

# Nordic Summerschool of Cancer Epidemiology

**Bendix Carstensen** Steno Diabetes Center  
Gentofte, Denmark  
<http://BendixCarstensen.com>

**Esa Läärä** University of Oulu  
Oulu, Finland

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

# Introduction

**Bendix Carstensen & Esa Läärä**

Nordic Summerschool of  
Cancer Epidemiology

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

intro-measures

## Key references

- IS: dos Santos Silva, I. (1999).  
*Cancer Epidemiology: Principles and Methods*. IARC, Lyon.

## Key references

IS: dos Santos Silva, I. (1999).

*Cancer Epidemiology: Principles and Methods*. IARC, Lyon.

B&D: Breslow, N.E., Day, N.E. (1987).

*Statistical Methods in Cancer Research Volume II – The Design and Analysis of Cohort Studies*.

IARC Scientific Publications No. 82, IARC, Lyon.

## Key references

IS: dos Santos Silva, I. (1999).

*Cancer Epidemiology: Principles and Methods*. IARC, Lyon.

B&D: Breslow, N.E., Day, N.E. (1987).

*Statistical Methods in Cancer Research Volume II – The Design and Analysis of Cohort Studies*.

IARC Scientific Publications No. 82, IARC, Lyon.

C&H: Clayton, D., Hills, M. (1993).

*Statistical Models in Epidemiology*. OUP, Oxford.

## Key references

- IS: dos Santos Silva, I. (1999).  
*Cancer Epidemiology: Principles and Methods*. IARC, Lyon.
- B&D: Breslow, N.E., Day, N.E. (1987).  
*Statistical Methods in Cancer Research Volume II – The Design and Analysis of Cohort Studies*.  
IARC Scientific Publications No. 82, IARC, Lyon.
- C&H: Clayton, D., Hills, M. (1993).  
*Statistical Models in Epidemiology*. OUP, Oxford.
- BxC: Carstensen, B (2021).  
*Epidemiology with R*. OUP, Oxford.  
<http://bendixcarstensen.com/EwR>

## Internet resources on cancer statistics

**NORDCAN** : Cancer Incidence and Mortality in the Nordic Countries, Version 4.0. Association of Nordic Cancer Registries, Danish Cancer Society, 2002. <http://www-dep.iarc.fr/nordcan.htm>

NORDCAN is a graphical package providing data on the incidence of, and mortality from 40 major cancers for 80 regions of the Nordic countries (Denmark, Finland, Iceland, Norway and Sweden). Using NORDCAN, these data can be presented as a variety of tables and graphs that can be easily exported or printed. NORDCAN allows countries and cancer sites to be grouped and compared as desired.

## Internet resources on cancer statistics

**NORDCAN** : Cancer Incidence and Mortality in the Nordic Countries, Version 4.0. Association of Nordic Cancer Registries, Danish Cancer Society, 2002. <http://www-dep.iarc.fr/nordcan.htm>

NORDCAN is a graphical package providing data on the incidence of, and mortality from 40 major cancers for 80 regions of the Nordic countries (Denmark, Finland, Iceland, Norway and Sweden). Using NORDCAN, these data can be presented as a variety of tables and graphs that can be easily exported or printed. NORDCAN allows countries and cancer sites to be grouped and compared as desired.

**GLOBOCAN 2008** : Cancer Incidence and Mortality Worldwide in 2008  
<http://globocan.iarc.fr/>



# Basic Concepts

**Bendix Carstensen & Esa Läärä**

Nordic Summerschool of  
Cancer Epidemiology

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

basic-concepts

# What is Epidemiology?

Some textbook definitions of epidemiology:

Greek: *epi* = upon, *demos* = people

- ▶ “study of the **distribution** and **determinants** of disease **frequency** in man” (MacMahon and Pugh, 1970)

# What is Epidemiology?

Some textbook definitions of epidemiology:

Greek: *epi* = upon, *demos* = people

- ▶ “study of the **distribution** and **determinants** of disease **frequency** in man” (MacMahon and Pugh, 1970)
- ▶ “study of the distribution and determinants of health related **states** and **events** in specified populations, . . .” (Last (ed.) Dictionary of Epidemiology, 2000)

# What is Epidemiology?

Some textbook definitions of epidemiology:

Greek: *epi* = upon, *demos* = people

- ▶ “study of the **distribution** and **determinants** of disease **frequency** in man” (MacMahon and Pugh, 1970)
- ▶ “study of the distribution and determinants of health related **states** and **events** in specified populations, . . .” (Last (ed.) Dictionary of Epidemiology, 2000)
- ▶ “discipline on principles of **occurrence** research in medicine” (Miettinen, 1985)

# Different epidemiologies

- ▶ descriptive epidemiology

# Different epidemiologies

- ▶ **descriptive** epidemiology
  - ▶ monitoring & surveillance of diseases for planning of health services

# Different epidemiologies

- ▶ **descriptive** epidemiology
  - ▶ monitoring & surveillance of diseases for planning of health services
  - ▶ a major activity of cancer registries, and other health data collectors

# Different epidemiologies

- ▶ **descriptive** epidemiology
  - ▶ monitoring & surveillance of diseases for planning of health services
  - ▶ a major activity of cancer registries, and other health data collectors
- ▶ **etiologic** or “analytic” epidemiology:



# Different epidemiologies

- ▶ **descriptive** epidemiology
  - ▶ monitoring & surveillance of diseases for planning of health services
  - ▶ a major activity of cancer registries, and other health data collectors
- ▶ **etiologic** or “analytic” epidemiology:
  - ▶ study of cause-effect relationships

## Different epidemiologies

- ▶ **disease** epidemiologies — e.g. of cancer, cardiovascular diseases, infectious diseases, musculoskeletal disorders, mental health, ...

## Different epidemiologies

- ▶ **disease** epidemiologies — *e.g.* of cancer, cardiovascular diseases, infectious diseases, musculoskeletal disorders, mental health, ...
- ▶ **determinant-based** epidemiologies

## Different epidemiologies

- ▶ **disease** epidemiologies — e.g. of cancer, cardiovascular diseases, infectious diseases, musculoskeletal disorders, mental health, ...
- ▶ **determinant-based** epidemiologies
  - ▶ occupational epidemiology

## Different epidemiologies

- ▶ **disease** epidemiologies — e.g. of cancer, cardiovascular diseases, infectious diseases, musculoskeletal disorders, mental health, ...
- ▶ **determinant-based** epidemiologies
  - ▶ occupational epidemiology
  - ▶ nutritional epidemiology

## Different epidemiologies

- ▶ **disease** epidemiologies — e.g. of cancer, cardiovascular diseases, infectious diseases, musculoskeletal disorders, mental health, ...
- ▶ **determinant-based** epidemiologies
  - ▶ occupational epidemiology
  - ▶ nutritional epidemiology
  - ▶ ...

# Different epidemiologies

- ▶ **disease** epidemiologies — e.g. of cancer, cardiovascular diseases, infectious diseases, musculoskeletal disorders, mental health, ...
- ▶ **determinant-based** epidemiologies
  - ▶ occupational epidemiology
  - ▶ nutritional epidemiology
  - ▶ ...
- ▶ **clinical** epidemiology

# Different epidemiologies

- ▶ **disease** epidemiologies — e.g. of cancer, cardiovascular diseases, infectious diseases, musculoskeletal disorders, mental health, ...
- ▶ **determinant-based** epidemiologies
  - ▶ occupational epidemiology
  - ▶ nutritional epidemiology
  - ▶ ...
- ▶ **clinical** epidemiology
  - ▶ study of diagnosis, prognosis and effectiveness of therapies in patient populations



# Different epidemiologies

- ▶ **disease** epidemiologies — e.g. of cancer, cardiovascular diseases, infectious diseases, musculoskeletal disorders, mental health, ...
- ▶ **determinant-based** epidemiologies
  - ▶ occupational epidemiology
  - ▶ nutritional epidemiology
  - ▶ ...
- ▶ **clinical** epidemiology
  - ▶ study of diagnosis, prognosis and effectiveness of therapies in patient populations
  - ▶ basis of evidence-based medicine

# Different epidemiologies

- ▶ **disease** epidemiologies — e.g. of cancer, cardiovascular diseases, infectious diseases, musculoskeletal disorders, mental health, ...
- ▶ **determinant-based** epidemiologies
  - ▶ occupational epidemiology
  - ▶ nutritional epidemiology
  - ▶ ...
- ▶ **clinical** epidemiology
  - ▶ study of diagnosis, prognosis and effectiveness of therapies in patient populations
  - ▶ basis of evidence-based medicine
  - ▶ essential in health-services research

## Cancer i Norden 1997 (NORDCAN)

Frequency of cancer (all sites excl. non-melanoma skin) in Nordic male populations expressed by different measures:

|         | New cases     | Crude rate | ASR (World) | Cumul. risk | SIR        |
|---------|---------------|------------|-------------|-------------|------------|
| Denmark | 11,787        | 452        | 281         | 27.8        | 104        |
| Finland | 10,058        | <u>401</u> | 269         | 26.5        | 101        |
| Iceland | <u>633</u>    | 464        | <b>347</b>  | <b>32.6</b> | <b>132</b> |
| Norway  | 10,246        | <b>469</b> | 294         | 29.4        | 109        |
| Sweden  | <b>19 908</b> | 455        | <u>249</u>  | <u>25.4</u> | <u>93</u>  |

## Cancer i Norden 1997 (NORDCAN)

Frequency of cancer (all sites excl. non-melanoma skin) in Nordic male populations expressed by different measures:

|         | New cases     | Crude rate | ASR (World) | Cumul. risk | SIR        |
|---------|---------------|------------|-------------|-------------|------------|
| Denmark | 11,787        | 452        | 281         | 27.8        | 104        |
| Finland | 10,058        | <u>401</u> | 269         | 26.5        | 101        |
| Iceland | <u>633</u>    | 464        | <b>347</b>  | <b>32.6</b> | <b>132</b> |
| Norway  | 10,246        | <b>469</b> | 294         | 29.4        | 109        |
| Sweden  | <b>19 908</b> | 455        | <u>249</u>  | <u>25.4</u> | <u>93</u>  |

- ▶ Where is the frequency truly **highest**, where **lowest**?

## Questions on frequency & occurrence

How many women in Denmark:

- ▶ are carriers of breast cancer today? — **prevalence**

## Questions on frequency & occurrence

How many women in Denmark:

- ▶ are carriers of breast cancer today? — **prevalence**
- ▶ will contract a new breast ca. during 2007? — **incidence**

## Questions on frequency & occurrence

How many women in Denmark:

- ▶ are carriers of breast cancer today? — **prevalence**
- ▶ will contract a new breast ca. during 2007? — **incidence**
- ▶ die from breast ca. in 2007? — **mortality**

## Questions on frequency & occurrence

How many women in Denmark:

- ▶ are carriers of breast cancer today? — **prevalence**
- ▶ will contract a new breast ca. during 2007? — **incidence**
- ▶ die from breast ca. in 2007? — **mortality**
- ▶ will be alive after 5 years since diagnosis among those getting breast ca. in 2007? — **survival**



## Questions on frequency & occurrence

How many women in Denmark:

- ▶ are carriers of breast cancer today? — **prevalence**
- ▶ will contract a new breast ca. during 2007? — **incidence**
- ▶ die from breast ca. in 2007? — **mortality**
- ▶ will be alive after 5 years since diagnosis among those getting breast ca. in 2007? — **survival**
- ▶ are cured from breast cancer during 2007? — **cure**

## Questions on frequency & occurrence

How many women in Denmark:

- ▶ are carriers of breast cancer today? — **prevalence**
- ▶ will contract a new breast ca. during 2007? — **incidence**
- ▶ die from breast ca. in 2007? — **mortality**
- ▶ will be alive after 5 years since diagnosis among those getting breast ca. in 2007? — **survival**
- ▶ are cured from breast cancer during 2007? — **cure**

## Questions on frequency & occurrence

How many women in Denmark:

- ▶ are carriers of breast cancer today? — **prevalence**
- ▶ will contract a new breast ca. during 2007? — **incidence**
- ▶ die from breast ca. in 2007? — **mortality**
- ▶ will be alive after 5 years since diagnosis among those getting breast ca. in 2007? — **survival**
- ▶ are cured from breast cancer during 2007? — **cure**

What are the proportions / rates of these?

What is the **dimension** (units) of these measures?

## What is risk?

What do we mean by “risk of disease  $S$ ”?

- (a) **probability** of *getting*  $S$  during a given **risk period**  
→ **incidence** probability, (cumulative risk over the period)

## What is risk?

What do we mean by “risk of disease  $S$ ”?

- (a) **probability** of *getting*  $S$  during a given **risk period**  
→ **incidence** probability, (cumulative risk over the period)
- (b) **rate** of that probability (relative to risk period)  
→ **hazard** or intensity,

## What is risk?

What do we mean by “risk of disease  $S$ ”?

- (a) **probability** of *getting*  $S$  during a given **risk period**  
→ **incidence** probability, (cumulative risk over the period)
- (b) **rate** of that probability (relative to risk period)  
→ **hazard** or intensity,
- (c) **probability** of *carrying*  $S$  at a given *time point*

## What is risk?

What do we mean by “risk of disease  $S$ ”?

- (a) **probability** of *getting*  $S$  during a given **risk period**  
→ **incidence** probability, (cumulative risk over the period)
- (b) **rate** of that probability (relative to risk period)  
→ **hazard** or intensity,
- (c) **probability** of *carrying*  $S$  at a given *time point*

## What is risk?

What do we mean by “risk of disease  $S$ ”?

- (a) **probability** of *getting*  $S$  during a given **risk period**  
→ **incidence** probability, (cumulative risk over the period)
- (b) **rate** of that probability (relative to risk period)  
→ **hazard** or intensity,
- (c) **probability** of *carrying*  $S$  at a given *time point*  
→ **prevalence** probability.



## What is risk?

What do we mean by “risk of disease  $S$ ”?

- (a) **probability** of *getting*  $S$  during a given **risk period**  
→ **incidence** probability, (cumulative risk over the period)
- (b) **rate** of that probability (relative to risk period)  
→ **hazard** or intensity,
- (c) **probability** of *carrying*  $S$  at a given *time point*  
→ **prevalence** probability.

Most common use of “risk” is (a)

## What is risk?

What do we mean by “risk of disease  $S$ ”?

- (a) **probability** of *getting*  $S$  during a given **risk period**  
→ **incidence** probability, (cumulative risk over the period)
- (b) **rate** of that probability (relative to risk period)  
→ **hazard** or intensity,
- (c) **probability** of *carrying*  $S$  at a given *time point*  
→ **prevalence** probability.

Most common use of “risk” is (a)

**NB:** “Risk” should not be used in the meaning of **risk factor**

In statistics, “hazard” mostly refers to notion probability per unit time.

# Risks are conditional probabilities

## Risks are conditional probabilities

- ▶ All risks are conditional on a multitude of factors, like:

## Risks are conditional probabilities

- ▶ All risks are conditional on a multitude of factors, like:
  - ▶ length of risk period (e.g. next week or lifetime),

# Risks are conditional probabilities

- ▶ All risks are conditional on a multitude of factors, like:
  - ▶ length of risk period (e.g. next week or lifetime),
  - ▶ age and sex,

# Risks are conditional probabilities

- ▶ All risks are conditional on a multitude of factors, like:
  - ▶ length of risk period (e.g. next week or lifetime),
  - ▶ age and sex,
  - ▶ genetic constitution,

# Risks are conditional probabilities

- ▶ All risks are conditional on a multitude of factors, like:
  - ▶ length of risk period (e.g. next week or lifetime),
  - ▶ age and sex,
  - ▶ genetic constitution,
  - ▶ health behaviour & environmental exposures.



# Risks are conditional probabilities

- ▶ All risks are conditional on a multitude of factors, like:
  - ▶ length of risk period (e.g. next week or lifetime),
  - ▶ age and sex,
  - ▶ genetic constitution,
  - ▶ health behaviour & environmental exposures.
- ▶ In principle each individual has a “personal” value for the risk of given disease in any defined risk period, depending on his/her own risk factor profile—not estimable from data

## Risks are conditional probabilities

- ▶ All risks are conditional on a multitude of factors, like:
  - ▶ length of risk period (e.g. next week or lifetime),
  - ▶ age and sex,
  - ▶ genetic constitution,
  - ▶ health behaviour & environmental exposures.
- ▶ In principle each individual has a “personal” value for the risk of given disease in any defined risk period, depending on his/her own risk factor profile—not estimable from data
- ▶ **Average risks** of disease in large groups sharing common characteristics (like gender, age, smoking status) are estimable through **measures of occurrence**.

# Mathematical reminder

**Bendix Carstensen & Esa Läärä**

Nordic Summerschool of  
Cancer Epidemiology

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

math

## Logarithms and exponentials

$$10^2 = 10 \times 10$$

$$10^3 = 10 \times 10 \times 10$$

$$10^2 \times 10^3 = 10^5$$

$$10^3/10^2 = 10^1$$

$$(10^3)^2 = 10^6$$

---

$$10^2/10^2 = 10^0 = 1$$

$$10^2/10^3 = 10^{-1} = 1/10$$

$$10^{1/2} \times 10^{1/2} = 10^1$$

$$10^{1/3} \times 10^{1/3} \times 10^{1/3} = 10^1$$

---

$$10^{0.3010} = 2$$

$$\log_{10}(2) = 0.3010$$

---

$$\text{Mathematical reminder (math)} \quad 10^{0.4771} = 3$$

## Multiplication and division

$$2 \times 3 = 6$$

---

$$10^{0.3010} \times 10^{0.4771} = 10^{0.7781}$$

$$10^{0.7781} = 6$$

---

$$\log_{10}(2) = 0.3010$$

$$\log_{10}(3) = 0.4771$$

$$0.3010 + 0.4771 = 0.7781$$

$$10^{0.7781} = 6$$

---

In general:  $\log(xy) = \log(x) + \log(y)$

$$\log(x/y) = \log(x) - \log(y)$$

$$\log(x^a) = a\log(x)$$

## Natural logarithms $e = 2.7183$

$$\log_e(e) = 1$$

---

$$e^{0.6931} = 2$$

$$\log_e(2) = 0.6931$$

---

$$e^{1.0986} = 3$$

$$\log_e(3) = 1.0986$$

---

$$2 \times 3 = 6$$

---

$$e^{0.6931} \times e^{1.0986} = e^{1.7918}$$

$$e^{1.7918} = 6$$

---

In general:  $e^x \times e^y = e^{x+y}$

$$e^x / e^y = e^{x-y}$$

$$(e^x)^y = e^{x \times y}$$

## Names for the logarithms

### Engineers and calculators:

$\log$  is the logarithm to base 10.

$\ln$  is the logarithm to base  $e$ , the natural log

### Mathematicians:

$\log$  is the logarithm to base  $e$ , the natural log

$\log_{10}$  is the logarithm to base 10.

We use  $\log$  for the natural logarithm, and explicitly  $\log_{10}$  when this is needed.

## Why natural logarithms?

For small values of  $x$ :  $e^x \approx 1 + x$

$$e^{-x} \approx 1 - x$$

$$\ln(1 + x) \approx x$$

$$\ln(1 - x) \approx -x$$

---

For example:  $\ln(1.01) = 0.01$

$$\ln(0.99) = -0.01$$

---

**But:**  $\log_{10}(1.01) = 0.4343 \times 0.01$

$$\log_{10}(0.99) = 0.4343 \times -0.01$$

---

In general:  $\log_{10}(x) = 0.4343 \times \ln(x)$



# Frequency measures

**Bendix Carstensen & Esa Läärä**

Nordic Summerschool of  
Cancer Epidemiology

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

ELmeasures

## Basic measures of frequency or occurrence

Quantification of the occurrence of disease (or any other health-related state or event) requires specification of:

1. what is meant by a **case**, *i.e.*, an individual in a population who has or gets the disease
  - more generally: possesses the state or undergoes the event of interest.
  - ⇒ challenge to accurate diagnosis and classification!

## Basic measures of frequency or occurrence

Quantification of the occurrence of disease (or any other health-related state or event) requires specification of:

1. what is meant by a **case**, *i.e.*, an individual in a population who has or gets the disease  
— more generally: possesses the state or undergoes the event of interest.  
⇒ challenge to accurate diagnosis and classification!
2. the **population** from which the cases originate.

## Basic measures of frequency or occurrence

Quantification of the occurrence of disease (or any other health-related state or event) requires specification of:

1. what is meant by a **case**, *i.e.*, an individual in a population who has or gets the disease  
— more generally: possesses the state or undergoes the event of interest.  
⇒ challenge to accurate diagnosis and classification!
2. the **population** from which the cases originate.
3. the **time point** or **period** of observation.

## Types of occurrence measures

- ▶ Longitudinal – **incidence** measures: incidence rate & incidence proportion

## Types of occurrence measures

- ▶ Longitudinal – **incidence** measures: incidence rate & incidence proportion
- ▶ Cross-sectional – **prevalence** measures.

## Types of occurrence measures

- ▶ Longitudinal – **incidence** measures: incidence rate & incidence proportion
- ▶ Cross-sectional – **prevalence** measures.

## Types of occurrence measures

- ▶ Longitudinal – **incidence** measures: incidence rate & incidence proportion
- ▶ Cross-sectional – **prevalence** measures.

General form of frequency or occurrence measures

$$\frac{\text{numerator}}{\text{denominator}}$$

**Numerator:** number of cases observed in the population.



## Types of occurrence measures

- ▶ Longitudinal – **incidence** measures: incidence rate & incidence proportion
- ▶ Cross-sectional – **prevalence** measures.

General form of frequency or occurrence measures

$$\frac{\text{numerator}}{\text{denominator}}$$

**Numerator:** number of cases observed in the population.

**Denominator:** generally proportional to the size of the population from which the cases emerge.

## Types of occurrence measures

- ▶ Longitudinal – **incidence** measures: incidence rate & incidence proportion
- ▶ Cross-sectional – **prevalence** measures.

General form of frequency or occurrence measures

$$\frac{\text{numerator}}{\text{denominator}}$$

**Numerator:** number of cases observed in the population.

**Denominator:** generally proportional to the size of the population from which the cases emerge.

Numerator and denominator must cover the same **population**, and the same **period** or same **time point**.

## Incidence measures

- ▶ **Incidence proportion** (cumulative risk) ( $Q$ ) over a fixed *risk period*:

$$Q = \frac{\text{number of incident (new) cases during period}}{\text{size of pop'n at risk at start of the period}}$$

Also called **cumulative incidence** (even “risk”; e.g. in **IS**).

**NB.** “Cumulative incidence” has other meanings, too.

## Incidence measures

- ▶ **Incidence proportion** (cumulative risk) ( $Q$ ) over a fixed *risk period*:

$$Q = \frac{\text{number of incident (new) cases during period}}{\text{size of pop'n at risk at start of the period}}$$

Also called **cumulative incidence** (even “risk”; e.g. in **IS**).

**NB.** “Cumulative incidence” has other meanings, too.

- ▶ **Incidence rate** ( $I$ ) over a defined observation period:

## Incidence measures

- ▶ **Incidence proportion** (cumulative risk) ( $Q$ ) over a fixed *risk period*:

$$Q = \frac{\text{number of incident (new) cases during period}}{\text{size of pop'n at risk at start of the period}}$$

Also called **cumulative incidence** (even “risk”; e.g. in **IS**).

**NB.** “Cumulative incidence” has other meanings, too.

- ▶ **Incidence rate** ( $I$ ) over a defined observation period:

## Incidence measures

- ▶ **Incidence proportion** (cumulative risk) ( $Q$ ) over a fixed *risk period*:

$$Q = \frac{\text{number of incident (new) cases during period}}{\text{size of pop'n at risk at start of the period}}$$

Also called **cumulative incidence** (even “risk”; e.g. in **IS**).

**NB.** “Cumulative incidence” has other meanings, too.

- ▶ **Incidence rate** ( $I$ ) over a defined observation period:

$$I = \frac{\text{number of incident (new) cases during period}}{\text{sum of follow-up times of pop'n at risk}}$$

## Incidence measures

- ▶ **Incidence proportion** (cumulative risk) ( $Q$ ) over a fixed *risk period*:

$$Q = \frac{\text{number of incident (new) cases during period}}{\text{size of pop'n at risk at start of the period}}$$

Also called **cumulative incidence** (even “risk”; e.g. in **IS**).

**NB.** “Cumulative incidence” has other meanings, too.

- ▶ **Incidence rate** ( $I$ ) over a defined observation period:

$$I = \frac{\text{number of incident (new) cases during period}}{\text{sum of follow-up times of pop'n at risk}}$$

Also called **incidence density** or **hazard**.

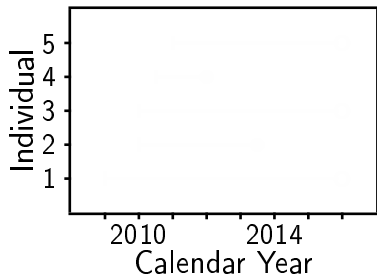
## Example: Follow-up of a small cohort

- | = entry, ○ = exit with censoring; outcome not observed,
- = exit with outcome event (disease onset) observed



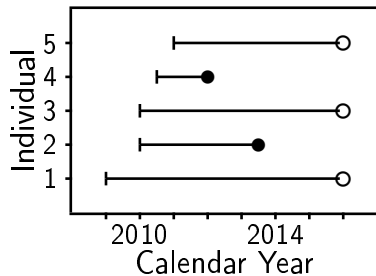
## Example: Follow-up of a small cohort

- | = entry, ○ = exit with censoring; outcome not observed,
- = exit with outcome event (disease onset) observed



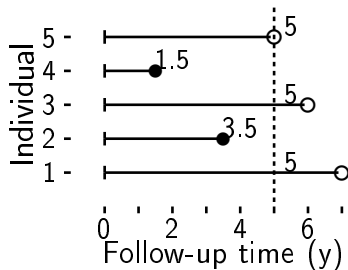
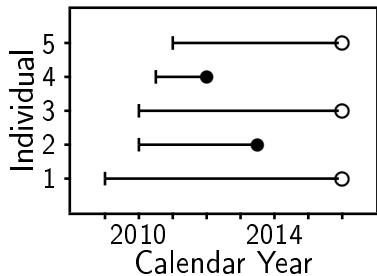
## Example: Follow-up of a small cohort

- | = entry, ○ = exit with censoring; outcome not observed,  
● = exit with outcome event (disease onset) observed



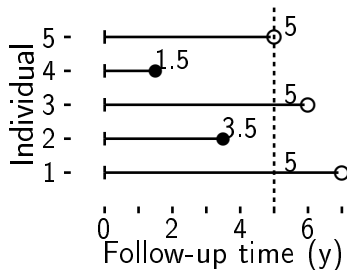
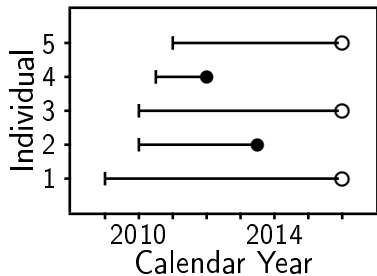
## Example: Follow-up of a small cohort

- | = entry, ○ = exit with censoring; outcome not observed,  
● = exit with outcome event (disease onset) observed



## Example: Follow-up of a small cohort

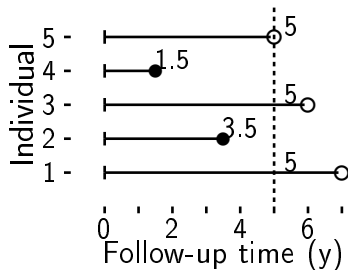
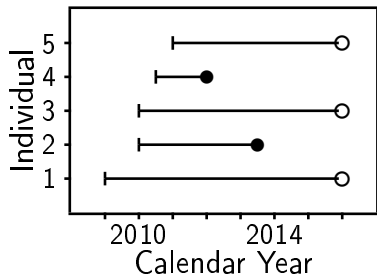
- | = entry, ○ = exit with censoring; outcome not observed,  
● = exit with outcome event (disease onset) observed



Complete follow-up in the 5-year risk period

## Example: Follow-up of a small cohort

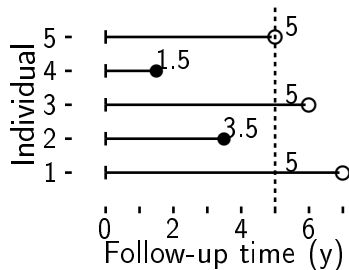
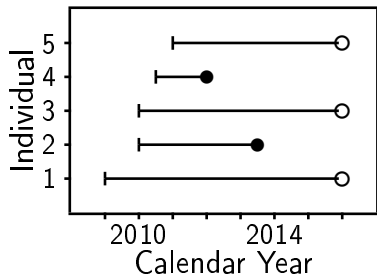
- | = entry, ○ = exit with censoring; outcome not observed,  
● = exit with outcome event (disease onset) observed



**Complete follow-up** in the 5-year risk period  $\Rightarrow$  can calculate both:

## Example: Follow-up of a small cohort

- | = entry, ○ = exit with censoring; outcome not observed,  
● = exit with outcome event (disease onset) observed

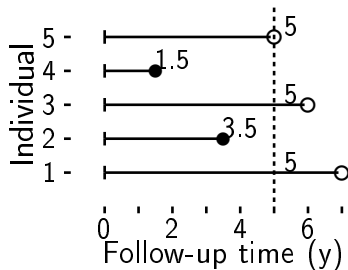
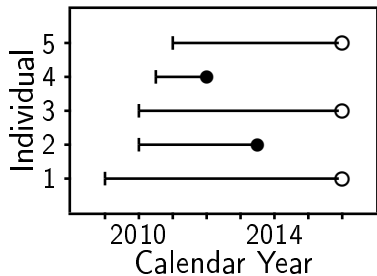


**Complete follow-up** in the 5-year risk period  $\Rightarrow$  can calculate both:

$$\text{Inc. rate} = \frac{2 \text{ cases}}{5 + 3.5 + 5 + 1.5 + 5 \text{ years}} = 10 \text{ per 100 years,}$$

## Example: Follow-up of a small cohort

- | = entry, ○ = exit with censoring; outcome not observed,  
● = exit with outcome event (disease onset) observed



**Complete follow-up** in the 5-year risk period  $\Rightarrow$  can calculate both:

$$\text{Inc. rate} = \frac{2 \text{ cases}}{5 + 3.5 + 5 + 1.5 + 5 \text{ years}} = 10 \text{ per } 100 \text{ years,}$$

$$\text{Inc. prop.} = 2/5 = 0.4 \text{ or } 40 \text{ per cent.}$$

## Properties of incidence proportion (cumulative risk)

- ▶ Dimensionless quantity ranging from 0 to 1  
(0% to 100%) = *relative frequency*,



## Properties of incidence proportion (cumulative risk)

- ▶ Dimensionless quantity ranging from 0 to 1 (0% to 100%) = *relative frequency*,
- ▶ Estimates the average theoretical **risk**, *i.e.* the probability of the outcome occurring during the risk period, in the **population at risk**

## Properties of incidence proportion (cumulative risk)

- ▶ Dimensionless quantity ranging from 0 to 1 (0% to 100%) = *relative frequency*,
- ▶ Estimates the average theoretical **risk**, *i.e.* the probability of the outcome occurring during the risk period, in the **population at risk**

## Properties of incidence proportion (cumulative risk)

- ▶ Dimensionless quantity ranging from 0 to 1 (0% to 100%) = *relative frequency*,
- ▶ Estimates the average theoretical **risk**, *i.e.* the probability of the outcome occurring during the risk period, in the **population at risk** – *i.e.* among those who are still free from the outcome at the start of the period.
- ▶ Simple formula valid when the follow-up time is fixed & equals the risk period, and when there are no **competing events** or **censoring**.

## Properties of incidence proportion (cumulative risk)

- ▶ Dimensionless quantity ranging from 0 to 1 (0% to 100%) = *relative frequency*,
- ▶ Estimates the average theoretical **risk**, *i.e.* the probability of the outcome occurring during the risk period, in the **population at risk** – *i.e.* among those who are still free from the outcome at the start of the period.
- ▶ Simple formula valid when the follow-up time is fixed & equals the risk period, and when there are no **competing events** or **censoring**.
- ▶ Competing events & censoring  $\Rightarrow$   
Calculations need to be corrected using special methods of survival analysis.

## Properties of incidence rate

- ▶ Like a *frequency* quantity in physics; measurement scale is  $\text{time}^{-1}$ :  
e.g. Hz = 1/second, 1/year, or 1/1000 y.

## Properties of incidence rate

- ▶ Like a *frequency* quantity in physics; measurement scale is  $\text{time}^{-1}$ :  
e.g. Hz = 1/second, 1/year, or 1/1000 y.
- ▶ Estimates the average underlying **intensity** or **hazard rate** of the outcome in a population,

## Properties of incidence rate

- ▶ Like a *frequency* quantity in physics; measurement scale is  $\text{time}^{-1}$ :  
e.g. Hz = 1/second, 1/year, or 1/1000 y.
- ▶ Estimates the average underlying **intensity** or **hazard rate** of the outcome in a population,
- ▶ Estimation accurate in the **constant hazard model**,

## Properties of incidence rate

- ▶ Like a *frequency* quantity in physics; measurement scale is  $\text{time}^{-1}$ :  
e.g. Hz = 1/second, 1/year, or 1/1000 y.
- ▶ Estimates the average underlying **intensity** or **hazard rate** of the outcome in a population,
- ▶ Estimation accurate in the **constant hazard model**,
- ▶ Calculation straightforward also with competing events and censored observations.



## Properties of incidence rate

- ▶ Like a *frequency* quantity in physics; measurement scale is  $\text{time}^{-1}$ :  
e.g. Hz = 1/second, 1/year, or 1/1000 y.
- ▶ Estimates the average underlying **intensity** or **hazard rate** of the outcome in a population,
- ▶ Estimation accurate in the **constant hazard model**,
- ▶ Calculation straightforward also with competing events and censored observations.
- ▶ Hazard usually depends on age  $\Rightarrow$  **age-specific** rates needed.

## Properties of incidence rate

- ▶ Like a *frequency* quantity in physics; measurement scale is  $\text{time}^{-1}$ :  
e.g. Hz = 1/second, 1/year, or 1/1000 y.
- ▶ Estimates the average underlying **intensity** or **hazard rate** of the outcome in a population,
- ▶ Estimation accurate in the **constant hazard model**,
- ▶ Calculation straightforward also with competing events and censored observations.
- ▶ Hazard usually depends on age  $\Rightarrow$  **age-specific** rates needed.

## Properties of incidence rate

- ▶ Like a *frequency* quantity in physics; measurement scale is  $\text{time}^{-1}$ :  
e.g. Hz = 1/second, 1/year, or 1/1000 y.
- ▶ Estimates the average underlying **intensity** or **hazard rate** of the outcome in a population,
- ▶ Estimation accurate in the **constant hazard model**,
- ▶ Calculation straightforward also with competing events and censored observations.
- ▶ Hazard usually depends on age  $\Rightarrow$  **age-specific** rates needed.
- ▶ Incidence proportions can be estimated from rates.

## Properties of incidence rate

- ▶ Like a *frequency* quantity in physics; measurement scale is  $\text{time}^{-1}$ :  
e.g. Hz = 1/second, 1/year, or 1/1000 y.
- ▶ Estimates the average underlying **intensity** or **hazard rate** of the outcome in a population,
- ▶ Estimation accurate in the **constant hazard model**,
- ▶ Calculation straightforward also with competing events and censored observations.
- ▶ Hazard usually depends on age  $\Rightarrow$  **age-specific** rates needed.
- ▶ Incidence proportions can be estimated from rates.

## Properties of incidence rate

- ▶ Like a *frequency* quantity in physics; measurement scale is time<sup>-1</sup>:  
e.g. Hz = 1/second, 1/year, or 1/1000 y.
- ▶ Estimates the average underlying **intensity** or **hazard rate** of the outcome in a population,
- ▶ Estimation accurate in the **constant hazard model**,
- ▶ Calculation straightforward also with competing events and censored observations.
- ▶ Hazard usually depends on age  $\Rightarrow$  **age-specific** rates needed.
- ▶ Incidence proportions can be estimated from rates.  
In the constant hazard model with no competing risks:

$$Q = 1 - \exp(-I \times \Delta) \approx I \times \Delta$$

## Competing events and censoring

The outcome event of interest (e.g. onset of disease) is not always observed for all subjects during the chosen risk period.

## Competing events and censoring

The outcome event of interest (e.g. onset of disease) is not always observed for all subjects during the chosen risk period.

- ▶ Some subjects die (from other causes) before the event.

## Competing events and censoring

The outcome event of interest (e.g. onset of disease) is not always observed for all subjects during the chosen risk period.

- ▶ Some subjects die (from other causes) before the event.
  - ▶ **Competing event**, after which the outcome can no more occur.



## Competing events and censoring

The outcome event of interest (e.g. onset of disease) is not always observed for all subjects during the chosen risk period.

- ▶ Some subjects die (from other causes) before the event.
  - ▶ **Competing event**, after which the outcome can no more occur.
- ▶ Others emigrate and escape national disease registration, or the whole study is closed “now”, prematurely interrupting the follow-up of them.

## Competing events and censoring

The outcome event of interest (e.g. onset of disease) is not always observed for all subjects during the chosen risk period.

- ▶ Some subjects die (from other causes) before the event.
  - ▶ **Competing event**, after which the outcome can no more occur.
- ▶ Others emigrate and escape national disease registration, or the whole study is closed “now”, prematurely interrupting the follow-up of them.
  - ▶ **Censoring, withdrawal, or loss to follow-up**

## Competing events and censoring

The outcome event of interest (e.g. onset of disease) is not always observed for all subjects during the chosen risk period.

- ▶ Some subjects die (from other causes) before the event.
  - ▶ **Competing event**, after which the outcome can no more occur.
- ▶ Others emigrate and escape national disease registration, or the whole study is closed “now”, prematurely interrupting the follow-up of them.
  - ▶ **Censoring, withdrawal, or loss to follow-up**

## Competing events and censoring

The outcome event of interest (e.g. onset of disease) is not always observed for all subjects during the chosen risk period.

- ▶ Some subjects die (from other causes) before the event.
  - ▶ **Competing event**, after which the outcome can no more occur.
- ▶ Others emigrate and escape national disease registration, or the whole study is closed “now”, prematurely interrupting the follow-up of them.
  - ▶ **Censoring, withdrawal, or loss to follow-up**

In both cases persons are **removed** from the risk population

## Person-years in dynamic populations

With a **dynamic** study population, individual follow-up times are always variable and impossible to measure accurately.

## Person-years in dynamic populations

With a **dynamic** study population, individual follow-up times are always variable and impossible to measure accurately.

Common approximation – **mid-population** principle:

## Person-years in dynamic populations

With a **dynamic** study population, individual follow-up times are always variable and impossible to measure accurately.

Common approximation – **mid-population** principle:

1. Let the population size be  $N_b$  at beginning ( $t_b$ ) and  $N_e$  at the end ( $t_e$ ) of the observation period of length  $t_e - t_b = u_t$

## Person-years in dynamic populations

With a **dynamic** study population, individual follow-up times are always variable and impossible to measure accurately.

Common approximation – **mid-population** principle:

1. Let the population size be  $N_b$  at beginning ( $t_b$ ) and  $N_e$  at the end ( $t_e$ ) of the observation period of length  $t_e - t_b = u_t$
2. Mid-population for the period around  $t = (t_b + t_e)/2$ :  $\bar{N}_t = \frac{1}{2} \times (N_b + N_e)$ .



## Person-years in dynamic populations

With a **dynamic** study population, individual follow-up times are always variable and impossible to measure accurately.

Common approximation – **mid-population** principle:

1. Let the population size be  $N_b$  at beginning ( $t_b$ ) and  $N_e$  at the end ( $t_e$ ) of the observation period of length  $t_e - t_b = u_t$
2. Mid-population for the period around  $t = (t_b + t_e)/2$ :  $\bar{N}_t = \frac{1}{2} \times (N_b + N_e)$ .
3. Approximate person-years:  $\tilde{Y}_t = \bar{N}_t \times u_t$ .

## Person-years in dynamic populations

With a **dynamic** study population, individual follow-up times are always variable and impossible to measure accurately.

Common approximation – **mid-population** principle:

1. Let the population size be  $N_b$  at beginning ( $t_b$ ) and  $N_e$  at the end ( $t_e$ ) of the observation period of length  $t_e - t_b = u_t$
2. Mid-population for the period around  $t = (t_b + t_e)/2$ :  $\bar{N}_t = \frac{1}{2} \times (N_b + N_e)$ .
3. Approximate person-years:  $\tilde{Y}_t = \bar{N}_t \times u_t$ .

## Person-years in dynamic populations

With a **dynamic** study population, individual follow-up times are always variable and impossible to measure accurately.

Common approximation – **mid-population** principle:

1. Let the population size be  $N_b$  at beginning ( $t_b$ ) and  $N_e$  at the end ( $t_e$ ) of the observation period of length  $t_e - t_b = u_t$
2. Mid-population for the period around  $t = (t_b + t_e)/2$ :  $\bar{N}_t = \frac{1}{2} \times (N_b + N_e)$ .
3. Approximate person-years:  $\tilde{Y}_t = \bar{N}_t \times u_t$ .

**NB:** The actual study population often contains some already affected, thus not belonging to the population at risk. With rare outcomes their influence is small.

## Male person-years in Finland 1991-95

Total male population (1000s) on 31 December by year:

| 1990  | 1991  | 1992  | 1993  | 1994  | 1995  |
|-------|-------|-------|-------|-------|-------|
| 2,431 | 2,443 | 2,457 | 2,470 | 2,482 | 2,492 |

## Male person-years in Finland 1991-95

Total male population (1000s) on 31 December by year:

| 1990  | 1991  | 1992  | 1993  | 1994  | 1995  |
|-------|-------|-------|-------|-------|-------|
| 2,431 | 2,443 | 2,457 | 2,470 | 2,482 | 2,492 |

Approximate person-years (1000s) in various periods:

$$1992: \quad \frac{1}{2} \times (2,443 + 2,457) \times 1 = \quad 2450$$

## Male person-years in Finland 1991-95

Total male population (1000s) on 31 December by year:

| 1990  | 1991  | 1992  | 1993  | 1994  | 1995  |
|-------|-------|-------|-------|-------|-------|
| 2,431 | 2,443 | 2,457 | 2,470 | 2,482 | 2,492 |

Approximate person-years (1000s) in various periods:

$$1992: \quad \frac{1}{2} \times (2,443 + 2,457) \times 1 = \quad 2450$$

$$1993-94: \quad \frac{1}{2} \times (2,457 + 2,482) \times 2 = \quad 4937$$

## Male person-years in Finland 1991-95

Total male population (1000s) on 31 December by year:

| 1990  | 1991  | 1992  | 1993  | 1994  | 1995  |
|-------|-------|-------|-------|-------|-------|
| 2,431 | 2,443 | 2,457 | 2,470 | 2,482 | 2,492 |

Approximate person-years (1000s) in various periods:

|          |   |         |
|----------|---|---------|
| 1992:    | $\frac{1}{2} \times (2,443 + 2,457) \times 1 =$ | 2450    |
| 1993-94: | $\frac{1}{2} \times (2,457 + 2,482) \times 2 =$ | 4937    |
| 1991-95: | $\frac{1}{2} \times (2,431 + 2,492) \times 5 =$ | 12307.5 |

## Male person-years in Finland 1991-95

Total male population (1000s) on 31 December by year:

| 1990  | 1991  | 1992  | 1993  | 1994  | 1995  |
|-------|-------|-------|-------|-------|-------|
| 2,431 | 2,443 | 2,457 | 2,470 | 2,482 | 2,492 |

Approximate person-years (1000s) in various periods:

$$1992: \quad \frac{1}{2} \times (2,443 + 2,457) \times 1 = \quad 2450$$

$$1993-94: \quad \frac{1}{2} \times (2,457 + 2,482) \times 2 = \quad 4937$$

$$1991-95: \quad \frac{1}{2} \times (2,431 + 2,492) \times 5 = \quad 12307.5$$

$$1991-95: \quad (2,431/2 + 2,443 + 2,457 + 2,470 + 2,482 + 2,492/2) = \quad 12313.5$$



## Mortality

**Cause-specific mortality** from cause  $C$  is described by **mortality rates** defined like incidence rates, but

the numerator is the number of deaths from cause  $C$  during the time period of interest, and the denominator is the number of people at risk of dying from cause  $C$  during the time period of interest.

Mortality depends on incidence and the prognosis or case-fatality of an illness. The number of deaths is the product of the number of cases and the case-fatality rate.

## Mortality

**Cause-specific mortality** from cause  $C$  is described by **mortality rates** defined like incidence rates, but

- ▶ cases are *deaths* from cause  $C$ , and

# Mortality

**Cause-specific mortality** from cause  $C$  is described by **mortality rates** defined like incidence rates, but

- ▶ cases are *deaths* from cause  $C$ , and
- ▶ follow-up is extended until death or censoring.

# Mortality

**Cause-specific mortality** from cause  $C$  is described by **mortality rates** defined like incidence rates, but

- ▶ cases are *deaths* from cause  $C$ , and
- ▶ follow-up is extended until death or censoring.

# Mortality

**Cause-specific mortality** from cause  $C$  is described by **mortality rates** defined like incidence rates, but

- ▶ cases are *deaths* from cause  $C$ , and
- ▶ follow-up is extended until death or censoring.

Cause-specific **mortality proportion** (cumulative risk) from cause depends on **all** cause-specific rates, not **only** on the  $C$ -specific mortality rate.

# Mortality

**Cause-specific mortality** from cause  $C$  is described by **mortality rates** defined like incidence rates, but

- ▶ cases are *deaths* from cause  $C$ , and
- ▶ follow-up is extended until death or censoring.

Cause-specific **mortality proportion** (cumulative risk) from cause depends on **all** cause-specific rates, not **only** on the  $C$ -specific mortality rate.

**Total mortality:**

- ▶ cases are deaths from any cause.

# Mortality

**Cause-specific mortality** from cause  $C$  is described by **mortality rates** defined like incidence rates, but

- ▶ cases are *deaths* from cause  $C$ , and
- ▶ follow-up is extended until death or censoring.

Cause-specific **mortality proportion** (cumulative risk) from cause depends on **all** cause-specific rates, not **only** on the  $C$ -specific mortality rate.

**Total mortality:**

- ▶ cases are deaths from any cause.

# Mortality

**Cause-specific mortality** from cause  $C$  is described by **mortality rates** defined like incidence rates, but

- ▶ cases are *deaths* from cause  $C$ , and
- ▶ follow-up is extended until death or censoring.

Cause-specific **mortality proportion** (cumulative risk) from cause depends on **all** cause-specific rates, not **only** on the  $C$ -specific mortality rate.

**Total mortality:**

- ▶ cases are deaths from any cause.

Mortality depends on the incidence and the **prognosis** or **case fatality** of the disease, *i.e.* the **survival** of those affected by it.



## Prevalence measures

- ▶ **Point prevalence** or simply **prevalence**  $P$  of a health state  $C$  in a population at a given **time point**  $t$  is defined as:

$$P = \frac{\text{number of existing or prevalent cases of } C}{\text{size of the whole population}}$$

## Prevalence measures

- ▶ **Point prevalence** or simply **prevalence**  $P$  of a health state  $C$  in a population at a given **time point**  $t$  is defined as:

$$P = \frac{\text{number of existing or prevalent cases of } C}{\text{size of the whole population}}$$

- ▶ Can be computed from a cross-sectional study base.

## Prevalence measures

- ▶ **Point prevalence** or simply **prevalence**  $P$  of a health state  $C$  in a population at a given **time point**  $t$  is defined as:

$$P = \frac{\text{number of existing or prevalent cases of } C}{\text{size of the whole population}}$$

- ▶ Can be computed from a cross-sectional study base.
- ▶ Empirical counterpart of  $P\{\text{random person has disease } C\}$

## Prevalence measures

- ▶ **Point prevalence** or simply **prevalence**  $P$  of a health state  $C$  in a population at a given **time point**  $t$  is defined as:

$$P = \frac{\text{number of existing or prevalent cases of } C}{\text{size of the whole population}}$$

- ▶ Can be computed from a cross-sectional study base.
- ▶ Empirical counterpart of  $P\{\text{random person has disease } C\}$
- ▶ **Period prevalence** for period from  $t_1$  to  $t_2$  is like  $P$  but

## Prevalence measures

- ▶ **Point prevalence** or simply **prevalence**  $P$  of a health state  $C$  in a population at a given **time point**  $t$  is defined as:

$$P = \frac{\text{number of existing or prevalent cases of } C}{\text{size of the whole population}}$$

- ▶ Can be computed from a cross-sectional study base.
  - ▶ Empirical counterpart of  $P\{\text{random person has disease } C\}$
- ▶ **Period prevalence** for period from  $t_1$  to  $t_2$  is like  $P$  but
  - ▶ numerator refers to all cases prevalent already at  $t_1$  plus new cases occurring during the period, and

## Prevalence measures

- ▶ **Point prevalence** or simply **prevalence**  $P$  of a health state  $C$  in a population at a given **time point**  $t$  is defined as:

$$P = \frac{\text{number of existing or prevalent cases of } C}{\text{size of the whole population}}$$

- ▶ Can be computed from a cross-sectional study base.
  - ▶ Empirical counterpart of  $P\{\text{random person has disease } C\}$
- ▶ **Period prevalence** for period from  $t_1$  to  $t_2$  is like  $P$  but
  - ▶ numerator refers to all cases prevalent already at  $t_1$  plus new cases occurring during the period, and
  - ▶ denominator is the population size at  $t_2$

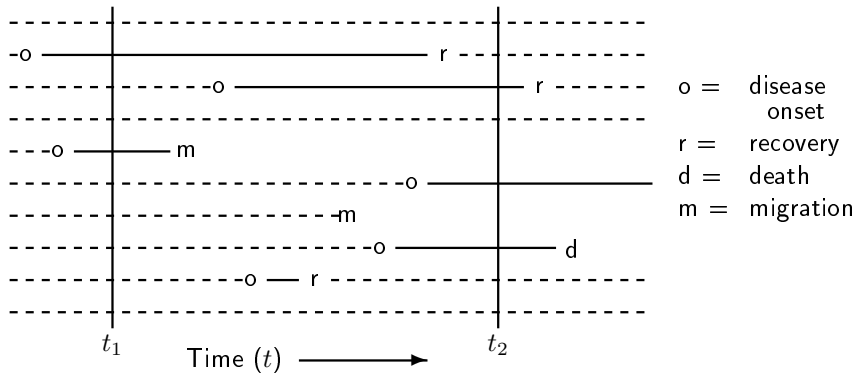
## Prevalence measures

- ▶ **Point prevalence** or simply **prevalence**  $P$  of a health state  $C$  in a population at a given **time point**  $t$  is defined as:

$$P = \frac{\text{number of existing or prevalent cases of } C}{\text{size of the whole population}}$$

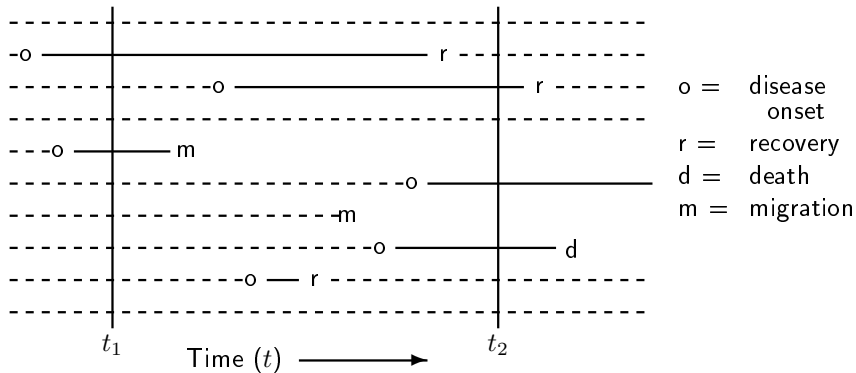
- ▶ Can be computed from a cross-sectional study base.
- ▶ Empirical counterpart of  $P\{\text{random person has disease } C\}$
- ▶ **Period prevalence** for period from  $t_1$  to  $t_2$  is like  $P$  but
  - ▶ numerator refers to all cases prevalent already at  $t_1$  plus new cases occurring during the period, and
  - ▶ denominator is the population size at  $t_2$
  - ▶ Has no simple probability counterpart

## Example 4.1 (IS: p. 59)



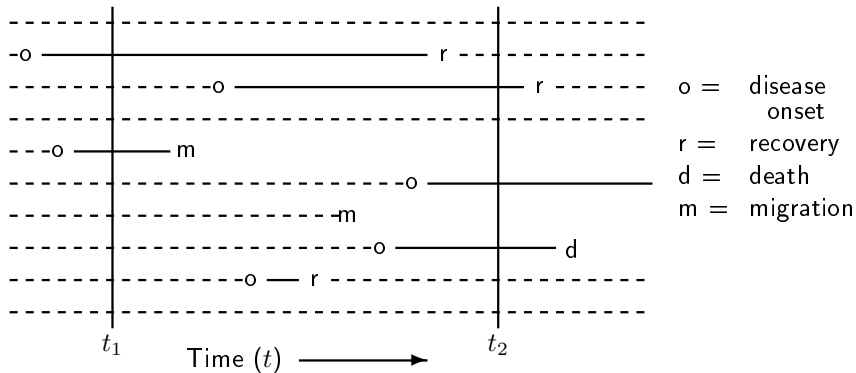


## Example 4.1 (IS: p. 59)



Prevalence at time  $t_1$  :  $2/10 = 0.2 = 20\%$

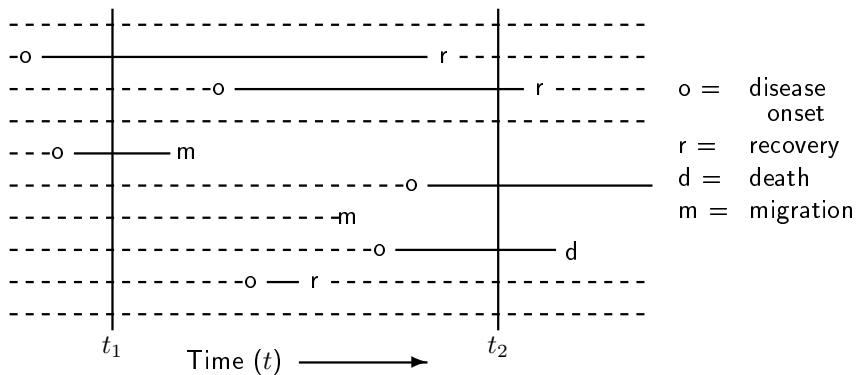
## Example 4.1 (IS: p. 59)



Prevalence at time  $t_1$  :  $2/10 = 0.2 = 20\%$

Prevalence at time  $t_2$  :  $3/8 = 0.38 = 38\%$

## Example 4.1 (IS: p. 59)



Prevalence at time  $t_1$  :  $2/10 = 0.2 = 20\%$

Prevalence at time  $t_2$  :  $3/8 = 0.38 = 38\%$

Period prevalence:  $6/8 = 0.75 = 75\%$

## Prevalence and incidence are related

Point prevalence of  $C$  at given time point  $t$  depends on the

total cases, the fatality rate ( $F$ ) and on the average duration of the disease ( $D$ ).  
Point prevalence ( $P$ ), incidence ( $I$ ) and average duration ( $D$ ) of  $C$  are related as follows:  
$$P = I \times D$$
  
Prevalence = incidence  $\times$  duration  
Prevalence is not the same as incidence.

## Prevalence and incidence are related

Point prevalence of  $C$  at given time point  $t$  depends on the

- ▶ **incidence** of new cases of  $C$  before  $t$ , and the

## Prevalence and incidence are related

Point prevalence of  $C$  at given time point  $t$  depends on the

- ▶ **incidence** of new cases of  $C$  before  $t$ , and the
- ▶ **duration** of  $C$ , depending in turn on the probability of

## Prevalence and incidence are related

Point prevalence of  $C$  at given time point  $t$  depends on the

- ▶ **incidence** of new cases of  $C$  before  $t$ , and the
- ▶ **duration** of  $C$ , depending in turn on the probability of
  - ▶ **cure** or recovery from  $C$ , or

## Prevalence and incidence are related

Point prevalence of  $C$  at given time point  $t$  depends on the

- ▶ **incidence** of new cases of  $C$  before  $t$ , and the
- ▶ **duration** of  $C$ , depending in turn on the probability of
  - ▶ **cure** or recovery from  $C$ , or
  - ▶ **survival** of those affected



## Prevalence and incidence are related

Point prevalence of  $C$  at given time point  $t$  depends on the

- ▶ **incidence** of new cases of  $C$  before  $t$ , and the
- ▶ **duration** of  $C$ , depending in turn on the probability of
  - ▶ **cure** or recovery from  $C$ , or
  - ▶ **survival** of those affected

## Prevalence and incidence are related

Point prevalence of  $C$  at given time point  $t$  depends on the

- ▶ **incidence** of new cases of  $C$  before  $t$ , and the
- ▶ **duration** of  $C$ , depending in turn on the probability of
  - ▶ **cure** or recovery from  $C$ , or
  - ▶ **survival** of those affected

Simple case: In a **stationary** (“stable”) population, the

prevalence ( $P$ ), incidence ( $I$ ), and average duration ( $\bar{d}$ ) of  $S$

are related as

$$P = I \times \bar{d}$$

prevalence = incidence  $\times$  duration

The approximation works well, when  $P < 0.1$  (10%).

# Prevalence of cancer?

- ▶ How do we know, whether and when cancer is cured?

## Prevalence of cancer?

- ▶ How do we know, whether and when cancer is cured?
- ▶  $\Rightarrow$  Existing or prevalent case problematic to define.

## Prevalence of cancer?

- ▶ How do we know, whether and when cancer is cured?
- ▶  $\Rightarrow$  Existing or prevalent case problematic to define.
- ▶ NORDCAN: Prevalence of cancer  $C$  at time point  $t$  in the target population refers to the • number & proportion of population members who

## Prevalence of cancer?

- ▶ How do we know, whether and when cancer is cured?
- ▶  $\Rightarrow$  Existing or prevalent case problematic to define.
- ▶ NORDCAN: Prevalence of cancer  $C$  at time point  $t$  in the target population refers to the • number & proportion of population members who
  - ▶ are alive and resident in the population at  $t$ , and

## Prevalence of cancer?

- ▶ How do we know, whether and when cancer is cured?
- ▶  $\Rightarrow$  Existing or prevalent case problematic to define.
- ▶ NORDCAN: Prevalence of cancer  $C$  at time point  $t$  in the target population refers to the • number & proportion of population members who
  - ▶ are alive and resident in the population at  $t$ , and
  - ▶ have a record of an incident cancer  $C$  diagnosed before  $t$ .

## Prevalence of cancer?

- ▶ How do we know, whether and when cancer is cured?
- ▶  $\Rightarrow$  Existing or prevalent case problematic to define.
- ▶ NORDCAN: Prevalence of cancer  $C$  at time point  $t$  in the target population refers to the • number & proportion of population members who
  - ▶ are alive and resident in the population at  $t$ , and
  - ▶ have a record of an incident cancer  $C$  diagnosed before  $t$ .
- ▶ **Partial prevalence:** Cases limited to those diagnosed during a fixed time in the past; e.g. within last



## Prevalence of cancer?

- ▶ How do we know, whether and when cancer is cured?
- ▶  $\Rightarrow$  Existing or prevalent case problematic to define.
- ▶ NORDCAN: Prevalence of cancer  $C$  at time point  $t$  in the target population refers to the • number & proportion of population members who
  - ▶ are alive and resident in the population at  $t$ , and
  - ▶ have a record of an incident cancer  $C$  diagnosed before  $t$ .
- ▶ **Partial prevalence:** Cases limited to those diagnosed during a fixed time in the past; e.g. within last
  - ▶ 1 y (initial treatment period),

## Prevalence of cancer?

- ▶ How do we know, whether and when cancer is cured?
- ▶  $\Rightarrow$  Existing or prevalent case problematic to define.
- ▶ NORDCAN: Prevalence of cancer  $C$  at time point  $t$  in the target population refers to the • number & proportion of population members who
  - ▶ are alive and resident in the population at  $t$ , and
  - ▶ have a record of an incident cancer  $C$  diagnosed before  $t$ .
- ▶ **Partial prevalence:** Cases limited to those diagnosed during a fixed time in the past; e.g. within last
  - ▶ 1 y (initial treatment period),
  - ▶ 3 y (clinical follow-up),

## Prevalence of cancer?

- ▶ How do we know, whether and when cancer is cured?
- ▶  $\Rightarrow$  Existing or prevalent case problematic to define.
- ▶ NORDCAN: Prevalence of cancer  $C$  at time point  $t$  in the target population refers to the • number & proportion of population members who
  - ▶ are alive and resident in the population at  $t$ , and
  - ▶ have a record of an incident cancer  $C$  diagnosed before  $t$ .
- ▶ **Partial prevalence:** Cases limited to those diagnosed during a fixed time in the past; e.g. within last
  - ▶ 1 y (initial treatment period),
  - ▶ 3 y (clinical follow-up),
  - ▶ 5 y (cure?)

## Ex: Cancer with poor and with good prognosis

Age-standardized<sup>a</sup> incidence, mortality, prevalence, and survival for cancers of kidney and thyroid in women of Finland:

## Ex: Cancer with poor and with good prognosis

Age-standardized<sup>a</sup> incidence, mortality, prevalence, and survival for cancers of kidney and thyroid in women of Finland:

|  | Kidney | Thyroid |
|--|--------|---------|
| Incidence rate in 2011 (per 10 <sup>5</sup> y) | 12     | 11      |

## Ex: Cancer with poor and with good prognosis

Age-standardized<sup>a</sup> incidence, mortality, prevalence, and survival for cancers of kidney and thyroid in women of Finland:

|  | Kidney | Thyroid |
|--|--------|---------|
| Incidence rate in 2011 (per 10 <sup>5</sup> y) | 12     | 11      |
| Mortality rate in 2011 (per 10 <sup>5</sup> y) | 5      | 1       |

## Ex: Cancer with poor and with good prognosis

Age-standardized<sup>a</sup> incidence, mortality, prevalence, and survival for cancers of kidney and thyroid in women of Finland:

|   | Kidney | Thyroid |
|---|--------|---------|
| Incidence rate in 2011 (per 10 <sup>5</sup> y)  | 12     | 11      |
| Mortality rate in 2011 (per 10 <sup>5</sup> y)  | 5      | 1       |
| Prevalence on 31.12.2011 (per 10 <sup>5</sup> ) | 92     | 198     |

## Ex: Cancer with poor and with good prognosis

Age-standardized<sup>a</sup> incidence, mortality, prevalence, and survival for cancers of kidney and thyroid in women of Finland:

|   | Kidney | Thyroid |
|---|--------|---------|
| Incidence rate in 2011 (per 10 <sup>5</sup> y)  | 12     | 11      |
| Mortality rate in 2011 (per 10 <sup>5</sup> y)  | 5      | 1       |
| Prevalence on 31.12.2011 (per 10 <sup>5</sup> ) | 92     | 198     |
| – diagnosed < 1 y ago                           | 9      | 10      |
| – diagnosed < 3 y ago                           | 24     | 29      |
| – diagnosed < 5 y ago                           | 35     | 47      |
| – diagnosed > 5 y ago                           | 57     | 151     |



## Ex: Cancer with poor and with good prognosis

Age-standardized<sup>a</sup> incidence, mortality, prevalence, and survival for cancers of kidney and thyroid in women of Finland:

|   | Kidney | Thyroid |
|---|--------|---------|
| Incidence rate in 2011 (per 10 <sup>5</sup> y)  | 12     | 11      |
| Mortality rate in 2011 (per 10 <sup>5</sup> y)  | 5      | 1       |
| Prevalence on 31.12.2011 (per 10 <sup>5</sup> ) | 92     | 198     |
| – diagnosed < 1 y ago                           | 9      | 10      |
| – diagnosed < 3 y ago                           | 24     | 29      |
| – diagnosed < 5 y ago                           | 35     | 47      |
| – diagnosed > 5 y ago                           | 57     | 151     |
| 5-y relative survival; cases 2004–8 (%)         | 64     | 90      |

<sup>a</sup> Standard: Nordic population in 2000

# Comparative measures

**Bendix Carstensen & Esa Läärä**

Nordic Summerschool of  
Cancer Epidemiology

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

ELcomparative

## Measures of effect — comparative measures

- ▶ Quantification of the **association** between a determinant (risk factor) and an outcome (disease) is based on

## Measures of effect — comparative measures

- ▶ Quantification of the **association** between a determinant (risk factor) and an outcome (disease) is based on

## Measures of effect — comparative measures

- ▶ Quantification of the **association** between a determinant (risk factor) and an outcome (disease) is based on
- ▶ **comparison of occurrence** between the *index* (“exposed”) and the *reference* (“unexposed”) groups by

## Measures of effect — comparative measures

- ▶ Quantification of the **association** between a determinant (risk factor) and an outcome (disease) is based on
- ▶ **comparison of occurrence** between the *index* (“exposed”) and the *reference* (“unexposed”) groups by
  - ▶ relative comparative measures (ratio)

## Measures of effect — comparative measures

- ▶ Quantification of the **association** between a determinant (risk factor) and an outcome (disease) is based on
- ▶ **comparison of occurrence** between the *index* (“exposed”) and the *reference* (“unexposed”) groups by
  - ▶ relative comparative measures (ratio)
  - ▶ absolute comparative measures (difference)

## Measures of effect — comparative measures

- ▶ Quantification of the **association** between a determinant (risk factor) and an outcome (disease) is based on
- ▶ **comparison of occurrence** between the *index* (“exposed”) and the *reference* (“unexposed”) groups by
  - ▶ relative comparative measures (ratio)
  - ▶ absolute comparative measures (difference)
- ▶ Interpreted as the **effect** of exposure  
—using the **assumption** that the direction of causality is from exposure to outcome.



## Measures of effect — comparative measures

- ▶ Quantification of the **association** between a determinant (risk factor) and an outcome (disease) is based on
- ▶ **comparison of occurrence** between the *index* (“exposed”) and the *reference* (“unexposed”) groups by
  - ▶ relative comparative measures (ratio)
  - ▶ absolute comparative measures (difference)
- ▶ Interpreted as the **effect** of exposure  
—using the **assumption** that the direction of causality is from exposure to outcome.
- ▶ Causality is assumed; effect is measured (estimated from data)

## Measures of effect — comparative measures

- ▶ Quantification of the **association** between a determinant (risk factor) and an outcome (disease) is based on
- ▶ **comparison of occurrence** between the *index* (“exposed”) and the *reference* (“unexposed”) groups by
  - ▶ relative comparative measures (ratio)
  - ▶ absolute comparative measures (difference)
- ▶ Interpreted as the **effect** of exposure  
—using the **assumption** that the direction of causality is from exposure to outcome.
- ▶ Causality is assumed; effect is measured (estimated from data)
- ▶ Yet, caution is needed in inferences on causal effects, as often the groups to be compared suffer from **poor comparability**  $\Leftrightarrow$  **Confounding**.

## Relative comparative measures

Generic name “**relative risk**” (RR) comparing occurrences between exposed (1) and unexposed (0) groups can refer to:

- ▶ incidence rate ratio  $IR = I_1/I_0$ ,

## Relative comparative measures

Generic name “**relative risk**” (RR) comparing occurrences between exposed (1) and unexposed (0) groups can refer to:

- ▶ incidence rate ratio  $IR = I_1/I_0$ ,
- ▶ incidence proportion ratio  $IPR = Q_1/Q_0$ ,

## Relative comparative measures

Generic name “**relative risk**” (RR) comparing occurrences between exposed (1) and unexposed (0) groups can refer to:

- ▶ incidence rate ratio  $IR = I_1/I_0$ ,
- ▶ incidence proportion ratio  $IPR = Q_1/Q_0$ ,
- ▶ incidence odds ratio  $IOR = [Q_1/(1 - Q_1)]/[Q_0/(1 - Q_0)]$ ,

## Relative comparative measures

Generic name “**relative risk**” (RR) comparing occurrences between exposed (1) and unexposed (0) groups can refer to:

- ▶ incidence rate ratio  $IR = I_1/I_0$ ,
- ▶ incidence proportion ratio  $IPR = Q_1/Q_0$ ,
- ▶ incidence odds ratio  $IOR = [Q_1/(1 - Q_1)]/[Q_0/(1 - Q_0)]$ ,
- ▶ prevalence ratio  $PR = P_1/P_0$ , or

## Relative comparative measures

Generic name “**relative risk**” (RR) comparing occurrences between exposed (1) and unexposed (0) groups can refer to:

- ▶ incidence rate ratio  $IR = I_1/I_0$ ,
- ▶ incidence proportion ratio  $IPR = Q_1/Q_0$ ,
- ▶ incidence odds ratio  $IOR = [Q_1/(1 - Q_1)]/[Q_0/(1 - Q_0)]$ ,
- ▶ prevalence ratio  $PR = P_1/P_0$ , or
- ▶ prevalence odds ratio  $POR = [P_1/(1 - P_1)]/[P_0/(1 - P_0)]$ ,

## Relative comparative measures

Generic name “**relative risk**” (RR) comparing occurrences between exposed (1) and unexposed (0) groups can refer to:

- ▶ incidence rate ratio  $IR = I_1/I_0$ ,
- ▶ incidence proportion ratio  $IPR = Q_1/Q_0$ ,
- ▶ incidence odds ratio  $IOR = [Q_1/(1 - Q_1)]/[Q_0/(1 - Q_0)]$ ,
- ▶ prevalence ratio  $PR = P_1/P_0$ , or
- ▶ prevalence odds ratio  $POR = [P_1/(1 - P_1)]/[P_0/(1 - P_0)]$ ,



## Relative comparative measures

Generic name “**relative risk**” (RR) comparing occurrences between exposed (1) and unexposed (0) groups can refer to:

- ▶ incidence rate ratio  $IR = I_1/I_0$ ,
- ▶ incidence proportion ratio  $IPR = Q_1/Q_0$ ,
- ▶ incidence odds ratio  $IOR = [Q_1/(1 - Q_1)]/[Q_0/(1 - Q_0)]$ ,
- ▶ prevalence ratio  $PR = P_1/P_0$ , or
- ▶ prevalence odds ratio  $POR = [P_1/(1 - P_1)]/[P_0/(1 - P_0)]$ ,

depending on study base and details of its design.

## Relative comparative measures

Generic name “**relative risk**” (RR) comparing occurrences between exposed (1) and unexposed (0) groups can refer to:

- ▶ incidence rate ratio  $IR = I_1/I_0$ ,
- ▶ incidence proportion ratio  $IPR = Q_1/Q_0$ ,
- ▶ incidence odds ratio  $IOR = [Q_1/(1 - Q_1)]/[Q_0/(1 - Q_0)]$ ,
- ▶ prevalence ratio  $PR = P_1/P_0$ , or
- ▶ prevalence odds ratio  $POR = [P_1/(1 - P_1)]/[P_0/(1 - P_0)]$ ,

depending on study base and details of its design.

Incidence rate ratio  $IR = I_1/I_0$  is the most commonly used comparative measure in cancer epidemiology.

## Absolute comparative measures

Generic term “**excess risk**” or “**risk difference**” (RD) btw exposed and unexposed can refer to

## Absolute comparative measures

Generic term “**excess risk**” or “**risk difference**” (RD) btw exposed and unexposed can refer to

- ▶ incidence rate difference  $ID = I_1 - I_0$ ,

## Absolute comparative measures

Generic term “**excess risk**” or “**risk difference**” (RD) btw exposed and unexposed can refer to

- ▶ incidence rate difference  $ID = I_1 - I_0$ ,
- ▶ incidence proportion difference  $IPD = Q_1 - Q_0$ , or

## Absolute comparative measures

Generic term “**excess risk**” or “**risk difference**” (RD) btw exposed and unexposed can refer to

- ▶ incidence rate difference  $ID = I_1 - I_0$ ,
- ▶ incidence proportion difference  $IPD = Q_1 - Q_0$ , or
- ▶ prevalence difference  $PD = P_1 - P_0$ .

## Absolute comparative measures

Generic term “**excess risk**” or “**risk difference**” (RD) btw exposed and unexposed can refer to

- ▶ incidence rate difference  $ID = I_1 - I_0$ ,
- ▶ incidence proportion difference  $IPD = Q_1 - Q_0$ , or
- ▶ prevalence difference  $PD = P_1 - P_0$ .

## Absolute comparative measures

Generic term “**excess risk**” or “**risk difference**” (RD) btw exposed and unexposed can refer to

- ▶ incidence rate difference  $ID = I_1 - I_0$ ,
- ▶ incidence proportion difference  $IPD = Q_1 - Q_0$ , or
- ▶ prevalence difference  $PD = P_1 - P_0$ .

Use of relative and absolute comparisons

- ▶ Ratios — describe the **biological strength** of the exposure



## Absolute comparative measures

Generic term “**excess risk**” or “**risk difference**” (RD) btw exposed and unexposed can refer to

- ▶ incidence rate difference  $ID = I_1 - I_0$ ,
- ▶ incidence proportion difference  $IPD = Q_1 - Q_0$ , or
- ▶ prevalence difference  $PD = P_1 - P_0$ .

Use of relative and absolute comparisons

- ▶ Ratios — describe the **biological strength** of the exposure
- ▶ Differences — inform about its **public health importance**.

## Example (IS, Table 5.2, p.97)

Relative and absolute comparisons between the exposed and the unexposed to risk factor  $X$  in two diseases.

## Example (IS, Table 5.2, p.97)

Relative and absolute comparisons between the exposed and the unexposed to risk factor  $X$  in two diseases.

|   | Disease A | Disease B |
|---|-----------|-----------|
| Incidence rate among exposed <sup>a</sup>   | 20        | 80        |
| Incidence rate among unexposed <sup>a</sup> | 5         | 40        |
| Rate ratio                                  | 4.0       | 2.0       |
| Rate difference <sup>a</sup>                | 15        | 40        |

<sup>a</sup> Rates per 100 000 pyrs.

## Example (IS, Table 5.2, p.97)

Relative and absolute comparisons between the exposed and the unexposed to risk factor  $X$  in two diseases.

|   | Disease A | Disease B |
|---|-----------|-----------|
| Incidence rate among exposed <sup>a</sup>   | 20        | 80        |
| Incidence rate among unexposed <sup>a</sup> | 5         | 40        |
| Rate ratio                                  | 4.0       | 2.0       |
| Rate difference <sup>a</sup>                | 15        | 40        |

<sup>a</sup> Rates per 100 000 pyrs.

Factor  $X$  has a stronger biological potency for disease A, but it has a greater public health importance for disease B.

## Attributable fraction (excess fraction)

- ▶ **Measures of potential impact:**  
Combination of absolute and relative comparisons.

## Attributable fraction (excess fraction)

- ▶ **Measures of potential impact:**  
Combination of absolute and relative comparisons.
- ▶ This measure estimates the fraction out of all new cases of disease among those exposed, which are attributable to (or “caused” by) the exposure itself, and which thus could be avoided if the exposure were absent.

## Attributable fraction (excess fraction)

- ▶ **Measures of potential impact:**  
Combination of absolute and relative comparisons.
- ▶ This measure estimates the fraction out of all new cases of disease among those exposed, which are attributable to (or “caused” by) the exposure itself, and which thus could be avoided if the exposure were absent.
- ▶ When the incidence is higher in the exposed, the **attributable fraction** (AF) for the exposure or risk factor is defined as:

$$AF = \frac{I_1 - I_0}{I_1} = \frac{RR - 1}{RR}.$$

Also called **excess fraction** (or even “attributable risk” in old texts).

## Population attributable fraction

- ▶ Suppose we ask instead:

How large a fraction of all cases in the population would be prevented, if the exposure were eliminated?

▶ The answer to the question depends in addition on

the prevalence of exposure in the population.

▶ The **population attributable (excess) fraction (PAF)** is defined

$$\text{PAF} = \frac{I_{\text{pop}} - I_{\text{pop}}(RR=1)}{I_{\text{pop}}(RR=1)}$$

▶ It is a useful type of measure

▶ It is a **relative** measure on the population level



## Population attributable fraction

- ▶ Suppose we ask instead:
- ▶ *“How large a fraction of all cases in the population would be prevented, if the exposure were eliminated?”*

## Population attributable fraction

- ▶ Suppose we ask instead:
- ▶ *“How large a fraction of all cases in the population would be prevented, if the exposure were eliminated?”*
- ▶ The answer to this question depends in addition on

$p_E =$  proportion of exposed in the population.

## Population attributable fraction

- ▶ Suppose we ask instead:
- ▶ *“How large a fraction of all cases in the population would be prevented, if the exposure were eliminated?”*
- ▶ The answer to this question depends in addition on

$p_E$  = proportion of exposed in the population.

- ▶ **Population attributable (excess) fraction (PAF)** is defined:

$$\text{PAF} = \frac{I - I_0}{I} = \frac{p_E(\text{RR} - 1)}{1 + p_E(\text{RR} - 1)}$$

## Population attributable fraction

- ▶ Suppose we ask instead:
- ▶ *“How large a fraction of all cases in the population would be prevented, if the exposure were eliminated?”*
- ▶ The answer to this question depends in addition on

$p_E$  = proportion of exposed in the population.

- ▶ **Population attributable (excess) fraction (PAF)** is defined:

$$\text{PAF} = \frac{I - I_0}{I} = \frac{p_E(\text{RR} - 1)}{1 + p_E(\text{RR} - 1)}$$

- ▶ AF: biological impact of exposure,

## Population attributable fraction

- ▶ Suppose we ask instead:
- ▶ *“How large a fraction of all cases in the population would be prevented, if the exposure were eliminated?”*
- ▶ The answer to this question depends in addition on

$p_E$  = proportion of exposed in the population.

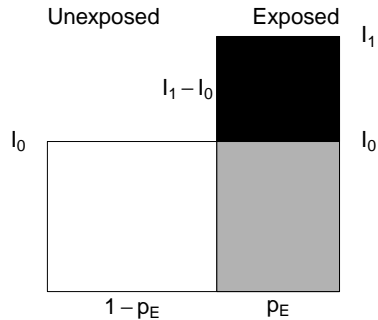
- ▶ **Population attributable (excess) fraction (PAF)** is defined:

$$\text{PAF} = \frac{I - I_0}{I} = \frac{p_E(\text{RR} - 1)}{1 + p_E(\text{RR} - 1)}$$

- ▶ AF: biological impact of exposure,
- ▶ PAF: impact of exposure on the population level.

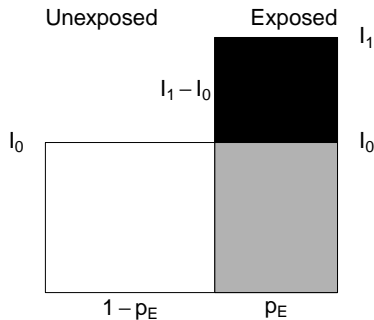
# Attributable fraction illustrated

- ▶ The population is divided into exposed and unexposed.



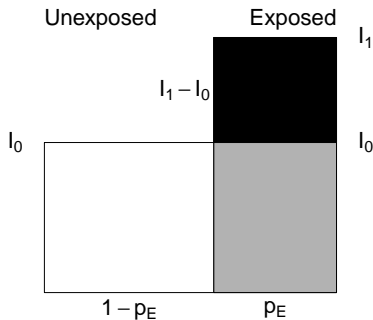
## Attributable fraction illustrated

- ▶ The population is divided into exposed and unexposed.
- ▶ The rate  $I_1$  among the exposed would be  $I_0$ , *i.e.* the same as in the unexposed, if the exposure had no effect.



## Attributable fraction illustrated

- ▶ The population is divided into exposed and unexposed.
- ▶ The rate  $I_1$  among the exposed would be  $I_0$ , *i.e.* the same as in the unexposed, if the exposure had no effect.
- ▶ The excess incidence  $I_1 - I_0$  is caused by the exposure.

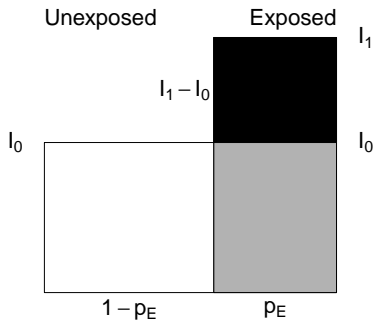




## Attributable fraction illustrated

- ▶ The population is divided into exposed and unexposed.
- ▶ The rate  $I_1$  among the exposed would be  $I_0$ , *i.e.* the same as in the unexposed, if the exposure had no effect.
- ▶ The excess incidence  $I_1 - I_0$  is caused by the exposure.

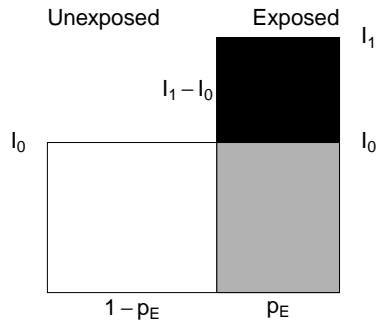
- ▶ 
$$AF = \frac{I_1 - I_0}{I_1}$$



## Attributable fraction illustrated

- ▶ The population is divided into exposed and unexposed.
- ▶ The rate  $I_1$  among the exposed would be  $I_0$ , *i.e.* the same as in the unexposed, if the exposure had no effect.
- ▶ The excess incidence  $I_1 - I_0$  is caused by the exposure.

- ▶  $AF = \frac{I_1 - I_0}{I_1}$ ,  
= fraction of  
black area  
out of total  
black + gray area.

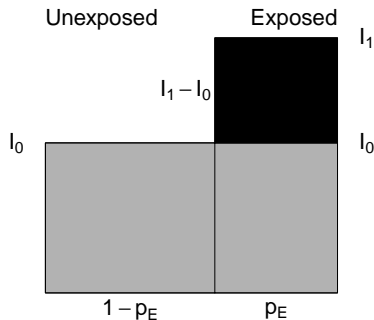


## PAF illustrated

- ▶ Total incidence  $I$  in the population – a weighted average:

$$I = p_E \times I_1 + (1 - p_E) \times I_0 \quad (\text{total area})$$

would equal  $I_0$ , if exposure had no effect



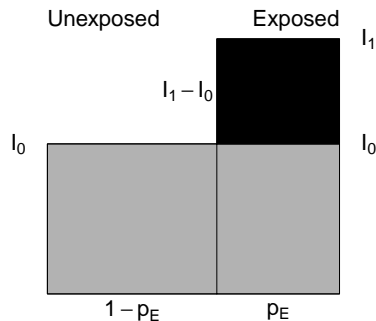
## PAF illustrated

- ▶ Total incidence  $I$  in the population – a weighted average:

$$I = p_E \times I_1 + (1 - p_E) \times I_0 \quad (\text{total area})$$

would equal  $I_0$ , if exposure had no effect

- ▶ Excess incidence caused by exposure:  $I - I_0 = p_E \times (I_1 - I_0)$  (black area).



## PAF illustrated

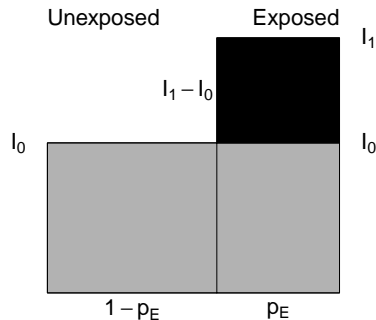
- ▶ Total incidence  $I$  in the population – a weighted average:

$$I = p_E \times I_1 + (1 - p_E) \times I_0 \quad (\text{total area})$$

would equal  $I_0$ , if exposure had no effect

- ▶ Excess incidence caused by exposure:  $I - I_0 = p_E \times (I_1 - I_0)$  (black area).

- ▶  $PAF = \frac{I - I_0}{I}$ ,



## PAF illustrated

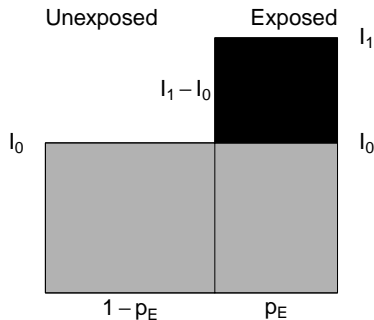
- ▶ Total incidence  $I$  in the population – a weighted average:

$$I = p_E \times I_1 + (1 - p_E) \times I_0 \quad (\text{total area})$$

would equal  $I_0$ , if exposure had no effect

- ▶ Excess incidence caused by exposure:  $I - I_0 = p_E \times (I_1 - I_0)$  (black area).

- ▶  $PAF = \frac{I - I_0}{I}$ ,  
= fraction of  
black area  
out of total  
black + gray area.



# Prevented fractions

## Prevented fractions

- ▶ When the incidence in exposed is lower, we define the **prevented fraction**:

$$PF = \frac{I_0 - I_1}{I_0} = 1 - RR$$



## Prevented fractions

- ▶ When the incidence in exposed is lower, we define the **prevented fraction**:

$$PF = \frac{I_0 - I_1}{I_0} = 1 - RR$$

## Prevented fractions

- ▶ When the incidence in exposed is lower, we define the **prevented fraction**:

$$PF = \frac{I_0 - I_1}{I_0} = 1 - RR$$

also called **relative risk reduction**

= percentage of cases prevented among the exposed due to the exposure.

- ▶ Used to evaluate the relative effect of a preventive intervention (“exposure”) vs. no intervention.

## Prevented fractions

- ▶ When the incidence in exposed is lower, we define the **prevented fraction**:

$$PF = \frac{I_0 - I_1}{I_0} = 1 - RR$$

also called **relative risk reduction**

= percentage of cases prevented among the exposed due to the exposure.

- ▶ Used to evaluate the relative effect of a preventive intervention (“exposure”) vs. no intervention.

## Prevented fractions

- ▶ When the incidence in exposed is lower, we define the **prevented fraction**:

$$PF = \frac{I_0 - I_1}{I_0} = 1 - RR$$

also called **relative risk reduction**

= percentage of cases prevented among the exposed due to the exposure.

- ▶ Used to evaluate the relative effect of a preventive intervention (“exposure”) vs. no intervention.
- ▶ **Population prevented fraction (PPF)** combines this with the prevalence of exposure in the population:

$$PPF = \frac{I_0 - I}{I_0} = p_E \times (1 - RR),$$

measuring the relative reduction in caseload attributable to the presence of preventive factor in the population.

## Smoking on mortality by cause (IS: Ex 5.14, p. 98)

| Underlying cause of death | Never smoked regularly<br>Rate <sup>b</sup><br>(1) | Current cigarette smoker<br>Rate <sup>b</sup><br>(2) | Rate ratio<br>(2)/(1) | Rate difference <sup>b</sup><br>(2) - (1) | Attributable fraction (%)<br>$\frac{(2) - (1)}{(2)} \times 100$ |
|---------------------------|--|--|-----------------------|---|---|
| Cancer type               |  |  |                       |   |   |
| All sites                 | 305  | 656  | 2.2                   | 351                                       | 54  |
| Lung                      | 14   | 209  | 14.9                  | 195                                       | 93  |
| Oesophagus                | 4  | 30   | 7.5                   | 26  | 87  |
| Bladder                   | 13   | 30   | 2.3                   | 17  | 57  |
| Respiratory diseases      | 107  | 313  | 2.9                   | 206                                       | 66  |
| Vascular diseases         | 1037   | 1643   | 1.6                   | 606                                       | 37  |
| All causes                | 1706   | 3038   | 1.8                   | 1332                                      | 44  |

<sup>a</sup> Data from Doll *et al.*, 1994a.

<sup>b</sup> Age-adjusted rates per 100 000 pyrs.

# Time scales

**Bendix Carstensen & Esa Läärä**

Nordic Summerschool of  
Cancer Epidemiology

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

ELrates

## Rates on several time scales

can be studied on various distinct time scales, e.g.

| Time scale          | Origin: date of ... |
|---------------------|---------------------|
| age                 | birth               |
| calendar time       | 1900-1-1            |
| exposure time       | first exposure      |
| follow-up time      | entry to study      |
| duration of disease | diagnosis           |

- ▶ Age is usually the strongest time-dependent determinant of health outcomes.

## Rates on several time scales

can be studied on various distinct time scales, e.g.

| Time scale          | Origin: date of ... |
|---------------------|---------------------|
| age                 | birth               |
| calendar time       | 1900-1-1            |
| exposure time       | first exposure      |
| follow-up time      | entry to study      |
| duration of disease | diagnosis           |

- ▶ Age is usually the strongest time-dependent determinant of health outcomes.



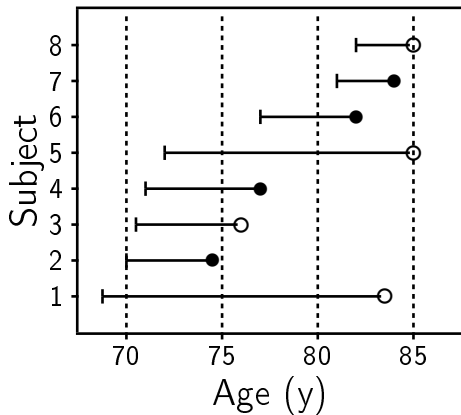
## Rates on several time scales

can be studied on various distinct time scales, e.g.

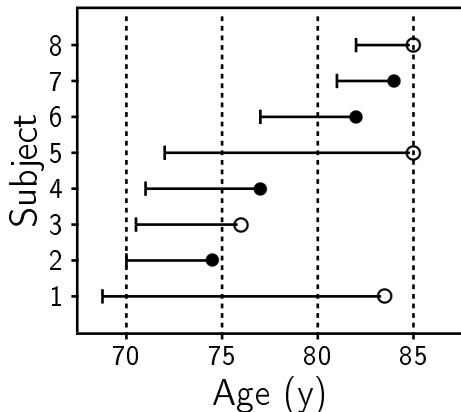
| Time scale          | Origin: date of ... |
|---------------------|---------------------|
| age                 | birth               |
| calendar time       | 1900-1-1            |
| exposure time       | first exposure      |
| follow-up time      | entry to study      |
| duration of disease | diagnosis           |

- ▶ Age is usually the strongest time-dependent determinant of health outcomes.
- ▶ Age is also often correlated with cumulative exposure (e.g. years of smoking).

## Follow-up of a small geriatric cohort

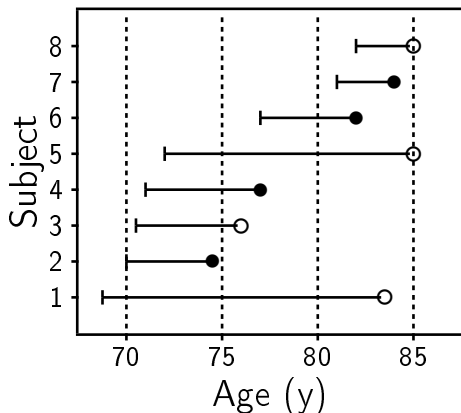


## Follow-up of a small geriatric cohort



Overall rate: 4 cases/53.5 person-years = 7.5 per 100 y.

## Follow-up of a small geriatric cohort



Overall rate: 4 cases/53.5 person-years = 7.5 per 100 y.

But the “true” rate varies by age, it is higher among the old.

# Splitting follow-up into agebands

## Splitting follow-up into agebands

- ▶ To describe, how incidence varies by age, individual person-years from age of entry to age of exit must first be split or divided into narrower agebands.

## Splitting follow-up into agebands

- ▶ To describe, how incidence varies by age, individual person-years from age of entry to age of exit must first be split or divided into narrower agebands.

## Splitting follow-up into agebands

- ▶ To describe, how incidence varies by age, individual person-years from age of entry to age of exit must first be split or divided into narrower agebands.
- ▶ Usually these are based on common 5-year age grouping.



## Splitting follow-up into agebands

- ▶ To describe, how incidence varies by age, individual person-years from age of entry to age of exit must first be split or divided into narrower agebands.
- ▶ Usually these are based on common 5-year age grouping.

## Splitting follow-up into agebands

- ▶ To describe, how incidence varies by age, individual person-years from age of entry to age of exit must first be split or divided into narrower agebands.
- ▶ Usually these are based on common 5-year age grouping.
- ▶ Numbers of cases are equally divided into same agebands.

## Splitting follow-up into agebands

- ▶ To describe, how incidence varies by age, individual person-years from age of entry to age of exit must first be split or divided into narrower agebands.
- ▶ Usually these are based on common 5-year age grouping.
- ▶ Numbers of cases are equally divided into same agebands.

## Splitting follow-up into agebands

- ▶ To describe, how incidence varies by age, individual person-years from age of entry to age of exit must first be split or divided into narrower agebands.
- ▶ Usually these are based on common 5-year age grouping.
- ▶ Numbers of cases are equally divided into same agebands.
- ▶ **Age-specific incidence rate** for age group  $k$  is

## Splitting follow-up into agebands

- ▶ To describe, how incidence varies by age, individual person-years from age of entry to age of exit must first be split or divided into narrower agebands.
- ▶ Usually these are based on common 5-year age grouping.
- ▶ Numbers of cases are equally divided into same agebands.
- ▶ **Age-specific incidence rate** for age group  $k$  is

## Splitting follow-up into agebands

- ▶ To describe, how incidence varies by age, individual person-years from age of entry to age of exit must first be split or divided into narrower agebands.
- ▶ Usually these are based on common 5-year age grouping.
- ▶ Numbers of cases are equally divided into same agebands.
- ▶ **Age-specific incidence rate** for age group  $k$  is

$$I_k = \frac{\text{number of cases observed in ageband}}{\text{person-years contained in ageband}}$$

- ▶ Underlying assumption: **piecewise constant rates** (in each age band)

## P-years and cases in agebands: age-specific rates

| Subject             | Ageband |       |       |     |
|---------------------|---------|-------|-------|-----|
|                     | 70-75   | 75-79 | 80-84 | 85+ |
| Sum of person-years | 100     | 100   | 100   | 100 |
| Cases               | 1       | 2     | 4     | 5   |
| Rate (100 y)        | 1%      | 2%    | 4%    | 5%  |

## P-years and cases in agebands: age-specific rates

| Subject | Ageband |       |       | Total |
|---------|---------|-------|-------|-------|
|         | 70-74   | 75-79 | 80-84 |       |



## P-years and cases in agebands: age-specific rates

| Subject | Ageband |       |       | Total |
|---------|---------|-------|-------|-------|
|         | 70-74   | 75-79 | 80-84 |       |
| 1       | 5.0     | 5.0   | 3.5   | 13.5  |

## P-years and cases in agebands: age-specific rates

| Subject | Ageband |       |       | Total |
|---------|---------|-------|-------|-------|
|         | 70-74   | 75-79 | 80-84 |       |
| 1       | 5.0     | 5.0   | 3.5   | 13.5  |
| 2       | 4.5     | -     | -     | 4.5   |

## P-years and cases in agebands: age-specific rates

| Subject | Ageband |       |       | Total |
|---------|---------|-------|-------|-------|
|         | 70-74   | 75-79 | 80-84 |       |
| 1       | 5.0     | 5.0   | 3.5   | 13.5  |
| 2       | 4.5     | -     | -     | 4.5   |
| 3       | 4.5     | 1.0   | -     | 5.5   |

## P-years and cases in agebands: age-specific rates

| Subject | Ageband |       |       | Total |
|---------|---------|-------|-------|-------|
|         | 70-74   | 75-79 | 80-84 |       |
| 1       | 5.0     | 5.0   | 3.5   | 13.5  |
| 2       | 4.5     | -     | -     | 4.5   |
| 3       | 4.5     | 1.0   | -     | 5.5   |
| 4       | 4.0     | 2.0   | -     | 6.0   |
| 5       | 3.0     | 5.0   | 5.0   | 13.0  |
| 6       | -       | 3.0   | 2.0   | 5.0   |
| 7       | -       | -     | 3.0   | 3.0   |
| 8       | -       | -     | 3.0   | 3.0   |

## P-years and cases in agebands: age-specific rates

| Subject             | Ageband |       |       | Total |
|---------------------|---------|-------|-------|-------|
|                     | 70-74   | 75-79 | 80-84 |       |
| 1                   | 5.0     | 5.0   | 3.5   | 13.5  |
| 2                   | 4.5     | -     | -     | 4.5   |
| 3                   | 4.5     | 1.0   | -     | 5.5   |
| 4                   | 4.0     | 2.0   | -     | 6.0   |
| 5                   | 3.0     | 5.0   | 5.0   | 13.0  |
| 6                   | -       | 3.0   | 2.0   | 5.0   |
| 7                   | -       | -     | 3.0   | 3.0   |
| 8                   | -       | -     | 3.0   | 3.0   |
| Sum of person-years | 21.0    | 16.0  | 16.5  | 53.5  |

## P-years and cases in agebands: age-specific rates

| Subject             | Ageband |       |       | Total |
|---------------------|---------|-------|-------|-------|
|                     | 70-74   | 75-79 | 80-84 |       |
| 1                   | 5.0     | 5.0   | 3.5   | 13.5  |
| 2                   | 4.5     | -     | -     | 4.5   |
| 3                   | 4.5     | 1.0   | -     | 5.5   |
| 4                   | 4.0     | 2.0   | -     | 6.0   |
| 5                   | 3.0     | 5.0   | 5.0   | 13.0  |
| 6                   | -       | 3.0   | 2.0   | 5.0   |
| 7                   | -       | -     | 3.0   | 3.0   |
| 8                   | -       | -     | 3.0   | 3.0   |
| Sum of person-years | 21.0    | 16.0  | 16.5  | 53.5  |
| Cases               | 1       | 1     | 2     | 4     |

## P-years and cases in agebands: age-specific rates

| Subject             | Ageband            |       |       | Total   |
|---------------------|--------------------|-------|-------|---------|
|                     | 70-74              | 75-79 | 80-84 |         |
| 1                   | 5.0                | 5.0   | 3.5   | 13.5    |
| 2                   | 4.5                | -     | -     | 4.5     |
| 3                   | 4.5                | 1.0   | -     | 5.5     |
| 4                   | 4.0                | 2.0   | -     | 6.0     |
| 5                   | 3.0                | 5.0   | 5.0   | 13.0    |
| 6                   | -                  | 3.0   | 2.0   | 5.0     |
| 7                   | -                  | -     | 3.0   | 3.0     |
| 8                   | -                  | -     | 3.0   | 3.0     |
| Sum of person-years | 21.0               | 16.0  | 16.5  | 53.5    |
| Cases               | 1                  | 1     | 2     | 4       |
| Rate (/100 y)       | 4.8                | 6.2   | 12.1  | 7.5     |
|                     | Age-specific rates |       |       | overall |

## Ex. Lung cancer incidence in Finland by age and period (compare IS, Table 4.1)

| Calendar<br>period | Age group (y) |       |       |       |       |       |       |       |       |     |
|--------------------|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-----|
|                    | 40-44         | 45-49 | 50-54 | 55-59 | 60-64 | 65-69 | 70-74 | 75-79 | 80-84 | 85+ |
| 1953-57            | 21            | 61    | 119   | 209   | 276   | 340   | 295   | 279   | 193   | 93  |
| 1958-62            | 22            | 65    | 135   | 243   | 360   | 405   | 429   | 368   | 265   | 224 |
| 1963-67            | 24            | 61    | 143   | 258   | 395   | 487   | 509   | 479   | 430   | 280 |
| 1968-72            | 21            | 61    | 134   | 278   | 424   | 529   | 614   | 563   | 471   | 358 |
| 1973-77            | 16            | 50    | 134   | 251   | 413   | 541   | 629   | 580   | 490   | 392 |
| 1978-82            | 13            | 36    | 115   | 234   | 369   | 514   | 621   | 653   | 593   | 442 |
| 1983-87            | 11            | 31    | 74    | 186   | 347   | 450   | 566   | 635   | 592   | 447 |
| 1988-92            | 9             | 25    | 57    | 128   | 262   | 411   | 506   | 507   | 471   | 441 |
| 1993-97            | 7             | 22    | 48    | 106   | 188   | 329   | 467   | 533   | 487   | 367 |
| 1998-02            | 5             | 14    | 46    | 77    | 150   | 239   | 358   | 445   | 396   | 346 |



## Ex. Lung cancer incidence in Finland by age and period (compare IS, Table 4.1)

| Calendar<br>period | Age group (y) |       |       |       |       |       |       |       |       |     |
|--------------------|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-----|
|                    | 40-44         | 45-49 | 50-54 | 55-59 | 60-64 | 65-69 | 70-74 | 75-79 | 80-84 | 85+ |
| 1953-57            | 21            | 61    | 119   | 209   | 276   | 340   | 295   | 279   | 193   | 93  |
| 1958-62            | 22            | 65    | 135   | 243   | 360   | 405   | 429   | 368   | 265   | 224 |
| 1963-67            | 24            | 61    | 143   | 258   | 395   | 487   | 509   | 479   | 430   | 280 |
| 1968-72            | 21            | 61    | 134   | 278   | 424   | 529   | 614   | 563   | 471   | 358 |
| 1973-77            | 16            | 50    | 134   | 251   | 413   | 541   | 629   | 580   | 490   | 392 |
| 1978-82            | 13            | 36    | 115   | 234   | 369   | 514   | 621   | 653   | 593   | 442 |
| 1983-87            | 11            | 31    | 74    | 186   | 347   | 450   | 566   | 635   | 592   | 447 |
| 1988-92            | 9             | 25    | 57    | 128   | 262   | 411   | 506   | 507   | 471   | 441 |
| 1993-97            | 7             | 22    | 48    | 106   | 188   | 329   | 467   | 533   | 487   | 367 |
| 1998-02            | 5             | 14    | 46    | 77    | 150   | 239   | 358   | 445   | 396   | 346 |

► Rows: age-incidence pattern in different calendar periods.

## Ex. Lung cancer incidence in Finland by age and period (compare IS, Table 4.1)

| Calendar<br>period | Age group (y) |       |       |       |       |       |       |       |       |     |
|--------------------|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-----|
|                    | 40-44         | 45-49 | 50-54 | 55-59 | 60-64 | 65-69 | 70-74 | 75-79 | 80-84 | 85+ |
| 1953-57            | 21            | 61    | 119   | 209   | 276   | 340   | 295   | 279   | 193   | 93  |
| 1958-62            | 22            | 65    | 135   | 243   | 360   | 405   | 429   | 368   | 265   | 224 |
| 1963-67            | 24            | 61    | 143   | 258   | 395   | 487   | 509   | 479   | 430   | 280 |
| 1968-72            | 21            | 61    | 134   | 278   | 424   | 529   | 614   | 563   | 471   | 358 |
| 1973-77            | 16            | 50    | 134   | 251   | 413   | 541   | 629   | 580   | 490   | 392 |
| 1978-82            | 13            | 36    | 115   | 234   | 369   | 514   | 621   | 653   | 593   | 442 |
| 1983-87            | 11            | 31    | 74    | 186   | 347   | 450   | 566   | 635   | 592   | 447 |
| 1988-92            | 9             | 25    | 57    | 128   | 262   | 411   | 506   | 507   | 471   | 441 |
| 1993-97            | 7             | 22    | 48    | 106   | 188   | 329   | 467   | 533   | 487   | 367 |
| 1998-02            | 5             | 14    | 46    | 77    | 150   | 239   | 358   | 445   | 396   | 346 |

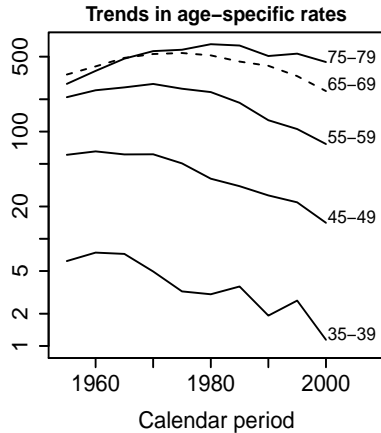
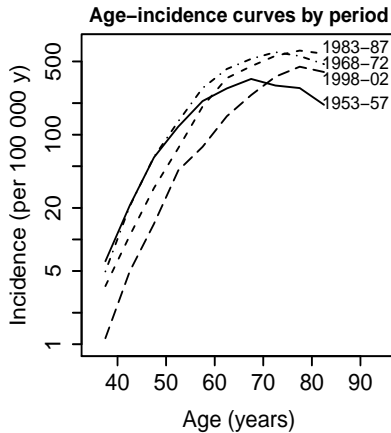
► Rows: age-incidence pattern in different calendar periods.

## Ex. Lung cancer incidence in Finland by age and period (compare IS, Table 4.1)

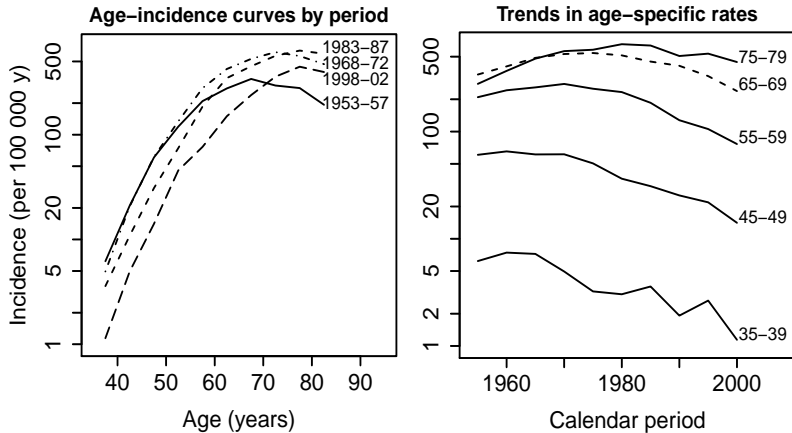
| Calendar<br>period | Age group (y) |       |       |       |       |       |       |       |       |     |
|--------------------|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-----|
|                    | 40-44         | 45-49 | 50-54 | 55-59 | 60-64 | 65-69 | 70-74 | 75-79 | 80-84 | 85+ |
| 1953-57            | 21            | 61    | 119   | 209   | 276   | 340   | 295   | 279   | 193   | 93  |
| 1958-62            | 22            | 65    | 135   | 243   | 360   | 405   | 429   | 368   | 265   | 224 |
| 1963-67            | 24            | 61    | 143   | 258   | 395   | 487   | 509   | 479   | 430   | 280 |
| 1968-72            | 21            | 61    | 134   | 278   | 424   | 529   | 614   | 563   | 471   | 358 |
| 1973-77            | 16            | 50    | 134   | 251   | 413   | 541   | 629   | 580   | 490   | 392 |
| 1978-82            | 13            | 36    | 115   | 234   | 369   | 514   | 621   | 653   | 593   | 442 |
| 1983-87            | 11            | 31    | 74    | 186   | 347   | 450   | 566   | 635   | 592   | 447 |
| 1988-92            | 9             | 25    | 57    | 128   | 262   | 411   | 506   | 507   | 471   | 441 |
| 1993-97            | 7             | 22    | 48    | 106   | 188   | 329   | 467   | 533   | 487   | 367 |
| 1998-02            | 5             | 14    | 46    | 77    | 150   | 239   | 358   | 445   | 396   | 346 |

- ▶ Rows: age-incidence pattern in different calendar periods.
- ▶ Columns: Trends of age-specific rates over calendar time.

# Lung cancer rates by age and period

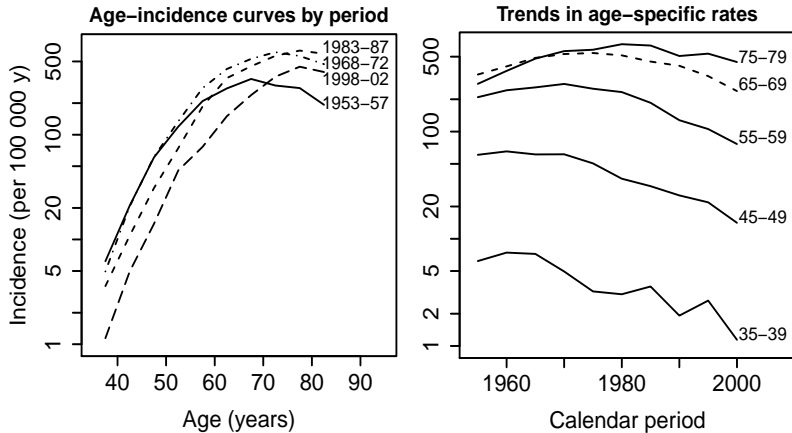


# Lung cancer rates by age and period



- ▶ Age-incidence curves: overall level and peak age variable across periods.

# Lung cancer rates by age and period



- ▶ Age-incidence curves: overall level and peak age variable across periods.
- ▶ Time trends inconsistent across age groups: decline onset at different dates.

# Incidence by age, period & cohort

Birth cohort = people born during the same defined time interval (eg. 1940-1945)  
calendar year, or 5 years period

## Incidence by age, period & cohort

- ▶ **Secular trends** of specific and adjusted rates show, how the “cancer burden” has developed over periods of calendar time.



## Incidence by age, period & cohort

- ▶ **Secular trends** of specific and adjusted rates show, how the “cancer burden” has developed over periods of calendar time.

## Incidence by age, period & cohort

- ▶ **Secular trends** of specific and adjusted rates show, how the “cancer burden” has developed over periods of calendar time.

**Birth cohort** = people born during the same limited time interval, *e.g.* single calendar year, or 5 years period.

## Incidence by age, period & cohort

- ▶ **Secular trends** of specific and adjusted rates show, how the “cancer burden” has developed over periods of calendar time.

**Birth cohort** = people born during the same limited time interval, *e.g.* single calendar year, or 5 years period.

- ▶ Analysis of rates by birth cohort reveals, how the level of incidence (or mortality) differs between successive generations.
  - May reflect differences in risk factor levels across birth cohorts.

## Incidence by age, period & cohort

- ▶ **Secular trends** of specific and adjusted rates show, how the “cancer burden” has developed over periods of calendar time.

**Birth cohort** = people born during the same limited time interval, *e.g.* single calendar year, or 5 years period.

- ▶ Analysis of rates by birth cohort reveals, how the level of incidence (or mortality) differs between successive generations.
  - May reflect differences in risk factor levels across birth cohorts.
- ▶ Often more informative about “true” age-incidence pattern than age-specific incidences of single calendar period.

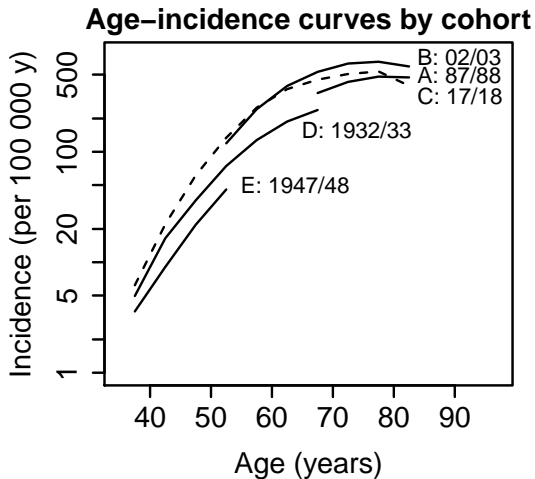
## Age-specific rates by birth cohort

| Calendar period | Age group (y) |       |       |            |       |       |       |       |  |
|-----------------|---------------|-------|-------|------------|-------|-------|-------|-------|--|
|                 | 40-44         | 45-49 | 50-54 | 55-59      | 60-64 | 65-69 | 70-74 | 75-79 |  |
| 1953-57         | 21            | 61    | 119   | 209        | 276   | 340   | 295   | 279   |  |
| 1958-62         | 22            | 65    | 135   | 243        | 360   | 405   | 429   | 368   |  |
| 1963-67         | 24            | 61    | 143   | 258        | 395   | 487   | 509   | 479   |  |
| 1968-72         | 21            | 61    | 134   | 278        | 424   | 529   | 614   | 563   |  |
| 1973-77         | 16            | 50    | 134   | 251        | 413   | 541   | 629   | 580   |  |
| 1978-82         | 13            | 36    | 115   | 234        | 369   | 514   | 621   | 653   |  |
| 1983-87         | 11            | 31    | 74    | 186        | 347   | 450   | 566   | 635   |  |
| 1988-92         | 9             | 25    | 57    | 128        | 262   | 411   | 506   | 507   |  |
| 1993-97         | 7             | 22    | 48    | 106        | 188   | 329   | 467   | 533   |  |
| 1998-02         | 5             | 14    | 46    | 77         | 150   | 239   | 358   | 445   |  |
| E: 1947/48      |               |       |       | D: 1932/33 |       |       |       |       |  |

A = synthetic cohort born around 1887/88, B: 1902/03, C: 1917/18

Diagonals reflect age-incidence patterns in various birth cohorts.

## Age-incidence curves in 5 birth cohorts



Variable overall rates but fairly consistent form and similar peak age across different birth cohorts.

## Split of follow-up by age and period

- ▶ Incidence of (or mortality from) disease  $C$  in special **cohort of exposed** (e.g. occupational group, patients on certain treatment)

## Split of follow-up by age and period

- ▶ Incidence of (or mortality from) disease  $C$  in special **cohort of exposed** (e.g. occupational group, patients on certain treatment)



## Split of follow-up by age and period

- ▶ Incidence of (or mortality from) disease  $C$  in special **cohort of exposed** (e.g. occupational group, patients on certain treatment)
  - often compared to incidence in an external **reference** or “general” population.

## Split of follow-up by age and period

- ▶ Incidence of (or mortality from) disease  $C$  in special **cohort of exposed** (e.g. occupational group, patients on certain treatment)
  - often compared to incidence in an external **reference** or “general” population.
- ▶ Adjustment for age and calendar time needed, e.g. by comparing **observed** to **expected** cases with SIR (see p. 70-74).

## Split of follow-up by age and period

- ▶ Incidence of (or mortality from) disease  $C$  in special **cohort of exposed** (e.g. occupational group, patients on certain treatment)
  - often compared to incidence in an external **reference** or “general” population.
- ▶ Adjustment for age and calendar time needed, e.g. by comparing **observed** to **expected** cases with SIR (see p. 70-74).

## Split of follow-up by age and period

- ▶ Incidence of (or mortality from) disease  $C$  in special **cohort of exposed** (e.g. occupational group, patients on certain treatment)
  - often compared to incidence in an external **reference** or “general” population.
- ▶ Adjustment for age and calendar time needed, e.g. by comparing **observed** to **expected** cases with SIR (see p. 70-74).
  - ⇒ Cases and person-years in the study cohort must be split by more than one time scale (age).

## Example (adapted from C&H, Tables 6.2 & 6.3, p. 54)

Entry and exit dates for a small cohort of four subjects

| Subject | Born | Entry | Exit | Age at entry | Outcome         |
|---------|------|-------|------|--------------|-----------------|
| 1       | 1954 | 1993  | 2002 | 39           | Migrated        |
| 2       | 1974 | 1998  | 2005 | 24           | Disease $C$     |
| 3       | 1964 | 1995  | 2011 | 31           | Study ends      |
| 4       | 1970 | 1998  | 2006 | 28           | Unrelated death |

## Example (adapted from C&H, Tables 6.2 & 6.3, p. 54)

Entry and exit dates for a small cohort of four subjects

| Subject | Born | Entry | Exit | Age at entry | Outcome          |
|---------|------|-------|------|--------------|------------------|
| 1       | 1954 | 1993  | 2002 | 39           | Migrated         |
| 2       | 1974 | 1998  | 2005 | 24           | Disease <i>C</i> |
| 3       | 1964 | 1995  | 2011 | 31           | Study ends       |
| 4       | 1970 | 1998  | 2006 | 28           | Unrelated death  |

Subject 1: Follow-up time spent in each ageband

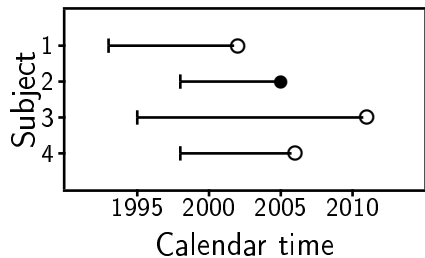
| Age band | Date in | Date out | Time (years) |
|----------|---------|----------|--------------|
| 35–39    | 1993    | 1994     | 1            |
| 40–44    | 1994    | 1999     | 5            |
| 45–49    | 1999    | 2002     | 3            |

## Example: (cf. C&H, Figures 6.1 & 6.2, p. 55)

Follow-up of cohort members by calendar time and age:

| entry

- exit because of disease onset (outcome of interest)
- exit due to other reason (censoring)



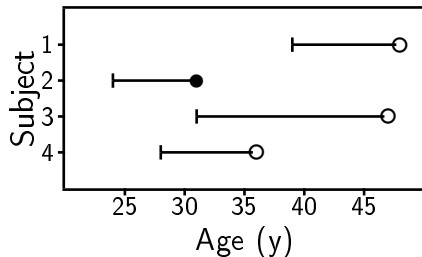
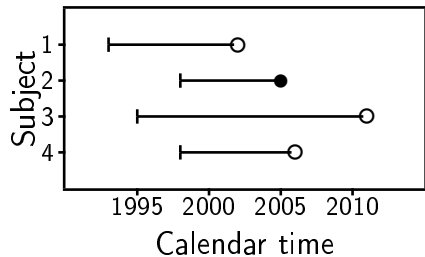
## Example: (cf. C&H, Figures 6.1 & 6.2, p. 55)

Follow-up of cohort members by calendar time and age:

| entry

● exit because of disease onset (outcome of interest)

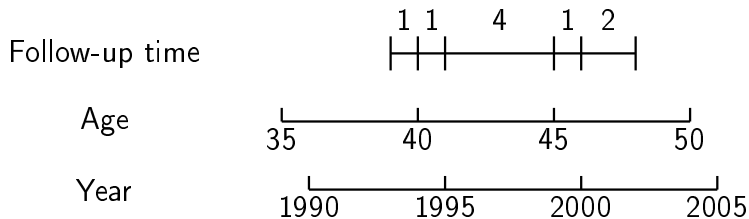
○ exit due to other reason (censoring)





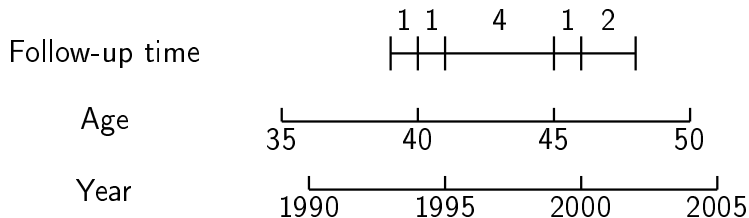
## Person-years by age and period (cf. C&H, Figure 6.4)

Subject 1: Follow-up jointly split by age and calendar time:



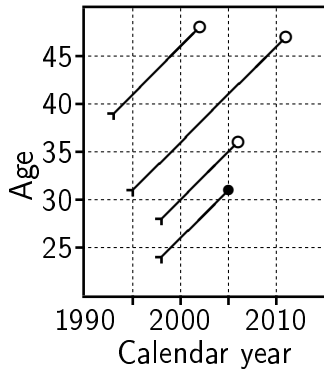
## Person-years by age and period (cf. C&H, Figure 6.4)

Subject 1: Follow-up jointly split by age and calendar time:

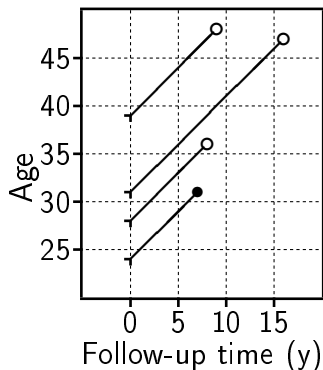
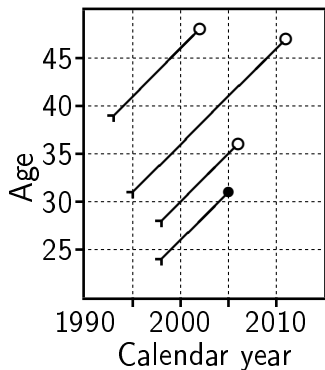


This person contributes person-time into 5 different cells defined by ageband & calendar period

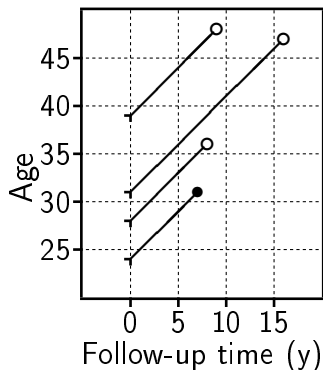
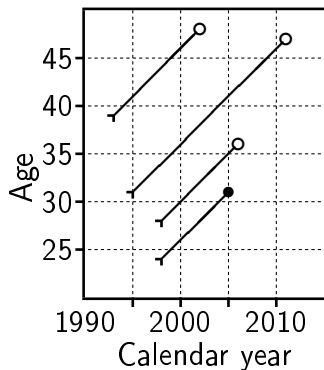
## Follow-up in Lexis-diagrams (cf. C&H, pp. 58-59)



## Follow-up in Lexis-diagrams (cf. C&H, pp. 58-59)

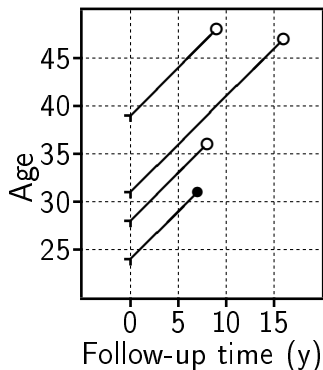
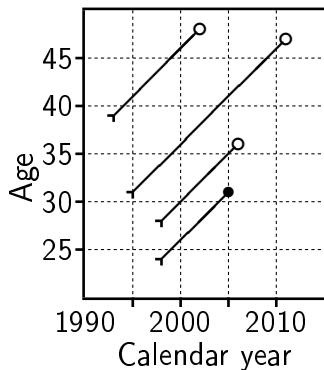


## Follow-up in Lexis-diagrams (cf. C&H, pp. 58-59)



Follow-up lines run diagonally through different ages and calendar periods.

## Follow-up in Lexis-diagrams (cf. C&H, pp. 58-59)



Follow-up lines run diagonally through different ages and calendar periods.

Incidence rates can depend **both** on age and calendar time:

See the `Lexis` functions in the `Epi` package.

# Standardization

**Bendix Carstensen & Esa Läärä**

Nordic Summerschool of  
Cancer Epidemiology

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

ELstd

## Standardization of rates

- ▶ Incidence of most cancers (and many other diseases) increases strongly by age in all populations.



## Standardization of rates

- ▶ Incidence of most cancers (and many other diseases) increases strongly by age in all populations.

## Standardization of rates

- ▶ Incidence of most cancers (and many other diseases) increases strongly by age in all populations.
  - ⇒ Most of the caseload comes from older age groups.

## Standardization of rates

- ▶ Incidence of most cancers (and many other diseases) increases strongly by age in all populations.  
⇒ Most of the caseload comes from older age groups.
- ▶ **Crude incidence rate** =  $\frac{\text{total no. of new cases}}{\text{total person-years}}$ ,

## Standardization of rates

- ▶ Incidence of most cancers (and many other diseases) increases strongly by age in all populations.
  - ⇒ Most of the caseload comes from older age groups.
- ▶ **Crude incidence rate** =  $\frac{\text{total no. of new cases}}{\text{total person-years}}$ ,
  - numerator = sum of age-specific numbers of cases,

## Standardization of rates

- ▶ Incidence of most cancers (and many other diseases) increases strongly by age in all populations.
  - ⇒ Most of the caseload comes from older age groups.
- ▶ **Crude incidence rate** =  $\frac{\text{total no. of new cases}}{\text{total person-years}}$ ,
  - numerator = sum of age-specific numbers of cases,
  - denominator = sum of age-specific person-years.

## Standardization of rates

- ▶ Incidence of most cancers (and many other diseases) increases strongly by age in all populations.
  - ⇒ Most of the caseload comes from older age groups.
- ▶ **Crude incidence rate** =  $\frac{\text{total no. of new cases}}{\text{total person-years}}$ ,
  - numerator = sum of age-specific numbers of cases,
  - denominator = sum of age-specific person-years.
- ▶ This is generally a poor summary measure.

## Standardization of rates

- ▶ Incidence of most cancers (and many other diseases) increases strongly by age in all populations.
  - ⇒ Most of the caseload comes from older age groups.
- ▶ **Crude incidence rate** =  $\frac{\text{total no. of new cases}}{\text{total person-years}}$ ,
  - numerator = sum of age-specific numbers of cases,
  - denominator = sum of age-specific person-years.
- ▶ This is generally a poor summary measure.
- ▶ Comparisons of crude incidences between populations can be very misleading, when the age structures differ.

## Standardization of rates

- ▶ Incidence of most cancers (and many other diseases) increases strongly by age in all populations.
  - ⇒ Most of the caseload comes from older age groups.
- ▶ **Crude incidence rate** =  $\frac{\text{total no. of new cases}}{\text{total person-years}}$ ,
  - numerator = sum of age-specific numbers of cases,
  - denominator = sum of age-specific person-years.
- ▶ This is generally a poor summary measure.
- ▶ Comparisons of crude incidences between populations can be very misleading, when the age structures differ.



## Standardization of rates

- ▶ Incidence of most cancers (and many other diseases) increases strongly by age in all populations.
  - ⇒ Most of the caseload comes from older age groups.
- ▶ **Crude incidence rate** =  $\frac{\text{total no. of new cases}}{\text{total person-years}}$ ,
  - numerator = sum of age-specific numbers of cases,
  - denominator = sum of age-specific person-years.
- ▶ This is generally a poor summary measure.
- ▶ Comparisons of crude incidences between populations can be very misleading, when the age structures differ.
- ▶ **Adjustment** or **standardization** for age needed!

## Ex. Male stomach cancer in Cali and Birmingham

(IS, Table 4.2, p. 71)

|     | Cali  |                   |                      | Birmingham |                   |                      |       |
|-----|-------|-------------------|----------------------|------------|-------------------|----------------------|-------|
|     | Male  | Male              | Incid.               | Male       | Male              | Incid.               |       |
|     | cases | Popu-             | Rate                 | cases      | Popu-             | Rate                 |       |
| Age | 1982  | lation            | (/10 <sup>5</sup> y) | 1983       | lation            | (/10 <sup>5</sup> y) |       |
| (y) | -86   | 1984              | 1982                 | -86        | 1985              | 1983                 | Rate  |
|     |       | ( $\times 10^3$ ) | -86                  |            | ( $\times 10^3$ ) | -86                  | ratio |

## Ex. Male stomach cancer in Cali and Birmingham

(IS, Table 4.2, p. 71)

| Age<br>(y) | Cali                         |  |   | Birmingham                   |  |   | Rate<br>ratio |
|------------|------------------------------|--|---|------------------------------|--|---|---------------|
|            | Male<br>cases<br>1982<br>-86 | Male<br>Popu-<br>lation<br>1984<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1982<br>-86 | Male<br>cases<br>1983<br>-86 | Male<br>Popu-<br>lation<br>1985<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1983<br>-86 |               |
| 0-44       | 39                           | 524.2  | <b>1.5</b>  | 79                           | 1 683.6  | <b>1.2</b>  | <i>1.25</i>   |

## Ex. Male stomach cancer in Cali and Birmingham

(IS, Table 4.2, p. 71)

|         | Cali               |  |  | Birmingham         |  |  | Rate ratio  |
|---------|--------------------|--|--|--------------------|--|--|-------------|
|         | Male cases 1982-86 | Male Population 1984 ( $\times 10^3$ ) | Incid. Rate (/10 <sup>5</sup> y) 1982-86 | Male cases 1983-86 | Male Population 1985 ( $\times 10^3$ ) | Incid. Rate (/10 <sup>5</sup> y) 1983-86 |             |
| Age (y) |                    |  |  |                    |  |  |             |
| 0-44    | 39                 | 524.2                                  | <b>1.5</b>                               | 79                 | 1 683.6                                | <b>1.2</b>                               | <i>1.25</i> |
| 45-64   | 266                | 76.3                                   | <b>69.7</b>                              | 1037               | 581.5                                  | <b>44.6</b>                              | <i>1.56</i> |

## Ex. Male stomach cancer in Cali and Birmingham

(IS, Table 4.2, p. 71)

| Age<br>(y) | Cali                         |  |   | Birmingham                   |  |   | Rate<br>ratio |
|------------|------------------------------|--|---|------------------------------|--|---|---------------|
|            | Male<br>cases<br>1982<br>-86 | Male<br>Popu-<br>lation<br>1984<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1982<br>-86 | Male<br>cases<br>1983<br>-86 | Male<br>Popu-<br>lation<br>1985<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1983<br>-86 |               |
| 0-44       | 39                           | 524.2  | <b>1.5</b>  | 79                           | 1 683.6  | <b>1.2</b>  | <i>1.25</i>   |
| 45-64      | 266                          | 76.3   | <b>69.7</b>   | 1037                         | 581.5  | <b>44.6</b>   | <i>1.56</i>   |
| 65+        | 315                          | 22.4   | <b>281.3</b>  | 2352                         | 291.1  | <b>202.0</b>  | <i>1.39</i>   |

## Ex. Male stomach cancer in Cali and Birmingham

(IS, Table 4.2, p. 71)

| Age<br>(y) | Cali                         |  |   | Birmingham                   |  |   | Rate<br>ratio |
|------------|------------------------------|--|---|------------------------------|--|---|---------------|
|            | Male<br>cases<br>1982<br>-86 | Male<br>Popu-<br>lation<br>1984<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1982<br>-86 | Male<br>cases<br>1983<br>-86 | Male<br>Popu-<br>lation<br>1985<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1983<br>-86 |               |
| 0-44       | 39                           | 524.2  | <b>1.5</b>  | 79                           | 1 683.6  | <b>1.2</b>  | <i>1.25</i>   |
| 45-64      | 266                          | 76.3   | <b>69.7</b>   | 1037                         | 581.5  | <b>44.6</b>   | <i>1.56</i>   |
| 65+        | 315                          | 22.4   | <b>281.3</b>  | 2352                         | 291.1  | <b>202.0</b>  | <i>1.39</i>   |
| Total      | 620                          | 622.9  | <b>19.9</b>   | 3468                         | 2 556.2  | <b>33.9</b>   | <i>0.59</i>   |

## Ex. Male stomach cancer in Cali and Birmingham

(IS, Table 4.2, p. 71)

| Age<br>(y) | Cali                         |  |   | Birmingham                   |  |   | Rate<br>ratio |
|------------|------------------------------|--|---|------------------------------|--|---|---------------|
|            | Male<br>cases<br>1982<br>-86 | Male<br>Popu-<br>lation<br>1984<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1982<br>-86 | Male<br>cases<br>1983<br>-86 | Male<br>Popu-<br>lation<br>1985<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1983<br>-86 |               |
| 0-44       | 39                           | 524.2  | <b>1.5</b>  | 79                           | 1 683.6  | <b>1.2</b>  | <i>1.25</i>   |
| 45-64      | 266                          | 76.3   | <b>69.7</b>   | 1037                         | 581.5  | <b>44.6</b>   | <i>1.56</i>   |
| 65+        | 315                          | 22.4   | <b>281.3</b>  | 2352                         | 291.1  | <b>202.0</b>  | <i>1.39</i>   |
| Total      | 620                          | 622.9  | <b>19.9</b>   | 3468                         | 2 556.2  | <b>33.9</b>   | <i>0.59</i>   |

- ▶ In each age group Cali has a higher incidence
  - but the crude incidence is higher in Birmingham.

## Ex. Male stomach cancer in Cali and Birmingham

(IS, Table 4.2, p. 71)

| Age<br>(y) | Cali                         |  |   | Birmingham                   |  |   | Rate<br>ratio |
|------------|------------------------------|--|---|------------------------------|--|---|---------------|
|            | Male<br>cases<br>1982<br>-86 | Male<br>Popu-<br>lation<br>1984<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1982<br>-86 | Male<br>cases<br>1983<br>-86 | Male<br>Popu-<br>lation<br>1985<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1983<br>-86 |               |
| 0-44       | 39                           | 524.2  | <b>1.5</b>  | 79                           | 1 683.6  | <b>1.2</b>  | <i>1.25</i>   |
| 45-64      | 266                          | 76.3   | <b>69.7</b>   | 1037                         | 581.5  | <b>44.6</b>   | <i>1.56</i>   |
| 65+        | 315                          | 22.4   | <b>281.3</b>  | 2352                         | 291.1  | <b>202.0</b>  | <i>1.39</i>   |
| Total      | 620                          | 622.9  | <b>19.9</b>   | 3468                         | 2 556.2  | <b>33.9</b>   | <i>0.59</i>   |

- ▶ In each age group Cali has a higher incidence
  - but the crude incidence is higher in Birmingham.



## Ex. Male stomach cancer in Cali and Birmingham

(IS, Table 4.2, p. 71)

| Age<br>(y) | Cali                         |  |   | Birmingham                   |  |   | Rate<br>ratio |
|------------|------------------------------|--|---|------------------------------|--|---|---------------|
|            | Male<br>cases<br>1982<br>-86 | Male<br>Popu-<br>lation<br>1984<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1982<br>-86 | Male<br>cases<br>1983<br>-86 | Male<br>Popu-<br>lation<br>1985<br>( $\times 10^3$ ) | Incid.<br>Rate<br>(/10 <sup>5</sup> y)<br>1983<br>-86 |               |
| 0-44       | 39                           | 524.2  | <b>1.5</b>  | 79                           | 1 683.6  | <b>1.2</b>  | 1.25          |
| 45-64      | 266                          | 76.3   | <b>69.7</b>   | 1037                         | 581.5  | <b>44.6</b>   | 1.56          |
| 65+        | 315                          | 22.4   | <b>281.3</b>  | 2352                         | 291.1  | <b>202.0</b>  | 1.39          |
| Total      | 620                          | 622.9  | <b>19.9</b>   | 3468                         | 2 556.2  | <b>33.9</b>   | 0.59          |

- ▶ In each age group Cali has a higher incidence  
– but the crude incidence is higher in Birmingham.
- ▶ *Is there a paradox?*

## Comparison of age structures (IS, Tables 4.3, 4.4)

| Age<br>(years) | % of male population |               |                 |                 |
|----------------|----------------------|---------------|-----------------|-----------------|
|                | Cali<br>1984         | B'ham<br>1985 | Finland<br>2011 | World<br>Stand. |
| 0-44           | 84                   | 66            | 56              | 74              |
| 45-64          | 12                   | 23            | 29              | 19              |
| 65+            | 4                    | 11            | 15              | 7               |
| All ages       | 100                  | 100           | 100             | 100             |

## Comparison of age structures (IS, Tables 4.3, 4.4)

| Age<br>(years) | % of male population |               |                 |                 |
|----------------|----------------------|---------------|-----------------|-----------------|
|                | Cali<br>1984         | B'ham<br>1985 | Finland<br>2011 | World<br>Stand. |
| 0-44           | 84                   | 66            | 56              | 74              |
| 45-64          | 12                   | 23            | 29              | 19              |
| 65+            | 4                    | 11            | 15              | 7               |
| All ages       | 100                  | 100           | 100             | 100             |

The fraction of old men greater in Birmingham than in Cali.

⇒ Crude rates are **confounded** by age.

## Comparison of age structures (IS, Tables 4.3, 4.4)

| Age<br>(years) | % of male population |               |                 |                 |
|----------------|----------------------|---------------|-----------------|-----------------|
|                | Cali<br>1984         | B'ham<br>1985 | Finland<br>2011 | World<br>Stand. |
| 0-44           | 84                   | 66            | 56              | 74              |
| 45-64          | 12                   | 23            | 29              | 19              |
| 65+            | 4                    | 11            | 15              | 7               |
| All ages       | 100                  | 100           | 100             | 100             |

The fraction of old men greater in Birmingham than in Cali.

⇒ Crude rates are **confounded** by age.

⇒ Any summary rate must be **adjusted for age**.

## Adjustment by direct standardisation

Age-standardised incidence rate (ASR):

$$\text{ASR} = \sum_{k=1}^K \text{weight}_k \times \text{rate}_k / \text{sum of weights} = \sum_{k=1}^K w_k \times \text{rate}_k, \quad \left( \sum_{k=1}^K w_k = 1 \right)$$

- ▶ A **weighted average** of age-specific rates over the age-groups  $k = 1, \dots, K$ .

## Adjustment by direct standardisation

Age-standardised incidence rate (ASR):

$$\text{ASR} = \sum_{k=1}^K \text{weight}_k \times \text{rate}_k / \text{sum of weights} = \sum_{k=1}^K w_k \times \text{rate}_k, \quad \left( \sum_{k=1}^K w_k = 1 \right)$$

- ▶ A **weighted average** of age-specific rates over the age-groups  $k = 1, \dots, K$ .
- ▶ Weights describe the age distribution of some **standard population**.

## Adjustment by direct standardisation

Age-standardised incidence rate (ASR):

$$\text{ASR} = \sum_{k=1}^K \text{weight}_k \times \text{rate}_k / \text{sum of weights} = \sum_{k=1}^K w_k \times \text{rate}_k, \quad \left( \sum_{k=1}^K w_k = 1 \right)$$

- ▶ A **weighted average** of age-specific rates over the age-groups  $k = 1, \dots, K$ .
- ▶ Weights describe the age distribution of some **standard population**.
- ▶ Standard population can be

## Adjustment by direct standardisation

Age-standardised incidence rate (ASR):

$$\text{ASR} = \sum_{k=1}^K \text{weight}_k \times \text{rate}_k / \text{sum of weights} = \sum_{k=1}^K w_k \times \text{rate}_k, \quad \left( \sum_{k=1}^K w_k = 1 \right)$$

- ▶ A **weighted average** of age-specific rates over the age-groups  $k = 1, \dots, K$ .
- ▶ Weights describe the age distribution of some **standard population**.
- ▶ Standard population can be
  - ▶ real (e.g. one of the populations compared, or their total), or



## Adjustment by direct standardisation

Age-standardised incidence rate (ASR):

$$\text{ASR} = \sum_{k=1}^K \text{weight}_k \times \text{rate}_k / \text{sum of weights} = \sum_{k=1}^K w_k \times \text{rate}_k, \quad \left( \sum_{k=1}^K w_k = 1 \right)$$

- ▶ A **weighted average** of age-specific rates over the age-groups  $k = 1, \dots, K$ .
- ▶ Weights describe the age distribution of some **standard population**.
- ▶ Standard population can be
  - ▶ real (e.g. one of the populations compared, or their total), or
  - ▶ fictitious (e.g. World Standard Population, WSP)

## Adjustment by direct standardisation

Age-standardised incidence rate (ASR):

$$\text{ASR} = \sum_{k=1}^K \text{weight}_k \times \text{rate}_k / \text{sum of weights} = \sum_{k=1}^K w_k \times \text{rate}_k, \quad \left( \sum_{k=1}^K w_k = 1 \right)$$

- ▶ A **weighted average** of age-specific rates over the age-groups  $k = 1, \dots, K$ .
- ▶ Weights describe the age distribution of some **standard population**.
- ▶ Standard population can be
  - ▶ real (e.g. one of the populations compared, or their total), or
  - ▶ fictitious (e.g. World Standard Population, WSP)
- ▶ Choice of standard population always more or less arbitrary.

## Some standard populations:

| Age group (years) | African | World   | European | NORDCAN (2000) |
|-------------------|---------|---------|----------|----------------|
| 0-4               | 10 000  | 12 000  | 8 000    | 5 900          |
| 5-9               | 10 000  | 10 000  | 7 000    | 6 600          |
| 10-14             | 10 000  | 9 000   | 7 000    | 6 200          |
| 15-19             | 10 000  | 9 000   | 7 000    | 5 800          |
| 20-24             | 10 000  | 8 000   | 7 000    | 6 100          |
| 25-29             | 10 000  | 8 000   | 7 000    | 6 800          |
| 30-34             | 10 000  | 6 000   | 7 000    | 7 300          |
| 35-39             | 10 000  | 6 000   | 7 000    | 7 300          |
| 40-44             | 5 000   | 6 000   | 7 000    | 7 000          |
| 45-49             | 5 000   | 6 000   | 7 000    | 6 900          |
| 50-54             | 3 000   | 5 000   | 7 000    | 7 400          |
| 55-59             | 2 000   | 4 000   | 6 000    | 6 100          |
| 60-64             | 2 000   | 4 000   | 5 000    | 4 800          |
| 65-69             | 1 000   | 3 000   | 4 000    | 4 100          |
| 70-74             | 1 000   | 2 000   | 3 000    | 3 900          |
| 75-79             | 500     | 1 000   | 2 000    | 3 500          |
| 80-84             | 300     | 500     | 1 000    | 2 400          |
| 85+               | 200     | 500     | 1 000    | 1 900          |
| Total             | 100 000 | 100 000 | 100 000  | 100 000        |

## Stomach cancer in Cali & Birmingham

Age-standardized rates by the World Standard Population:

| Age                          | Cali              |              | Birmingham        |              |
|------------------------------|-------------------|--------------|-------------------|--------------|
|                              | Rate <sup>a</sup> | Weight       | Rate <sup>a</sup> | Weight       |
| 0-44                         | 1.5 ×             | 0.74 = 1.11  | 1.2 ×             | 0.74 = 0.89  |
| 45-64                        | 69.7 ×            | 0.19 = 13.24 | 44.6 ×            | 0.19 = 8.47  |
| 65+                          | 281.3 ×           | 0.07 = 19.69 | 202.0 ×           | 0.07 = 14.14 |
| <b>Age-standardised rate</b> |                   | <b>34.04</b> |                   | <b>23.50</b> |

## Stomach cancer in Cali & Birmingham

Age-standardized rates by the World Standard Population:

| Age                          | Cali              |              | Birmingham        |              |
|------------------------------|-------------------|--------------|-------------------|--------------|
|                              | Rate <sup>a</sup> | Weight       | Rate <sup>a</sup> | Weight       |
| 0-44                         | 1.5 ×             | 0.74 = 1.11  | 1.2 ×             | 0.74 = 0.89  |
| 45-64                        | 69.7 ×            | 0.19 = 13.24 | 44.6 ×            | 0.19 = 8.47  |
| 65+                          | 281.3 ×           | 0.07 = 19.69 | 202.0 ×           | 0.07 = 14.14 |
| <b>Age-standardised rate</b> |                   | <b>34.04</b> |                   | <b>23.50</b> |

- ▶ ASR in Cali higher – coherent with the age-specific rates.

## Stomach cancer in Cali & Birmingham

Age-standardized rates by the World Standard Population:

| Age                          | Cali              |              | Birmingham        |              |
|------------------------------|-------------------|--------------|-------------------|--------------|
|                              | Rate <sup>a</sup> | Weight       | Rate <sup>a</sup> | Weight       |
| 0-44                         | 1.5 ×             | 0.74 = 1.11  | 1.2 ×             | 0.74 = 0.89  |
| 45-64                        | 69.7 ×            | 0.19 = 13.24 | 44.6 ×            | 0.19 = 8.47  |
| 65+                          | 281.3 ×           | 0.07 = 19.69 | 202.0 ×           | 0.07 = 14.14 |
| <b>Age-standardised rate</b> |                   | <b>34.04</b> |                   | <b>23.50</b> |

- ▶ ASR in Cali higher – coherent with the age-specific rates.
- ▶ Summary rate ratio estimate: **standardized rate ratio**

$$\text{SRR} = 34.0/23.5 = 1.44.$$

## Stomach cancer in Cali & Birmingham

Age-standardized rates by the World Standard Population:

| Age                          | Cali              |              | Birmingham        |              |
|------------------------------|-------------------|--------------|-------------------|--------------|
|                              | Rate <sup>a</sup> | Weight       | Rate <sup>a</sup> | Weight       |
| 0-44                         | 1.5 ×             | 0.74 = 1.11  | 1.2 ×             | 0.74 = 0.89  |
| 45-64                        | 69.7 ×            | 0.19 = 13.24 | 44.6 ×            | 0.19 = 8.47  |
| 65+                          | 281.3 ×           | 0.07 = 19.69 | 202.0 ×           | 0.07 = 14.14 |
| <b>Age-standardised rate</b> | <b>34.04</b>      |              | <b>23.50</b>      |              |

- ▶ ASR in Cali higher – coherent with the age-specific rates.
- ▶ Summary rate ratio estimate: **standardized rate ratio**

$$\text{SRR} = 34.0/23.5 = 1.44.$$

## Stomach cancer in Cali & Birmingham

Age-standardized rates by the World Standard Population:

| Age                          | Cali              |              | Birmingham        |              |
|------------------------------|-------------------|--------------|-------------------|--------------|
|                              | Rate <sup>a</sup> | Weight       | Rate <sup>a</sup> | Weight       |
| 0-44                         | 1.5 ×             | 0.74 = 1.11  | 1.2 ×             | 0.74 = 0.89  |
| 45-64                        | 69.7 ×            | 0.19 = 13.24 | 44.6 ×            | 0.19 = 8.47  |
| 65+                          | 281.3 ×           | 0.07 = 19.69 | 202.0 ×           | 0.07 = 14.14 |
| <b>Age-standardised rate</b> |                   | <b>34.04</b> |                   | <b>23.50</b> |

- ▶ ASR in Cali higher – coherent with the age-specific rates.
- ▶ Summary rate ratio estimate: **standardized rate ratio**  
$$\text{SRR} = 34.0/23.5 = 1.44.$$
- ▶ This is also called as **comparative mortality figure (CMF)**, when the outcome is death (from cause  $C$  or from all causes).



## Cumulative rate and “cumulative risk”

- ▶ A neutral alternative to arbitrary standard population for age-adjustment is provided by **cumulative rate**:

$$\text{CumRate} = \sum_{k=1}^K \text{width}_k \times \text{rate}_k,$$

## Cumulative rate and “cumulative risk”

- ▶ A neutral alternative to arbitrary standard population for age-adjustment is provided by **cumulative rate**:

$$\text{CumRate} = \sum_{k=1}^K \text{width}_k \times \text{rate}_k,$$

- ▶ Weights are now widths of the agebands to be included, usually up to 75 y.

## Cumulative rate and “cumulative risk”

- ▶ A neutral alternative to arbitrary standard population for age-adjustment is provided by **cumulative rate**:

$$\text{CumRate} = \sum_{k=1}^K \text{width}_k \times \text{rate}_k,$$

- ▶ Weights are now widths of the agebands to be included, usually up to 75 y.
- ▶ NORDCAN & GLOBOCAN use the transformation:

$$\text{CumRisk} = 1 - \exp(-\text{CumRate}),$$

calling it as the **cumulative risk** of getting the disease by given age, in the absence of competing causes.

## Cumulative rate and “cumulative risk”

- ▶ A neutral alternative to arbitrary standard population for age-adjustment is provided by **cumulative rate**:

$$\text{CumRate} = \sum_{k=1}^K \text{width}_k \times \text{rate}_k,$$

- ▶ Weights are now widths of the agebands to be included, usually up to 75 y.
- ▶ NORDCAN & GLOBOCAN use the transformation:

$$\text{CumRisk} = 1 - \exp(-\text{CumRate}),$$

calling it as the **cumulative risk** of getting the disease by given age, in the absence of competing causes.

- ▶ Since competing events are present, the probability interpretation of CumRisk is somewhat problematic.

## Stomach cancer in Cali & Birmingham

From age-specific rates of Table 4.2, the cumulative rates up to 65 years and their ratio are

$$\text{Cali: } 45 \text{ y} \times \frac{1.5}{10^5 \text{ y}} + 20 \text{ y} \times \frac{69.7}{10^5 \text{ y}} = 0.0146 = \mathbf{1.46} \text{ per } 100$$

## Stomach cancer in Cali & Birmingham

From age-specific rates of Table 4.2, the cumulative rates up to 65 years and their ratio are

$$\text{Cali: } 45 \text{ y} \times \frac{1.5}{10^5 \text{ y}} + 20 \text{ y} \times \frac{69.7}{10^5 \text{ y}} = 0.0146 = \mathbf{1.46} \text{ per } 100$$

$$\text{B'ham: } 45 \text{ y} \times \frac{1.2}{10^5 \text{ y}} + 20 \text{ y} \times \frac{44.6}{10^5 \text{ y}} = 0.0095 = \mathbf{0.95} \text{ per } 100$$

## Stomach cancer in Cali & Birmingham

From age-specific rates of Table 4.2, the cumulative rates up to 65 years and their ratio are

$$\text{Cali: } 45 \text{ y} \times \frac{1.5}{10^5 \text{ y}} + 20 \text{ y} \times \frac{69.7}{10^5 \text{ y}} = 0.0146 = \mathbf{1.46} \text{ per } 100$$

$$\text{B'ham: } 45 \text{ y} \times \frac{1.2}{10^5 \text{ y}} + 20 \text{ y} \times \frac{44.6}{10^5 \text{ y}} = 0.0095 = \mathbf{0.95} \text{ per } 100$$

$$\text{ratio: } \mathbf{1.46/0.95 = 1.54}$$

## Stomach cancer in Cali & Birmingham

From age-specific rates of Table 4.2, the cumulative rates up to 65 years and their ratio are

$$\text{Cali: } 45 \text{ y} \times \frac{1.5}{10^5 \text{ y}} + 20 \text{ y} \times \frac{69.7}{10^5 \text{ y}} = 0.0146 = \mathbf{1.46} \text{ per } 100$$

$$\text{B'ham: } 45 \text{ y} \times \frac{1.2}{10^5 \text{ y}} + 20 \text{ y} \times \frac{44.6}{10^5 \text{ y}} = 0.0095 = \mathbf{0.95} \text{ per } 100$$

$$\text{ratio: } \mathbf{1.46/0.95 = 1.54}$$

“Cumulative risks” & their ratio up to 65 y:



## Stomach cancer in Cali & Birmingham

From age-specific rates of Table 4.2, the cumulative rates up to 65 years and their ratio are

$$\text{Cali: } 45 \text{ y} \times \frac{1.5}{10^5 \text{ y}} + 20 \text{ y} \times \frac{69.7}{10^5 \text{ y}} = 0.0146 = \mathbf{1.46} \text{ per } 100$$

$$\text{B'ham: } 45 \text{ y} \times \frac{1.2}{10^5 \text{ y}} + 20 \text{ y} \times \frac{44.6}{10^5 \text{ y}} = 0.0095 = \mathbf{0.95} \text{ per } 100$$

$$\text{ratio: } \mathbf{1.46/0.95 = 1.54}$$

“Cumulative risks” & their ratio up to 65 y:

$$\text{Cali: } 1 - \exp(-0.0146) = 0.0145 = \mathbf{1.45\%}$$

## Stomach cancer in Cali & Birmingham

From age-specific rates of Table 4.2, the cumulative rates up to 65 years and their ratio are

$$\text{Cali: } 45 \text{ y} \times \frac{1.5}{10^5 \text{ y}} + 20 \text{ y} \times \frac{69.7}{10^5 \text{ y}} = 0.0146 = \mathbf{1.46} \text{ per } 100$$

$$\text{B'ham: } 45 \text{ y} \times \frac{1.2}{10^5 \text{ y}} + 20 \text{ y} \times \frac{44.6}{10^5 \text{ y}} = 0.0095 = \mathbf{0.95} \text{ per } 100$$

$$\text{ratio: } \mathbf{1.46/0.95 = 1.54}$$

“Cumulative risks” & their ratio up to 65 y:

$$\text{Cali: } 1 - \exp(-0.0146) = 0.0145 = \mathbf{1.45\%}$$

$$\text{B'ham: } 1 - \exp(-0.0095) = 0.0094 = \mathbf{0.94\%}$$

## Stomach cancer in Cali & Birmingham

From age-specific rates of Table 4.2, the cumulative rates up to 65 years and their ratio are

$$\text{Cali: } 45 \text{ y} \times \frac{1.5}{10^5 \text{ y}} + 20 \text{ y} \times \frac{69.7}{10^5 \text{ y}} = 0.0146 = \mathbf{1.46} \text{ per } 100$$

$$\text{B'ham: } 45 \text{ y} \times \frac{1.2}{10^5 \text{ y}} + 20 \text{ y} \times \frac{44.6}{10^5 \text{ y}} = 0.0095 = \mathbf{0.95} \text{ per } 100$$

$$\text{ratio: } \mathbf{1.46/0.95 = 1.54}$$

“Cumulative risks” & their ratio up to 65 y:

$$\text{Cali: } 1 - \exp(-0.0146) = 0.0145 = \mathbf{1.45\%}$$

$$\text{B'ham: } 1 - \exp(-0.0095) = 0.0094 = \mathbf{0.94\%}$$

$$\text{ratio: } \mathbf{1.45/0.94 = 1.54}$$

## Stomach cancer in Cali & Birmingham

From age-specific rates of Table 4.2, the cumulative rates up to 65 years and their ratio are

$$\text{Cali: } 45 \text{ y} \times \frac{1.5}{10^5 \text{ y}} + 20 \text{ y} \times \frac{69.7}{10^5 \text{ y}} = 0.0146 = \mathbf{1.46} \text{ per } 100$$

$$\text{B'ham: } 45 \text{ y} \times \frac{1.2}{10^5 \text{ y}} + 20 \text{ y} \times \frac{44.6}{10^5 \text{ y}} = 0.0095 = \mathbf{0.95} \text{ per } 100$$

$$\text{ratio: } \mathbf{1.46/0.95 = 1.54}$$

“Cumulative risks” & their ratio up to 65 y:

$$\text{Cali: } 1 - \exp(-0.0146) = 0.0145 = \mathbf{1.45\%}$$

$$\text{B'ham: } 1 - \exp(-0.0095) = 0.0094 = \mathbf{0.94\%}$$

$$\text{ratio: } \mathbf{1.45/0.94 = 1.54}$$

Cumulative rate and cumulative risk are roughly the same of  $< 0.05$ .

**NB:** For more appropriate estimates of cumulative risks, correction for total mortality (competing event) needed.

## Cum. measures in B'ham with 5-y groups (IS, Fig 4.11)

| Age-group (years)  | Incidence rate (per 100 000 pyrs) |
|--------------------|-----------------------------------|
| 0-4, . . . , 15-19 | 0.0                               |
| 20-24, 25-29       | 0.1                               |
| 30-34              | 0.9                               |
| 35-39              | 3.5                               |
| 40-44              | 6.7                               |
| 45-49              | 14.5                              |
| 50-54              | 26.8                              |
| 55-59              | 52.6                              |
| 60-64              | 87.2                              |
| 65-69              | 141.7                             |
| 70-74              | 190.8                             |
| Sum                | 524.9                             |

## Cum. measures in B'ham with 5-y groups (IS, Fig 4.11)

| Age-group (years)  | Incidence rate (per 100 000 pyrs) |
|--------------------|-----------------------------------|
| 0-4, . . . , 15-19 | 0.0                               |
| 20-24, 25-29       | 0.1                               |
| 30-34              | 0.9                               |
| 35-39              | 3.5                               |
| 40-44              | 6.7                               |
| 45-49              | 14.5                              |
| 50-54              | 26.8                              |
| 55-59              | 52.6                              |
| 60-64              | 87.2                              |
| 65-69              | 141.7                             |
| 70-74              | 190.8                             |
| Sum                | 524.9                             |

$$\text{Cumulative rate 0-75 y} = 5 \text{ y} \times \frac{524.9}{10^5 \text{ y}} = 0.0262 = \mathbf{2.6} \text{ per 100}$$

## Cum. measures in B'ham with 5-y groups (IS, Fig 4.11)

| Age-group (years)  | Incidence rate (per 100 000 pyrs) |
|--------------------|-----------------------------------|
| 0-4, . . . , 15-19 | 0.0                               |
| 20-24, 25-29       | 0.1                               |
| 30-34              | 0.9                               |
| 35-39              | 3.5                               |
| 40-44              | 6.7                               |
| 45-49              | 14.5                              |
| 50-54              | 26.8                              |
| 55-59              | 52.6                              |
| 60-64              | 87.2                              |
| 65-69              | 141.7                             |
| 70-74              | 190.8                             |
| Sum                | 524.9                             |

$$\text{Cumulative rate 0-75 y} = 5 \text{ y} \times \frac{524.9}{10^5 \text{ y}} = 0.0262 = \mathbf{2.6} \text{ per 100}$$

$$\text{"Cumulative risk" 0-75 y} = 1 - \exp(-0.0262) = 0.0259 = \mathbf{2.6\%}.$$

## Cumulative and life-time risks

It is, of course, an interesting and relevant question to ask:

*“What are my chances of getting cancer  $C$ , say, in the next 10 years, between ages 50 to 75 years, or during the whole lifetime?”*



## Cumulative and life-time risks

It is, of course, an interesting and relevant question to ask:

*“What are my chances of getting cancer  $C$ , say, in the next 10 years, between ages 50 to 75 years, or during the whole lifetime?”*

This is not easy to answer.

- ▶ Fully individualized risks are unidentifiable.

## Cumulative and life-time risks

It is, of course, an interesting and relevant question to ask:

*“What are my chances of getting cancer  $C$ , say, in the next 10 years, between ages 50 to 75 years, or during the whole lifetime?”*

This is not easy to answer.

- ▶ Fully individualized risks are unidentifiable.
- ▶ Age-specific and standardized rates are not very informative as such.

## Cumulative and life-time risks

It is, of course, an interesting and relevant question to ask:

*“What are my chances of getting cancer  $C$ , say, in the next 10 years, between ages 50 to 75 years, or during the whole lifetime?”*

This is not easy to answer.

- ▶ Fully individualized risks are unidentifiable.
- ▶ Age-specific and standardized rates are not very informative as such.
- ▶ Average cumulative risks are often estimated from cumulative rates using the simple formula above.

## Cumulative and life-time risks

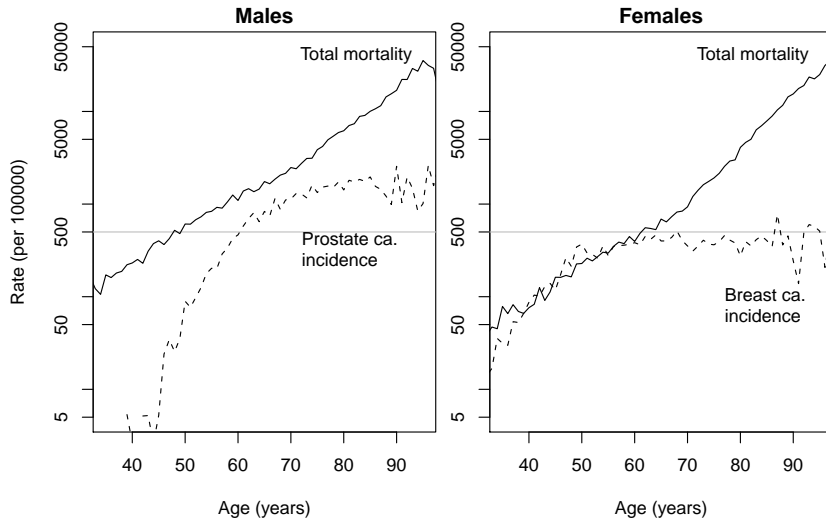
It is, of course, an interesting and relevant question to ask:

*“What are my chances of getting cancer  $C$ , say, in the next 10 years, between ages 50 to 75 years, or during the whole lifetime?”*

This is not easy to answer.

- ▶ Fully individualized risks are unidentifiable.
- ▶ Age-specific and standardized rates are not very informative as such.
- ▶ Average cumulative risks are often estimated from cumulative rates using the simple formula above.
- ▶ Yet, these naive estimates fictitiously presume that a person would not die from any cause before cancer hits him/her, but could even survive forever!

# Total mortality and incidence of two cancers, Finland 2005



## Estimation of cumulative risks

- ▶ The probability of contracting cancer during realistic lifespan or in any age range depends not only on age-specific hazard rates of cancer itself but also of probabilities of overall survival up to relevant ages,

## Estimation of cumulative risks

- ▶ The probability of contracting cancer during realistic lifespan or in any age range depends not only on age-specific hazard rates of cancer itself but also of probabilities of overall survival up to relevant ages,
- ▶ Hence, the dependence of total mortality by age in the population at risk must be incorporated in the estimation of cumulative risks of cancer.

## Estimation of cumulative risks

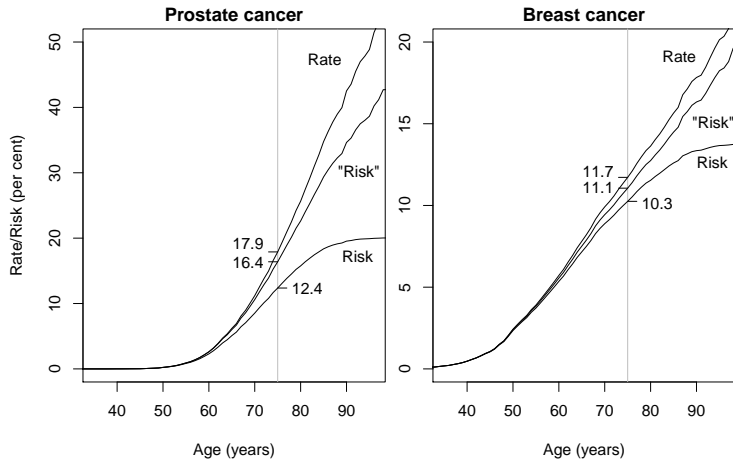
- ▶ The probability of contracting cancer during realistic lifespan or in any age range depends not only on age-specific hazard rates of cancer itself but also of probabilities of overall survival up to relevant ages,
- ▶ Hence, the dependence of total mortality by age in the population at risk must be incorporated in the estimation of cumulative risks of cancer.
- ▶ When this is properly done, the corrected estimates of cumulative risk will always be lower than the uncorrected “risks”.



## Estimation of cumulative risks

- ▶ The probability of contracting cancer during realistic lifespan or in any age range depends not only on age-specific hazard rates of cancer itself but also of probabilities of overall survival up to relevant ages,
- ▶ Hence, the dependence of total mortality by age in the population at risk must be incorporated in the estimation of cumulative risks of cancer.
- ▶ When this is properly done, the corrected estimates of cumulative risk will always be lower than the uncorrected “risks”.
- ▶ The magnitude of bias in the latter grows by age, but is reduced with increased life expectancy.

## Cumulative measures, Finland 2005



Greater differences in males reflect shorter life expectancy and relatively high rates of prostate ca. in old ages.

## Special cohorts of exposed subjects

- ▶ Occupational cohorts, exposed to potentially hazardous agents, e.g. asbestos workers, uranium miners (see Johanna's lecture on cohort studies)

## Special cohorts of exposed subjects

- ▶ Occupational cohorts, exposed to potentially hazardous agents, e.g. asbestos workers, uranium miners (see Johanna's lecture on cohort studies)
- ▶ Cohorts of patients on intensive treatment, which may have harmful long-term side-effects, e.g. people with a history of childhood cancer.

## Special cohorts of exposed subjects

- ▶ Occupational cohorts, exposed to potentially hazardous agents, e.g. asbestos workers, uranium miners (see Johanna's lecture on cohort studies)
- ▶ Cohorts of patients on intensive treatment, which may have harmful long-term side-effects, e.g. people with a history of childhood cancer.
- ▶ Often no internal comparison group of unexposed subjects available.

## Special cohorts of exposed subjects

- ▶ Occupational cohorts, exposed to potentially hazardous agents, e.g. asbestos workers, uranium miners (see Johanna's lecture on cohort studies)
- ▶ Cohorts of patients on intensive treatment, which may have harmful long-term side-effects, e.g. people with a history of childhood cancer.
- ▶ Often no internal comparison group of unexposed subjects available.

## Special cohorts of exposed subjects

- ▶ Occupational cohorts, exposed to potentially hazardous agents, e.g. asbestos workers, uranium miners (see Johanna's lecture on cohort studies)
- ▶ Cohorts of patients on intensive treatment, which may have harmful long-term side-effects, e.g. people with a history of childhood cancer.
- ▶ Often no internal comparison group of unexposed subjects available.

**Question:** Do incidence or mortality rates in the **exposed** target cohort differ from those of a roughly comparable **reference** population?

Reference rates obtained from:

- ▶ population statistics (mortality rates)

## Special cohorts of exposed subjects

- ▶ Occupational cohorts, exposed to potentially hazardous agents, e.g. asbestos workers, uranium miners (see Johanna's lecture on cohort studies)
- ▶ Cohorts of patients on intensive treatment, which may have harmful long-term side-effects, e.g. people with a history of childhood cancer.
- ▶ Often no internal comparison group of unexposed subjects available.

**Question:** Do incidence or mortality rates in the **exposed** target cohort differ from those of a roughly comparable **reference** population?

Reference rates obtained from:

- ▶ population statistics (mortality rates)
- ▶ disease & hospital discharge registers (incidence)



## Observed and expected cases—SIR (indirect standardization)

- ▶ Compare rates in a study cohort with a standard set of age-specific rates from the reference population.

## Observed and expected cases—SIR (indirect standardization)

- ▶ Compare rates in a study cohort with a standard set of age-specific rates from the reference population.
- ▶ Reference rates normally based on large numbers of cases, so they are assumed to be “known” without error.

## Observed and expected cases—SIR (indirect standardization)

- ▶ Compare rates in a study cohort with a standard set of age-specific rates from the reference population.
- ▶ Reference rates normally based on large numbers of cases, so they are assumed to be “known” without error.
- ▶ Calculate **expected** number of cases,  $E$ , if the standard age-specific rates had applied in our study cohort.

## Observed and expected cases—SIR (indirect standardization)

- ▶ Compare rates in a study cohort with a standard set of age-specific rates from the reference population.
- ▶ Reference rates normally based on large numbers of cases, so they are assumed to be “known” without error.
- ▶ Calculate **expected** number of cases,  $E$ , if the standard age-specific rates had applied in our study cohort.
- ▶ Compare this with the **observed** number of cases,  $D$ , by the **standardized incidence ratio** (SIR)

$$\text{SIR} = D/E, \quad \text{SE}(\log[\text{SIR}]) = 1/\sqrt{D}$$

## Observed and expected cases—SIR (indirect standardization)

- ▶ Compare rates in a study cohort with a standard set of age-specific rates from the reference population.
- ▶ Reference rates normally based on large numbers of cases, so they are assumed to be “known” without error.
- ▶ Calculate **expected** number of cases,  $E$ , if the standard age-specific rates had applied in our study cohort.
- ▶ Compare this with the **observed** number of cases,  $D$ , by the **standardized incidence ratio** (SIR)

$$\text{SIR} = D/E, \quad \text{SE}(\log[\text{SIR}]) = 1/\sqrt{D}$$

- ▶ Analogously, **standardized mortality ratio** (SMR) with death as outcome.

## Example: Hormone therapy and breast cancer

- ▶ A cohort of 974 women treated with hormone (replacement) therapy (HT) were followed up.

| Age      | $Y$  | $\lambda_a^*$ | $E$   |
|----------|------|---------------|-------|
| 40–44    | 975  | 113           | 1.10  |
| 45–49    | 1079 | 162           | 1.75  |
| 50–54    | 2161 | 151           | 3.26  |
| 55–59    | 2793 | 183           | 5.11  |
| 60–64    | 3096 | 179           | 5.54  |
| $\Sigma$ |      |               | 16.77 |

## Example: Hormone therapy and breast cancer

- ▶ A cohort of 974 women treated with hormone (replacement) therapy (HT) were followed up.
- ▶  $D = 15$  incident cases of breast cancer were observed.

| Age      | $Y$  | $\lambda_a^*$ | $E$   |
|----------|------|---------------|-------|
| 40–44    | 975  | 113           | 1.10  |
| 45–49    | 1079 | 162           | 1.75  |
| 50–54    | 2161 | 151           | 3.26  |
| 55–59    | 2793 | 183           | 5.11  |
| 60–64    | 3096 | 179           | 5.54  |
| $\Sigma$ |      |               | 16.77 |

## Example: Hormone therapy and breast cancer

- ▶ A cohort of 974 women treated with hormone (replacement) therapy (HT) were followed up.
- ▶  $D = 15$  incident cases of breast cancer were observed.
- ▶ Person-years ( $Y$ ) and reference rates ( $\lambda_a^*$ , per 100000 y) by age group:

| Age      | $Y$  | $\lambda_a^*$ | $E$   |
|----------|------|---------------|-------|
| 40–44    | 975  | 113           | 1.10  |
| 45–49    | 1079 | 162           | 1.75  |
| 50–54    | 2161 | 151           | 3.26  |
| 55–59    | 2793 | 183           | 5.11  |
| 60–64    | 3096 | 179           | 5.54  |
| $\Sigma$ |      |               | 16.77 |



## Ex: HT and breast ca. (cont'd)

- ▶ “Expected” cases at ages 40–44:

$$975 \times \frac{113}{100\,000} = 1.10$$

## Ex: HT and breast ca. (cont'd)

- ▶ “Expected” cases at ages 40–44:

$$975 \times \frac{113}{100\,000} = 1.10$$

- ▶ Total “expected” cases is  $E = 16.77$

## Ex: HT and breast ca. (cont'd)

- ▶ “Expected” cases at ages 40–44:

$$975 \times \frac{113}{100\,000} = 1.10$$

- ▶ Total “expected” cases is  $E = 16.77$
- ▶  $SIR = 15/16.77 = \mathbf{0.89}$ .

## Ex: HT and breast ca. (cont'd)

- ▶ “Expected” cases at ages 40–44:

$$975 \times \frac{113}{100\,000} = 1.10$$

- ▶ Total “expected” cases is  $E = 16.77$
- ▶  $SIR = 15/16.77 = \mathbf{0.89}$ .
- ▶ Error-factor:  $\exp(1.96 \times \sqrt{1/15}) = 1.66$

## Ex: HT and breast ca. (cont'd)

- ▶ “Expected” cases at ages 40–44:

$$975 \times \frac{113}{100\,000} = 1.10$$

- ▶ Total “expected” cases is  $E = 16.77$
- ▶  $SIR = 15/16.77 = \mathbf{0.89}$ .
- ▶ Error-factor:  $\exp(1.96 \times \sqrt{1/15}) = 1.66$
- ▶ 95% confidence interval is:

$$0.89 \overset{\times}{\div} 1.66 = (0.54, 1.48)$$

## SIR for Cali with Birmingham as reference (IS: Fig. 4.9)

Total person-years at risk and expected number of cases in Cali 1982-86 based on age-specific rates in Birmingham

| Age             | Person-years                      | Expected cases in Cali                |
|-----------------|-----------------------------------|---------------------------------------|
| 0-44            | $524\,220 \times 5 = 2\,621\,100$ | $0.000012 \times 2\,621\,100 = 31.45$ |
| 45-64           | $76\,304 \times 5 = 381\,520$     | $0.000446 \times 381\,520 = 170.15$   |
| 65+             | $22\,398 \times 5 = 111\,990$     | $0.002020 \times 111\,990 = 226.00$   |
| <b>All ages</b> | <b>= 3 114 610</b>                | <b>Total expected (E) 427.82</b>      |

## SIR for Cali with Birmingham as reference (IS: Fig. 4.9)

Total person-years at risk and expected number of cases in Cali 1982-86 based on age-specific rates in Birmingham

| Age             | Person-years                      | Expected cases in Cali                |
|-----------------|-----------------------------------|---------------------------------------|
| 0-44            | $524\,220 \times 5 = 2\,621\,100$ | $0.000012 \times 2\,621\,100 = 31.45$ |
| 45-64           | $76\,304 \times 5 = 381\,520$     | $0.000446 \times 381\,520 = 170.15$   |
| 65+             | $22\,398 \times 5 = 111\,990$     | $0.002020 \times 111\,990 = 226.00$   |
| <b>All ages</b> | <b>= 3 114 610</b>                | <b>Total expected (E) 427.82</b>      |

Total observed number  $O = 620$ .

## SIR for Cali with Birmingham as reference (IS: Fig. 4.9)

Total person-years at risk and expected number of cases in Cali 1982-86 based on age-specific rates in Birmingham

| Age             | Person-years                      | Expected cases in Cali                |
|-----------------|-----------------------------------|---------------------------------------|
| 0-44            | $524\,220 \times 5 = 2\,621\,100$ | $0.000012 \times 2\,621\,100 = 31.45$ |
| 45-64           | $76\,304 \times 5 = 381\,520$     | $0.000446 \times 381\,520 = 170.15$   |
| 65+             | $22\,398 \times 5 = 111\,990$     | $0.002020 \times 111\,990 = 226.00$   |
| <b>All ages</b> | <b>= 3 114 610</b>                | <b>Total expected (E) 427.82</b>      |

Total observed number  $O = 620$ .

Standardised incidence ratio:

$$\text{SIR} = \frac{O}{E} = \frac{620}{427.8} = 1.45 \quad (\text{or } 145 \text{ per } 100)$$



## Crude and adjusted rates compared (IS: Table 4.6)

|                                  | Cali,<br>1982-86 | B'ham,<br>1983-86 | Rate<br>ratio |
|----------------------------------|------------------|-------------------|---------------|
| Crude rates (/10 <sup>5</sup> y) | 19.9             | 33.9              | <i>0.59</i>   |

## Crude and adjusted rates compared (IS: Table 4.6)

|   | Cali,<br>1982-86 | B'ham,<br>1983-86 | Rate<br>ratio |
|---|------------------|-------------------|---------------|
| Crude rates (/10 <sup>5</sup> y)                              | 19.9             | 33.9              | <i>0.59</i>   |
| ASR (/10 <sup>5</sup> y) <sup>B</sup> with 3 broad age groups | 48.0             | 33.9              | <i>1.42</i>   |
| ASR (/10 <sup>5</sup> y) <sup>C</sup>                         | —                | 14.4              | <i>1.38</i>   |
| ASR (/10 <sup>5</sup> y) <sup>W</sup>                         | —                | 23.5              | <i>1.44</i>   |

## Crude and adjusted rates compared (IS: Table 4.6)

|   | Cali,<br>1982-86 | B'ham,<br>1983-86 | Rate<br>ratio |
|---|------------------|-------------------|---------------|
| Crude rates (/10 <sup>5</sup> y)                              | 19.9             | 33.9              | <i>0.59</i>   |
| ASR (/10 <sup>5</sup> y) <sup>B</sup> with 3 broad age groups | 48.0             | 33.9              | <i>1.42</i>   |
| ASR (/10 <sup>5</sup> y) <sup>C</sup> —"                      | 19.9             | 14.4              | <i>1.38</i>   |
| ASR (/10 <sup>5</sup> y) <sup>W</sup> —"                      | 34.0             | 23.5              | <i>1.44</i>   |
| Cum. rate < 65 y (per 1000) —"                                | 14.6             | 9.5               | <i>1.54</i>   |

## Crude and adjusted rates compared (IS: Table 4.6)

|   | Cali,<br>1982-86 | B'ham,<br>1983-86 | Rate<br>ratio |
|---|------------------|-------------------|---------------|
| Crude rates (/10 <sup>5</sup> y)                                | 19.9             | 33.9              | 0.59          |
| ASR (/10 <sup>5</sup> y) <sup>B</sup> with 3 broad age groups   | 48.0             | 33.9              | 1.42          |
| ASR (/10 <sup>5</sup> y) <sup>C</sup> —"—                       | 19.9             | 14.4              | 1.38          |
| ASR (/10 <sup>5</sup> y) <sup>W</sup> —"—                       | 34.0             | 23.5              | 1.44          |
| Cum. rate < 65 y (per 1000) —"—                                 | 14.6             | 9.5               | 1.54          |
| ASR (/10 <sup>5</sup> y) <sup>W</sup> with 18 5-year age groups | 36.3             | 21.2              | 1.71          |
| Cum. rate < 75 y (per 1000) —"—                                 | 46.0             | 26.0              | 1.77          |

Standard population: <sup>B</sup> Birmingham 1985, <sup>C</sup> Cali 1985, <sup>W</sup> World SP

## Crude and adjusted rates compared (IS: Table 4.6)

|   | Cali,<br>1982-86 | B'ham,<br>1983-86 | Rate<br>ratio |      |
|---|------------------|-------------------|---------------|------|
| Crude rates (/10 <sup>5</sup> y)                                | 19.9             | 33.9              | 0.59          |      |
| ASR (/10 <sup>5</sup> y) <sup>B</sup> with 3 broad age groups   | 48.0             | 33.9              | 1.42          |      |
| ASR (/10 <sup>5</sup> y) <sup>C</sup>                           | — <sup>1</sup> — | 19.9              | 14.4          | 1.38 |
| ASR (/10 <sup>5</sup> y) <sup>W</sup>                           | — <sup>1</sup> — | 34.0              | 23.5          | 1.44 |
| Cum. rate < 65 y (per 1000)                                     | — <sup>1</sup> — | 14.6              | 9.5           | 1.54 |
| ASR (/10 <sup>5</sup> y) <sup>W</sup> with 18 5-year age groups | 36.3             | 21.2              | 1.71          |      |
| Cum. rate < 75 y (per 1000)                                     | — <sup>1</sup> — | 46.0              | 26.0          | 1.77 |

Standard population: <sup>B</sup> Birmingham 1985, <sup>C</sup> Cali 1985, <sup>W</sup> World SP

**NB:** The ratios of age-adjusted rates appear less dependent on the choice of standard weights than on the coarseness of age grouping.

Narrow age groups are preferred, we do have computers...

# Survival

**Bendix Carstensen & Esa Läärä**

Nordic Summerschool of  
Cancer Epidemiology

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

ELsurv

# SURVIVAL ANALYSIS

Questions of interest in the prognosis of cancer

Survival analysis: in principle like incidence analysis but

# SURVIVAL ANALYSIS

Questions of interest on the **prognosis** of cancer:

- ▶ what are the patients' chances to **survive** at least 1 year, or 5 years *etc.*, since diagnosis?



# SURVIVAL ANALYSIS

Questions of interest on the **prognosis** of cancer:

- ▶ what are the patients' chances to **survive** at least 1 year, or 5 years *etc.*, since diagnosis?

# SURVIVAL ANALYSIS

Questions of interest on the **prognosis** of cancer:

- ▶ what are the patients' chances to **survive** at least 1 year, or 5 years *etc.*, since diagnosis?

**Survival analysis:** In principle like incidence analysis but

# SURVIVAL ANALYSIS

Questions of interest on the **prognosis** of cancer:

- ▶ what are the patients' chances to **survive** at least 1 year, or 5 years *etc.*, since diagnosis?

**Survival analysis:** In principle like incidence analysis but

- ▶ population at risk = patients with cancer,

# SURVIVAL ANALYSIS

Questions of interest on the **prognosis** of cancer:

- ▶ what are the patients' chances to **survive** at least 1 year, or 5 years *etc.*, since diagnosis?

**Survival analysis:** In principle like incidence analysis but

- ▶ population at risk = patients with cancer,

# SURVIVAL ANALYSIS

Questions of interest on the **prognosis** of cancer:

- ▶ what are the patients' chances to **survive** at least 1 year, or 5 years *etc.*, since diagnosis?

**Survival analysis:** In principle like incidence analysis but

- ▶ population at risk = patients with cancer,
- ▶ basic time variable = time since the date of diagnosis, on which the follow-up starts,

# SURVIVAL ANALYSIS

Questions of interest on the **prognosis** of cancer:

- ▶ what are the patients' chances to **survive** at least 1 year, or 5 years *etc.*, since diagnosis?

**Survival analysis:** In principle like incidence analysis but

- ▶ population at risk = patients with cancer,
- ▶ basic time variable = time since the date of diagnosis, on which the follow-up starts,

# SURVIVAL ANALYSIS

Questions of interest on the **prognosis** of cancer:

- ▶ what are the patients' chances to **survive** at least 1 year, or 5 years *etc.*, since diagnosis?

**Survival analysis:** In principle like incidence analysis but

- ▶ population at risk = patients with cancer,
- ▶ basic time variable = time since the date of diagnosis, on which the follow-up starts,
- ▶ outcome event of interest = death,

# SURVIVAL ANALYSIS

Questions of interest on the **prognosis** of cancer:

- ▶ what are the patients' chances to **survive** at least 1 year, or 5 years *etc.*, since diagnosis?

**Survival analysis:** In principle like incidence analysis but

- ▶ population at risk = patients with cancer,
- ▶ basic time variable = time since the date of diagnosis, on which the follow-up starts,
- ▶ outcome event of interest = death,



# SURVIVAL ANALYSIS

Questions of interest on the **prognosis** of cancer:

- ▶ what are the patients' chances to **survive** at least 1 year, or 5 years *etc.*, since diagnosis?

**Survival analysis:** In principle like incidence analysis but

- ▶ population at risk = patients with cancer,
- ▶ basic time variable = time since the date of diagnosis, on which the follow-up starts,
- ▶ outcome event of interest = death,
- ▶ measures and methods used somewhat different from those used in incidence analysis.

## Follow-up of 8 out of 40 breast cancer patients

(from IS, table 12.1., p. 264)

---

| No. | Age<br>(y) | Sta-<br>ge <sup>a</sup> | Date of<br>diag-<br>nosis | Date at<br>end of<br>follow-up | Vital<br>status<br>at end of<br>follow-up | Cause<br>of<br>death <sup>c</sup> | Full<br>years<br>from<br>diagn's<br>up to<br>end of<br>follow-up | Days<br>from<br>diagn's<br>up to<br>end of<br>follow-up |
|-----|------------|-------------------------|---------------------------|--------------------------------|---|-----------------------------------|--|---|
|-----|------------|-------------------------|---------------------------|--------------------------------|---|-----------------------------------|--|---|

---

## Follow-up of 8 out of 40 breast cancer patients

(from IS, table 12.1., p. 264)

| No. | Age (y) | Sta-ge <sup>a</sup> | Date of diag-nosis | Date at end of follow-up | Vital status at end of follow-up | Cause of death <sup>c</sup> | Full years from diagn's up to end of follow-up | Days from diagn's up to end of follow-up |
|-----|---------|---------------------|--------------------|--------------------------|----------------------------------|-----------------------------|--|--|
| 1   | 39      | 1                   | 01/02/89           | 23/10/92                 | A                                | -                           | 3  | 1360                                     |
| 3   | 56      | 2                   | 16/04/89           | 05/09/89                 | D                                | BC                          | 0  | 142                                      |
| 5   | 62      | 2                   | 12/06/89           | 28/12/95                 | A                                | -                           | 6  | 2390                                     |
| 15  | 60      | 2                   | 03/08/90           | 27/11/94                 | A                                | -                           | 4  | 1577                                     |
| 22  | 64      | 2                   | 17/02/91           | 06/09/94                 | D                                | O                           | 3  | 1297                                     |
| 25  | 42      | 2                   | 20/06/91           | 15/03/92                 | D                                | BC                          | 0  | 269                                      |
| 30  | 77      | 1                   | 05/05/92           | 10/05/95                 | A                                | -                           | 3  | 1100                                     |
| 37  | 45      | 1                   | 11/05/93           | 07/02/94                 | D                                | BC                          | 0  | 272                                      |

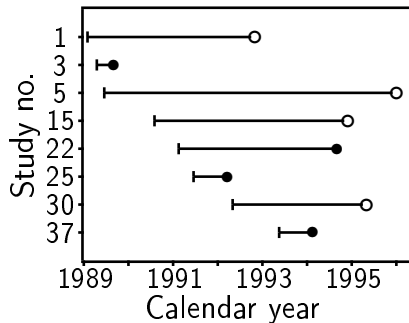
<sup>a</sup> 1 = absence of regional lymph node involment and metastases

2 = involvement of regional lymph node and/or presence of metastases

<sup>b</sup> A = alive; D = dead; <sup>c</sup> BC = breast cancer; O = other causes

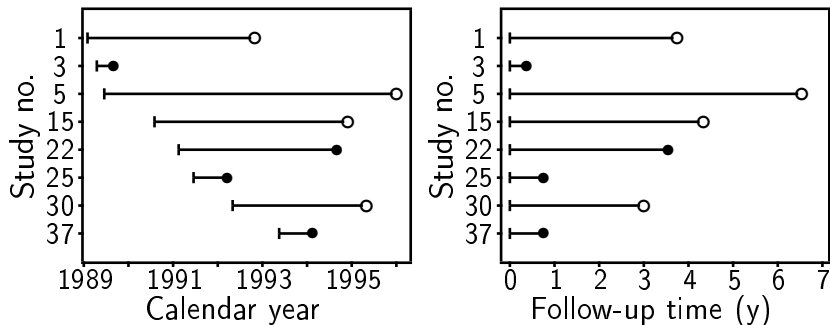
## Follow-up of breast ca. patients (cont'd)

| entry = diagnosis; ● exit = death; ○ exit = censoring



## Follow-up of breast ca. patients (cont'd)

| entry = diagnosis; ● exit = death; ○ exit = censoring



(IS: Figure 12.1, p. 265)

# Life table or actuarial method

Commonly used in population based survival analysis by cancer registries.  
(In clinical applications the Kaplan-Meier method is more popular.)

## Life table or actuarial method

Commonly used in population-based survival analysis by cancer registries.  
(In clinical applications the **Kaplan-Meier** method is more popular.)

- (1) Divide the follow-up time into subintervals  $k = 1, \dots, K$ ;  
most of these having width of 1 year.

Often the first year is divided into monthly intervals, or at two intervals with widths of 3 mo and 9 mo, respectively.

## Life table or actuarial method

Commonly used in population-based survival analysis by cancer registries.  
(In clinical applications the **Kaplan-Meier** method is more popular.)

- (1) Divide the follow-up time into subintervals  $k = 1, \dots, K$ ;  
most of these having width of 1 year.

Often the first year is divided into monthly intervals, or at two intervals with widths of 3 mo and 9 mo, respectively.



## Life table or actuarial method

Commonly used in population-based survival analysis by cancer registries.  
(In clinical applications the **Kaplan-Meier** method is more popular.)

- (1) Divide the follow-up time into subintervals  $k = 1, \dots, K$ ;  
most of these having width of 1 year.

Often the first year is divided into monthly intervals, or at two intervals with widths of 3 mo and 9 mo, respectively.

- (2) Tabulate from original data for each interval

## Life table or actuarial method

Commonly used in population-based survival analysis by cancer registries.  
(In clinical applications the **Kaplan-Meier** method is more popular.)

- (1) Divide the follow-up time into subintervals  $k = 1, \dots, K$ ;  
most of these having width of 1 year.

Often the first year is divided into monthly intervals, or at two intervals with widths of 3 mo and 9 mo, respectively.

- (2) Tabulate from original data for each interval

## Life table or actuarial method

Commonly used in population-based survival analysis by cancer registries.  
(In clinical applications the **Kaplan-Meier** method is more popular.)

- (1) Divide the follow-up time into subintervals  $k = 1, \dots, K$ ;  
most of these having width of 1 year.

Often the first year is divided into monthly intervals, or at two intervals with widths of 3 mo and 9 mo, respectively.

- (2) Tabulate from original data for each interval

$N_k =$  size of the **risk set**, *i.e.* the no. of subjects still alive and under follow-up at the start of interval,

## Life table or actuarial method

Commonly used in population-based survival analysis by cancer registries.  
(In clinical applications the **Kaplan-Meier** method is more popular.)

- (1) Divide the follow-up time into subintervals  $k = 1, \dots, K$ ;  
most of these having width of 1 year.

Often the first year is divided into monthly intervals, or at two intervals with widths of 3 mo and 9 mo, respectively.

- (2) Tabulate from original data for each interval

$N_k =$  size of the **risk set**, *i.e.* the no. of subjects still alive and under follow-up at the start of interval,

## Life table or actuarial method

Commonly used in population-based survival analysis by cancer registries.  
(In clinical applications the **Kaplan-Meier** method is more popular.)

- (1) Divide the follow-up time into subintervals  $k = 1, \dots, K$ ;  
most of these having width of 1 year.

Often the first year is divided into monthly intervals, or at two intervals with widths of 3 mo and 9 mo, respectively.

- (2) Tabulate from original data for each interval

$N_k$  = size of the **risk set**, *i.e.* the no. of subjects still alive and under follow-up at the start of interval,

$D_k$  = no. of **cases**, *i.e.* deaths observed in the interval,

## Life table or actuarial method

Commonly used in population-based survival analysis by cancer registries.  
(In clinical applications the **Kaplan-Meier** method is more popular.)

- (1) Divide the follow-up time into subintervals  $k = 1, \dots, K$ ;  
most of these having width of 1 year.

Often the first year is divided into monthly intervals, or at two intervals with widths of 3 mo and 9 mo, respectively.

- (2) Tabulate from original data for each interval

$N_k$  = size of the **risk set**, *i.e.* the no. of subjects still alive and under follow-up at the start of interval,

$D_k$  = no. of **cases**, *i.e.* deaths observed in the interval,

## Life table or actuarial method

Commonly used in population-based survival analysis by cancer registries.  
(In clinical applications the **Kaplan-Meier** method is more popular.)

- (1) Divide the follow-up time into subintervals  $k = 1, \dots, K$ ;  
most of these having width of 1 year.

Often the first year is divided into monthly intervals, or at two intervals with widths of 3 mo and 9 mo, respectively.

- (2) Tabulate from original data for each interval

$N_k$  = size of the **risk set**, *i.e.* the no. of subjects still alive and under follow-up at the start of interval,

$D_k$  = no. of **cases**, *i.e.* deaths observed in the interval,

$L_k$  = no. of **losses**, *i.e.* individuals **censored** during the interval before being observed to die.

# Life table items in a tree diagram

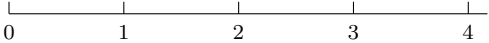


Following time (years) divided into 5-year intervals

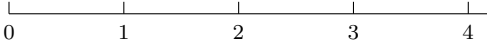
ELsurv is the probability of survival at the start of the 5th year interval



# Life table items in a tree diagram

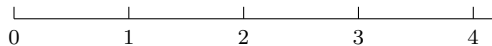


# Life table items in a tree diagram



Follow-up time (years), divided into 1-y subintervals

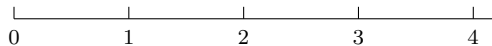
# Life table items in a tree diagram



Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

# Life table items in a tree diagram



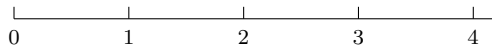
Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

# Life table items in a tree diagram

at risk  $\square$   
 $N_1$

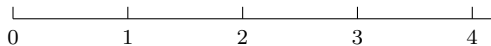
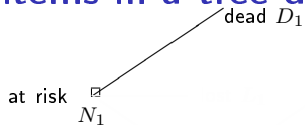


Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

# Life table items in a tree diagram

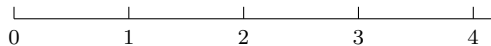
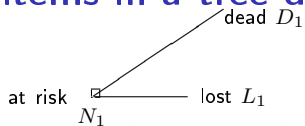


Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

# Life table items in a tree diagram

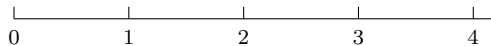
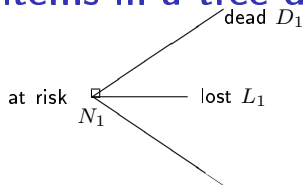


Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

# Life table items in a tree diagram



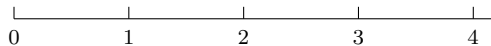
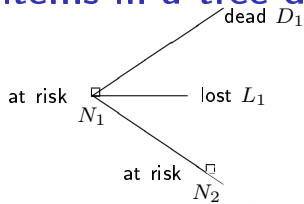
Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$



# Life table items in a tree diagram

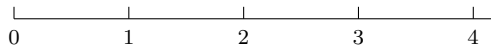
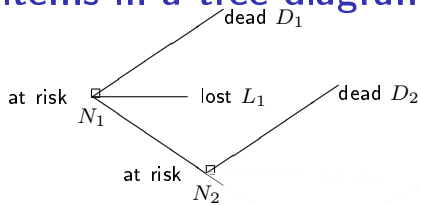


Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

# Life table items in a tree diagram

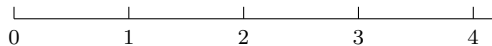
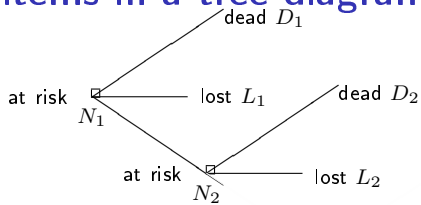


Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

# Life table items in a tree diagram

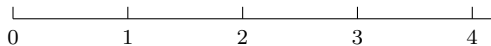
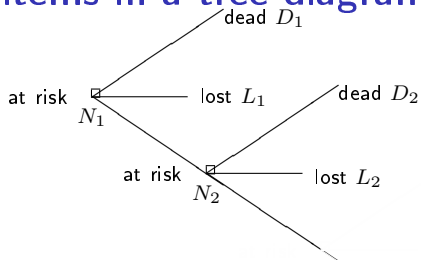


Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

# Life table items in a tree diagram

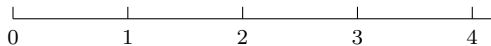
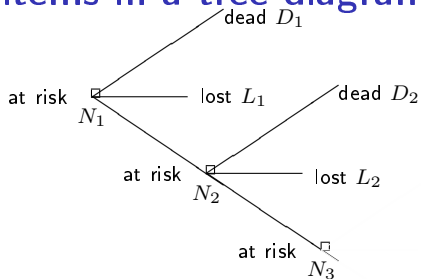


Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

# Life table items in a tree diagram

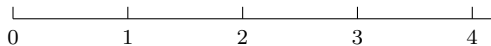
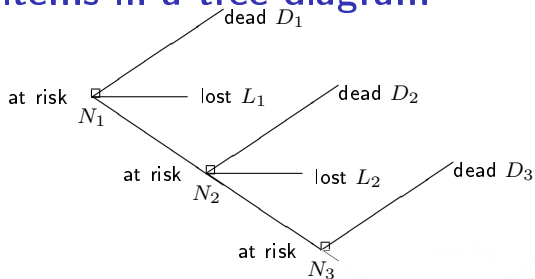


Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

## Life table items in a tree diagram

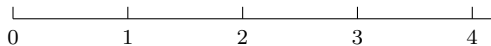
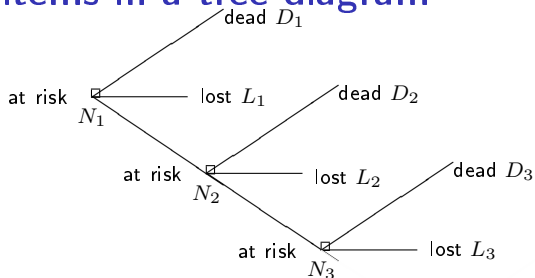


Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

## Life table items in a tree diagram

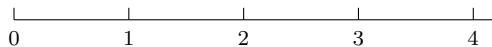
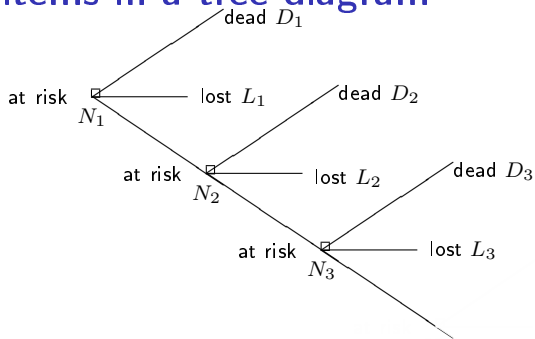


Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

# Life table items in a tree diagram



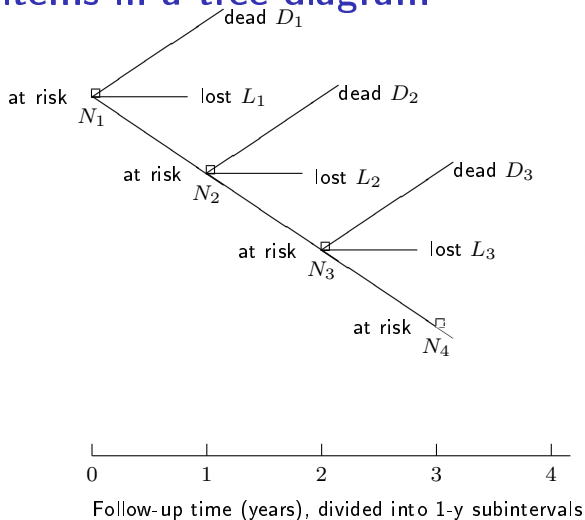
Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$



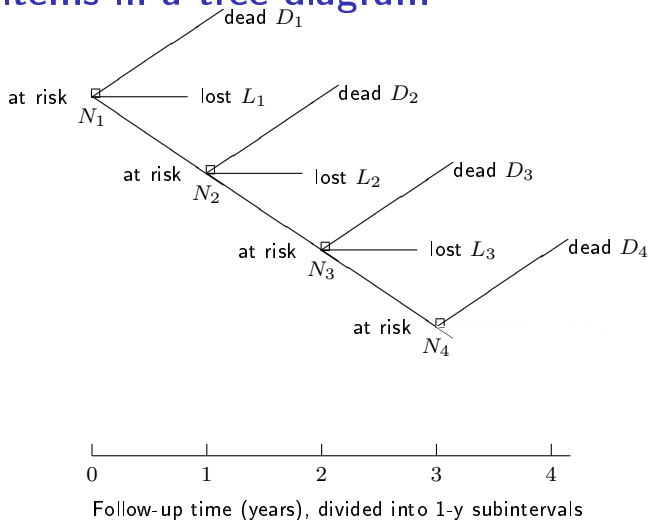
## Life table items in a tree diagram



$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

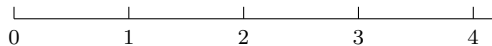
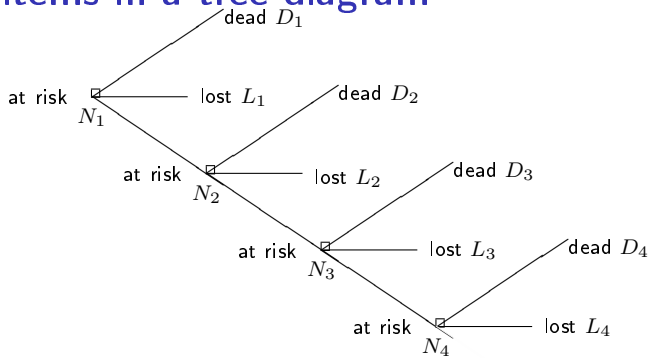
## Life table items in a tree diagram



$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

## Life table items in a tree diagram

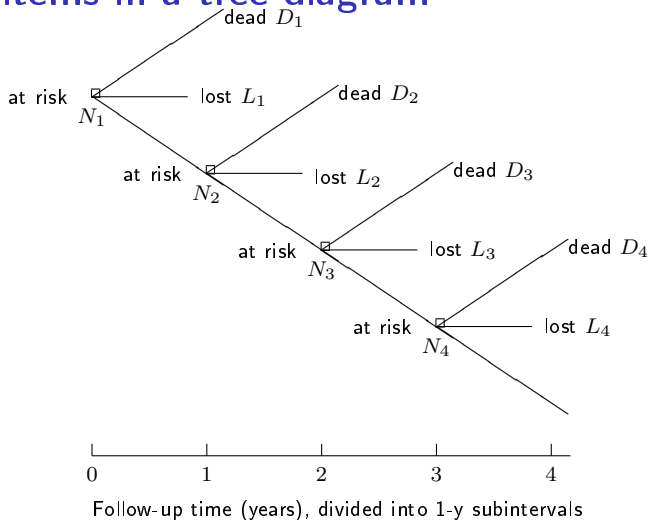


Follow-up time (years), divided into 1-y subintervals

$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

# Life table items in a tree diagram



$N_k$  = population at risk at the start of the  $k$ th subinterval

$D_k$  = no. of deaths,  $L_k$  = no. of losses or censorings in interval  $k$

## Life table items for breast ca. patients

(IS: Table 12.2., p. 273, first 4 columns)

| Inter-<br>val | Years<br>since<br>diagnosis | No. at<br>start of<br>interval | No. of<br>deaths | No. of<br>losses |
|---------------|-----------------------------|--------------------------------|------------------|------------------|
| $(k)$         |                             | $(N_k)$                        | $(D_k)$          | $(L_k)$          |
| 1             | 0- < 1                      | 40                             | 7                | 0                |
| 2             | 1- < 2                      | 33                             | 3                | 6                |
| 3             | 2- < 3                      | 24                             | 4                | 3                |
| 4             | 3- < 4                      | 17                             | 4                | 4                |
| 5             | 4- < 5                      | 9                              | 2                | 3                |
| 6             | 5- < 6                      | 4                              | 1                | 2                |
| 7             | 6- < 7                      | 1                              | 0                | 1                |
| Total         |                             |                                | 21               | 19               |

## Life table calculations (cont'd)

(3) Calculate and tabulate for each interval

## Life table calculations (cont'd)

(3) Calculate and tabulate for each interval

## Life table calculations (cont'd)

(3) Calculate and tabulate for each interval

$N'_k = N_k - L_k/2 =$  corrected size of the risk set, or  
“effective denominator” at start of the interval,



## Life table calculations (cont'd)

(3) Calculate and tabulate for each interval

$N'_k = N_k - L_k/2 =$  corrected size of the risk set, or  
“effective denominator” at start of the interval,

## Life table calculations (cont'd)

(3) Calculate and tabulate for each interval

$N'_k = N_k - L_k/2 =$  corrected size of the risk set, or  
“effective denominator” at start of the interval,

$q_k = D_k/N'_k =$  estimated conditional probability of dying  
during the interval given survival up to its start,

## Life table calculations (cont'd)

(3) Calculate and tabulate for each interval

$N'_k = N_k - L_k/2 =$  corrected size of the risk set, or  
“effective denominator” at start of the interval,

$q_k = D_k/N'_k =$  estimated conditional probability of dying  
during the interval given survival up to its start,

## Life table calculations (cont'd)

(3) Calculate and tabulate for each interval

$N'_k = N_k - L_k/2 =$  corrected size of the risk set, or  
“effective denominator” at start of the interval,

$q_k = D_k/N'_k =$  estimated conditional probability of dying  
during the interval given survival up to its start,

$p_k = 1 - q_k =$  conditional survival proportion over the int'l,

## Life table calculations (cont'd)

(3) Calculate and tabulate for each interval

$N'_k = N_k - L_k/2 =$  corrected size of the risk set, or  
“effective denominator” at start of the interval,

$q_k = D_k/N'_k =$  estimated conditional probability of dying  
during the interval given survival up to its start,

$p_k = 1 - q_k =$  conditional survival proportion over the int'l,

## Life table calculations (cont'd)

(3) Calculate and tabulate for each interval

$N'_k = N_k - L_k/2 =$  corrected size of the risk set, or  
“effective denominator” at start of the interval,

$q_k = D_k/N'_k =$  estimated conditional probability of dying  
during the interval given survival up to its start,

$p_k = 1 - q_k =$  conditional survival proportion over the int'l,

$S_k = p_1 \times \cdots \times p_k =$  **cumulative survival proportion** from  
date of diagnosis until the end of the  $k$ th interval

## Life table calculations (cont'd)

### (3) Calculate and tabulate for each interval

$N'_k = N_k - L_k/2 =$  corrected size of the risk set, or  
“effective denominator” at start of the interval,

$q_k = D_k/N'_k =$  estimated conditional probability of dying  
during the interval given survival up to its start,

$p_k = 1 - q_k =$  conditional survival proportion over the int'l,

$S_k = p_1 \times \cdots \times p_k =$  **cumulative survival proportion** from  
date of diagnosis until the end of the  $k$ th interval

## Life table calculations (cont'd)

(3) Calculate and tabulate for each interval

$N'_k = N_k - L_k/2 =$  corrected size of the risk set, or  
“effective denominator” at start of the interval,

$q_k = D_k/N'_k =$  estimated conditional probability of dying  
during the interval given survival up to its start,

$p_k = 1 - q_k =$  conditional survival proportion over the int'l,

$S_k = p_1 \times \cdots \times p_k =$  **cumulative survival proportion** from  
date of diagnosis until the end of the  $k$ th interval

$=$  estimate of **survival probability** up to this time point.



## Follow-up of breast ca. patients (cont'd)

Actuarial life table completed (IS, table 12.2, p. 273)

| Interval | Years since diagnosis | No. at start of interval | No. of deaths | No. of losses | Effective denominator | Cond'l prop'n of deaths during int'l | Survival prop'n over int'l | Cumul. survival; est'd survival prob'ty |
|----------|-----------------------|--------------------------|---------------|---------------|-----------------------|--------------------------------------|----------------------------|---|
| ( $k$ )  |                       | ( $N_k$ )                | ( $D_k$ )     | ( $L_k$ )     | ( $N'_k$ )            | ( $q_k$ )                            | ( $p_k$ )                  | ( $S_k$ )                               |
| 1        | 0- < 1                | 40                       | 7             | 0             | 40.0                  | 0.175                                | 0.825                      | 0.825                                   |
| 2        | 1- < 2                | 33                       | 3             | 6             | 30.0                  | 0.100                                | 0.900                      | 0.743                                   |
| 3        | 2- < 3                | 24                       | 4             | 3             | 22.5                  | 0.178                                | 0.822                      | 0.610                                   |
| 4        | 3- < 4                | 17                       | 4             | 4             | 15.0                  | 0.267                                | 0.733                      | 0.447                                   |
| 5        | 4- < 5                | 9                        | 2             | 3             | 7.5                   | 0.267                                | 0.733                      | 0.328                                   |
| 6        | 5- < 6                | 4                        | 1             | 2             | 3.0                   | 0.333                                | 0.667                      | 0.219                                   |
| 7        | 6- < 7                | 1                        | 0             | 1             | 0.5                   | 0.0                                  | 1.0                        | 0.219                                   |

## Follow-up of breast ca. patients (cont'd)

Actuarial life table completed (IS, table 12.2, p. 273)

| Interval | Years since diagnosis | No. at start of interval | No. of deaths | No. of losses | Effective denominator | Cond'l prop'n of deaths during int'l | Survival prop'n over int'l | Cumul. survival; est'd survival prob'ty |
|----------|-----------------------|--------------------------|---------------|---------------|-----------------------|--------------------------------------|----------------------------|---|
| ( $k$ )  |                       | ( $N_k$ )                | ( $D_k$ )     | ( $L_k$ )     | ( $N'_k$ )            | ( $q_k$ )                            | ( $p_k$ )                  | ( $S_k$ )                               |
| 1        | 0- < 1                | 40                       | 7             | 0             | 40.0                  | 0.175                                | 0.825                      | 0.825                                   |
| 2        | 1- < 2                | 33                       | 3             | 6             | 30.0                  | 0.100                                | 0.900                      | 0.743                                   |
| 3        | 2- < 3                | 24                       | 4             | 3             | 22.5                  | 0.178                                | 0.822                      | 0.610                                   |
| 4        | 3- < 4                | 17                       | 4             | 4             | 15.0                  | 0.267                                | 0.733                      | 0.447                                   |
| 5        | 4- < 5                | 9                        | 2             | 3             | 7.5                   | 0.267                                | 0.733                      | 0.328                                   |
| 6        | 5- < 6                | 4                        | 1             | 2             | 3.0                   | 0.333                                | 0.667                      | 0.219                                   |
| 7        | 6- < 7                | 1                        | 0             | 1             | 0.5                   | 0.0                                  | 1.0                        | 0.219                                   |

1-year survival probability is thus estimated 82.5% and 5-year probability 32.8%.

# Comparison to previous methods

## Comparison to previous methods

- ▶ Complement of survival proportion  $Q_k = 1 - S_k$   
= incidence proportion of deaths.

## Comparison to previous methods

- ▶ Complement of survival proportion  $Q_k = 1 - S_k$   
= incidence proportion of deaths.

## Comparison to previous methods

- ▶ Complement of survival proportion  $Q_k = 1 - S_k$   
= incidence proportion of deaths.

Estimates the cumulative risk of death from the start of follow-up till the end of  $k$ th interval.

## Comparison to previous methods

- ▶ Complement of survival proportion  $Q_k = 1 - S_k$   
= incidence proportion of deaths.

Estimates the cumulative risk of death from the start of follow-up till the end of  $k$ th interval.

- ▶ Incidence rate in the  $k$ th interval is computed as:

$$I_k = \frac{\text{number of cases } (D_k)}{\text{approximate person-time } (\tilde{Y}_k)}$$

## Comparison to previous methods

- ▶ Complement of survival proportion  $Q_k = 1 - S_k$   
= incidence proportion of deaths.

Estimates the cumulative risk of death from the start of follow-up till the end of  $k$ th interval.

- ▶ Incidence rate in the  $k$ th interval is computed as:

$$I_k = \frac{\text{number of cases } (D_k)}{\text{approximate person-time } (\tilde{Y}_k)}$$



## Comparison to previous methods

- ▶ Complement of survival proportion  $Q_k = 1 - S_k$   
= incidence proportion of deaths.

Estimates the cumulative risk of death from the start of follow-up till the end of  $k$ th interval.

- ▶ Incidence rate in the  $k$ th interval is computed as:

$$I_k = \frac{\text{number of cases } (D_k)}{\text{approximate person-time } (\tilde{Y}_k)}$$

where the approximate person-time is given by

$$\tilde{Y}_k = \left[ N_k - \frac{1}{2}(D_k + L_k) \right] \times \text{width of interval}$$

The dead and censored thus contribute half of the interval width.

## Survival curve and other measures

The diagram of survival proportions through interval endpoints provides graphical estimates of interesting parameters of the survival time distribution.

- **RR**: When the curve ends at higher level than 0%, in which case some measures cannot be calculated.

## Survival curve and other measures

Line diagram of survival proportions through interval endpoints

## Survival curve and other measures

Line diagram of survival proportions through interval endpoints provides graphical estimates of interesting parameters of the survival time distribution, e.g.:

## Survival curve and other measures

Line diagram of survival proportions through interval endpoints provides graphical estimates of interesting parameters of the survival time distribution, e.g.:

- ▶ **median** and **quartiles**: time points at which the curve crosses the 50%, 75%, and 25% levels

## Survival curve and other measures

Line diagram of survival proportions through interval endpoints provides graphical estimates of interesting parameters of the survival time distribution, e.g.:

- ▶ **median** and **quartiles**: time points at which the curve crosses the 50%, 75%, and 25% levels

## Survival curve and other measures

Line diagram of survival proportions through interval endpoints provides graphical estimates of interesting parameters of the survival time distribution, e.g.:

- ▶ **median and quartiles:** time points at which the curve crosses the 50%, 75%, and 25% levels
- ▶ **mean residual lifetime:** area under the curve, given that it decreases all the way down to the 0% level.

## Survival curve and other measures

Line diagram of survival proportions through interval endpoints provides graphical estimates of interesting parameters of the survival time distribution, e.g.:

- ▶ **median and quartiles:** time points at which the curve crosses the 50%, 75%, and 25% levels
- ▶ **mean residual lifetime:** area under the curve, given that it decreases all the way down to the 0% level.



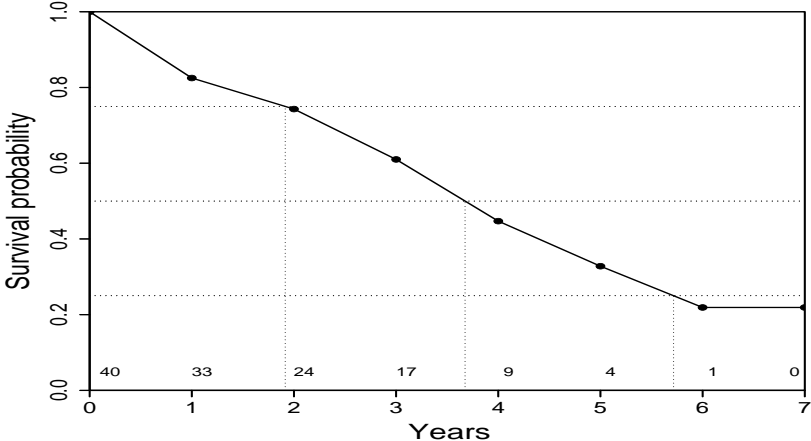
## Survival curve and other measures

Line diagram of survival proportions through interval endpoints provides graphical estimates of interesting parameters of the survival time distribution, e.g.:

- ▶ **median** and **quartiles**: time points at which the curve crosses the 50%, 75%, and 25% levels
- ▶ **mean residual lifetime**: area under the curve, given that it decreases all the way down to the 0% level.

**NB.** Often the curve ends at higher level than 0%, in which case some measures cannot be calculated.

# Survival curve of breast ca. patients (IS: Fig 12.8)



Numbers above  $x$ -axis show the size of population at risk.

# Relative survival analysis

# Relative survival analysis

## Relative survival analysis

- ▶ Another interesting and relevant question:

## Relative survival analysis

- ▶ Another interesting and relevant question:

*“How much worse are the chances of a cancer patient to survive, say, 5 years, as compared with a comparable person without the disease?”*

## Relative survival analysis

- ▶ Another interesting and relevant question:

*“How much worse are the chances of a cancer patient to survive, say, 5 years, as compared with a comparable person without the disease?”*

## Relative survival analysis

- ▶ Another interesting and relevant question:

*“How much worse are the chances of a cancer patient to survive, say, 5 years, as compared with a comparable person without the disease?”*

- ▶ An answer is provided by **relative survival proportions**:

$$R_k = S_k^{\text{obs}} / S_k^{\text{exp}}, \quad \text{where}$$



## Relative survival analysis

- ▶ Another interesting and relevant question:

*“How much worse are the chances of a cancer patient to survive, say, 5 years, as compared with a comparable person without the disease?”*

- ▶ An answer is provided by **relative survival proportions**:

$$R_k = S_k^{\text{obs}} / S_k^{\text{exp}}, \quad \text{where}$$

- $S_k^{\text{obs}}$  = **observed** survival proportion in cancer patient group  $k$  by age, gender and year of diagnosis,

## Relative survival analysis

- ▶ Another interesting and relevant question:

*“How much worse are the chances of a cancer patient to survive, say, 5 years, as compared with a comparable person without the disease?”*

- ▶ An answer is provided by **relative survival proportions**:

$$R_k = S_k^{\text{obs}} / S_k^{\text{exp}}, \quad \text{where}$$

- $S_k^{\text{obs}}$  = **observed** survival proportion in cancer patient group  $k$  by age, gender and year of diagnosis,

## Relative survival analysis

- ▶ Another interesting and relevant question:

*“How much worse are the chances of a cancer patient to survive, say, 5 years, as compared with a comparable person without the disease?”*

- ▶ An answer is provided by **relative survival proportions**:

$$R_k = S_k^{\text{obs}} / S_k^{\text{exp}}, \quad \text{where}$$

- $S_k^{\text{obs}}$  = **observed** survival proportion in cancer patient group  $k$  by age, gender and year of diagnosis,
- $S_k^{\text{exp}}$  = **expected** survival proportion

## Relative survival analysis

- ▶ Another interesting and relevant question:

*“How much worse are the chances of a cancer patient to survive, say, 5 years, as compared with a comparable person without the disease?”*

- ▶ An answer is provided by **relative survival proportions**:

$$R_k = S_k^{\text{obs}} / S_k^{\text{exp}}, \quad \text{where}$$

- $S_k^{\text{obs}}$  = **observed** survival proportion in cancer patient group  $k$  by age, gender and year of diagnosis,
- $S_k^{\text{exp}}$  = **expected** survival proportion

## Relative survival analysis

- ▶ Another interesting and relevant question:

*“How much worse are the chances of a cancer patient to survive, say, 5 years, as compared with a comparable person without the disease?”*

- ▶ An answer is provided by **relative survival proportions**:

$$R_k = S_k^{\text{obs}} / S_k^{\text{exp}}, \quad \text{where}$$

- $S_k^{\text{obs}}$  = **observed** survival proportion in cancer patient group  $k$  by age, gender and year of diagnosis,
- $S_k^{\text{exp}}$  = **expected** survival proportion based on the age-specific mortality rates of the same gender and calendar time in a reference population (compare with calculations of SIR!)

## Relative survival analysis

- ▶ Another interesting and relevant question:

*“How much worse are the chances of a cancer patient to survive, say, 5 years, as compared with a comparable person without the disease?”*

- ▶ An answer is provided by **relative survival proportions**:

$$R_k = S_k^{\text{obs}} / S_k^{\text{exp}}, \quad \text{where}$$

- $S_k^{\text{obs}}$  = **observed** survival proportion in cancer patient group  $k$  by age, gender and year of diagnosis,
  - $S_k^{\text{exp}}$  = **expected** survival proportion based on the age-specific mortality rates of the same gender and calendar time in a reference population (compare with calculations of SIR!)
- + No information on causes of death needed.

# Conclusion

**Bendix Carstensen & Esa Läärä**

Nordic Summerschool of  
Cancer Epidemiology

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

ELconcl

# CONCLUSION

Measuring and comparing disease frequencies

Challenges



# CONCLUSION

Measuring and comparing disease frequencies

# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but

# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but

# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but
- ▶ demands expert skills in epidemiologic methods.

# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but
- ▶ demands expert skills in epidemiologic methods.

# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but
- ▶ demands expert skills in epidemiologic methods.

Major challenges:

# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but
- ▶ demands expert skills in epidemiologic methods.

Major challenges:

- ▶ obtain the right denominator for each numerator,

# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but
- ▶ demands expert skills in epidemiologic methods.

Major challenges:

- ▶ obtain the right denominator for each numerator,



# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but
- ▶ demands expert skills in epidemiologic methods.

Major challenges:

- ▶ obtain the right denominator for each numerator,
- ▶ valid calculation of person-years,

# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but
- ▶ demands expert skills in epidemiologic methods.

Major challenges:

- ▶ obtain the right denominator for each numerator,
- ▶ valid calculation of person-years,

# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but
- ▶ demands expert skills in epidemiologic methods.

Major challenges:

- ▶ obtain the right denominator for each numerator,
- ▶ valid calculation of person-years,
- ▶ appropriate treatment of time and its various aspects,

# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but
- ▶ demands expert skills in epidemiologic methods.

Major challenges:

- ▶ obtain the right denominator for each numerator,
- ▶ valid calculation of person-years,
- ▶ appropriate treatment of time and its various aspects,

# CONCLUSION

Measuring and comparing disease frequencies

- ▶ not a trivial task but
- ▶ demands expert skills in epidemiologic methods.

Major challenges:

- ▶ obtain the right denominator for each numerator,
- ▶ valid calculation of person-years,
- ▶ appropriate treatment of time and its various aspects,
- ▶ removal of confounding from comparisons.

# R and how we use it

**Bendix Carstensen & Esa Läärä**

Nordic Summerschool of  
Cancer Epidemiology

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

R-start

# Introduction to R

What is **R**?

- ▶ A practical calculator:

How many times did you compute

$\frac{1}{2} + \frac{1}{3}$  by hand? (or a calculator?)

▶ A powerful language:

an environment for data analysis and graphics

▶ A free and open source software

Developed by international community of volunteers

▶ A growing user base

▶ A growing ecosystem

▶ A growing user community

# Introduction to R

What is **R**?

- ▶ A practical calculator:
  - You can see what you compute



# Introduction to R

What is **R**?

- ▶ A practical calculator:
  - You can see what you compute
  - ...and change easily to do similar calculations.

# Introduction to R

What is **R**?

- ▶ A practical calculator:
  - You can see what you compute
  - ...and change easily to do similar calculations.
- ▶ A statistical program.

# Introduction to R

What is **R**?

- ▶ A practical calculator:
  - You can see what you compute
  - ...and change easily to do similar calculations.
- ▶ A statistical program.
- ▶ An environment for data analysis and graphics.

# Introduction to R

What is **R**?

- ▶ A practical calculator:
  - You can see what you compute
  - ...and change easily to do similar calculations.
- ▶ A statistical program.
- ▶ An environment for data analysis and graphics.
- ▶ A programming language

# Introduction to R

What is **R**?

- ▶ A practical calculator:
  - You can see what you compute
  - ...and change easily to do similar calculations.
- ▶ A statistical program.
- ▶ An environment for data analysis and graphics.
- ▶ A programming language
- ▶ Developed by international community of volunteers.

# Introduction to R

What is **R**?

- ▶ A practical calculator:
  - You can see what you compute
  - ...and change easily to do similar calculations.
- ▶ A statistical program.
- ▶ An environment for data analysis and graphics.
- ▶ A programming language
- ▶ Developed by international community of volunteers.
- ▶ Free.

# Introduction to R

## What is R?

- ▶ A practical calculator:
  - You can see what you compute
  - ...and change easily to do similar calculations.
- ▶ A statistical program.
- ▶ An environment for data analysis and graphics.
- ▶ A programming language
- ▶ Developed by international community of volunteers.
- ▶ Free.
- ▶ Runs on any computer.

# Introduction to R

## What is R?

- ▶ A practical calculator:
  - You can see what you compute
  - ...and change easily to do similar calculations.
- ▶ A statistical program.
- ▶ An environment for data analysis and graphics.
- ▶ A programming language
- ▶ Developed by international community of volunteers.
- ▶ Free.
- ▶ Runs on any computer.
- ▶ Updated every 6 months.



# What does R offer for epidemiologists?

- ▶ Descriptive tools

▶ Descriptive statistics

▶ Descriptive graphics

▶ Descriptive tables

▶ Descriptive tools provided by a wide range of R and RStudio packages

▶ Many features of R are more appealing in the long run

# What does R offer for epidemiologists?

- ▶ Descriptive tools
  - ▶ Versatile tabulation

# What does R offer for epidemiologists?

- ▶ Descriptive tools
  - ▶ Versatile tabulation
  - ▶ High-quality graphics

# What does R offer for epidemiologists?

- ▶ Descriptive tools
  - ▶ Versatile tabulation
  - ▶ High-quality graphics
- ▶ Analytic methods

# What does R offer for epidemiologists?

- ▶ Descriptive tools
  - ▶ Versatile tabulation
  - ▶ High-quality graphics
- ▶ Analytic methods
  - ▶ Basic epidemiologic statistics

# What does R offer for epidemiologists?

- ▶ Descriptive tools
  - ▶ Versatile tabulation
  - ▶ High-quality graphics
- ▶ Analytic methods
  - ▶ Basic epidemiologic statistics
  - ▶ Survival analysis methods

# What does R offer for epidemiologists?

- ▶ Descriptive tools
  - ▶ Versatile tabulation
  - ▶ High-quality graphics
- ▶ Analytic methods
  - ▶ Basic epidemiologic statistics
  - ▶ Survival analysis methods
  - ▶ Common regression models and their extensions

# What does R offer for epidemiologists?

- ▶ Descriptive tools
  - ▶ Versatile tabulation
  - ▶ High-quality graphics
- ▶ Analytic methods
  - ▶ Basic epidemiologic statistics
  - ▶ Survival analysis methods
  - ▶ Common regression models and their extensions
  - ▶ Other...



# What does R offer for epidemiologists?

- ▶ Descriptive tools
  - ▶ Versatile tabulation
  - ▶ High-quality graphics
- ▶ Analytic methods
  - ▶ Basic epidemiologic statistics
  - ▶ Survival analysis methods
  - ▶ Common regression models and their extensions
  - ▶ Other...

# What does R offer for epidemiologists?

- ▶ Descriptive tools
  - ▶ Versatile tabulation
  - ▶ High-quality graphics
- ▶ Analytic methods
  - ▶ Basic epidemiologic statistics
  - ▶ Survival analysis methods
  - ▶ Common regression models and their extensions
  - ▶ Other...

But these are provided by *e.g.* SPSS, SAS and Stata too, so ...?

# What does R offer for epidemiologists?

- ▶ Descriptive tools
  - ▶ Versatile tabulation
  - ▶ High-quality graphics
- ▶ Analytic methods
  - ▶ Basic epidemiologic statistics
  - ▶ Survival analysis methods
  - ▶ Common regression models and their extensions
  - ▶ Other...

But these are provided by *e.g.* SPSS, SAS and Stata too, so ...?

Many features of **R** are more appealing in the long run.

# Graphics in R

- ▶ Versatile, flexible, high quality, . . .

- ▶ Many high-level graphic functions available

- ▶ Easy to add text (points, lines, text legends, etc.)

- ▶ Lots of plotting options

- ▶ Wide range of symbols, lines, axes, colours, etc. by

- ▶ graphics parameter (> 50 of them)

# Graphics in R

- ▶ Versatile, flexible, high quality, . . .
- ▶ Various **high-level** graphic functions available.

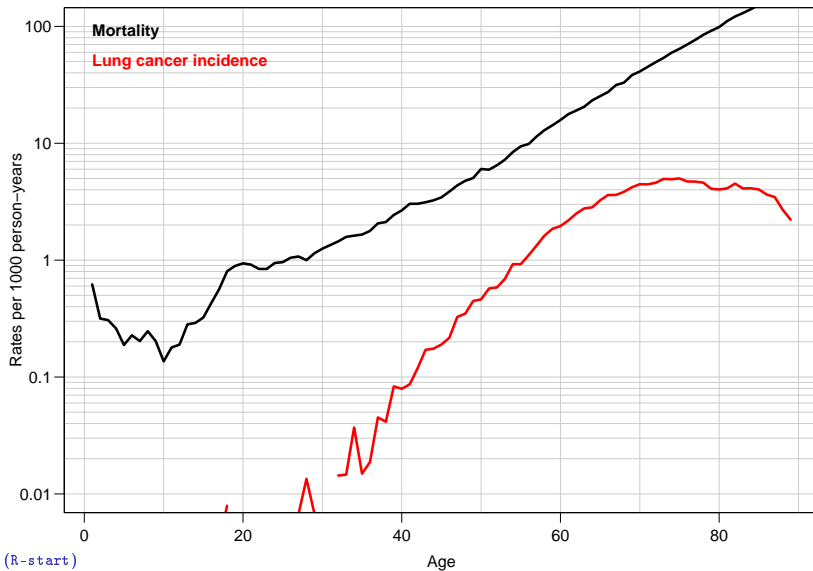
# Graphics in R

- ▶ Versatile, flexible, high quality, . . .
- ▶ Various **high-level** graphic functions available.
- ▶ Easy to add items (points, lines, text, legends . . . )  
to an existing graph.

## Graphics in R

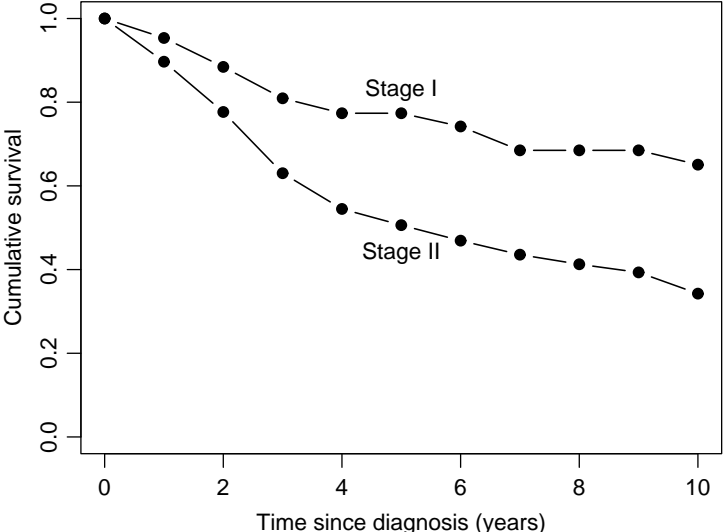
- ▶ Versatile, flexible, high quality, . . .
- ▶ Various **high-level** graphic functions available.
- ▶ Easy to add items (points, lines, text, legends . . . ) to an existing graph.
- ▶ Fine tuning of symbols, lines, axes, colours, etc. by *graphical parameters* (> 67 of them!)

# Total mortality and lung ca incidence in DK

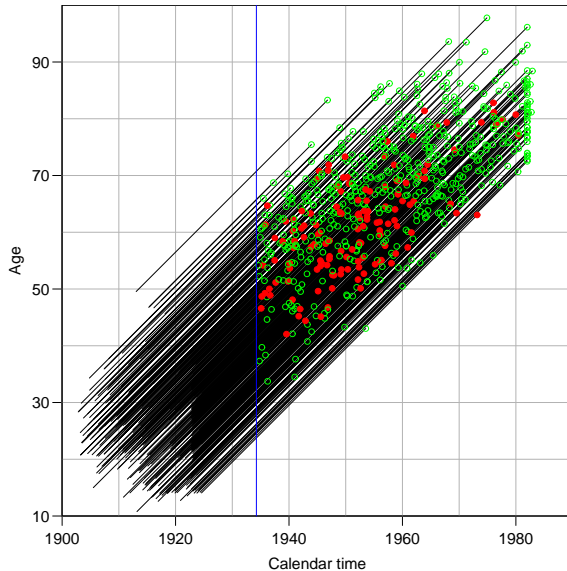




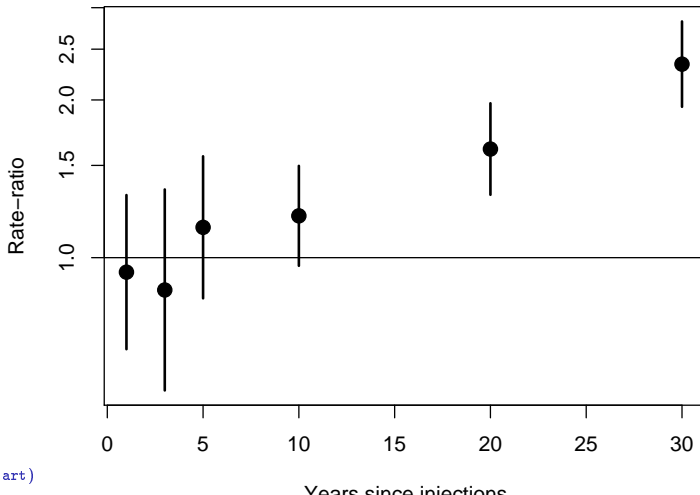
# Survival of cervix ca patients (C&H, 34)



# Lexis diagram of Welsh nickel cohort



# Rate ratios with confidence intervals



## Getting your graphs out

- ▶ Graphs can be saved to disk in almost any format

## Getting your graphs out

- ▶ Graphs can be saved to disk in almost any format
- ▶ .eps, .pdf, .bmp, .jpg, .png, ...

## Getting your graphs out

- ▶ Graphs can be saved to disk in almost any format
- ▶ .eps, .pdf, .bmp, .jpg, .png, ...
- ▶ Save graphs from the screen or write directly to a file.

## Getting your graphs out

- ▶ Graphs can be saved to disk in almost any format
- ▶ .eps, .pdf, .bmp, .jpg, .png, ...
- ▶ Save graphs from the screen or write directly to a file.
- ▶ You can also directly transport an R graph as a metafile into a Word document

## Tools for nearly anything!

- ▶ Thousands of add-on packages.

▶ Specialized tools for computational analysis

▶ Specialized tools for visualization

▶ Specialized tools for data management

▶ Specialized tools for workflow automation

▶ Specialized tools for reproducible research

▶ Specialized tools for teaching

▶ Specialized tools for documentation

▶ Specialized tools for collaboration

▶ Specialized tools for installation and updates from within R



## Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:

## Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:
  - ▶ Epi: focus on chronic disease epidemiology:

## Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:
  - ▶ Epi: focus on chronic disease epidemiology:
    - ▶ Cohort studies, splitting follow-up time

## Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:
  - ▶ Epi: focus on chronic disease epidemiology:
    - ▶ Cohort studies, splitting follow-up time
    - ▶ Lexis diagram, several timescales

## Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:
  - ▶ Epi: focus on chronic disease epidemiology:
    - ▶ Cohort studies, splitting follow-up time
    - ▶ Lexis diagram, several timescales
    - ▶ Multistate model support

# Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:
  - ▶ Epi: focus on chronic disease epidemiology:
    - ▶ Cohort studies, splitting follow-up time
    - ▶ Lexis diagram, several timescales
    - ▶ Multistate model support
    - ▶ Advanced tabulation

## Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:
  - ▶ Epi: focus on chronic disease epidemiology:
    - ▶ Cohort studies, splitting follow-up time
    - ▶ Lexis diagram, several timescales
    - ▶ Multistate model support
    - ▶ Advanced tabulation
    - ▶ Informative reporting of estimation results

## Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:
  - ▶ Epi: focus on chronic disease epidemiology:
    - ▶ Cohort studies, splitting follow-up time
    - ▶ Lexis diagram, several timescales
    - ▶ Multistate model support
    - ▶ Advanced tabulation
    - ▶ Informative reporting of estimation results
    - ▶ “Epidemiology with **R**”



## Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:
  - ▶ Epi: focus on chronic disease epidemiology:
    - ▶ Cohort studies, splitting follow-up time
    - ▶ Lexis diagram, several timescales
    - ▶ Multistate model support
    - ▶ Advanced tabulation
    - ▶ Informative reporting of estimation results
    - ▶ “Epidemiology with **R**”
  - ▶ `epicalc`:

## Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:
  - ▶ `Epi`: focus on chronic disease epidemiology:
    - ▶ Cohort studies, splitting follow-up time
    - ▶ Lexis diagram, several timescales
    - ▶ Multistate model support
    - ▶ Advanced tabulation
    - ▶ Informative reporting of estimation results
    - ▶ “Epidemiology with **R**”
  - ▶ `epicalc`:
  - ▶ `epitools`: Mostly infectious diseases.

## Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:
  - ▶ `Epi`: focus on chronic disease epidemiology:
    - ▶ Cohort studies, splitting follow-up time
    - ▶ Lexis diagram, several timescales
    - ▶ Multistate model support
    - ▶ Advanced tabulation
    - ▶ Informative reporting of estimation results
    - ▶ “Epidemiology with **R**”
  - ▶ `epicalc`:
  - ▶ `epitools`: Mostly infectious diseases.
  - ▶ `epiR`: Leaning towards veterinary epidemiology.

## Tools for nearly anything!

- ▶ Thousands of add-on packages.
- ▶ Several packages for epidemiological analyses:
  - ▶ Epi: focus on chronic disease epidemiology:
    - ▶ Cohort studies, splitting follow-up time
    - ▶ Lexis diagram, several timescales
    - ▶ Multistate model support
    - ▶ Advanced tabulation
    - ▶ Informative reporting of estimation results
    - ▶ “Epidemiology with **R**”
  - ▶ epicalc:
  - ▶ epitools: Mostly infectious diseases.
  - ▶ epiR: Leaning towards veterinary epidemiology.
- ▶ Packages can be installed and updated from within R.

# Running R

- ▶ Interactive but not mouse-driven!

*Difficult to learn & slow to use?*

## Running R

- ▶ Interactive but not mouse-driven!
- ▶ Commands typed from keyboard.

*Difficult to learn & slow to use?*

## Running R

- ▶ Interactive but not mouse-driven!
- ▶ Commands typed from keyboard.
- ▶ More practical: commands written and saved in a **script file** from which they are run.

*Difficult to learn & slow to use?*

## Running R

- ▶ Interactive but not mouse-driven!
- ▶ Commands typed from keyboard.
- ▶ More practical: commands written and saved in a **script file** from which they are run.
- ▶ Execution of tasks:

*Difficult to learn & slow to use?*



## Running R

- ▶ Interactive but not mouse-driven!
- ▶ Commands typed from keyboard.
- ▶ More practical: commands written and saved in a **script file** from which they are run.
- ▶ Execution of tasks:
  - ▶ evaluation of **expressions** contained in commands,

*Difficult to learn & slow to use?*

## Running R

- ▶ Interactive but not mouse-driven!
- ▶ Commands typed from keyboard.
- ▶ More practical: commands written and saved in a **script file** from which they are run.
- ▶ Execution of tasks:
  - ▶ evaluation of **expressions** contained in commands,
  - ▶ based on calls of **functions**.

*Difficult to learn & slow to use?*

## Running R

- ▶ Interactive but not mouse-driven!
- ▶ Commands typed from keyboard.
- ▶ More practical: commands written and saved in a **script file** from which they are run.
- ▶ Execution of tasks:
  - ▶ evaluation of **expressions** contained in commands,
  - ▶ based on calls of **functions**.

*Difficult to learn & slow to use?*

- ▶ Maybe in the beginning.

## Running R

- ▶ Interactive but not mouse-driven!
- ▶ Commands typed from keyboard.
- ▶ More practical: commands written and saved in a **script file** from which they are run.
- ▶ Execution of tasks:
  - ▶ evaluation of **expressions** contained in commands,
  - ▶ based on calls of **functions**.

*Difficult to learn & slow to use?*

- ▶ Maybe in the beginning.
- ▶ Versatility and flexibility rewarding in the long run.

## R as a simple calculator

Write the arithmetic expression on the empty line after the prompt and press Enter. The result is displayed immediately:

```
> 2+2
```

```
[1] 4
```

```
> 3*5 - 6/2
```

```
[1] 12
```

```
> (2+3)^2
```

```
[1] 25
```

```
> sqrt( 1/12 + 1/17 )
```

```
[1] 0.377037
```

```
> exp( 1.96 * sqrt( 1/12 + 1/17 ) )
```

```
[1] 2.093825
```

## R as a smart calculator

Simple summary of results from a cohort study:

|                           | Exposed | Unexposed |
|---------------------------|---------|-----------|
| No. of cases/Person-years | 20/2000 | 25/5000   |

## R as a smart calculator

- ▶ Numbers of cases and person-years are first assigned & saved into vectors D and Y;

```
> D <- c(20, 25) ; Y <- c(2000, 5000)
> rate <- 1000*D/Y ; rate
[1] 10 5
> ratio <- rate[1]/rate[2] ; diff <- rate[1]-rate[2]
> c(ratio, diff)
[1] 2 5
```

## R as a smart calculator

- ▶ Numbers of cases and person-years are first assigned & saved into vectors D and Y;
- ▶ Incidence rates in the two groups as well as their ratio and difference are then calculated and printed:

```
> D <- c(20, 25) ; Y <- c(2000, 5000)
> rate <- 1000*D/Y ; rate
[1] 10 5
> ratio <- rate[1]/rate[2] ; diff <- rate[1]-rate[2]
> c(ratio, diff)
[1] 2 5
```



## A couple of important things

- ▶ Names of **variables** (or any other **objects**)

## A couple of important things

- ▶ Names of **variables** (or any other **objects**)
  - ▶ Start with a letter from A, ..., Z or a, ..., z;  
lower case separated from upper case, e.g. 'x'  $\neq$  'X'

## A couple of important things

- ▶ Names of **variables** (or any other **objects**)
  - ▶ Start with a letter from A, ..., Z or a, ..., z;  
lower case separated from upper case, e.g. 'x'  $\neq$  'X'
  - ▶ Letters, integers 0, ..., 9, dots '.', and  
underlines '\_' allowed after 1st letter.

## A couple of important things

- ▶ Names of **variables** (or any other **objects**)
  - ▶ Start with a letter from A, ..., Z or a, ..., z;  
lower case separated from upper case, e.g. 'x'  $\neq$  'X'
  - ▶ Letters, integers 0, ..., 9, dots '.', and  
underlines '\_' allowed after 1st letter.
- ▶ **Assignment operator** '<-' (consists of '<' and '-')

## A couple of important things

- ▶ Names of **variables** (or any other **objects**)
  - ▶ Start with a letter from A, ..., Z or a, ..., z; lower case separated from upper case, e.g. 'x'  $\neq$  'X'
  - ▶ Letters, integers 0, ..., 9, dots '.', and underlines '\_' allowed after 1st letter.
- ▶ **Assignment operator** '<-' (consists of '<' and '-')
  - ▶ assigns a value to an object, for example

```
> A <- 5+2 ; A
[1] 7
```

means that a numeric variable **A** is given  $5+2 = 7$  as its value, and is then printed

## A couple of important things

- ▶ Names of **variables** (or any other **objects**)
  - ▶ Start with a letter from A, ..., Z or a, ..., z; lower case separated from upper case, e.g. 'x'  $\neq$  'X'
  - ▶ Letters, integers 0, ..., 9, dots '.', and underlines '\_' allowed after 1st letter.
- ▶ **Assignment operator** '<-' (consists of '<' and '-')
  - ▶ assigns a value to an object, for example

```
> A <- 5+2 ; A
[1] 7
```

means that a numeric variable **A** is given  $5+2 = 7$  as its value, and is then printed
  - ▶ the equal sign = is also allowed as assignment operator.

## Vectors and their arithmetics

A vector is ordered set of numbers (or other elements of the same type)

- ▶ Can be assigned values elementwise by function `c()`

## Vectors and their arithmetics

A vector is ordered set of numbers (or other elements of the same type)

- ▶ Can be assigned values elementwise by function `c()`
- ▶ Vector `x` with 4 elements 1, 2, 4, 7 assigned and printed:

```
> x <- c(1,2,4,7)
> x
[1] 1 2 4 7
```



## Vectors and their arithmetics

A vector is ordered set of numbers (or other elements of the same type)

- ▶ Can be assigned values elementwise by function `c()`
- ▶ Vector `x` with 4 elements 1, 2, 4, 7 assigned and printed:  

```
> x <- c(1,2,4,7)
> x
[1] 1 2 4 7
```
- ▶ Arithmetic operations `+`, `-`, `*`, `/`, `^` (power) for vectors of same **length**  
*i.e.* same number of

## Vectors and their arithmetics

A vector is ordered set of numbers (or other elements of the same type)

- ▶ Can be assigned values elementwise by function `c()`
- ▶ Vector `x` with 4 elements 1, 2, 4, 7 assigned and printed:  

```
> x <- c(1,2,4,7)
> x
[1] 1 2 4 7
```
- ▶ Arithmetic operations `+`, `-`, `*`, `/`, `^` (power) for vectors of same **length**  
*i.e.* same number of
- ▶ Outcome: a new vector whose elements are results of the operation on the corresponding elements in original vectors.

## Vectors and their arithmetics

A vector is ordered set of numbers (or other elements of the same type)

- ▶ Can be assigned values elementwise by function `c()`
- ▶ Vector `x` with 4 elements 1, 2, 4, 7 assigned and printed:

```
> x <- c(1,2,4,7)
> x
[1] 1 2 4 7
```
- ▶ Arithmetic operations `+`, `-`, `*`, `/`, `^` (power) for vectors of same **length** *i.e.* same number of
- ▶ Outcome: a new vector whose elements are results of the operation on the corresponding elements in original vectors.
- ▶ Function `seq( )` generates regular sequences.

## Vectors and their arithmetics

A vector is ordered set of numbers (or other elements of the same type)

- ▶ Can be assigned values elementwise by function `c()`
- ▶ Vector `x` with 4 elements 1, 2, 4, 7 assigned and printed:  

```
> x <- c(1,2,4,7)
> x
[1] 1 2 4 7
```
- ▶ Arithmetic operations `+`, `-`, `*`, `/`, `^` (power) for vectors of same **length** *i.e.* same number of
- ▶ Outcome: a new vector whose elements are results of the operation on the corresponding elements in original vectors.
- ▶ Function `seq()` generates regular sequences.
- ▶ Function `rep()` replicates same element(s).

## Vectors and their arithmetics

A vector is ordered set of numbers (or other elements of the same type)

- ▶ Can be assigned values elementwise by function `c()`
- ▶ Vector `x` with 4 elements 1, 2, 4, 7 assigned and printed:  

```
> x <- c(1,2,4,7)
> x
[1] 1 2 4 7
```
- ▶ Arithmetic operations `+`, `-`, `*`, `/`, `^` (power) for vectors of same **length** *i.e.* same number of
- ▶ Outcome: a new vector whose elements are results of the operation on the corresponding elements in original vectors.
- ▶ Function `seq()` generates regular sequences.
- ▶ Function `rep()` replicates same element(s).
- ▶ Common mathematical functions, like `sqrt()`, `log()`, `exp()` work in the same way for numeric vectors.

## R script – commands in a file

**R script file** is an ASCII file containing a sequence of **R** commands to be executed.

The **script editor** of R works as follows:

1. In RGui open the script editor window: *File - New script*, or when editing an existing script file: *File - Open script*,

## R script – commands in a file

**R script file** is an ASCII file containing a sequence of **R** commands to be executed.

The **script editor** of R works as follows:

1. In RGui open the script editor window: *File - New script*, or when editing an existing script file: *File - Open script*,
2. Write the command lines without prompt `>` or `+`.

## R script – commands in a file

**R script file** is an ASCII file containing a sequence of **R** commands to be executed.

The **script editor** of R works as follows:

1. In RGui open the script editor window: *File - New script*, or when editing an existing script file: *File - Open script*,
2. Write the command lines without prompt `>` or `+`.
3. Save the script file: *File - Save e.g. as c:\...\mycmds.R*  
or with some other file name having extension `.R`



## R script (cont'd)

4. Paint the lines to be executed and paste them on the console window using the third icon on the toolbar.

## R script (cont'd)

4. Paint the lines to be executed and paste them on the console window using the third icon on the toolbar.
5. Edit the file using *Edit* menu, save & continue.

## R script (cont'd)

4. Paint the lines to be executed and paste them on the console window using the third icon on the toolbar.
  5. Edit the file using *Edit* menu, save & continue.
- ▶ To run an entire script file, write in console window:
- ```
source("c:/.../mycmds.R", echo=TRUE)
```

## R script (cont'd)

4. Paint the lines to be executed and paste them on the console window using the third icon on the toolbar.
  5. Edit the file using *Edit* menu, save & continue.
- ▶ To run an entire script file, write in console window:  
`source("c:/.../mycmds.R", echo=TRUE)`
  - ▶ The script can also be written and edited by any external editor programs (like Notepad).

## R script (cont'd)

4. Paint the lines to be executed and paste them on the console window using the third icon on the toolbar.
  5. Edit the file using *Edit* menu, save & continue.
- ▶ To run an entire script file, write in console window:  
`source("c:/.../mycmds.R", echo=TRUE)`
  - ▶ The script can also be written and edited by any external editor programs (like Notepad).
  - ▶ *R Studio* — very versatile interface; see <https://www.rstudio.com/>. This may be what most of you have been introduced to.

## R in this course

- ▶ The main purpose is to inform you about the existence and potential of R, which you might find useful in any future work involving serious epidemiologic data analysis.

## R in this course

- ▶ The main purpose is to inform you about the existence and potential of R, which you might find useful in any future work involving serious epidemiologic data analysis.
- ▶ Here, **R** will be used only as a simple calculator.

## R in this course

- ▶ The main purpose is to inform you about the existence and potential of R, which you might find useful in any future work involving serious epidemiologic data analysis.
- ▶ Here, **R** will be used only as a simple calculator.
- ▶ No need for a lot of the more fancy stuff.



## R in this course

- ▶ The main purpose is to inform you about the existence and potential of R, which you might find useful in any future work involving serious epidemiologic data analysis.
- ▶ Here, **R** will be used only as a simple calculator.
- ▶ No need for a lot of the more fancy stuff.
- ▶ The script editor will help you keep your solutions for future reference.

## R in this course

- ▶ The main purpose is to inform you about the existence and potential of R, which you might find useful in any future work involving serious epidemiologic data analysis.
- ▶ Here, **R** will be used only as a simple calculator.
- ▶ No need for a lot of the more fancy stuff.
- ▶ The script editor will help you keep your solutions for future reference.
- ▶ After the course, solutions to all exercises will be provided.

## R in this course

- ▶ The main purpose is to inform you about the existence and potential of R, which you might find useful in any future work involving serious epidemiologic data analysis.
- ▶ Here, **R** will be used only as a simple calculator.
- ▶ No need for a lot of the more fancy stuff.
- ▶ The script editor will help you keep your solutions for future reference.
- ▶ After the course, solutions to all exercises will be provided.
- ▶ A good workbook introduction to R:  
<http://bendixcarstensen.com/Epi/R-intro.pdf>

# Practicals

**Bendix Carstensen & Esa Läärä**

Nordic Summerschool of  
Cancer Epidemiology

Danish Cancer Society / NCU, August 2022 / January 2023

<http://BendixCarstensen.com/NSCE/2022>

prac-seq

# How to do with practicals

- ▶ Read the text

▶ Find out what you want to do

▶ What is the question?

▶ Sources of practicals

## How to do with practicals

- ▶ Read the text
- ▶ Find out what you want to do

## How to do with practicals

- ▶ Read the text
- ▶ Find out what you want to do
- ▶ Then start using **R**

## How to do with practicals

- ▶ Read the text
- ▶ Find out what you want to do
- ▶ Then start using **R**
- ▶ Sequence of practicals:



## How to do with practicals

- ▶ Read the text
- ▶ Find out what you want to do
- ▶ Then start using **R**
- ▶ Sequence of practicals:
  1. Tuesday: 0, 1, 3, 4, 5, 7, 11, 12, 13

## How to do with practicals

- ▶ Read the text
- ▶ Find out what you want to do
- ▶ Then start using **R**
- ▶ Sequence of practicals:
  1. Tuesday: 0, 1, 3, 4, 5, 7, 11, 12, 13
  2. Monday: 7, 8, 2, 9, 10