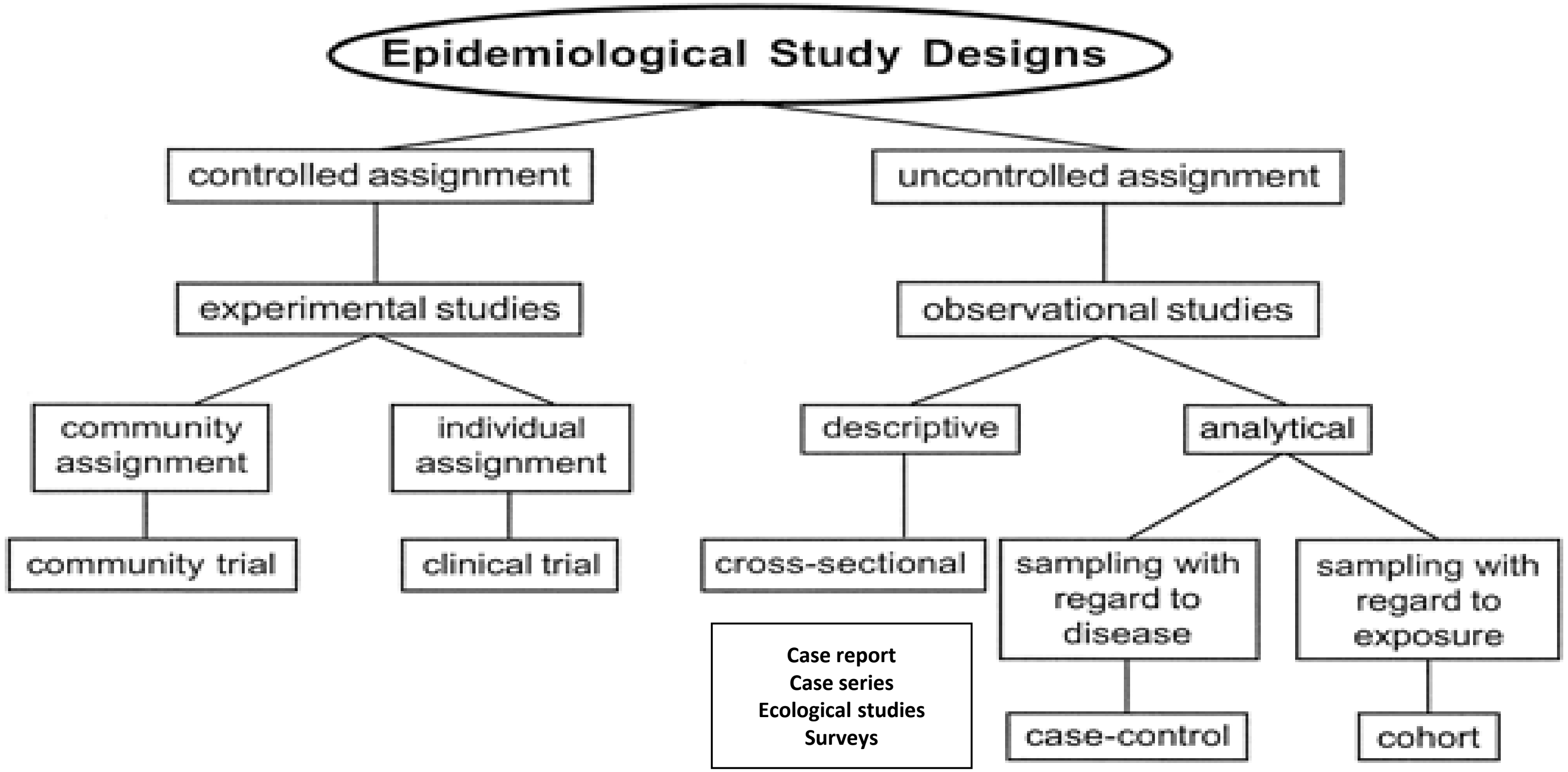
Overview of study design part 1: Analytical studies (cohort studies)

Dr Munir Abu-Helalah

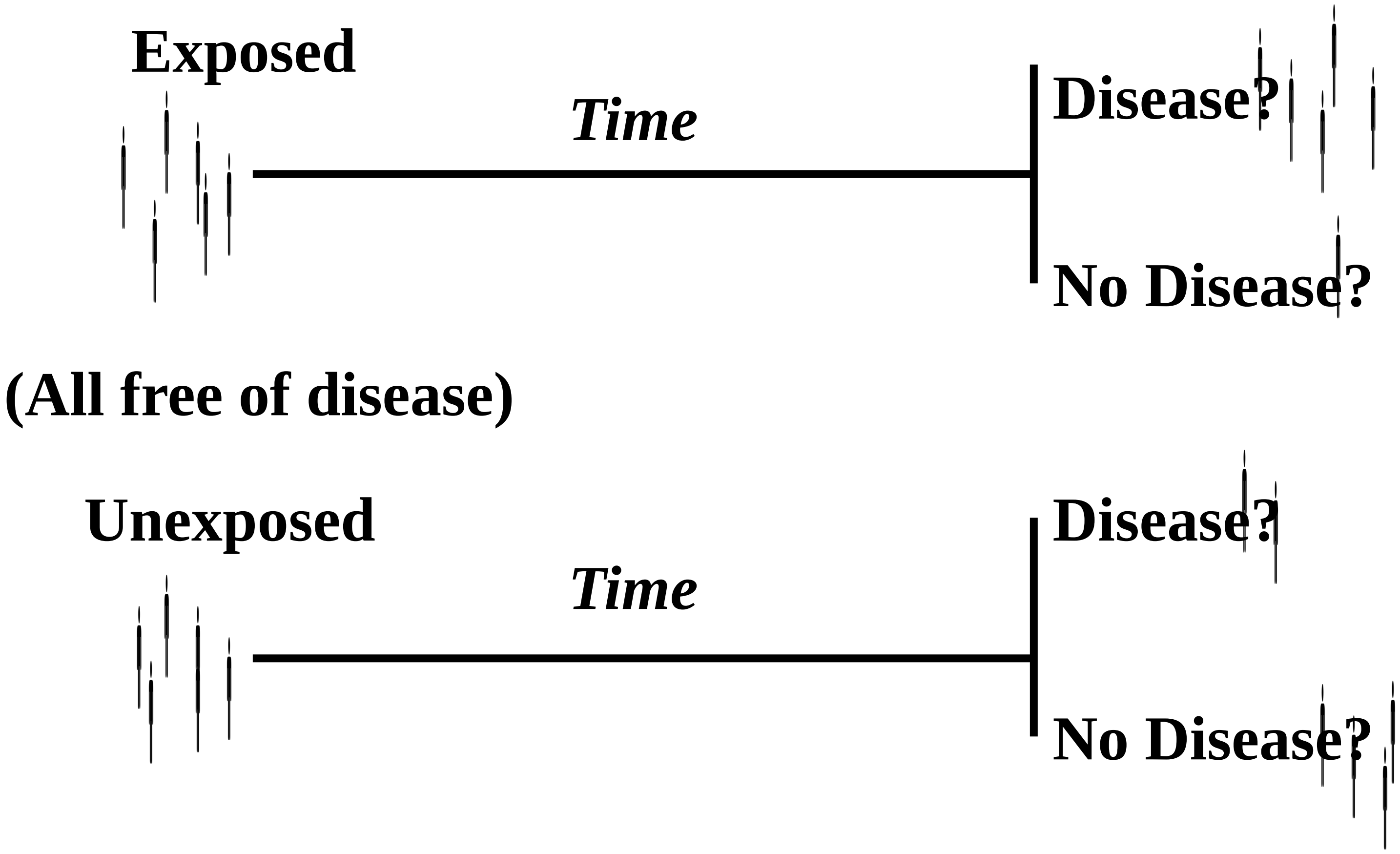
MD,MPH,PHD

Associate Professor of Epidemiology and Preventive Medicine



Source: Waning B, Montagne M: *Pharmacoepidemiology: Principles and Practice*: http://www.accesspharmacy.com

Cohort studies



Exposed

Time

Disease?

No Disease?

(All free of disease)

Unexposed

Time

Disease?

No Disease?

Cohort (or follow-up) studies

- **Are studies in which people are identified and grouped with respect to whether or not they have been exposed to a specific factor.**

- **The groups are followed up over time to determine whether the incidence of a particular disease is any greater (or less) in the exposed group than in the non-exposed group.**

- **The starting point is the risk factor!**

Cohort study examples:

- Life expectancy of cerebral palsy children
- Fine needle breast biopsy and breast cancer
- Aspirin intake and colorectal cancer

Cohort study:
Primary purposes

- **Descriptive (measures of frequency)**

– To describe the incidence rates of an outcome over time, or to describe the natural history of disease

- **Analytic (measures of association)**

– To analyze associations between the rates of the outcomes and risk factors or predictive factors

COHORT STUDY DESIGN

- This design is the best observational one for establishing cause–effect relationships.
- Prevention and intervention measures can be tested and affirmed or rejected.

- Cohort studies consider seasonal variation, fluctuations, or other changes over a longer period.

- Objective measures of exposure, such as biological markers, are preferred over subjective measures.

COHORT STUDY DESIGN
Strengths

- **We can measure incidence of disease in exposed and unexposed groups**
- **Can get a temporal (time related) sequence between exposure and outcome as all individuals must be free of disease at the beginning of the study.**
- **Good for looking at effects of rare exposures.**
- **Allows for examination of multiple effects/diseases of a single exposure.**
- **Not open to bias as much as other types of study**
- **Direct calculation of the risk ratio or relative risk is possible.**
- **Provide information on multiple exposures**

COHORT STUDY DESIGN

Limitations:

- Not efficient for rare diseases
- Can be expensive and time-cosuming
- Large sample
- Drop-out biases

  If study goes over many years, can get considerable loss to follow up. This can 'dilute' results or lead to bias, and therefore the validity of result can be seriously affected

- Locating subjects, developing tracking systems, and setting up examination and testing processes can be difficult.
- Changes over time in diagnostic methods, exposures, or study population may lead to biased results.

Cohort study: Example

Hypertension as a risk factor for spontaneous intracerebral hemorrhage

In study risk factors, we start with what is rare!

- Rare disease: we conduct case control study starting with cases

- Rare risk factor: we conduct a cohort study starting with rare risk factors

Physical Activity and Incident Cognitive Impairment in Elderly Persons

ARCH INTERN MED/VOL 170 (NO. 2), JAN 25, 2010

Background: Data regarding the relationship between physical activity and cognitive impairment are limited and controversial. We examined whether physical activity is associated with incident cognitive impairment during follow-up.

Methods: As part of a community-based prospective cohort study in southern Bavaria, Germany, 3903 participants older than 55 years were enrolled between 2001 and 2003 and followed up for 2 years. Physical activity (classified as no activity, moderate activity [<3 times/wk], and high activity [≥3 times/wk]), cognitive function (assessed by the 6-Item Cognitive Impairment Test), and potential confounders were evaluated. The main outcome measure was incident cognitive impairment after 2 years of follow-up.

Cohort study

Physical activity	**Cognitive impairment**		Total
	Yes	No	
Moderate	10	990	1000
None	100	900	1000
Total	110	1880	2000

Risk of outcome in exposed (not active) = 100/1000 = 10%

Risk of outcome in non-exposed (active)=10/1000 =1%

Relative risk 10%/1%=10 =

Measurement of risk

$$Risk\,(R) = \frac{\text{No of people becoming ill during the period of observation}}{\text{No of people exposed at the beginning of the period}}$$

It is proportion (0 - 1)

Hazards and the risks

- Hazards and the risks associated with them are everywhere, but when known measures can be taken to minimise or eliminate risk. When we go up or down stairs it is possible that we might fall, but the likelihood is that we will not.
- Stairs are a hazard, the likelihood of injury is known as the risk. The latter is often expressed as a fraction like 1 in 100 or 1 in a million.

Measuring the association between risk factor and diseases

Relative risk

$$\text{Relative Risk}\,(RR) = \frac{\text{Risk in the exposed}}{\text{Risk in the non exposed}}$$

- **RR=1**

  There is no association between exposure and disease.

- **RR>1**

  Exposure is associated with an *increase* of the frequency of the disease.

- **RR<1**

  Exposure is associated with a *decrease* of the frequency of the disease.

- **The value of the RR reflects the magnitude of the association between exposure and disease**

- **RR=5 means that the probability to develop the disease in the exposed is 5 times the probability to develop it in the non exposed**

Calculation of the relative risk

Cohort study

	Disease Present	Disease absent	
Exposure Present	a	b	a+b
Exposure absent	c	d	c+d
Total	a+c	b+d	a+b+c+d

	Disease Present	Disease absent	
Exposure Present	a	b	a+b
Exposure absent	c	d	c+d
Total	a+c	b+d	a+b+c+d

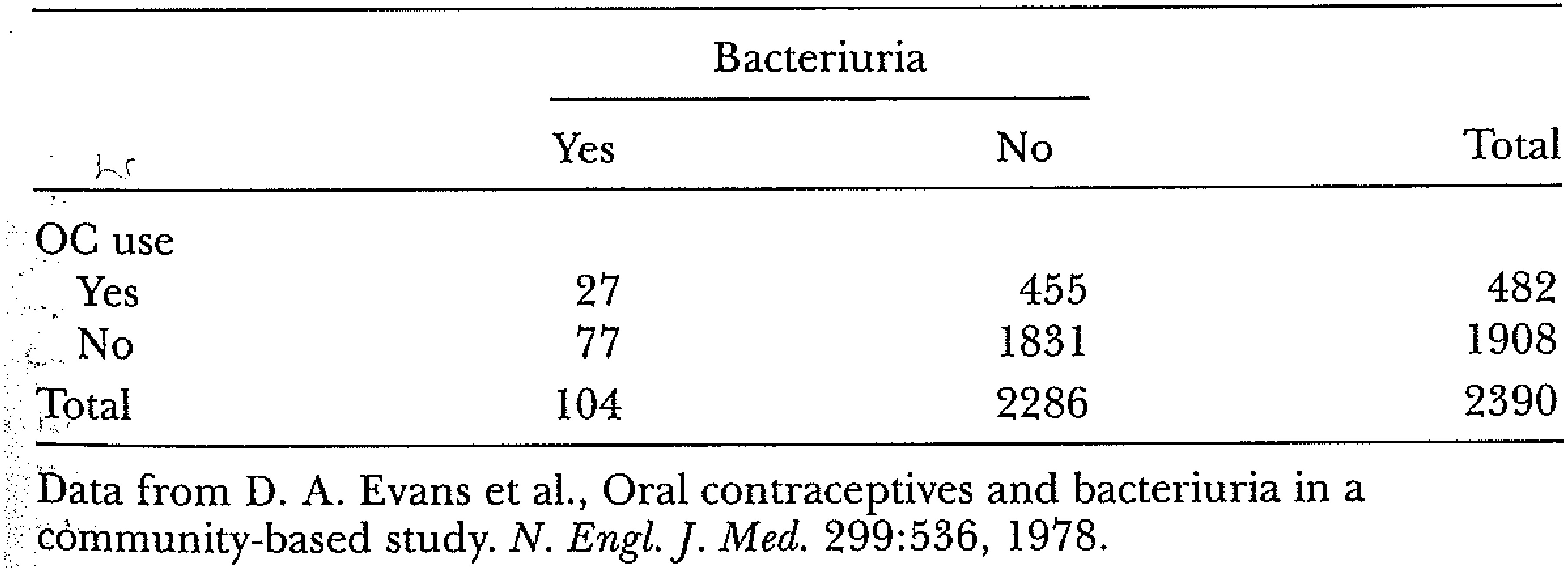
Risk in the exposed=(a)/(a+b)

Risk in the non exposed=(c)/(c+d)

$$Relative\ Risk\,(\,RR\,) = \frac{a\,/(\,a+b\,)}{c\,/(\,c+d\,)}$$

Example

Data from a cohort study of oral contraceptive (OC) use and bacteriuria among women aged 16-49 years



	Bacteriuria		
	Yes	No	Total
OC use			
Yes	27	455	482
No	77	1831	1908
Total	104	2286	2390

Data from D. A. Evans et al., Oral contraceptives and bacteriuria in a community-based study. *N. Engl. J. Med.* 299:536, 1978.

$$\text{Relative Risk}\,(RR) = \frac{27/482}{77/1908} = 1.4$$

Example

- **Rate of malaria among illiterate is 8/1000**
- **Rate of malaria among literate is 4/1000**
- **Rate ratio is 2**
- **This means that those who are illiterate have twice the rate of malaria than those who are literate**
- **Literacy is a marker rather than a causal risk**

Preventive fraction

If the exposure is preventive $I_{\text{exposed}} < I_{\text{unexposed}}$

$$PF = \frac{I_{\text{unexposed}} - I_{\text{exposed}}}{I_{\text{unexposed}}}$$

Example

Ischaemic heart disease (IHD) as a disease outcome and exercise as a preventative exposure.

	IHD risk
Exercise	**2/100**
No exercise	**8/100**

$$PF = \frac{8/100 - 2/100}{8/100} = 0.75$$

0.75 as a proportion can also be expressed as percentage, 75%. We can say that 75% of the cases of IHD in people who do not exercise could be prevented by exercise.

Design of cohort studies

1. Research question must be clear
2. Set the sample size
3. Set the follow-up period (immediate, short term and long term)
4. Specify study group Sample must be representative of the population you are studying
5. All participants should be free of the outcome (disease) at the beginning of the study
6. Must be able to get correct information about exposure status easily
7. Measure the outcome
8. Comparison group must be as similar as possible to exposed group
9. Put measures in place to reduce loss to follow up if possible

COHORT STUDY DESIGN

Selection of subjects for a cohort study

- Influenced by a variety of factors including:

1. Type of exposure being investigated
2. The frequency of the exposure in the population
3. The accessibility of subjects.

COHORT STUDY DESIGN

Selection of subjects for a cohort study

- Exposed and unexposed subjects must be free of the outcome of interest at the start of the study and equally susceptible to developing the outcome during the course of the study.
- If some subjects already have the outcome (e.g., disease) at the onset, then the temporal relationship between exposure and outcome becomes obscured.

COHORT STUDY DESIGN

Selection of subjects for a cohort study

- Each subject must rigidly satisfy the criteria for inclusion in the cohort study, and he or she should not be excluded from subsequent analysis because of any change in exposure status during follow-up.
- The degree of surveillance should be similar in exposed and unexposed groups.
- Frequency of examination and duration of follow-up depend on the type of exposure and the outcome under investigation.

COHORT STUDY DESIGN

Selection of subjects for a cohort study

- Both groups should be accessible and available for follow-up.
- Multiple comparison groups for exposed subjects chosen in different ways may reinforce the validity of findings.

Types of cohorts

- **Birth cohort : all individuals in a certain geographic area born in the same period (usually a year)**
- **Inception cohort: all individuals assembled at a given point based on some factor, e.g. where they live or work**
- **Exposure cohort: individuals assembled as a group based on some common exposure**
  - **e.g. smokers**
  - **e.g. radiation**

Healthy worker effect

phenomenon of workers usually exhibiting overall death rates lower than those of the general population due to the fact that the severely ill and disabled are ordinarily excluded from employment.

COHORT STUDY DESIGN

- Measurement of exposures should be based on intensity, duration, regularity, and variability.
- Some exposures are acute, one-time episodes never repeated in a subject's lifetime.
- Other exposures are long term, such as cigarette smoking or use of oral contraceptives.
- Exposures may also be intermittent.

COHORT STUDY DESIGN
Retrospective cohorts

- Uses information on prior exposure and disease status.
- All of the events in the study have occurred and conclusions can be drawn more rapidly.
- Costs can be lower
- May be the only feasible one for studying effects from exposures that no longer occur, such as discontinued medical treatments.
- The main disadvantage of a retrospective cohort study is that the investigator must rely on existing records or subject recall.

Retrospective cohort

- Smoking and type II DM

- We start from the year 2002 and follow up for 20 years until 2022.
- In the year 2002 we split the files into: Medical notes of smokers versus medical note for non-smokers
- Both groups should not have diabetes or impaired glucose profile at baseline
- Then, we measure the incidence of Type II DM in the smoking and no-smoking groups.
- The follow up was completed in the past, therefore, we call it a retrospective cohort study.

Ambidirectional Cohort

- **Data collected both retrospectively and prospectively on the same cohort to study short and long term effect of exposure**

- **If medical notes in the previous example were incomplete in 2002 but more complete and accurate data are available since 2015.**

- **From the year 2015 until date, the follow-up is in the past, if we continue for additional 12 year. This means a combination of retrospective and prospective data.**

COHORT STUDY DESIGN
Loss during follow-up

- Following subjects over a long period of time can lead to a variety of problems.
- Dropouts and losses of subjects to follow-up are major problems in cohort studies.
- Subjects may move away or leave the study for other reasons, including deaths from other causes than the disease under investigation.
- If losses to follow-up are significant during the study, then the validity of the results can be seriously affected.

COHORT STUDY DESIGN
Changes in exposure status

- It is also possible for exposure status to change during the course of the study.
- The exposure under study may be subject to variation over time.
- For example, cigarette smokers may quit, or employees may change jobs; therefore, their level of exposure to occupational hazards changes.

COHORT STUDY DESIGN
Analysis

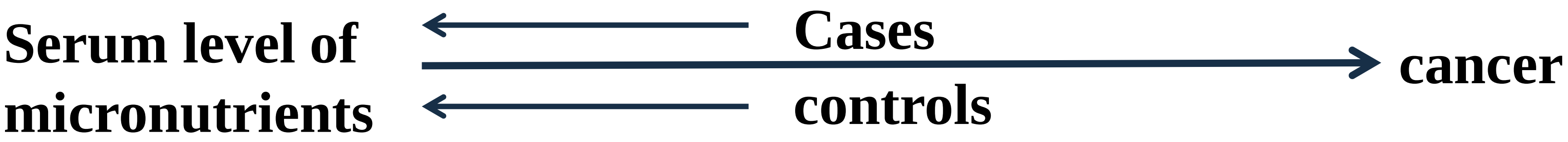
- Collection and analysis of data on the population subgroups, based on exposure, are divided according to variables of interest, like analysis in a cross-sectional study.
- Rates for subgroups are then calculated and compared.
- Data from cohort studies are analyzed in terms of relative risk and attributable risk fractions.

COHORT STUDY DESIGN
Midpoint analysis

- Occurs when, at a defined point in time in the study, all data collected to that point are analyzed so a decision can be made to stop or continue the study.

Nested case-control study

Case-control within a cohort study



Serum level of micronutrients

Cases

controls

cancer

Framingham Heart Study

Approximately 5100 residents of this Massachusetts community are followed for > 30 years.

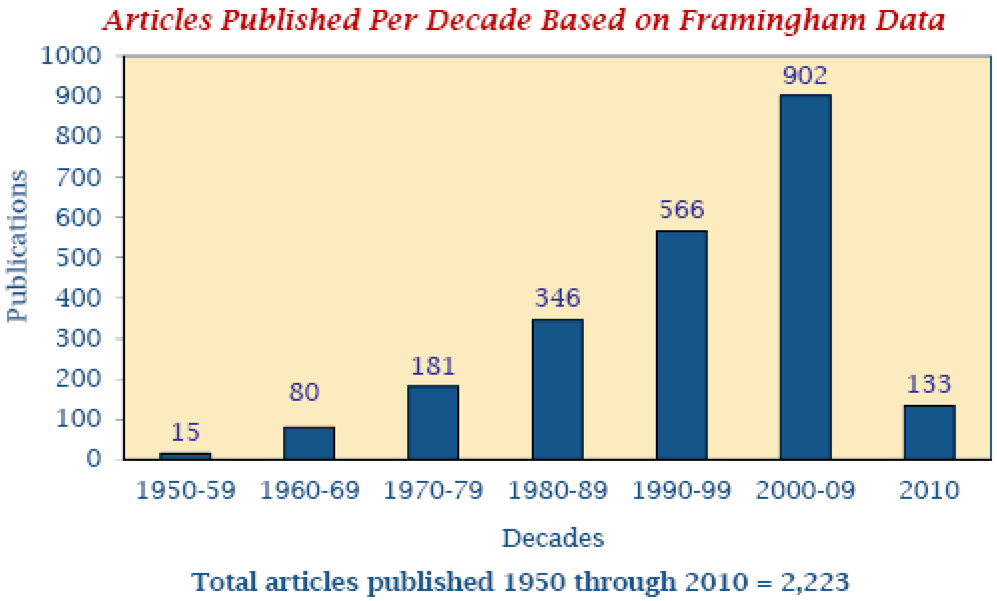
Selected because of a number of factors has permitted assessment of the effects of a wide variety of factors on the risk of numerous diseases

- **stable population,**
- **had a number of occupations and industries represented**
- **had a single, major hospital that was utilized by the vast majority of the population**
- **prepared annually updated population lists that would facilitate follow-up,**

Diseases studied included:

- **coronary heart disease**
- **rheumatic heart disease**
- **congestive heart failure**
- **angina pectoris**
- **intermittent claudication**
- **stroke**
- **gout**
- **gallbladder disease**
- **a number of eye conditions**

The Framingham Heart Study



http://www.framinghamheartstudy.org/risk/index.html

- http://www.ajconline.org/article/S0002-9149(00)00726-8/abstract

COHORT STUDY DESIGN: Summary

- In general, can investigate the effect of only a limited number of exposure
- Useful for investigating a range of outcomes associated with only one exposure
- Useful for study of rare exposure
- Not suitable for the study of rare diseases
- Follow-up studies are often large and expensive
- May take many years to complete
- Cannot test current hypotheses
- Can measure disease incidence

Bradford Hill Criteria

1. Strength of the evidence
2. Order in time
3. Consistency
4. Plausibility
5. Specificity
6. Biological gradient
7. Coherence
8. Experiment
9. Analogy

Week 6

Analytical Studies part 2

Case control studies

Dr Munir Abu-Helalah

Case-control studies

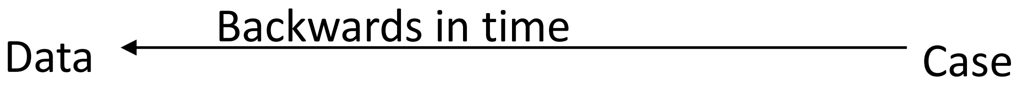
Are studies in which a group of people with a particular disease (the cases) are compared with a group of people without the disease (the controls). The purpose of the comparison is to determine whether, in the past, the cases have been exposed more (or less) often to a specific factor than the controls

- This type of study is done to identify factors that could be responsible for the development of a disease or drug use problem.

CASE-CONTROL STUDIES

- The direction of time

- Cases identified now
- Data on past events collected



Backwards in time

Data

Case

CASE-CONTROL STUDY DESIGN

- Designed to assess association between disease occurrence and exposures (e.g., causative agents, risk factors) suspected of causing or preventing the disease.

Case-control studies

- A group of people with a disease are compared to a group without the disease from the same population.
- Compare exposure to risk factors in both groups
- Able to look at many different possible risk factors
- Able to study diseases with a long latency period
- Most common analytic study design seen in the medical literature today

Case-control studies

- In general, the cases included in a case-control study include people with one specific disease only
- But, a case-control study can provide information on a wide range of possible exposures that could be associated with that particular disease
- Useful for the study of rare diseases
- Not suitable for the study of rare exposure
- Relatively small and inexpensive
- Takes a relatively short time to complete
- Can test current hypotheses
- Cannot measure disease incidence

CASE-CONTROL STUDIES

- Cases have the disease of interest

Eg. Cerebral palsy

- Controls do not have the disease

Eg. Healthy babies born at the same time

Case-control study: challenges

- Selecting cases
  - Eligibility
- Selecting controls
  - Representativeness
- Exposure assessment
  - Accurate

CASE-CONTROL STUDY DESIGN

- More efficient than a cohort study because a smaller sample size is required.
- One key feature of a case-control study, which distinguishes it from a cohort study, is the selection of subjects based on disease status.
- Controls are chosen from the same population yielding the cases

Design of case control studies

- **Comparability:Two groups must be as similar to each other as possible so selection of controls is very important. Controls must be as similar as possible to cases – except that they do not have the outcome (disease).**

- **Outcome (disease) must be very clearly defined. (Diagnostic criteria must be clear)**

- **Use objective data about exposure status wherever possible, to reduce the risk of bias**

CASE-CONTROL STUDIES

Strengths

- Suited to study disease with long latency periods, but can be used in outbreaks investigations
- Optimal for rare diseases
- Efficient in terms of time and costs: relatively quick and inexpensive
- Allows for evaluation of a wide range of possible causative factors that might relate to the disease being studied
- Odds ratio estimated

CASE-CONTROL STUDIES

Limitations

- Very susceptible to bias (especially selection and recall bias) as both the disease and the exposure have already occurred when participants enter the study. Cases and controls might not be representative of the whole population
- We cannot calculate incidence or prevalence rate of disease
- We cannot be certain that exposure came before disease
- Choice of controls difficult
- Controls do not usually represent non-exposed population
- Past records incomplete
- No absolute risk estimates

CASE-CONTROL STUDY DESIGN

- **Data Analysis**
- Data collection and analysis are based on whether the case-control study involves a matched or unmatched design. The measure used typically in case-control studies is the odds ratio.
- **Odds ratio (OR):** odds of a particular exposure among people with a specific condition divided by the corresponding odds of exposure among people without the condition under study

Odds Ratio

The word "odds" means the chances of an event to happen. The Odds of an event is the *ratio* of the event to happen over the event not to happen.

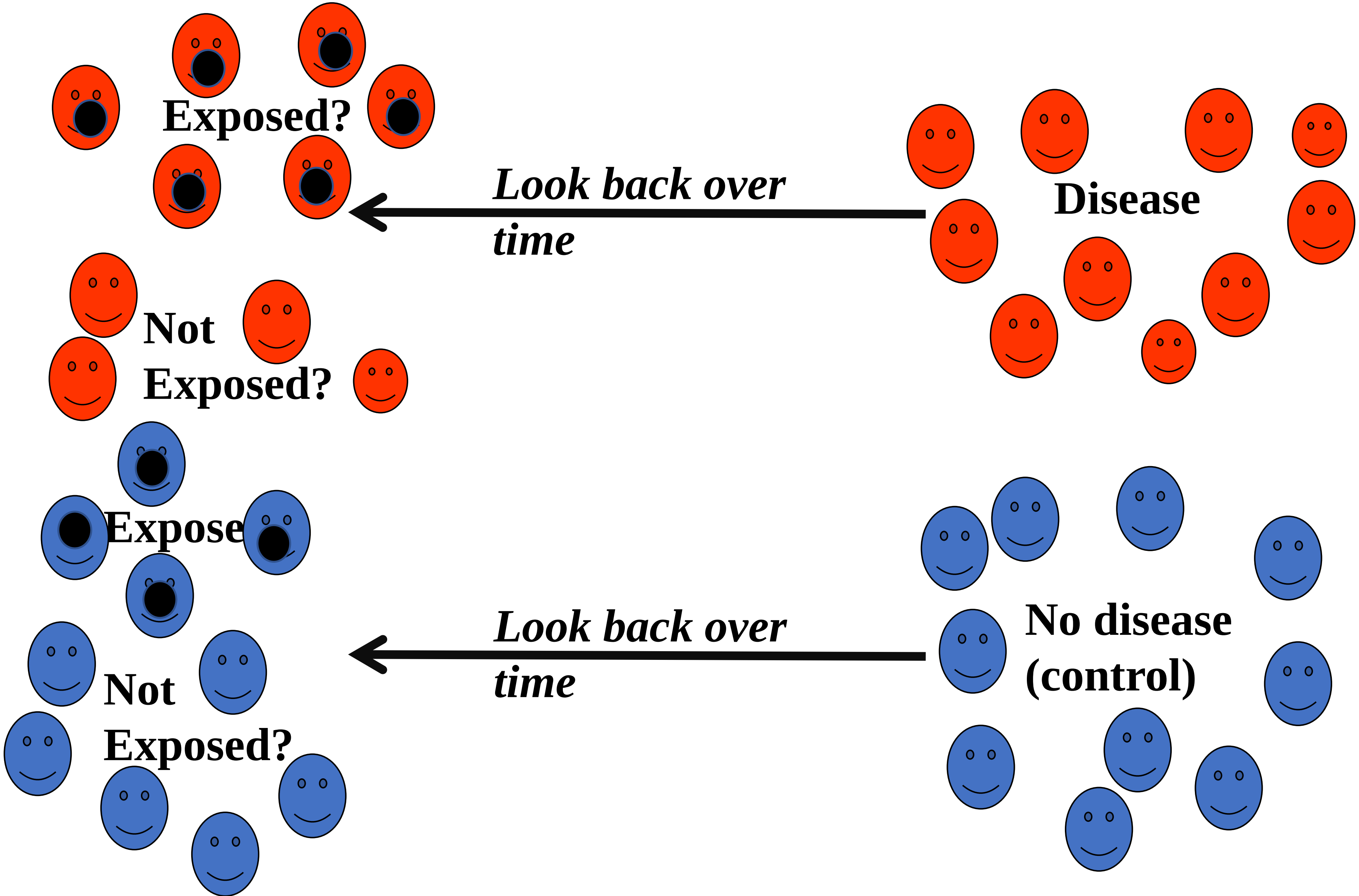
$$Odds(A) = \frac{probability(A\ happens)}{probability(A\ does\ not\ happen)} = \frac{prob(A)}{1 - prob(A)}$$

$$prob(A) = \frac{Odds(A)}{1 + Odds(A)}$$

Odds Ratio (OR)

$$OR = \frac{\text{Odds of exposure}_{\text{cases}}}{\text{Odds of exposure}_{\text{controls}}}$$

Case control studies



Exposed?

Not Exposed?

Look back over time

Disease

Expose

Not Exposed?

Look back over time

No disease (control)

Case-control study

	Disease Present	Disease absent	
Exposure Present	a	b	a+b
Exposure absent	c	d	c+d
Total	a+c	b+d	a+b+c+d

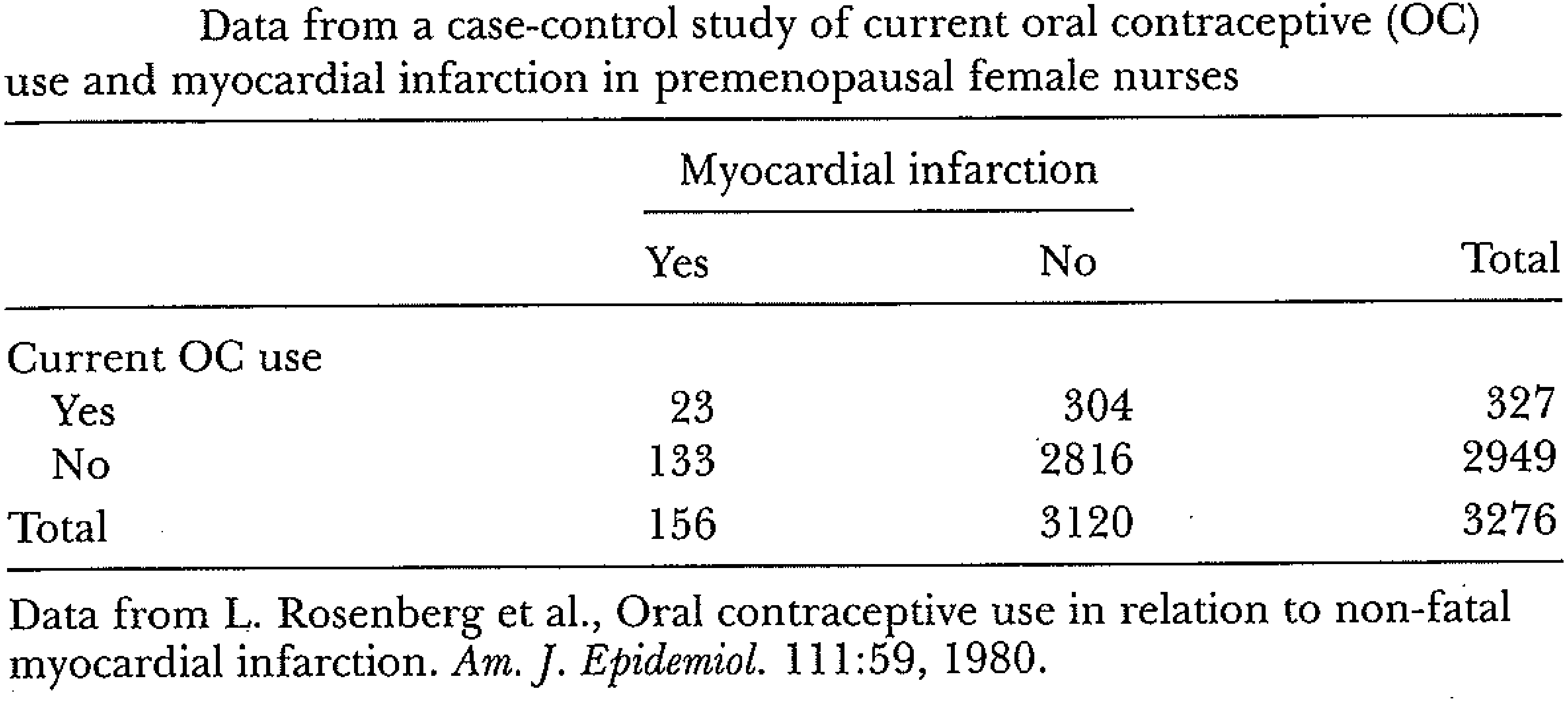
Odds of being ill in exposed=a/b

Odds of being ill in non exposed =c/d

Odds ratio (OR)=Odds in exposed/Odds in non exposed

= OR=(a/b)/(c/d)

$$Odds\ Ratio\,(OR) = \frac{ad}{cb}$$



Data from a case-control study of current oral contraceptive (OC) use and myocardial infarction in premenopausal female nurses

	Myocardial infarction		
	Yes	No	Total
Current OC use			
Yes	23	304	327
No	133	2816	2949
Total	156	3120	3276

Data from L. Rosenberg et al., Oral contraceptive use in relation to non-fatal myocardial infarction. *Am. J. Epidemiol.* 111:59, 1980.

$$OR = \frac{23 \times 2816}{304 \times 133} = 1.6$$

Women who were current OC users had a risk of MI 1.6 times that of nonusers

Two by two table

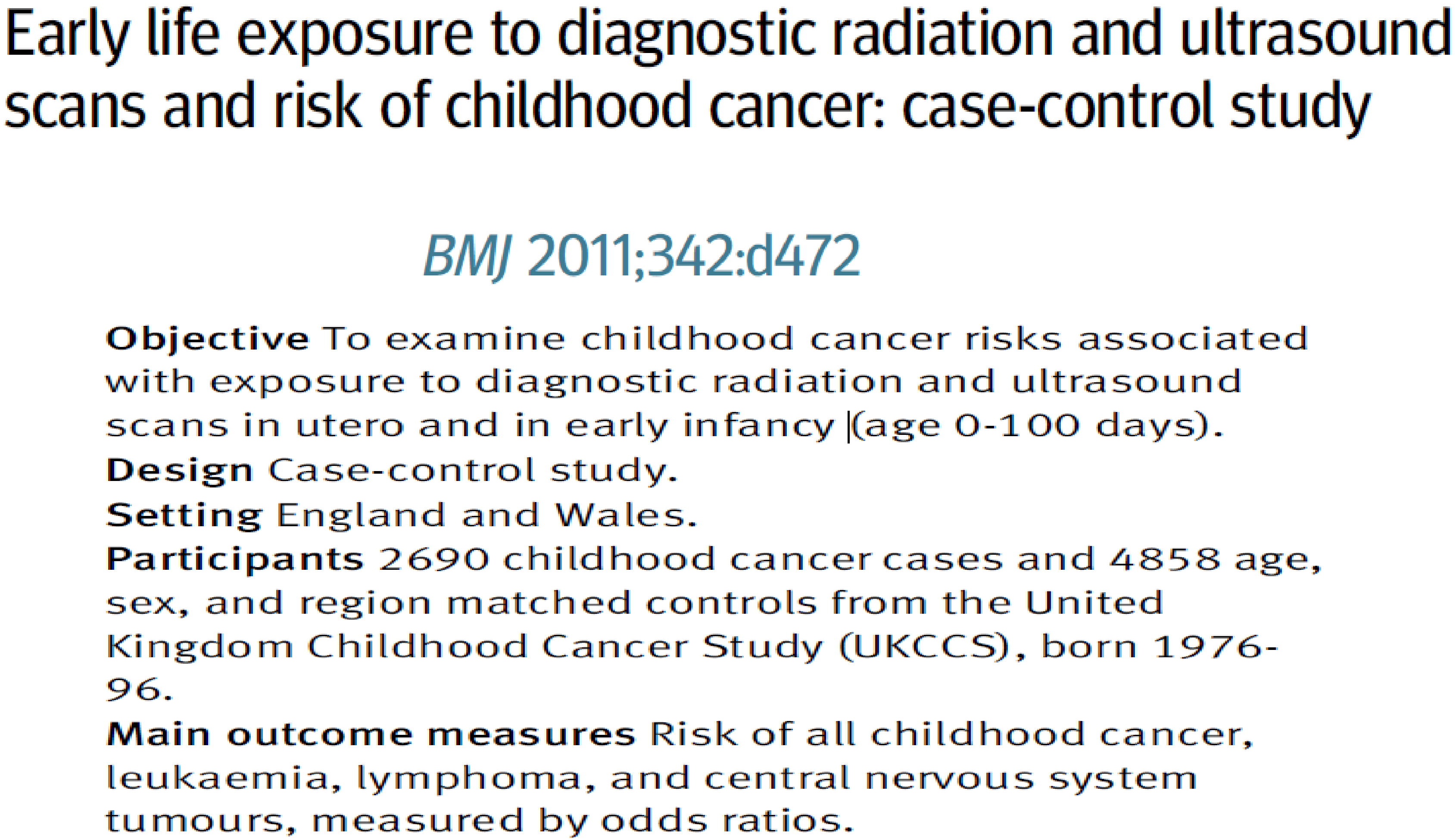
Exposure	**Outcome**		
	Yes	No	Total
Yes	a	b	a + b
No	c	d	c + d
Total	a + c	b + d	a + b + c + d

Odds of outcome in exposed = a / b

Odds of outcome in non- exposed = c / d

Outcome odds ratio = (a/b) / (c/d) = ad / bc

Case-control study



Early life exposure to diagnostic radiation and ultrasound scans and risk of childhood cancer: case-control study

BMJ 2011;342:d472

Objective To examine childhood cancer risks associated with exposure to diagnostic radiation and ultrasound scans in utero and in early infancy (age 0-100 days).
Design Case-control study.
Setting England and Wales.
Participants 2690 childhood cancer cases and 4858 age, sex, and region matched controls from the United Kingdom Childhood Cancer Study (UKCCS), born 1976-96.
Main outcome measures Risk of all childhood cancer, leukaemia, lymphoma, and central nervous system tumours, measured by odds ratios.

Case-control study: example

Radiation	**Case**	**Control**	Total
Yes	140	165	305
No	1550	5693	7243
Total	1690	5858	7548

Odds of outcome in exposed = 140 / 165 = 0.85

Odds of outcome in non-exposed = 1550 / 5693 = 0.27

Outcome odds ratio = (a/b) / (c/d) = 0.85/0.27=3.1

CASE-CONTROL STUDIES

Methods of data collection

Case-note review: Completeness

Postal questionnaire: response rate

Interview: Detailed information

How many controls?

- **control-to-case ratio is 1 : 1**

  is the optimal when the number of available cases and controls is large and the cost of obtaining information from both groups is comparable

- **control-to-case ratio is 1 : n**

  When the number of cases is limited or when the cost of obtaining information is greater for cases or controls

- **As the number of controls per case increases, the power of the study also increase**

- **It is not recommended that this ratio increase beyond 4 : 1**

CASE-CONTROL STUDY DESIGN

- **Selecting Cases and Controls**
- Identification and collection of cases involves specifying the criteria for defining a person as a case—in other words, as having the disease (also called *case definition*).
- This definition consists of a set of criteria, also called *eligibility criteria,* for inclusion in the study. There also are criteria for exclusion from the study.

CASE-CONTROL STUDY DESIGN

- The next step is selection of the controls.
- Controls are chosen from the source population.
- The source population is usually defined by geographic area. It is important to select controls so that participation does not depend on exposure.

CASE-CONTROL STUDY DESIGN

Source of controls

- The ideal situation is a random sample from the same source population as the cases.
- Investigators may use more than one control group.
- Controls can be selected by sampling:

  The general population in the same community; the hospital community (patients in the same hospital); individuals who reside in the same block or neighborhood; and spouses, siblings, or associates (schoolmates, co-workers) of the cases.

Obtaining cases and controls for case control studies

Study	Source of cases	Source of controls
PROM (premature rupture of membrane)	Hospital patients	Hospital patients
Rheumatoid arthritis	Outpatient clinic	Other outpatient clinic
Cervical screening	GP register	GP register

CASE-CONTROL STUDY DESIGN

Matching Cases and Controls

- **Matching is a popular approach to control for confounding and selection bias in case-control studies.**
- **Matching cases and controls helps to ensure that these groups are similar with respect to important risk factors, thereby making case-control comparisons less subject to confounding or selection bias.**

CASE-CONTROL STUDY DESIGN

Prior exposure to the risk factor(s) of interest

- Once cases and controls are selected, information must be collected on prior exposure to the risk factor(s) of interest.
- Interviews and questionnaires are the most common means of determining a subject's exposure history and medical records review is another source
- The most objective means for characterizing exposure is the use of a biological marker.

Bias

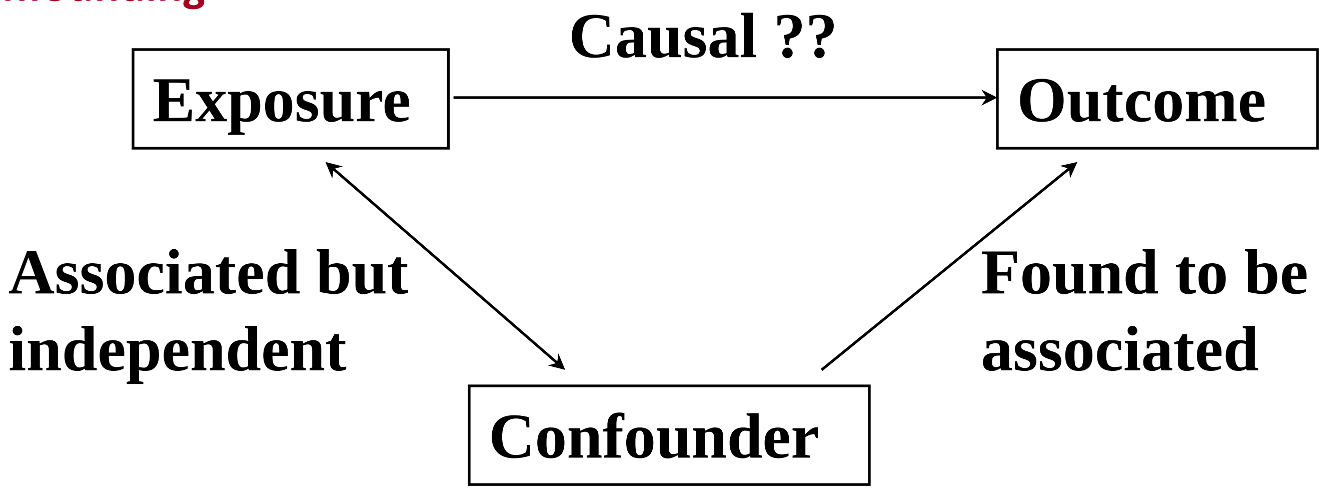
Bias is any systematic error in an epidemiological study that results in an incorrect estimate of the association between exposure and risk of the outcome

- **Selection bias:** inappropriate controls
- **Observation bias**
  - **Subject and recall bias:** eg recall bias of mothers with cerebral palsy babies
  - **Interviewer bias: blind if possible**
  - **Misclassification**

Confounding

A confounding factor is one that is associated with the exposure and that independently affects the risk of developing the outcome, but that is not an intermediate link in the causal chain between the exposure and the outcome under study

Matching - often used in case-control studies to decrease confounding



Causal ??

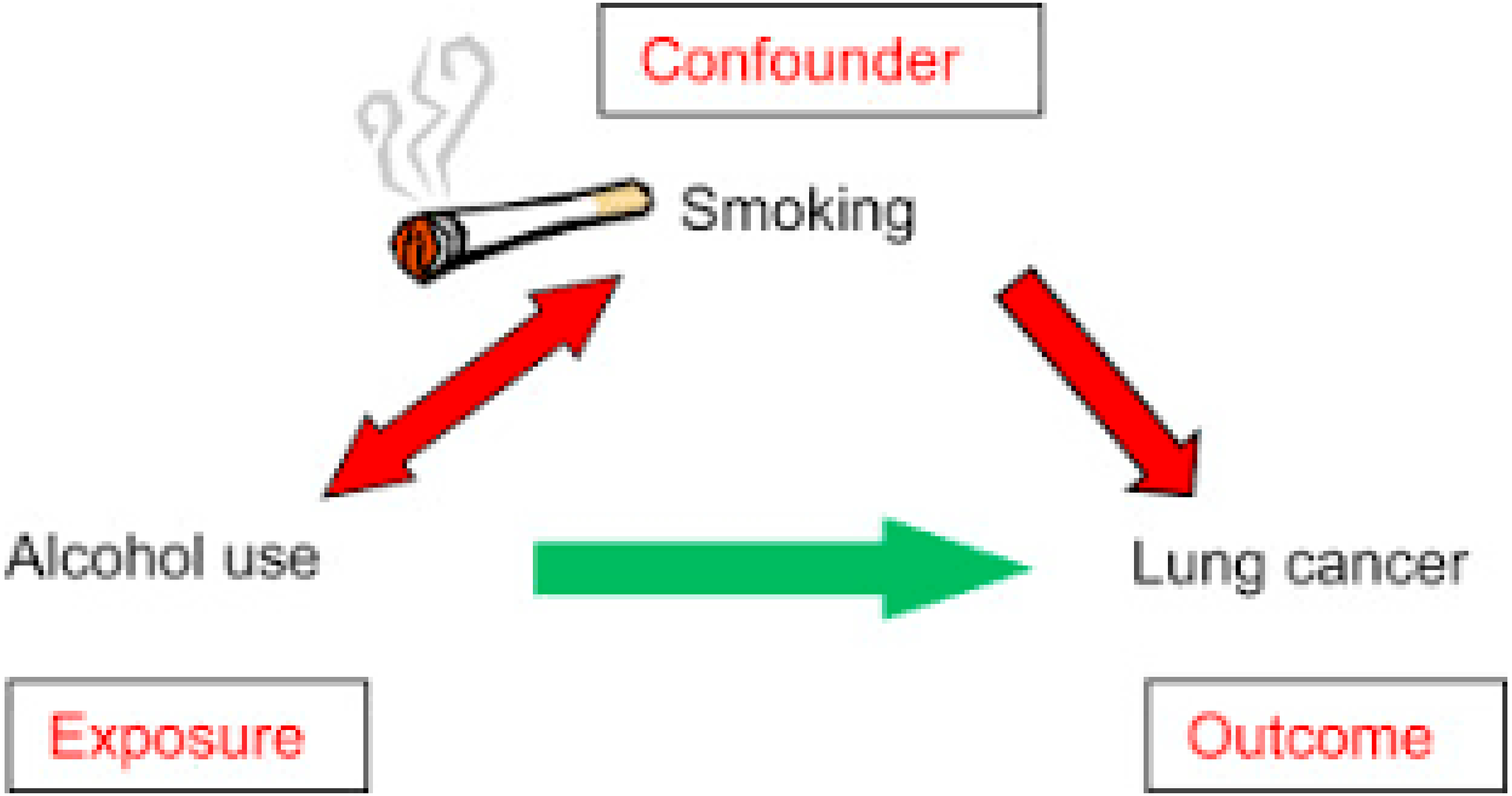
Exposure

Outcome

Associated but independent

Found to be associated

Confounder



Confounder
Smoking
Alcohol use
Lung cancer
Exposure
Outcome

Confounding

Matching Cases and Controls

- For example, if age and sex are the matching variables, then a 35 year old male case is matched to a 35 year old male control
  - Pair matching (one to one individual matching)
- The use of matching usually requires special analysis techniques (e.g. matched pair analyses and conditional logistic regression)