**Nordic Summer School in Cancer Epidemiology**
14 to 25 August, 2017
Danish Cancer Society, Copenhagen

# Measures of disease frequency and effects

**Esa Läärä**

Unit of Mathematical Sciences, University of Oulu, Finland
esa.laara@oulu.fi http://math.oulu.fi/

**& Bendix Carstensen**

Steno Diabetes Center, Denmark
& Department of Biostatistics, University of Copenhagen
bendix.carstensen@regionh.dk
www.bendixcarstensen.com

## Outline

Introduction

Basic measures of frequency or occurrence

Measures of effect – comparative measures

Rates in many time scales

Standardization of rates

Survival analysis

Conclusion

Appendix: Introduction to R



## Key references

IS: dos Santos Silva, I. (1999). *Cancer Epidemiology: Principles and Methods.* International Agency for Research on Cancer, Lyon.

B&D: Breslow, N.E., Day, N.E. (1987). *Statistical Methods in Cancer Research Vol. II – The Design and Analysis of Cohort Studies.* IARC, Lyon.

C&H: Clayton, D., Hills, M. (1993). *Statistical Models in Epidemiology.* OUP, Oxford.



## Internet resources on cancer statistics

- **NORDCAN:** Incidence, mortality, prevalence and survival statistics from 41 major cancers in the Nordic countries.

  Association of the Nordic Cancer Registries (ANCR), Danish Cancer Society
  http://www-dep.iarc.fr/nordcan/English/frame.asp

  Reference: Engholm, G. *et al.* (2010) NORDCAN – a Nordic tool for cancer information, planning, quality control and research. *Acta Oncologica* **49**: 725-736.

- **GLOBOCAN:** Estimates of the incidence of, mortality, prevalence and disability-adjusted life years (DALYs) from major type of cancers, at national level, for 184 countries of the world in 2008.

  International Agency for Research on Cancer (IARC);
  http://globocan.iarc.fr/



## INTRODUCTION

### What is epidemiology?

Some textbook definitions:

- "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, . . . " (Porta (ed.) Dictionary of Epidemiology, 2014)
- "discipline on principles of **occurrence** research in medicine" (Miettinen 1985)



## Different epidemiologies

- **descriptive** epidemiology – monitoring & surveillance of diseases for planning of health services – a major activity of cancer registries.
- **etiologic** or "analytic" epidemiology – study of cause-effect relationships
- **disease** epidemiologies – *e.g.* of cancer, cardiovascular diseases, infectious diseases, musculoskeletal disorders, mental health, . . .
- **determinant-based** epidemiologies – *e.g.* occupational epidemiology, nutritional epidemiology, . . .
- **clinical** epidemiology – study of diagnosis, prognosis and effectiveness of therapies in patient populations – basis of evidence-based medicine



## **Frequency** (from Webster's Dictionary)

Etymology: < L *frequentia* = assembly, multitude, crowd.

2. rate of occurrence
3. *Physics.* number of . . . regularly occurring events . . . in unit of time,
5. *Statistics.* the number of items occurring in a given category. Cf. **relative frequency**.

These meanings are all relevant in epidemiology.

But what are **rate** and **occurrence**?



## Cancer in 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 | 401 | 269 | 26.5 | 101 |
| Iceland | 633 | 464 | **347** | **32.6** | **132** |
| Norway | 10 246 | **469** | 294 | 29.4 | 109 |
| Sweden | **19 908** | 455 | 249 | 25.4 | 93 |

- Where is the frequency truly **highest**, where lowest?
- What do these measures mean?

## Questions on frequency & occurrence

How many women in Denmark

- are carriers of breast cancer today at 12? – **prevalence**
- will contract a new breast ca. during 2015? – **incidence**
- die from breast ca. in 2015? – **mortality**
- will be alive after 5 years since diagnosis among those getting breast ca. in 2015? – **survival**
- are cured of breast cancer during 2015? – **cure**

What are the **proportions** or/and **rates** of occurrence of these states and events?



## Questions on risk

- How great are the **risks** of these events?
- Is the risk of breast ca. among nulliparous **greater than** among parous women?
- What are the **excess** and **relative risks** for nulliparous compared to parous women?
- What is the **dose-response relationship** between occupational exposure to crystalline silica and the risk of getting lung cancer in terms of level and length of exposure?



## Descriptive and causal questions

- **Descriptive:** What is the occurrence of lung cancer workers exposed to silica dust as compared to that in subjects of other occupations?
- **Causal:** What is the risk of lung cancer among silica dust workers *as compared to* . . . what the risk in these same men *would be, had they not been* exposed to silica?

**NB.** Causal question – **counterfactual conditional**!

Challenge: *How to find a* **comparable** *group of unexposed?*



## What is risk?

Phrase "Risk of disease $S$" may refer to different concepts:

(i) **probability** of *getting* $S$ during a given **risk period** → **incidence** probability,

(ii) **rate** of change of that probability → **hazard** or intensity, or

(iii) **probability** of *carrying* $S$ at a given *time point* → **prevalence** probability.

Most commonly meaning (i) is attached with risk.

**NB.** "Risk" should not be used in the meaning of **risk factor**.

However, in **risk assessment** literature: "hazard" is often used in that meaning. In statistics, though, hazard refers to notion (ii): change of probability per unit time.



## Risks are conditional probabilities

- There are no "absolute risks".
- All risks are conditional on a multitude of factors, like
  - length of risk period (*e.g.* next week or lifetime),
  - age and gender,
  - genetic constitution,
  - health behaviour & environmental exposures.
- In principle each individual has an own quantitative value for the risk of given disease in any defined risk period, depending on his/her own risk factor profile.
- Yet, these individual risks are latent and unmeasurable.
- **Average risks** of disease in large groups sharing common characteristics (like gender, age, smoking status) are estimable from appropriate epidemiologic studies by pertinent **measures of occurrence**.



## BASIC MEASURES OF FREQUENCY OR OCCURRENCE

Quantification of the occurence 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
- 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** ($Q$) 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.* **IS**).

**NB.** "Cumulative incidence" has other meanings, too.

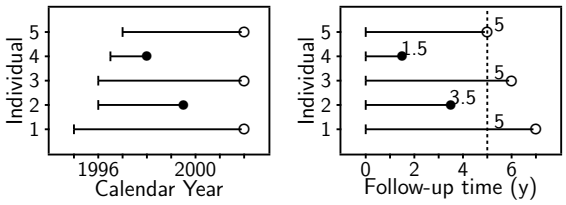
- **Indidence rate** ($I$) over a defined observation period:

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

Also called **incidence density**.

## Example: Follow-up of a small cohort

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



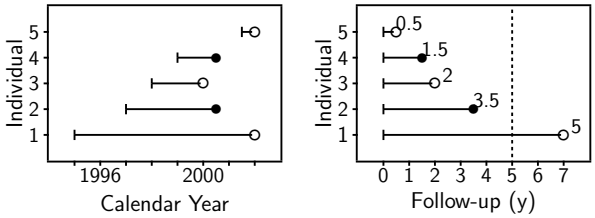

**Complete follow-up** in the 5-year risk period
⇒ can calculate both measures:

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

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



## Follow-up of another small cohort




Two censored observations ⇒ the rate can be calculated:

$$I = 2/12.5 \text{ y} = 16 \text{ per } 100 \text{ years}$$

but the 5-year incidence proportion **IS NO MORE** $2/5$ !

However, under the constant rate model and in the absence of competing risks, the incidence proportion is obtained:

$$Q = 1 - \exp(-5 \times 2/12.5) = 0.55 \text{ (or 55\%)}$$



## Properties of incidence proportion

- Dimensionless quantity ranging from 0 to 1 (0% to 100%) = *relative frequency*,
- Estimates the average theoretical **risk** or 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 ⇒ Calculations need to be corrected using special methods of survival analysis.



## Person-years in dynamic populations

With 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 $N_{t-1}$ at start and $N_t$ at the end of the observation period $t$ with length $u_t$ years,

(2) Mid-population for the period: $\bar{N}_t = \frac{1}{2} \times (N_{t-1} + N_t)$.

(3) Approximate person-years: $\widetilde{Y}_t = \bar{N}_t \times u_t$.

**NB.** The actual study population often contains also some already affected, who thus do not belong to the population at risk. With rare outcomes the influence of this is small.



## Properties of incidence rate

- Like *a frequency* quantity in physics; measurement unit: *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 depends on age (& other time variables) ⇒ rates *specific to age group etc.* needed,
- Incidence proportions can be estimated from rates. In the constant hazard model with no competing risks:

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



## Male person-years in Finland 1991-95

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

| 1990 | 1991 | 1992 | 1993 | 1994 | 1995 |
|---|---|---|---|---|---|
| 2431 | 2443 | 2457 | 2470 | 2482 | 2492 |

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

| | | |
|---|---|---|
| 1992: | $\frac{1}{2} \times (2443 + 2457) \times 1 =$ | 2450 |
| 1993-94: | $\frac{1}{2} \times (2457 + 2482) \times 2 =$ | 4937 |
| 1991-95: | $\frac{1}{2} \times (2431 + 2492) \times 5 =$ | 12307.5 |



## Competing events and censoring

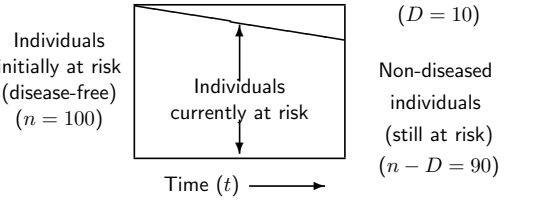
The outcome event of interest (*e.g.* onset of disease) is not always observed for all subjects during the chosen risk period.

- Some subjects die (from other causes) before the event.
  ⇒ Death is a **competing event** after which the outcome cannot occur any more.
- Others emigrate and escape national disease registration, or the whole study is closed "now", which prematurely interrupts the follow-up of some individuals
  ⇒ **censoring**, **withdrawal**, or **loss to follow-up**

Competing events and censorings require special statistical treatment in estimation of incidence and risk.



## Incidence proportion, rate, and odds (IS, Ex 4.5)




Assuming a risk period of 1 year with complete follow-up:

$$\text{Incidence proportion } Q = 10/100 = 0.10 = 10\%$$

$$\text{Incidence rate } I = 10/95 \text{ y} = 10.5 \text{ per 100 y}$$

$$\textbf{Incidence odds } Q/(1-Q) = 10/90 = 0.11 = 11 \text{ per } 100$$

## Approximate relations btw measures

With sufficiently

- "short" length $\Delta$ of risk period and
- "low" risk (say $Q < 5\%$)

the incidence proportion $Q$, rate $I$ and odds are approximately related as follows:

$$\frac{Q}{1-Q} \approx Q \approx I \times \Delta$$

The "**rare disease assumption**".



## Mortality

**Cause-specific mortality** from disease $S$ is described by **mortality rates** defined like $I$ but

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

Cause-specific **mortality proportions** must be corrected for the incidence of **competing causes of death**

**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

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

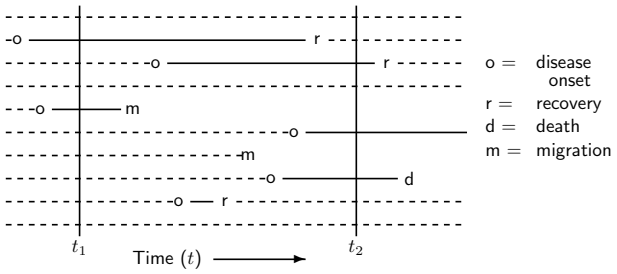
This is calculable from a cross-sectional study base.

**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 $t_2$.



## 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: $5/8 = 0.62 = 62\%$



## Prevalence and incidence are related

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

(a) *incidence* of new cases of $S$ before $t$, and the
(b) *duration* of $S$, depending in turn on the probability of *cure* or recovery from $S$, or *survival* of those affected

typically in a complicated way.

Simple special case: In a **stationary** population, the prevalence ($P$), incidence ($I$), and average duration ($\bar{d}$) of $S$ have a simple relationship:

$$P = \frac{I \times \bar{d}}{I \times \bar{d} + 1} \quad \approx \quad I \times \bar{d}$$

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



## 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
  (a) are alive and resident in the population at $t$, and
  (b) 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 1 y (initial treatment period), 3 y (clinical follow-up), or 5 y (cure?).



## Ex: Cancers with poor and good prognosis

Age-standardized[a] incidence, mortality, prevalence, and survival for cancers of kidney and thyroid in women of Finland.

| | Kidney | Thyroid |
|---|---|---|
| Incidence rate in 2011 (per $10^5$ y) | 12 | 11 |
| Mortality rate in 2011 (per $10^5$ y) | 5 | 1 |
| Prevalence on 31.12.2011 (per $10^5$) | 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 |

[a] Standard: Nordic population in 2000



## 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)

- In causal studies these are used to estimate the **causal effect** of the factor on the disease risk.

  $\Rightarrow$ **comparative measure** $\approx$ **effect measure**

- 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 $I_1/I_0$,
- incidence proportion ratio $Q_1/Q_0$,
- incidence odds ratio $[Q_1/(1-Q_1)]/[Q_0/(1-Q_0)]$,
- prevalence ratio $P_1/P_0$, or
- prevalence odds ratio $[P_1/(1-P_1)]/[P_0/(1-P_0)]$,

depending on study base and details of its design.

Incidence rate ratio 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

- incidence rate difference $I_1 - I_0$,
- incidence proportion difference $Q_1 - Q_0$, or
- prevalence difference $P_1 - P_0$.

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.

| | Disease A | Disease B |
|---|---|---|
| Incidence rate among exposed[a] | 20 | 80 |
| Incidence rate among unexposed[a] | 5 | 40 |
| Rate ratio | 4.0 | 2.0 |
| Rate difference[a] | 15 | 40 |

[a] 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.



## Ratio measures in "rare diseases" (IS: Ex 5.13)

| | Exposure | |
|---|---|---|
| | Yes | No |
| No. initially at risk | 4 000 | 16 000 |
| No. of cases | 30 | 60 |
| Person-years at risk | 7 970 | 31 940 |

Inc. prop'n ratio $= \frac{30/4\,000}{60/16\,000} = \frac{7.5 \text{ per } 1\,000}{3.75 \text{ per } 1\,000} =$ **2.0000**

Inc. rate ratio $= \frac{30/7\,970 \text{ y}}{60/31\,940 \text{ y}} = \frac{3.76 \text{ per } 1\,000 \text{ y}}{1.88 \text{ per } 1\,000 \text{ y}} =$ **2.0038**

Inc. odds ratio $= \frac{30/(4\,000-30)}{60/(16\,000-60)} = \frac{0.00756}{0.00376} =$ **2.0076**

With low incidence these ratios are very similar.



## Attributable fraction (excess fraction)

- **Measures of potential impact**: Combination of absolute and relative comparisons.
- When the incidence is higher in the exposed, the **attributable fraction** (AF) for the exposure or risk factor is defined as:

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

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

- 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.



## 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

$$p_\text{E} = \text{ proportion of exposed in the population.}$$

- **Population excess fraction** (PAF) is now defined:

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

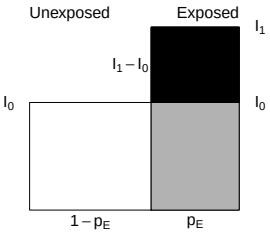
- AF: biological impact of exposure,
- PAF: impact of exposure on the population level.



## Excess fraction illustrated

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

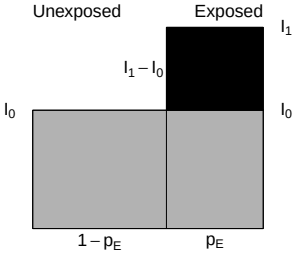
- Total incidence $I$ in population – weighted average:

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

would equal $I_0$, if exposure had no effect

- Excess incidence caused by exposure: $I - I_0 = p_\text{E} \times (I_1 - I_0)$ (black area).
- PAF $= \frac{I - I_0}{I}$,

= fraction of black area out of total black + gray area.



## Prevented fractions

- When the incidence in exposed is lower, we define the **prevented fraction** for such a preventive factor:

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

  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:

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

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



## Effect of smoking on mortality by cause

(**IS**: Example 5.14, p. 98)

| Underlying cause of death | Never smoked regularly Rate[b] | Current cigarette smoker Rate[b] | Rate ratio | Rate difference[b] | Excess fraction (%) |
|---|---|---|---|---|---|
| | (1) | (2) | (2)/(1) | (2) − (1) | $\frac{(2)-(1)}{(2)} \times 100$ |
| Cancer | | | | | |
| 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 (except cancer) | 107 | 313 | 2.9 | 206 | 66 |
| Vascular diseases | 1037 | 1643 | 1.6 | 606 | 37 |
| All causes | 1706 | 3038 | 1.8 | 1332 | 44 |

[a] Data from Doll *et al.*, 1994a.
[b] Age-adjusted rates per 100 000 pyrs.



## RATES IN MANY TIME SCALES

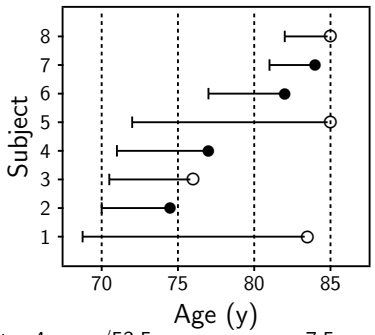
Incidence can be studied on various distinct time scales, *e.g.*

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

- Age is usully the strongest time-dependent determinant of health outcomes.
- Age is also often correlated with duration of "chronic" exposure (*e.g.* years of smoking).



## Follow-up of a small geriatric cohort




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

Hides the fact that the "true" rate probably varies by age, being higher among the old.



## 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 model**



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

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



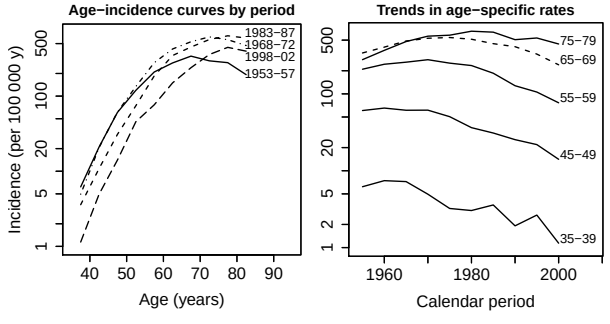
## Ex. Lung cancer incidence in Finland by age and period (compare IS, Table 4.1)

| Calendar | Age group (y) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| period | 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




- Age-incidence curves: overall level and peak age variable across periods.
- Time trends inconsistent across age groups.

## 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.
- Often more informative about "true" age-incidence pattern than age-specific incidences of single calendar period.



## Age-specific rates by birth cohort

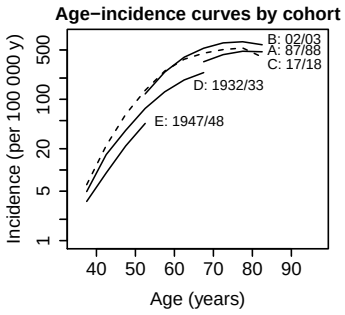
| Calendar | Age group (y) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| period | 40-44 | 45-49 | 50-54 | 55-59 | 60-64 | 65-69 | 70-74 | 75-79 | |
| 1953-57 | 21 | 61 | [119] | 209 | 276 | [340] | 295 | 279 | |
| 1958-62 | [22] | 65 | 135 | [243] | 360 | 405 | [429] | 368 | |
| 1963-67 | 24 | [61] | 143 | 258 | [395] | 487 | 509 | [479] | A |
| 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] | B |
| 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] | C |
| 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 pattern in birth cohorts.



## Age-incidence curves in 5 birth cohorts




Variable overall levels 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, users of certain medicine)
  - → often compared to incidence in a **reference** or "general" population.
- For examples, see Laufey's lecture on cohort studies (*e.g.* atomic bomb survivors, rubber workers, and those exposed to dyestaff)
- 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 (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 | 1904 | 1943 | 1952 | 39 | Migrated |
| 2 | 1924 | 1948 | 1955 | 24 | Disease $C$ |
| 3 | 1914 | 1945 | 1961 | 31 | Study ends |
| 4 | 1920 | 1948 | 1956 | 28 | Unrelated death |

Subject 1: Follow-up time spent in each ageband

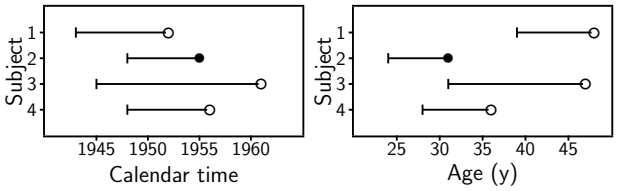
| Age band | Date in | Date out | Time (years) |
|---|---|---|---|
| 35–39 | 1943 | 1944 | 1 |
| 40–44 | 1944 | 1949 | 5 |
| 45–49 | 1949 | 1952 | 3 |



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

Follow-up of cohort members by calendar time and age

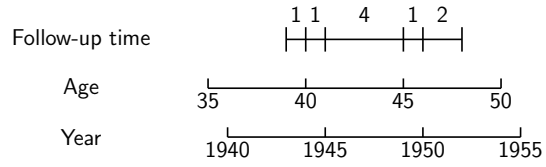
| entry
• exit because of disease onset (outcome of interest)
○ exit due to other reason (censoring)






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

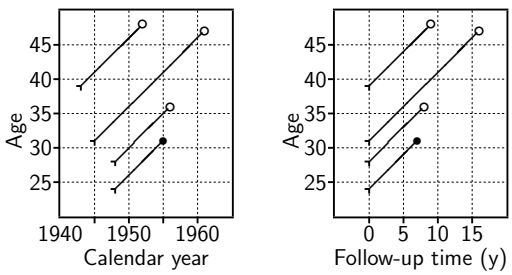
Subject 1: Follow-up jointly split by age and calendar time:




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



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




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

See also Laufey's lecture on cohort studies, slide 4.

# STANDARDIZATION OF RATES

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

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

| | % of male population | | | |
|---|---|---|---|---|
| Age (years) | Cali 1984 | B'ham 1985 | Finland 2011 | World 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.

$\Rightarrow$ Crude rates are **confounded** by age.

$\Rightarrow$ Any summary rate must be **adjusted for age**.



# Adjustment by standardisation

**Age-standardised incidence rate** (ASR):

$$\text{ASR} = \sum_{k=1}^{K} \text{weight}_k \times \text{rate}_k \;/\; \text{sum of weights}$$

$=$ **Weighted average** of age-specific rates over the age-groups $k = 1, \ldots, K$.

- Weights describe the age distribution of some **standard population**.
- Standard population can be real (*e.g.* one of the populations under comparison, or their average) 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 | Nordic[a] |
|---|---|---|---|---|
| 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 |

[a] NORDCAN population in 2000.



# Stomach cancer in Cali & B'ham

Age-standardized rates by the World Standard Population:

| | Cali | | | Birmingham | |
|---|---|---|---|---|---|
| Age | Rate[a] | Weight | | Rate[a] | Weight |
| 0–44 | 1.5 $\times$ | 0.74 = | 1.11 | 1.2 $\times$ | 0.74 = 0.89 |
| 45–64 | 69.7 $\times$ | 0.19 = | 13.24 | 44.6 $\times$ | 0.19 = 8.47 |
| 65+ | 281.3 $\times$ | 0.07 = | 19.69 | 202.0 $\times$ | 0.07 = 14.14 |
| **Age-standardised rate** | | | **34.04** | | **23.50** |

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

  $$\text{SRR} = 34.0/23.5 = 1.44.$$
- Known as **comparative mortality figure (CMF)** when the outcome is death (from cause $C$ or all causes).



# Cumulative rate and "cumulative risk"

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

  $$\text{CumRate} = \sum_{k=1}^{K} \text{width}_k \times \text{rate}_k,$$
- Weights are now widths of the agebands to be included, usually up to 65 or 75 y with 5-y bands.
- NORDCAN & GLOBOCAN use a 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.
- Yet, in reality competing events are present, so the probability interpretation of CumRisk is problematic.



# Stomach cancer in Cali & B'ham

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

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

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

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

"Cumulative risks" & their ratio up to 65 y:

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

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

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

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

## Cumulative measures using 5-y groups

(IS, Fig 4.11, p. 77)

| 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 |

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

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



## Cumulative and life-time risks

Of course, it is 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?*”

However, this is difficult 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 common cancers by age, Finland 2005

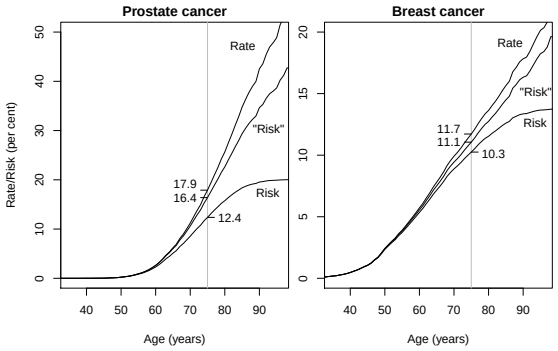



## Estimation of cumulative risks

- 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,
- 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.* rubber workers, see Laufey's lecture on cohort studies)
- Cohorts of patients on chronic medication, which may have harmful long-term side-effects
- No internal comparison group of unexposed subjects.

*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

- 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 (or st'zed mortality ratio SMR with death as outcome)

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



## Example: HT and breast ca.

- A cohort of 974 women treated with hormone (replacement) therapy 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 (a) were:

| 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 |
| $\sum$ | | | 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 $= 15/16.77 = 0.89$.
- 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 B'ham as reference

Total person-years at risk and expected number of cases in Cali 1982-86 based on age-specific rates in Birmingham (**IS**: Fig. 4.9, p. 74)

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

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, p. 78, extended)

| | Cali, 1982-86 | B'ham, 1983-86 | Rate ratio |
|---|---|---|---|
| Crude rates ($/10^5$ y) | 19.9 | 33.9 | *0.59* |
| ASR ($/10^5$ y)$^B$ with 3 broad age groups | 48.0 | 33.9 | *1.42* |
| ASR ($/10^5$ y)$^C$ –"– | 19.9 | 14.4 | *1.38* |
| ASR ($/10^5$ y)$^W$ –"– | 34.0 | 23.5 | *1.44* |
| Cum. rate < 65 y (per 1000) –"– | 14.6 | 9.5 | *1.54* |
| ASR ($/10^5$ y)$^W$ 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: $^B$ Birmingham 1985, $^C$ Cali 1985, $^W$ 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. 5-year age groups are preferred.



## 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)

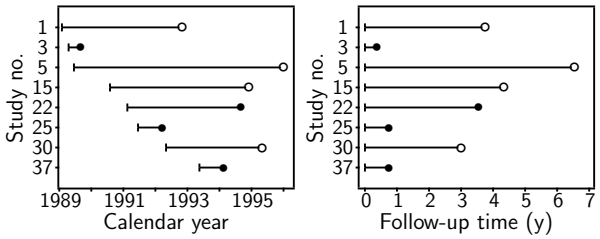
| No. | Age (y) | Sta-ge[a] | Date of diag-nosis | Date at end of follow -up | Vital status at end of follow -up | Cause of death[c] | Full years from diagn's up to end of follow -up | Days from diagn's up to end of follow -up |
|---|---|---|---|---|---|---|---|---|
| 1 | 39 | 1 | 01/02/89 | 23/10/92 | 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 |

[a] 1 = absence of regional lymph node involment and metastases
2 = involvment of regional lymph node and/or presence of metastases
[b] A = alive; D = dead; [c] BC = breast cancer; O = other causes



## Follow-up of breast ca. pts (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.)

**(1)** Divide the follow-up time into subintervals $k = 1, \ldots K$; most of these having width of 1 year.

Often the first year is divided into two intervals with widths of 3 mo and 9 mo, respectively.

**(2)** Tabulate from original data for each interval

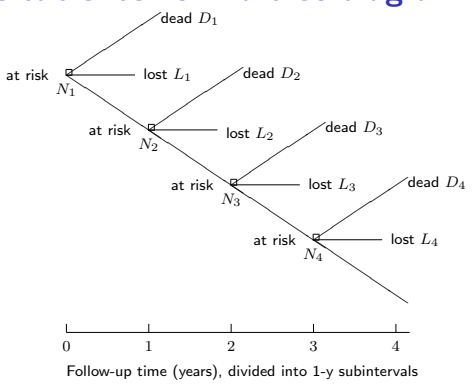
$N_k =$ size of the **risk set**, *i.e.* the no. of subjects still alive and under follow-up at the start of interval,

$D_k =$ 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




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 for breast ca. patients

(**IS**: Table 12.2., p. 273, first 4 columns)

| Interval ($k$) | Years since diagnosis | No. at start of interval ($N_k$) | No. of deaths ($D_k$) | No. of losses ($L_k$) |
|---|---|---|---|---|
| 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

$N'_k =$ $N_k - L_k/2 =$ 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 ($k$) | Years since diagnosis | No. at start of interval ($N_k$) | No. of deaths ($D_k$) | No. of losses ($L_k$) | Effective denominator ($N'_k$) | Cond'l prop'n of deaths during int'l ($q_k$) | Survival prop'n over int'l ($p_k$) | Cumul. survival; est'd survival prob'ty ($S_k$) |
|---|---|---|---|---|---|---|---|---|
| 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

- 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.
- Indidence rate in the $k$th interval is computed as:

$$I_k = \frac{\text{number of cases } (D_k)}{\text{approximate person-time } (\widetilde{Y}_k)}$$

where the approximate person-time is given by

$$\widetilde{Y}_k = \left[N_k - \frac{1}{2}(D_k + L_k)\right] \times \text{width of interval}$$

The dead and censored thus contribute half of the interval width.



## Survival curve and other measures

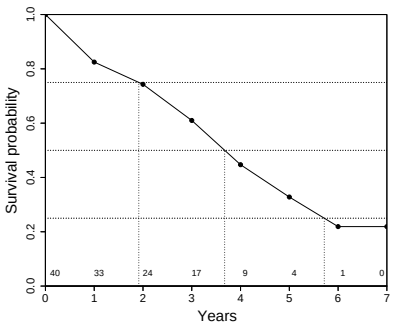
Line diagram of survival proportions through interval endpoints provides graphical estimates of interesting parameters of the survival time distribution, *e.g.*:

- **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.



## Survical curve of breast ca. patients (IS: Fig 12.8)




Numbers above $x$-axis show the size of population at risk.



## 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**:

$$R_k = S_k^{\text{obs}} / S_k^{\text{exp}}, \qquad \text{where}$$

- $S_k^{\text{obs}} =$ **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 (*cf.* SIR!)

$+$ No information on causes of death needed.



## 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.

# APPENDIX: 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
  - Versatile tabulation
  - High-quality graphics
- Analytic methods
  - Basic epidemiologic statistics
  - Survival analysis methods
  - Common regression models and their extensions
  - Other. . .

These are provided by *e.g.* SPSS, SAS and Stata, too, so . . . ?

Many features of R are more appealing in the long run.

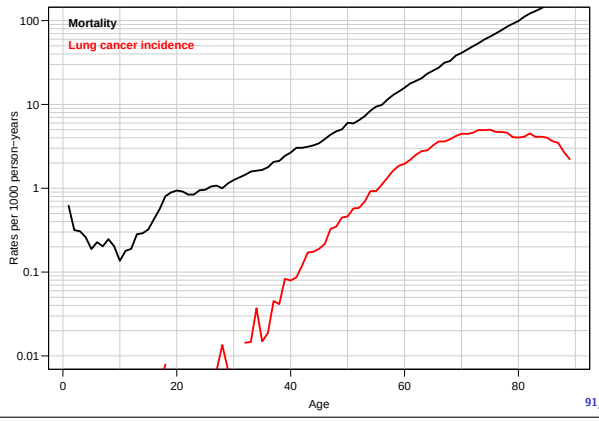


# Graphics in R

- Versatile, flexible, high quality, . . .
- 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!)
- Interactive tools using the mouse
  - Put new things on a graph
  - Identify points

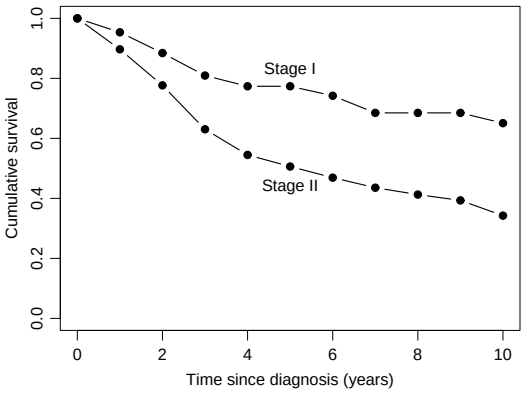


# Total mortality and lung ca incidence in DK



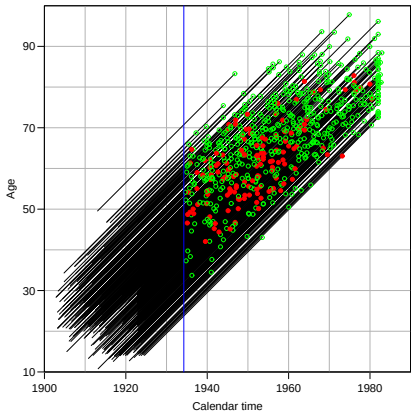



# Survival of cervix ca patients (C&H, 34)



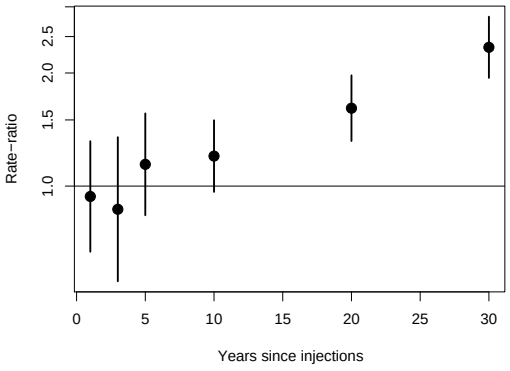



# Lexis diagram of Welsh nickel cohort






# Rate ratios with confidence intervals






# Getting your graphs out

- Graphs can be saved to disk in almost any format

  `.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.
- 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
  - `epicalc`:
  - `epitools`: Mostly infectious diseases.
  - `epiR`: Leaning towards veterinary epidemiology.
- Packages may be installed and updated from within R.



## 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.



## Running R on Windows

- Start by double-clicking the R-icon.
- `R Console`: the **console window**
  - command lines to be typed – or pasted from a script file – after prompt '>',
  - prompt '+' marks continuation of an incomplete command line,
  - output follows a completed command requesting it,
  - arrow key ↑ leads to previous command lines.
- Menu bar for a few useful pull-down menus.
- On-line help in HTML form.



## 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] 0.3770370
> 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 |

- 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**)
  - Start with a letter from `A,...,Z` or `a,...,z`; lower case separated from upper case, *e.g.* 'x' ≠ '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

**Vector** = ordered set of numbers (or other similar elements)

- 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 elements.

⇒ Outcome: a new vector whose elements are results of the operation on the corresponding elements in original vectors.

- 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*,
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 excecuted 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 a whole 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).
- Of these, *Tinn-R* provides nice facilities for editing, checking and running R scripts, see `http://www.sciviews.org/Tinn-R/`.
- *R Studio* – very versatile interface; see `http://www.rstudio.com/`.



## 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`