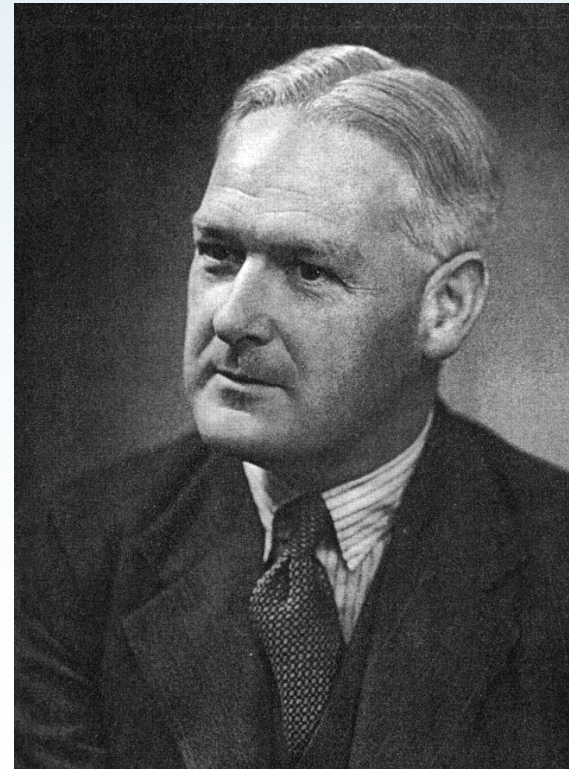
Association versus Causation

- Epidemiological research aims to discover aetiology of disease
- Epidemiology is the study of the association between a potential cause (risk factor/determinant) and a specific disease (outcome).
- Presence of a valid statistical association does not imply causality
- Association is not the same as causation
- Goes beyond association
- How do we decide whether a given association is causal or not?

Sir Austin BRADFORD HILL (1897-1991)

“Exposure and Disease: Association or Causation?”

1. Strength
2. Consistency
3. Specificity
4. Temporality
5. Dose-response
6. Biological plausibility
7. Coherence
8. Reversibility
9. Analogy



Guidelines for inferring causation

- The Bradford-Hill criteria of causation
(J Royal Soc Med 1965; 58: 295-300)

Strength of association

- Measured by RR, OR
- Strong association is less likely to be due to undetected confounding or bias
- Weak association may be causal
 - Measurement error dilutes association

Consistency of association

- Association observed in several different studies with different study designs and populations
- Less likely that same biases present in all of them
- Inconsistency between populations may reflect lack of association or differences in the prevalence of other causal complements

Specificity of association

- Occurs when a single factor is associated with a single outcome
- Increasingly irrelevant to current models of disease causation (single factor many outcomes)

Example

- asbestos and mesothelioma – shown
- HIV and AIDS – shown
- Low lead exposure and IQ – not clear. IQ is not a definable brain condition so there is the potential for confounding e.g. SES

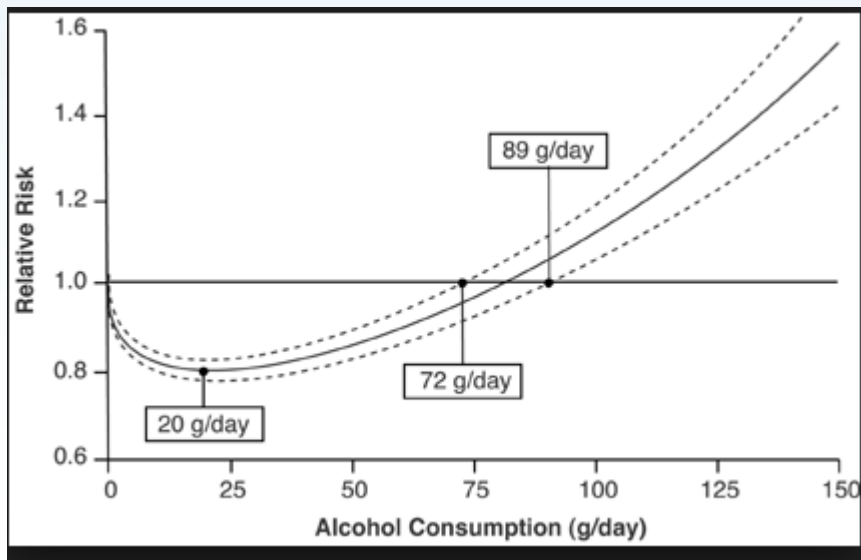
Temporal sequence of association

- The exposure must precede outcome
- Optimal study designs = randomised intervention study or prospective cohort study
- Weak designs for temporality: cross-sectional, case-control study
- Reverse causality may be problem in cohort or case-control study

Biological gradient (=dose response)

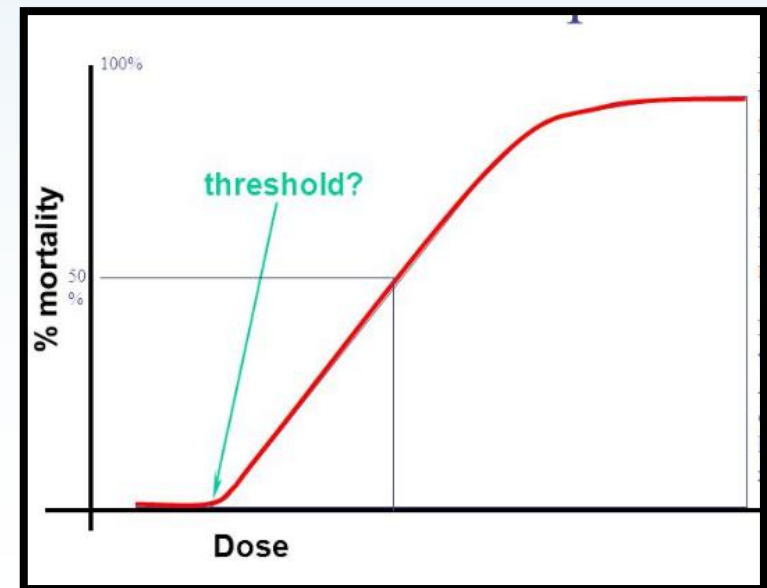
- Observe an increase in the magnitude of risk of outcome with magnitude of exposure
- Unlikely to be explained by bias or undetected confounding
- Lack of a biological gradient does not rule out causality

J- or U- shaped relationships



Source: pubs.niaaa.nih.gov

Threshold effect



Example to previous

Persons who have increasingly higher exposure levels have increasingly higher risks of disease

Smoking Status	Lung Cancer risk
None	1.0
Ex-smoker	1.1 (0.7-1.6)
1-20 per day	2.6 (1.7-4.0)
20-40 per day	4.4 (2.8-6.9)
40+ per day	6.8 (4.3-10.7)

Plausibility of association

- Practically we may accept a possible causal association even when there is no plausible mechanism or explanation
- Acceptance depends on how “unlikely association is”
- Reported association may stimulate search for mechanism

Example

- Cigarettes & lung cancer. Carcinogenic substances in cigarettes
- Low fibre diet & colon cancer. Dietary fibre increases intestinal motility and dilutes/absorbs fecal carcinogens

Coherence of association

- Reported association does not conflict with current knowledge
- Can lead to publication bias
- Can discourage search for alternative associations

Example

- Serum cholesterol lowering effect on heart attack, regardless of the means – diet or drug

Experiment (reversibility)

- Removal of exposure leads to a reduction in the risk of the outcome
- Currently perceived as the strongest type of evidence
- May be difficult to ascertain in diseases with long lag times between exposure and disease

Analogy

- Other similar demonstrated associations
- In practice may be limited by current knowledge

Bradford Hill Closing Remarks (1965)

“I do not believe ... that we can usefully lay down some hard-and-fast rules of evidence that must be observed before we accept cause and effect.

None ... can bring indisputable evidence for or against the cause and-effect hypothesis and none can be required ...

What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question - is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?”

Causal Inference

- Not just ticking boxes
- Weigh evidence of causal association against other explanations
- Understanding, judgement & interpretation are crucial
- Cannot prove a causal association
- Can only be inferred based on evidence
- May change in the light of new evidence



Reverse causality

Refers to the possibility that the link between exposure and outcome is a result of the disease or disease process being studied, not the exposure

Reverse causality is a type of confounding in the sense that it is 'real' and not an artefact of study design. It is relevant in some situations but not others

Example of potential reverse causality

Researchers are interested in the link between **blood levels of inflammatory markers** and **later CVD**

There are 4 possible explanations

1. Inflammation \rightarrow atherosclerosis (causal association)
2. Atherosclerosis \rightarrow inflammation (reverse causal association)
3. Inflammation \leftrightarrow atherosclerosis (association is bi-directional)
4. Other processes lead both to atherosclerosis and inflammation (confounding) e.g. diet

Public health policy

- Ideally based on ‘evidence’ - meta-analyses and systematic reviews
- Considerations of efficiency, cost-effectiveness and harm
- Eradication of poverty for improving health?
- Reduction in social inequality for reducing health inequality?

Causation and public health

- There is moment when action may be taken – it may vary from introduction of a new drug to advice to public on certain practice, or new legislation being introduced
- Complex process taking into account costs, benefits and harms
- Even when evidence become overwhelming, governments may be slow to act

Summary

- Epidemiology = the study of the distribution and determinants of disease in population
- Types of epidemiological studies =
 interventional, observational studies
- Measures of disease occurrence
- Bias, confounding, chance
- Causality