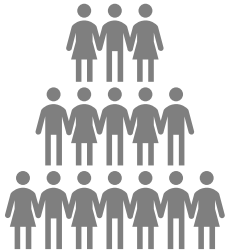
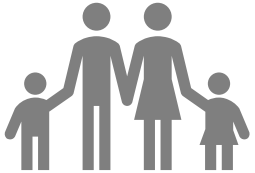
Statistics for Oncology

A Course for Scottish Trainees

by… The Edinburgh Cancer Informatics Research Group

https://edin.ac/oncology-statistics

Collection and use of Epidemiological Data


Clinical Oncology

Curriculum 2021

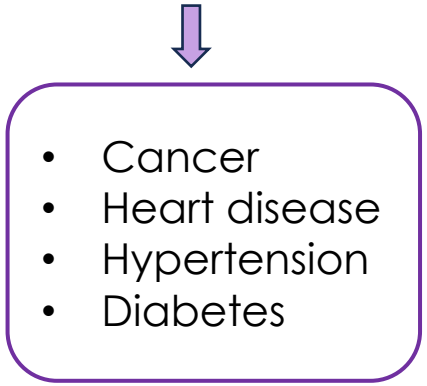
Clinical Oncology
The Royal College of Radiologists

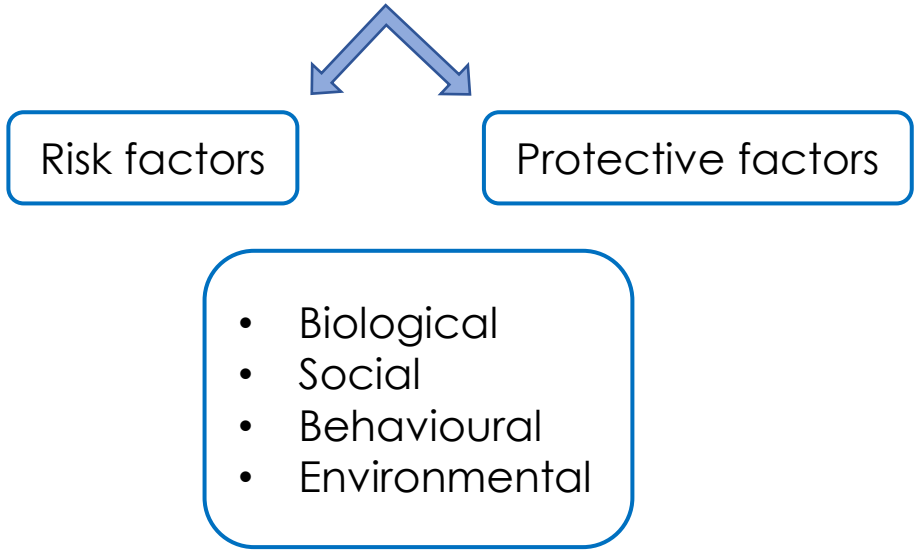
Medical statistics module

3.9 Collection and use of epidemiological data	▪ Contrast the design and interpretation of cross-sectional case control and cohort studies ▪ Define the principles, calculate and interpret odds ratios and risk ratios ▪ Define incidence, prevalence, mortality rates and standardised mortality rates

Epidemiological studies

- Epidemiology is the study of the occurrence and determinants of ill health in the population.
- Epidemiological studies, assess the relationship between **factors of interest** and the **occurrence of disease** in the population.

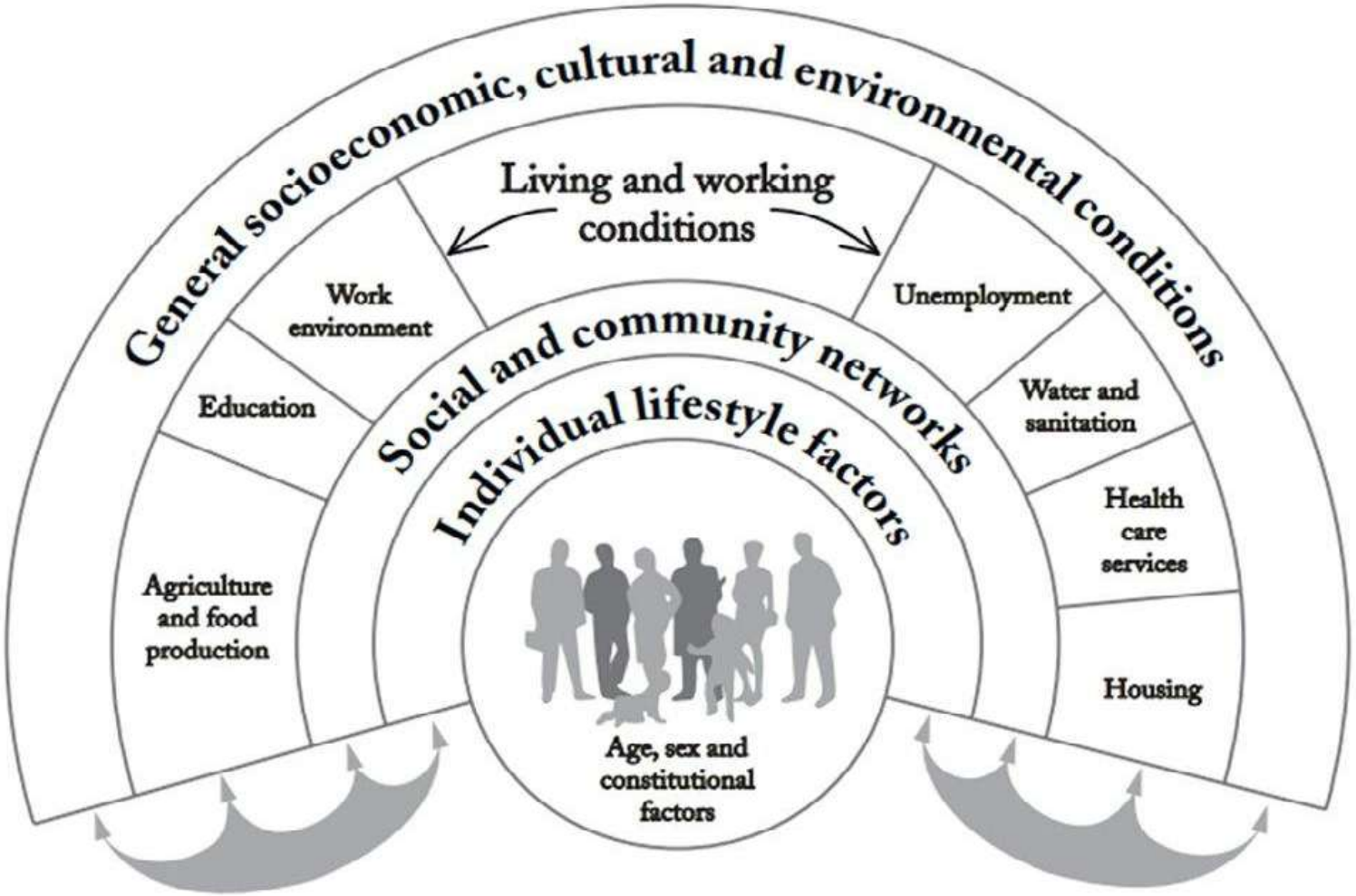


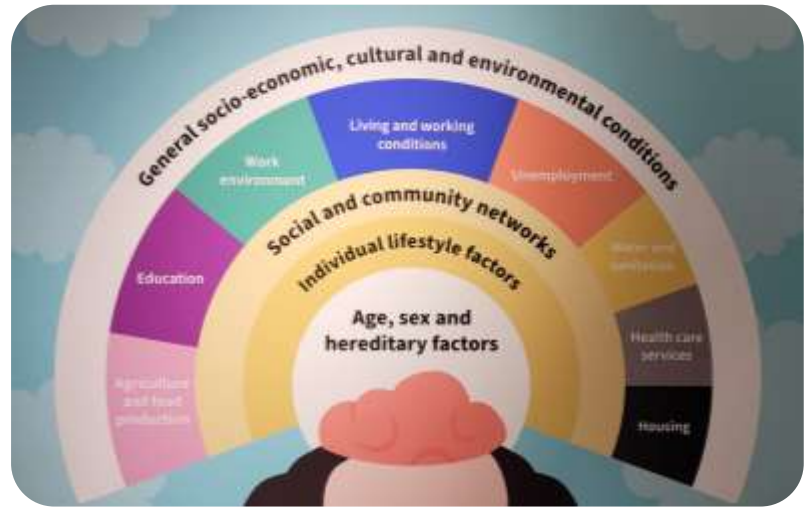


- Epidemiological studies are mostly **observational** in design

  (in contrast to **experimental** studies which involve interventions to affect an outcome).

One view of the determinants of health





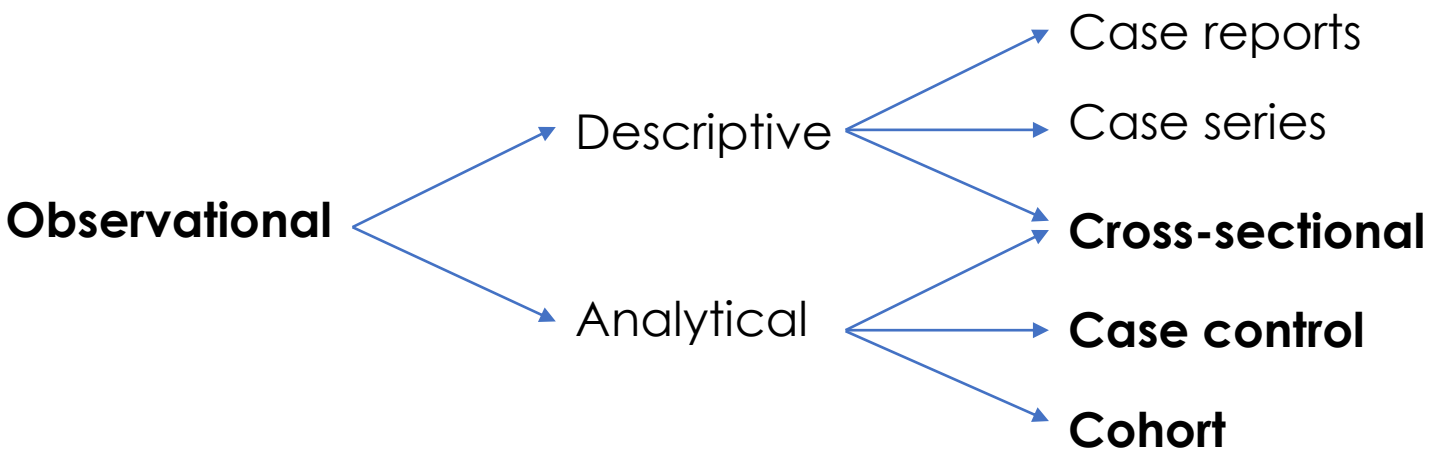
First developed to discover and understand possible causes of infectious diseases like smallpox, typhoid and polio.

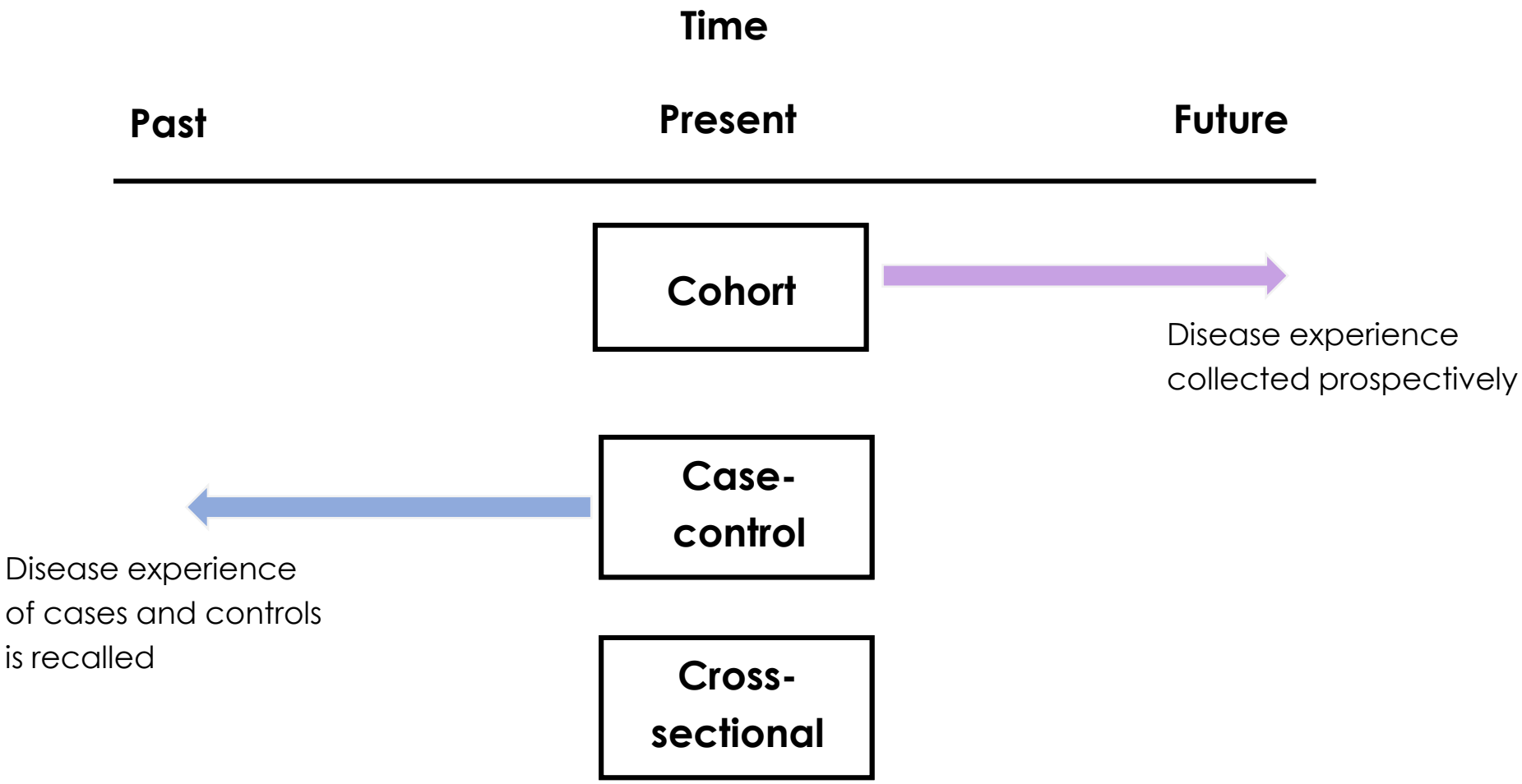
Expanded to include the study of societal factors associated with chronic diseases

Figure: The Dahlgren and Whitehead model of the main determinants of health

Epidemiological studies

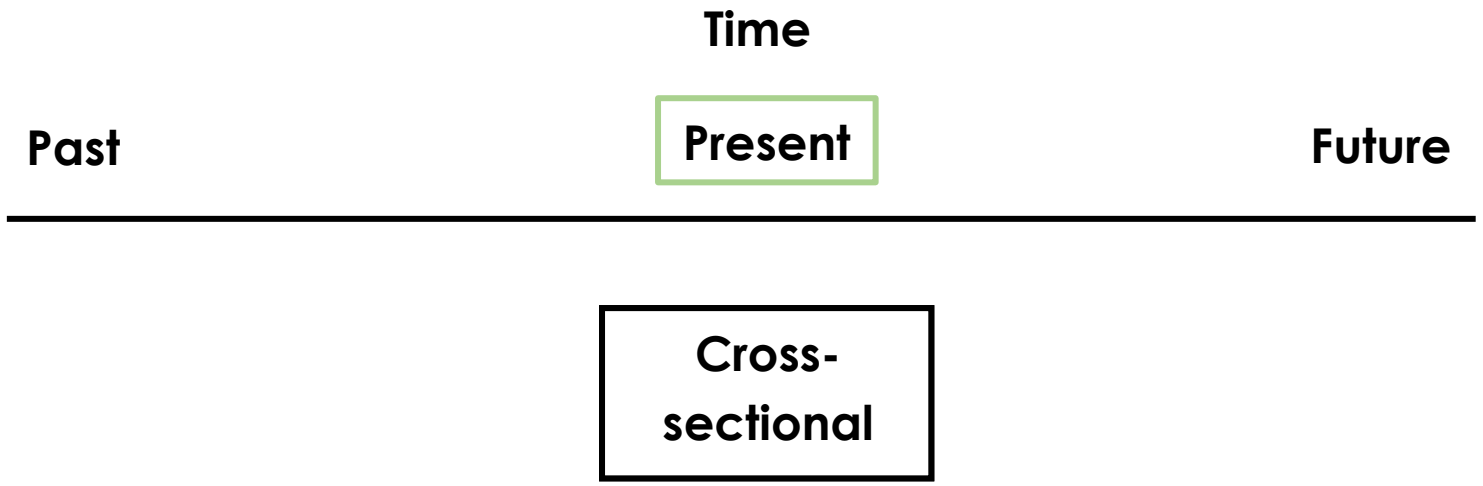
Epidemiological studies can be broadly divided into two types:





Prospective - Studies in which the health event of interest has yet to happen

Retrospective - Studies in which the health event has already occurred



- A cross-sectional study is carried out at a single point in time.
- **Health survey -** aim is to describe health behaviours or health status in a large sample of the population.
- **Census -** is a type of survey in which the entire target population is investigated
- Suitable for estimating the **prevalence** of a condition in the population.

Prevalence

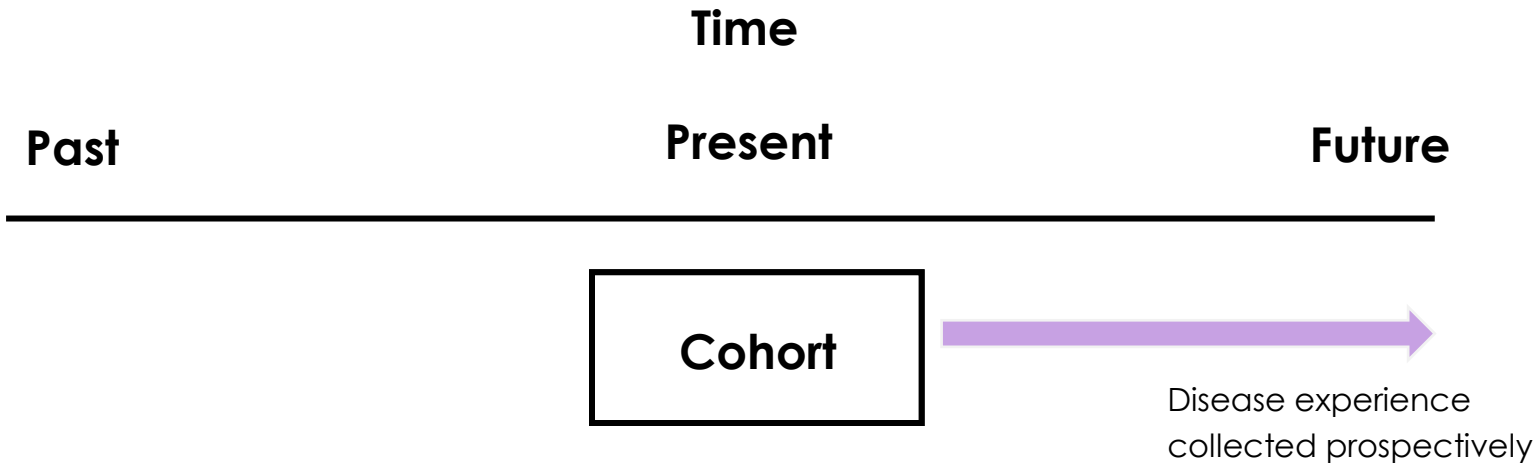
Prevalence is the proportion/ percent of individuals with a particular condition in the population <u>at a point in time</u>

$$\textbf{Point prevalence} = \frac{Number\ with\ the\ disease\ at\ a\ single\ time\ point}{Total\ number\ studied\ at\ the\ same\ time\ point}$$

$$\textbf{Period prevalence} = \frac{Number\ with\ the\ disease\ over\ a\ specified\ time\ period}{Total\ number\ studied\ during\ the\ same\ time\ period}$$

Q?. With cross-sectional studies:

1. Can we assess trends over time?
2. Can we estimate the **incidence** of a disease?



A cohort study takes a group of individuals and follows them forward in time.

Usually prospective.

The aim is to assess whether exposure to a particular factor affects the **incidence** of disease in the future.

Incidence

Incidence is the number of new cases of a condition occurring in a population over a set time period.

$$\textbf{Incidence risk} = \frac{\text{Number of new cases of disease in a specified period of time}}{\text{Number of persons at risk at the begining of the same time period}}$$

Incidence rate is the number of new cases divided by the **person-time** at risk

$$\textbf{Incidence rate} = \frac{\text{Number of new cases of disease in a specified period of time}}{\text{Total person} - \text{time at risk during the follow} - \text{up time period}}$$

Incidence

Incidence Risk

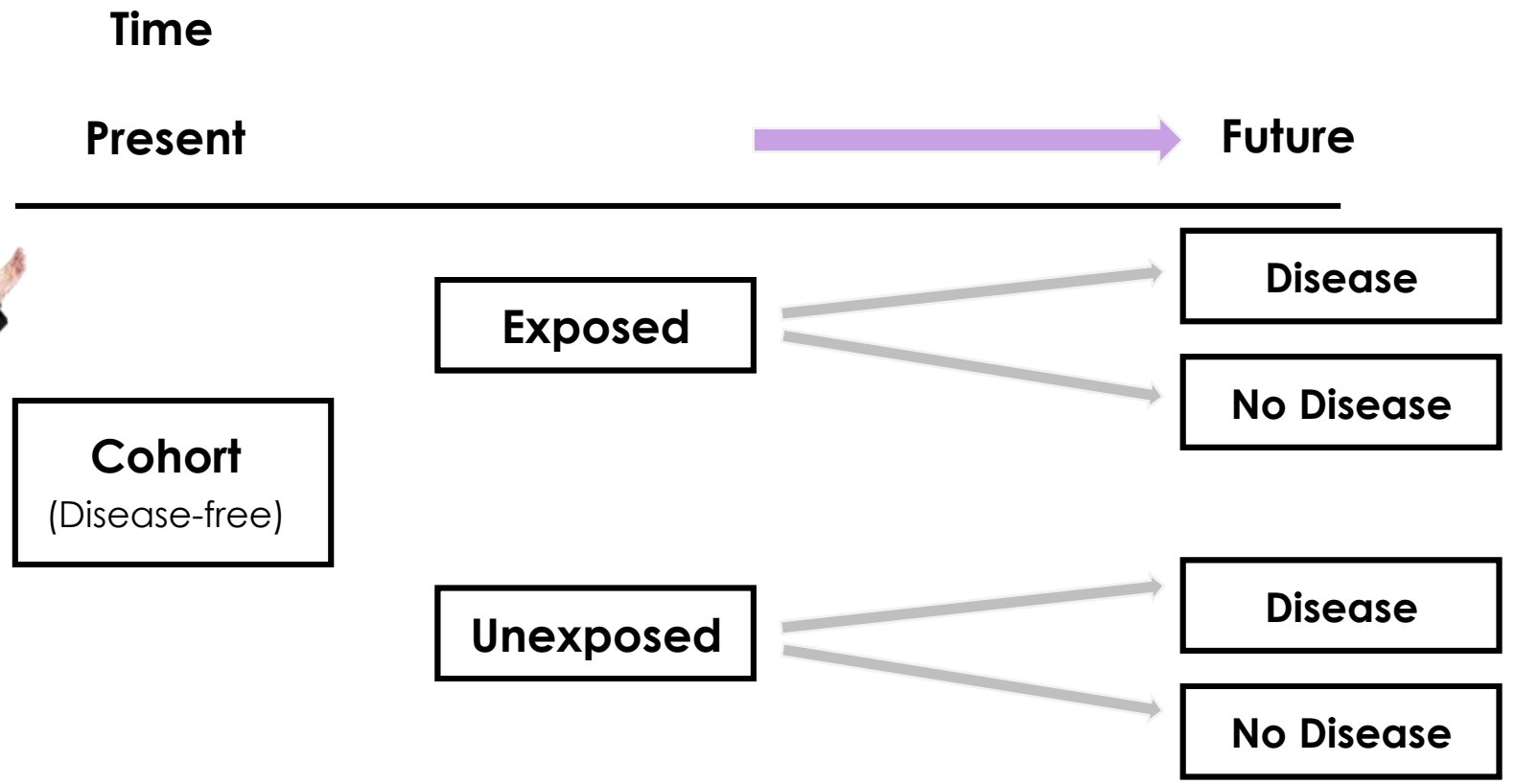
- The proportion of individuals in a population (initially free of disease) who develop the disease within a specified time interval.
- Incidence risk is expressed as a **percentage** (or if small as per 1000 persons).

- Assumes that the entire population at risk at the beginning of the study has been followed up for the specified period.
- However, in a cohort study participants may be lost during follow-up.
- To account for these variations in follow up, a more precise measure can be calculated.

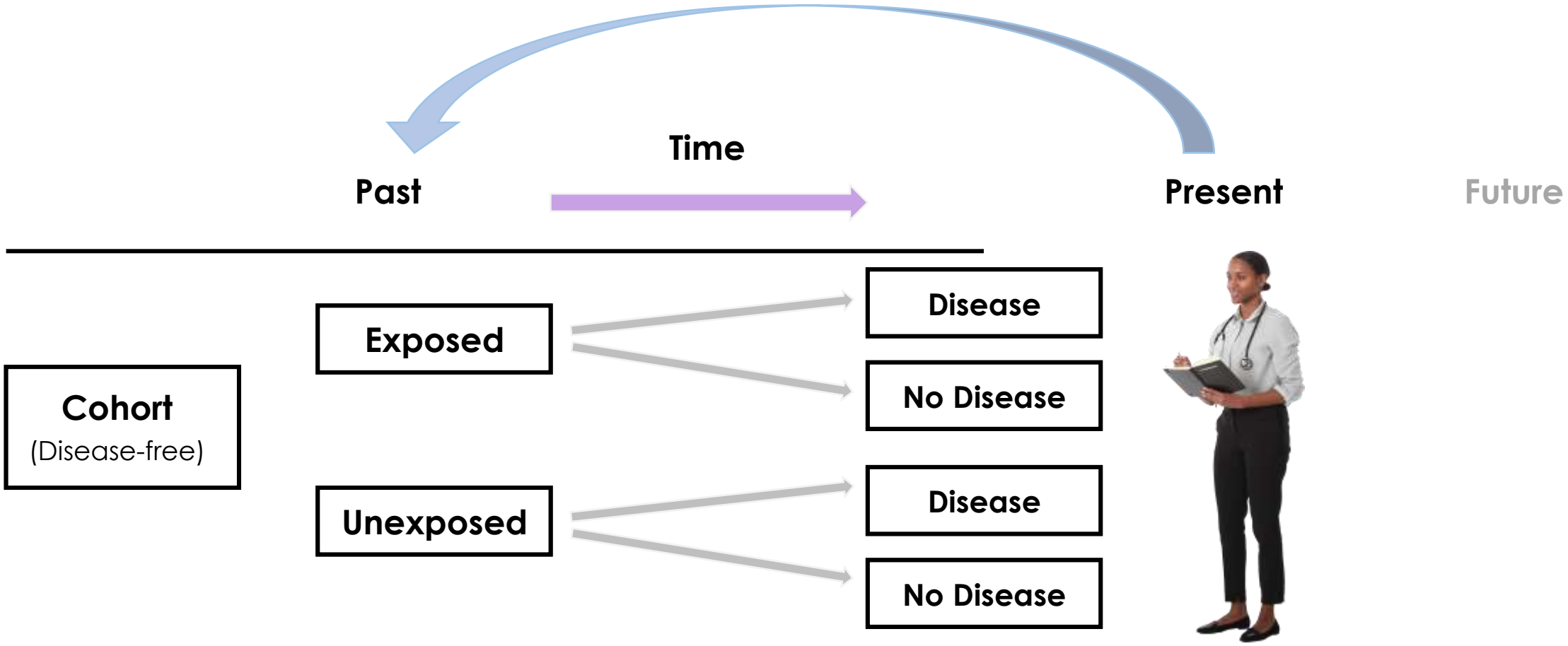
Incidence Rate

- Measures the frequency of new cases of disease in a population.
- However, incidence rates take into account the sum of the time that each person remained under observation and at risk of developing the outcome under investigation. **(person-time)**

Cohort study (Prospective)



Retrospective Cohort study



The analysis of cohort studies can be summarised by the ratio of incidence rates in the exposed and non-exposed groups (**incidence rate ratio** or **relative risk** or **risk ratio**).

		Disease of interest		
		Yes	No	Incidence
Exposed to factor	Yes	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{a}/(\boldsymbol{a}+\boldsymbol{b})$
	No	$\boldsymbol{c}$	$\boldsymbol{d}$	$\boldsymbol{c}/(\boldsymbol{c}+\boldsymbol{d})$

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

The **relative risk (RR)** indicates the increased (or decreased) risk of disease associated with exposure to the factor of interest

Relative Risk	Interpretation
>1	an increased risk in the exposed group
1	risk is the same in the exposed and unexposed groups
<1	a reduced risk in the exposed group

Cohort study investigating low serum ferritin and development of anaemia among women

		Anaemia		
		Yes	No	Incidence
Serum ferritin<20	Yes	7	8	7/15
	No	2	13	2/15

$$\boldsymbol{RR = \frac{7/15}{2/15} = 3.5}$$

Risk of developing anaemia among women with low serum ferritin is **3.5 times** the risk among women who do not have low serum ferritin

Risk is 250% higher (3.5 – 1)

Advantages

Disadvantages

- Multiple outcomes can be studied
- Incidence rates can be established
- Can provide an indication of the progression of disease over time
- Useful for relatively uncommon exposures

- Need to follow up over a long period of time
- Expensive
- Prone to loss to follow-up
- Not efficient for rare diseases
- Prone to confounding
- Participants may move between categories of exposure
- Knowledge of exposure status may bias classification of the outcome
- Being in the study may alter participant's behaviour

BRITISH MEDICAL JOURNAL

LONDON SATURDAY NOVEMBER 10 1956

LUNG CANCER AND OTHER CAUSES OF DEATH IN RELATION TO SMOKING

A SECOND REPORT ON THE MORTALITY OF BRITISH DOCTORS

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, C.B.E., F.R.S.

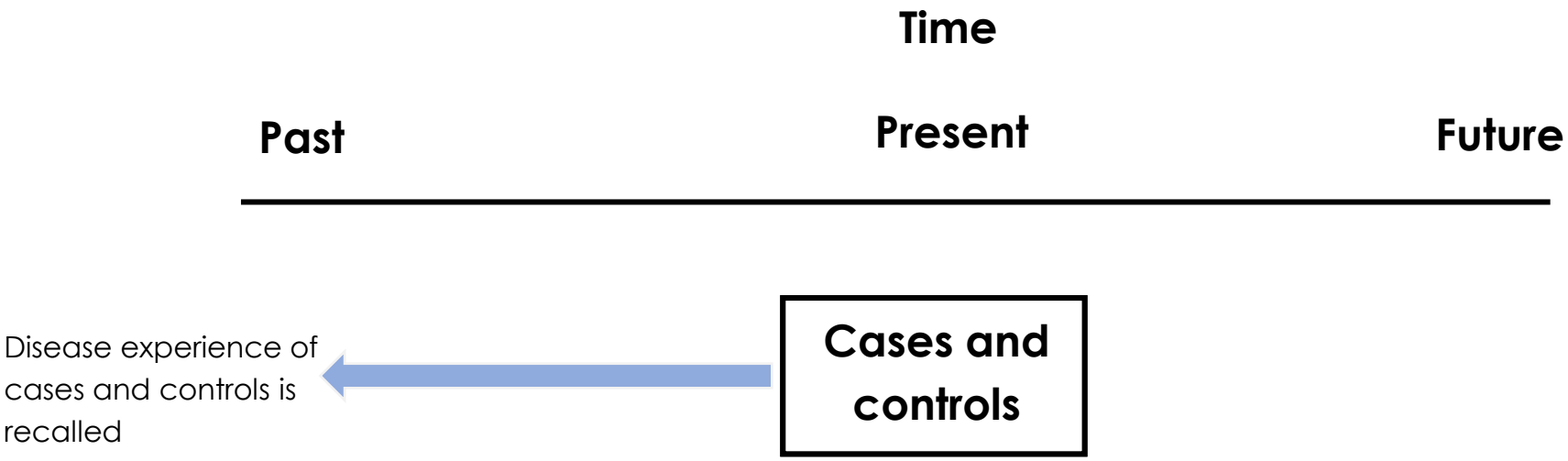
Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

On October 31, 1951, we sent a simple questionary to all members of the medical profession in the United Kingdom. In addition to giving their name, address, and age, they were asked to classify themselves into one of three groups—namely, (*a*) whether they were, at that time, smokers of tobacco; (*b*) whether they had smoked but had given up; or (*c*) whether they had never smoked regularly (which we defined as having never smoked as much as one cigarette a day, or its equivalent in pipe tobacco or cigars, for as long as one year). All smokers and ex-smokers were asked additional questions. The smokers were asked the ages at which they had started previously have been a light smoker or may since then have given up smoking altogether; we shall have continued to count him, or her, as a heavy smoker. If there is a differential death rate with smoking, we must by such errors tend to inflate the mortality among the light smokers and to reduce the mortality among the heavy smokers. In other words, the gradients we present in this paper may be understatements but (apart from sampling errors due to the play of chance) cannot be overstatements.

In 1954 we published a preliminary report on the results of this inquiry (Doll and Hill, 1954a). The num-

TABLE VII.—*Mortality From Lung Cancer in Relation to the Amount Smoked at Different Ages Above 35 Years: Annual Rates Per 1,000 Men*

Age in Years	No. of Deaths	Death Rate Among:			
		Non-smokers	Men Smoking a Daily Average of:		
			1–14 g.	15–24 g.	25 g. or More
35–54 ..	10	0·00	0·09	0·17	0·26
55–64 ..	24	0·00	0·32	0·52	3·10
65–74 ..	31	0·00	1·35	3·34	4·81
75 and over	19	0·70	2·78	2·07	4·16
All ages ..	84	0·07	0·47	0·86	1·66



- Compares the characteristics of a group with a particular disease (**cases**) to a group without the disease (**controls**)
- To see whether exposure to a factor occurred more or less frequently in the cases than the controls
- It is not possible to estimate the risk of disease.
  (Because patients are selected on the basis of their disease status)
- We can estimate the **odds** of being exposed to the risk factor for cases and controls.

The **odds ratio** (OR) gives an indication of the increased (or decreased) odds associated with exposure to the factor of interest.

		Disease status	
		Case	Control
Exposed to factor	Yes	a	b
	No	c	d
	Odds	a/c	b/d

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

The **odds ratio** (OR) gives an indication of the increased (or decreased) odds associated with exposure to the factor of interest.

Odds ratio	Interpretation
>1	an increased odds of disease in the exposed group
1	odds is same in the exposed and unexposed groups
<1	a reduced odds of disease in the exposed group

Case-control study of oral contraceptives and breast cancer

Cases → women diagnosed with breast cancer in a certain hospital
Controls → women inpatients in same hospital

		Breast cancer	
		Case	Control
Ever used oral contraceptive	Yes	537	554
	No	639	622
	Odds	537/639	554/622

$$OR = \frac{537/639}{554/622} = 0.94$$

Odds of contractive use among women with breast cancer is **0.94** (6% less) times the odds among women who do not have breast cancer.

BRITISH MEDICAL JOURNAL

LONDON SATURDAY SEPTEMBER 30 1950

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, Ph.D., D.Sc.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

TABLE IV.—*Proportion of Smokers and Non-smokers in Lung-carcinoma Patients and in Control Patients with Diseases Other Than Cancer*

Disease Group	No. of Non-smokers	No. of Smokers
Males:		
Lung-carcinoma patients (649)	2 (0·3%)	647
Control patients with diseases other than cancer (649) ..	27 (4·2%)	622
Females:		
Lung-carcinoma patients (60)	19 (31·7%)	41
Control patients with diseases other than cancer (60) ..	32 (53·3%)	28

Case-control study of smoking and lung cancer in men (Doll & Hill, 1950)

Cases → men diagnosed with lung cancer
Controls → men with diseases other than cancer

		Lung cancer	
		Cases	Controls
Smoker	Yes	647	622
	No	2	27
	Odds	647/2	622/27

$$\boldsymbol{OR = \frac{647 \times 27}{2 \times 622} = 14}$$

Odds of smoking use among men with lung cancer was **14 times** the odds among men who had other diseases.

Advantages | **Disadvantages**

- Useful for investigating:
  - rare diseases
  - diseases with a long time between exposure and disease
- Advantageous when studying dynamic populations in which follow-up is difficult
- No long follow-up period

- Selection of appropriated controls can be difficult
- Subject to selection bias
- Not very efficient for rare exposures
- Information on exposure is subject to observation/ recall bias
- Cannot establish incidence – because they are retrospective in nature

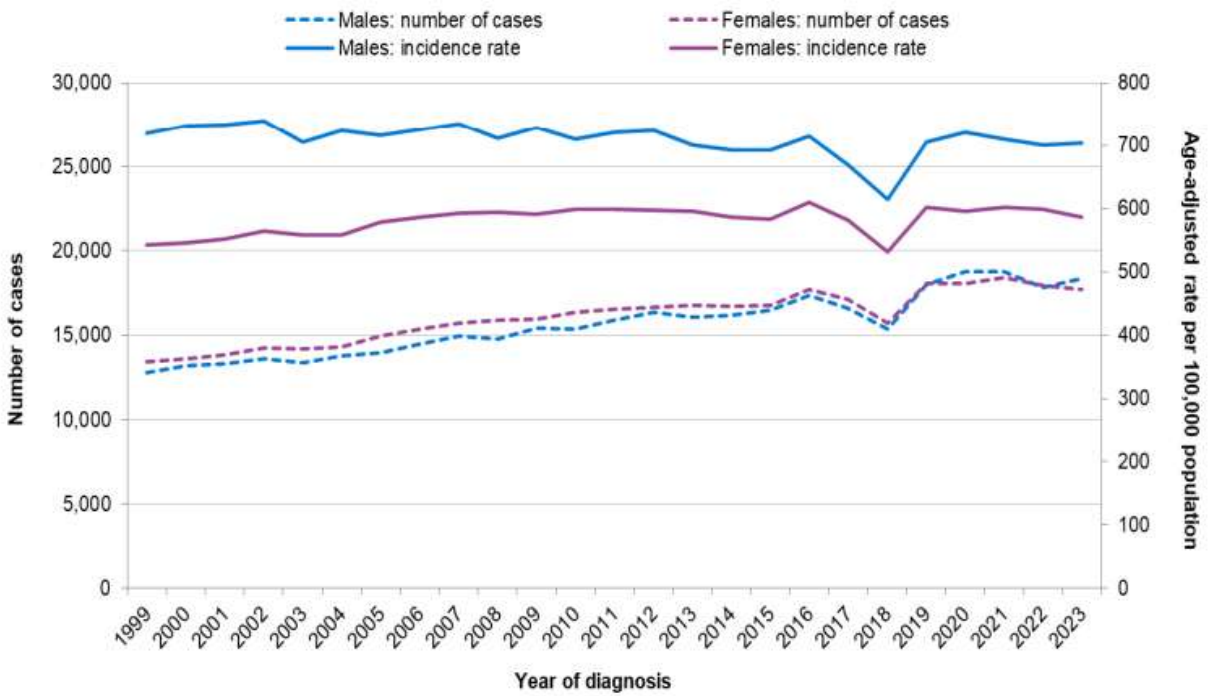
Routine Data

- Routinely collected administrative data:
  - Cancer registration (SMR06)
  - Hospital discharge records (SMR01)
  - Death certification (NRS/ GRO)

- Can be used for epidemiological purposes
- Provide insights into the health of the population
  - Cancer registration → Cancer Incidence
  - Death certification → Annual death rates, Survival rates

- These rates are usually age standardized to take account of differences in the age structure of the population over time or between places.

Figure 1: Number of cancer[1] registrations and age-adjusted incidence rate[2,3] in 1999-2023, by sex



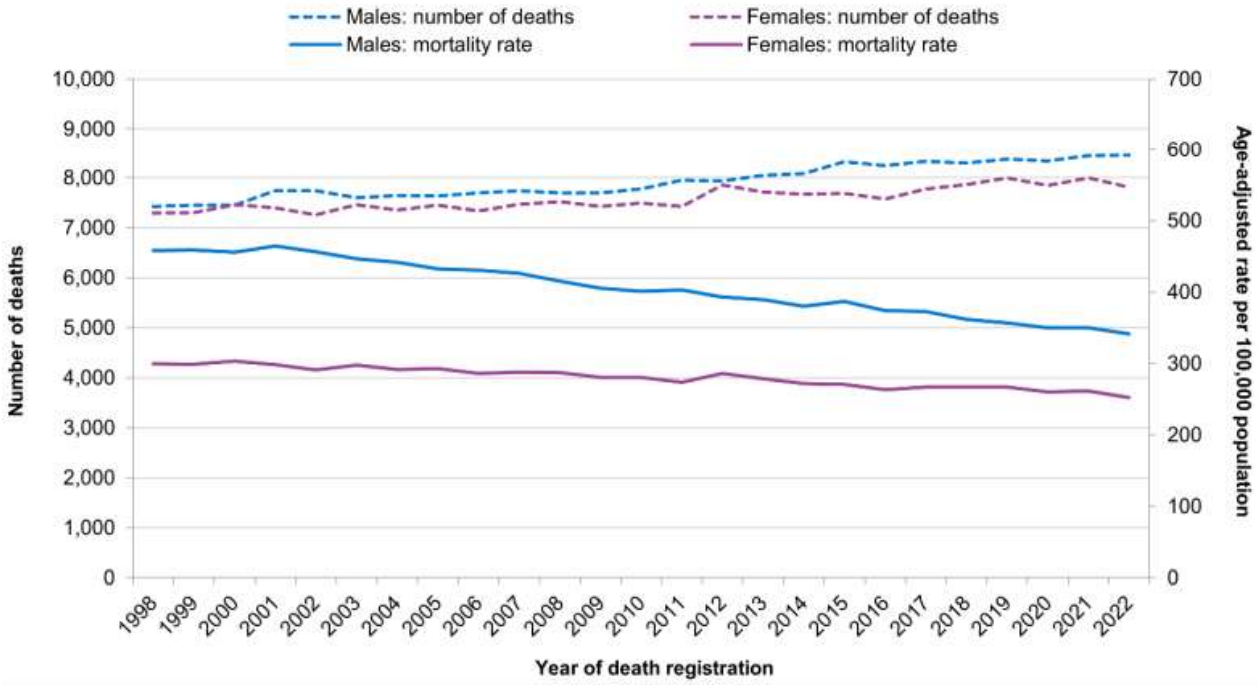
Source: Scottish Cancer Registry, Public Health Scotland (PHS)

1. All cancers excluding non-melanoma skin cancers (ICD-10 C00-C97 excluding C44).

2. EASR: European age- and sex-standardised incidence rate per 100,000 person years. See **Glossary** for further details.

https://publichealthscotland.scot/media/35152/20250930-cancer-incidence-2023-report-finalv.pdf

Figure 1: Cancer[1] mortality in Scotland, 1998-2022. Number of deaths and EASR[2] by sex.



Source: National Records of Scotland (NRS)

1. All cancers excluding non-melanoma skin cancer (ICD-10 C00-C97 excl. C44).

2. EASR: European age- and sex-standardised mortality rate per 100,000 population. See Glossary for further details.

https://publichealthscotland.scot/media/34451/2025-08-19-cancer-mortality-report-final.pdf

Mortality Rates

Infant mortality rate = $\dfrac{\text{number of deaths in year <1 year of age}}{\text{number of live births in year}} \times 1000$

Perinatal mortality rate = $\dfrac{\text{number of stillbirths + deaths <7 days in a year}}{\text{total number of births (live + still) in a year}} \times 1000$

Neonatal mortality rate = $\dfrac{\text{number of deaths in year < 28 days of age}}{\text{number of live births in a year}}$

Case fatality rate (%) = $\dfrac{\text{number of deaths in year (from specific disease}}{\text{number of cases of that disease in a year}} \times 100$

Proportional mortality rate due to TB = $\dfrac{\text{total number of deaths due to TB}}{\text{total number of deaths due to all causes}} \times 100$

Age/sex specific mortality rate = $\dfrac{\text{number of deaths in year in specific age/sex group}}{\text{mid-year population of age or sex group}} \times 1000$

Age Standardized Mortality Rates – why it is important?

- Is there a difference in the mortality rate in two separate groups of people?
- **Example:** We have data on a sample of people who had a specific treatment and some of whom who did not.
- We observe how many die within a specific time interval.

	Number of people	Number who die	**Mortality rate** (deaths per 100,000)
Treated	1000	40	4000
Not treated	1000	20	2000

- But this mortality rate does not tell the full story
- If the data is observational data – rather than from a RCT – there may have been proportionally more older people treated than not treated.
- We need to adjust the mortality rate to take account of the different age distributions in the two groups.

Total COVID-19 cases and deaths in June 2021 (*source:PHE*)

	COVID cases	Deaths	COVID Mortality rate (deaths per 100,000)
Vaccinated	27,197	70	257
Unvaccinated	53,822	44	82

Three times the rate in unvaccinated

Aged 50+

	COVID cases	Deaths	COVID Mortality rate (deaths per 100,000)
Vaccinated	7,499	68	907
Unvaccinated	976	38	3,893

Aged <50

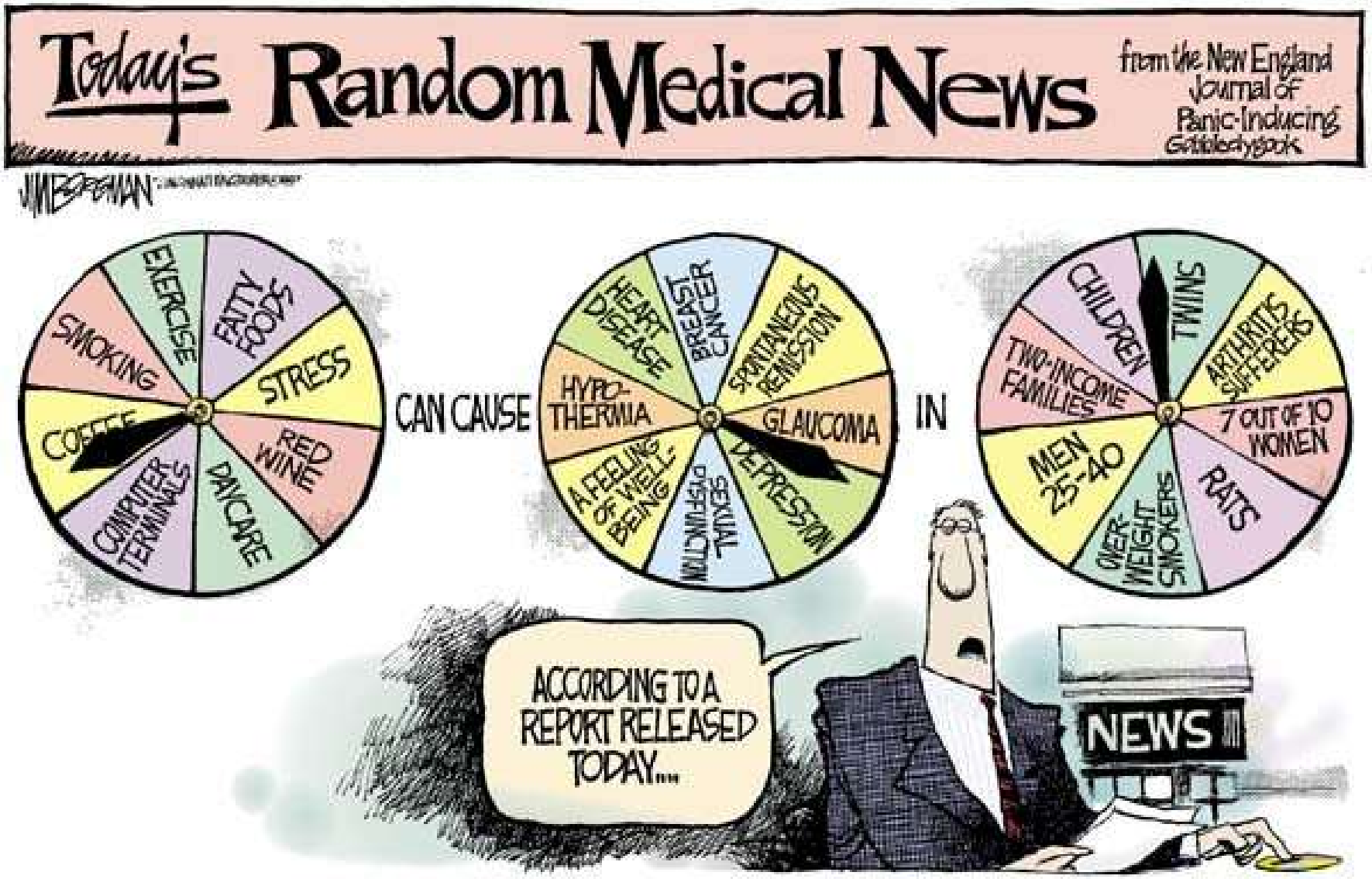
	COVID cases	Deaths	COVID Mortality rate (deaths per 100,000)
Vaccinated	19,693	2	10
Unvaccinated	52,846	6	11

In both age groups the mortality rate is higher in the unvaccinated

Initial observation is explained because the vaccinated group are older compared to the unvaccinated group

Age Standardized Mortality Rates

- Takes account of differences in the age structure of different groups.
- A weighted average of the different age categorised mortality rates.
- A weighted average of age specific rates in the COVID example will conclude that mortality rates in the unvaccinated are higher than in the vaccinated.
- Generally, weights are based on a ‘standard population’ size for each age category.
- For example, the ‘European standard population’
- Doing this will control for differences in the age structure of the two groups being compared



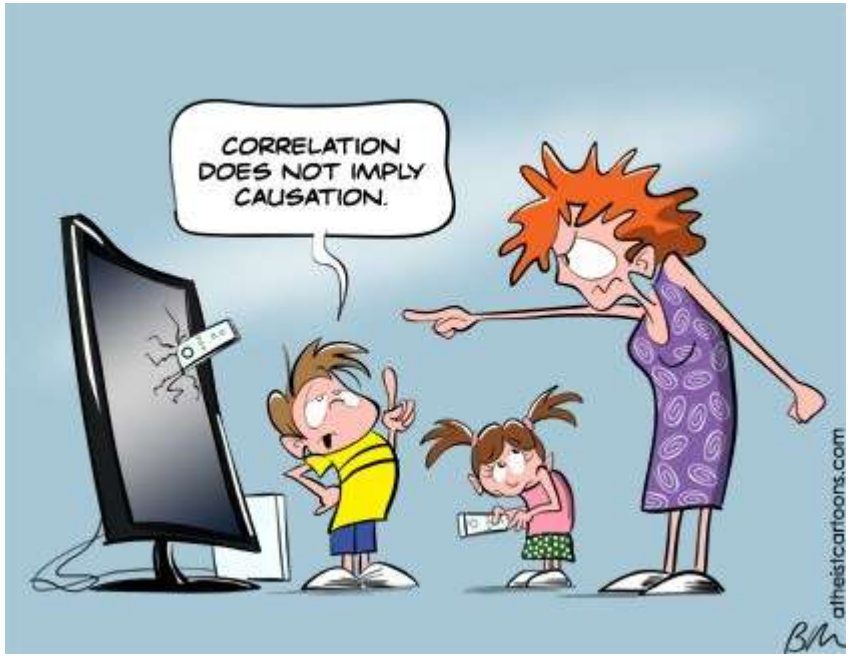
Cartoon by Jim Borgman, first published by the Cincinnati Inquirer and King Features Syndicate 1997 Apr 27; Forum section: 1 and reprinted in the New York Times, 27 April 1997, E4.

Causation

- Epidemiological studies cannot prove causation

- Cannot prove that a specific risk factor actually causes the disease being studied

- Can only show that a risk factor is **associated** with a higher incidence of disease in the population exposed to that risk factor



The Bradford-Hill criteria (J Roy Soc Med 1965:58:295-300)

1. Strength of the association. - The stronger the association between a risk factor and outcome, the more likely the relationship is to be causal. **(Effect size)**

2. Consistency of findings.- Have the same findings must be observed among different populations, in different study designs and different times? **(Reproducibility)**

3. Specificity of the association. - There must be a one-to-one relationship between factor (cause) and outcome (effect).

4. Temporal sequence of association. - Exposure must precede outcome.

5. Biological gradient. - Change in disease rates should follow from corresponding changes in exposure (dose-response). (Greater exposure → Greater effect)

6. Biological plausibility.- Presence of a potential biological mechanism.

7. Coherence.- Does the relationship agree with the current knowledge of the natural history/biology of the disease? (Coherence between epidemiological and laboratory findings)

8. Experiment.- Does the removal of the exposure alter the frequency of the outcome?

9. Analogy. - The effect of similar factors may be considered.


QUIZ TIME

Q1. Select all of the following statements which you believe to be true. Longitudinal studies:

A. Are either prospective or retrospective.

B. Are either experimental or observational.

C. Are particularly suitable for estimating the point prevalence of a condition.

D. Cannot be used to estimate the incidence of a disease.

E. Can be used for assessing causality.

Q2. Select all of the following studies that are repeated cross-sectional studies.

A. The UK national census, which takes place every 10 years.

B. A natural history study of individuals infected with hepatitis C virus followed from the time of diagnosis for 5 years.

C. A study of dietary patterns of first year medical students in the first week of October carried out for five consecutive years.

D. A study to consider the incidence of AIDS events in patients infected with HIV.

Q3. A cohort study has the following advantages:

A. It can be used to study the exposure to factors that are rare.

B. It is relatively cheap to perform because it follows a defined group of individuals.

C. It is unusual to have losses to follow-up because it follows a defined group of individuals.

D. It requires a reasonably small sample size if the outcome is rare.

E. The time sequence of events can be assessed.

Q4. The relative risk of a disease:

A. Always lies between zero and one.

B. Is always positive.

C. Measures the increased (or decreased) risk of the factor when the individual has the disease.

D. Measures the risk of the disease in the population.

E. Takes the value zero when the risk is equally likely in those exposed and unexposed to the factor of interest.

Q5. A case-control study may suffer from the following disadvantages:

A. It is not suitable for rare disease outcomes.

B. It is not suitable when the exposures to the risk factor are rare.

C. It is relatively expensive to perform.

D. It is limited to investigating only one risk factor.

E. It does not allow the direct evaluation of the relative risk.

Q6. The odds ratio:

A. Is an estimate of the relative risk when the incidence of the disease is rare.

B. Is calculated in a case-control study because the relative risk cannot be estimated directly.

C. Is equal to zero when the odds of being a case in the exposed and unexposed groups are equal.

D. Is the ratio of the probability of being a case in the exposed group to the probability of not being a case in exposed group.

E. Cannot be negative.

References

- Previous year's course materials by Philip McLoone
- Introduction to Medical Statistics - The Beatson West of Scotland Cancer Centre
- Medical Statistics at a Glance – Petrie & Sabin 3rd Ed 2009 (Wiley-Blackwell)
- Medical Statistics at a Glance – WORKBOOK (Quiz)