1.1 Disease occurrence and risk

review of classical sampling designs

What is Design?

Miettinen* (1982):

"a vision of the end product of a study on one hand

and

a scheme for carrying out a study on the other"

In this course:

"end product"

a measure of occurrence or risk of an outcome (disease)

"schemes" for sampling and analysis

** Miettinen O. Design options in epidemiologic research: an update Scan. J Work and Environ Heath. 1982.*

Measures of disease occurrence and risk

Measures of disease occurrence

Proportions

e.g. prevalence

The number of cases of the disease in a population at a specific time divided by the number of members in the population

Rates

e.g. incidence

The number of new cases of the disease in a population in a specified time, often reported as cases per 100,000 persons per year

Lot of terminology but only **two** types of measures (proportions or rates)!
only **two** types of epidemiology study (prevalence or incidence, Pearce*)

** Pearce N. Classification of eidemiological study designs. Int. Jour Epi. 2012*

Disease occurrence (proportions)

Prevalence: The proportion of people having the disease at a specified time, π (t)

Cumulative incidence: The proportion of people who get the disease during the follow up period Π.

Quiz

Another way of presenting proportions (Odds)

Prevalence: The proportion of people having the disease at a specified time, π (t)

Prevalence Odds: $\frac{\pi(t)}{1-\pi(t)} = \frac{No.of\ cases}{No.of\ non-cases}$ **at time t.**

Cumulative incidence: The proportion of people who get the disease during the follow up period Π

Cumulative odds : $\frac{\Pi}{1-\Pi} = \frac{No.of cases}{No.of\ non-cases}$ **at the end of follow-up**

Another way of presenting proportions (Odds)*

Table 1. *Examples of risks (given as fractions or percentages) and their corresponding odds (given as fractions)*

Risk	Corresponding Odds
1/1000 (.1%)	1/999
1/100 (1%)	1/99
1/50 (2%)	1/49
1/10 (10%)	1/9
1/4 (25%)	1/3
1/2 (50%)	1/1
9/10 (90%)	9/1
99/100 (99%)	99/1

* From Sainani, Physical Med and Rehab 2011, *Understanding odds ratios.*

Disease occurrence (rates)

A **rate** is a the number of events occurring *per unit of time* **r(t)**

Incidence rate of disease commonly expressed as number of new cases per 100,000 individuals in a specified time

- e.g. If 32 children develop diabetes in a population of 200,000 children in a year. The incidence is ”32 per 200,000 persons per year”, or 32/200,000 person-years = 16/100,000 person-years.

- Sometimes other denominators used: e.g. per 100 person years or 1000py or 10,000 py

Incidence rate vs. mortality rate

When death is the event whose incidence we are measuring, we refer to the *mortality rate*.

Other terms for Incidence Rate

- "Incidence density"
- "person-time incidence rate"
- "force of morbidity"
- …

Incidence rate vs. hazard rate

Hazard rate is the **instantaneous** incidence rate

Denote outcome as Y: event as Y=1, then

Hazard at time t, $h\,(t)\; = Prob(Y_{it} = 1 | Y_i \geq t)$

i.e. the probability of the event at time t for an individual who has not had the event before t.

Commonly used for mortality (i.e. "survival" studies)

Often we wish to study a *risk factor*

For a simple dichotomous “exposure”, compare those exposed and unexposed for their:

- prevalence (relative risk, **RR**)
- Odds (odds ratio, **OR**)
- Incidence (Incidence rate ratio, **IRR**)
- Hazard (Hazard ratio, **HR**)

Model how the risk depends on the level of an exposure X

$$\mathrm{r}\,(\mathrm{t}) = f\,(X)$$

and on other ”determinants” (including confounders and modifiers)

$$\mathrm{r}\,(\mathrm{t}) = f\,(X, D)$$

Population and Sample

Population: the whole collection of individuals about whom information is desired

In research studies, samples of individuals are taken from the population of interest, and used to make generalisations about the larger population

For these generalisations to be valid, it is important that the sample can validly represent the population

There are various prescribed ways of choosing a representative sample (i.e. the sampling “scheme” or “design”)

Target & Study population

Target Population:

the population to whom we wish to generalise our findings

Study Population:

the population from which we sampled, also called the "study base" (especially where we wish to denote the individuals and time)

Clinical Trials are often conducted on special subgroups of patients who meet some "inclusion criteria", so that we may only be able to generalise the results to a restricted "study population" represented by these patients, and not to the target population we would wish.

For an observational study, the target and study population should ideally be the same, but in practice, there may be some differences due to logistic constraints, incomplete data,…. (more in lecture 1.2)

Example: target and study population

we may wish to investigate the prevalence of diabetes in men age 40-50 years in a specific country (target population).

the available (electronic) data might consist only of diabetes cases in the inpatient and outpatient records of the hospitals

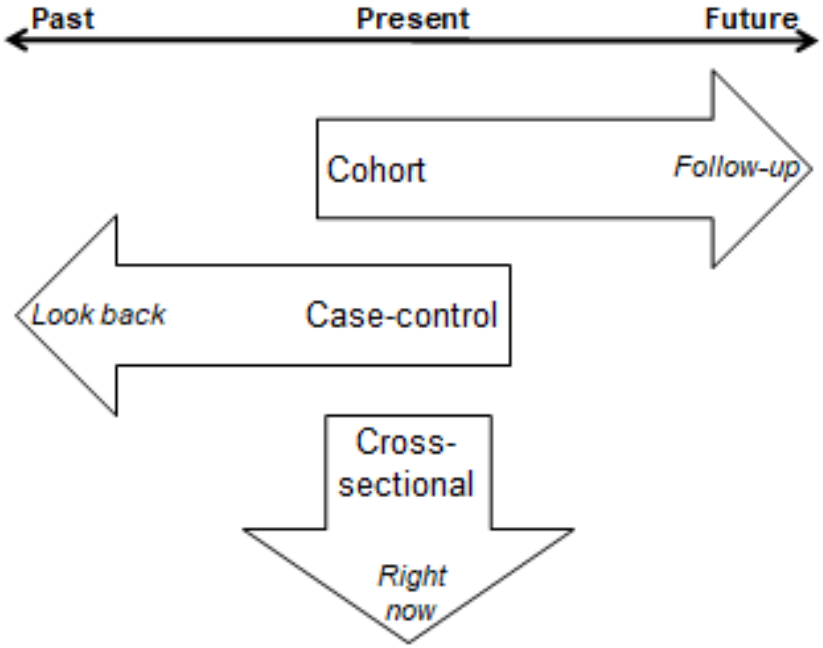
or we may only have data from

…… all the large/teaching hospitals

…… or only those in the capital city

If there are major differences, we may not be justified in generalising our findings

Observational Study Designs



Cross-sectional studies (Surveys)

- Like taking a “snapshot” of the population
- ascertain outcome (Y) and exposure (X)
- measure: prevalence, RR

Cohort studies

- Enrol a well-defined group of individuals at a given time (“time” can be age/date/other start)
- “follow” the experience of those individuals over time
- Like taking a “video”
- measure:incidence, IRR, HR

Case-control studies

- Start by identifying cases (Y=1) and reference/controls (Y=0).
- Measure: odds and OR of having the exposure (X=1 vs X=0)

OR of exposure in cases vs controls ≡ OR of disease in exposed vs. unexposed

Advantages of cohort studies

- Suitable for studying rare exposure
- Can assess multiple outcomes (effects) of a single exposure
- Can demonstrate temporal relationship between exposure and disease
- Allows direct measurement of incidence of disease in exposed and unexposed populations

Open cohort
Example: OCP use and CHD in a town*

On average, 120 000 women age 15 and 45 years without CHD are living in the town. Each day, new women turn15, others turn 46, some leave town and others come, some develop CHD and are replaced by others who do not

…… a dynamic cohort

on average, 40 000 use OCP and 80 000 do not

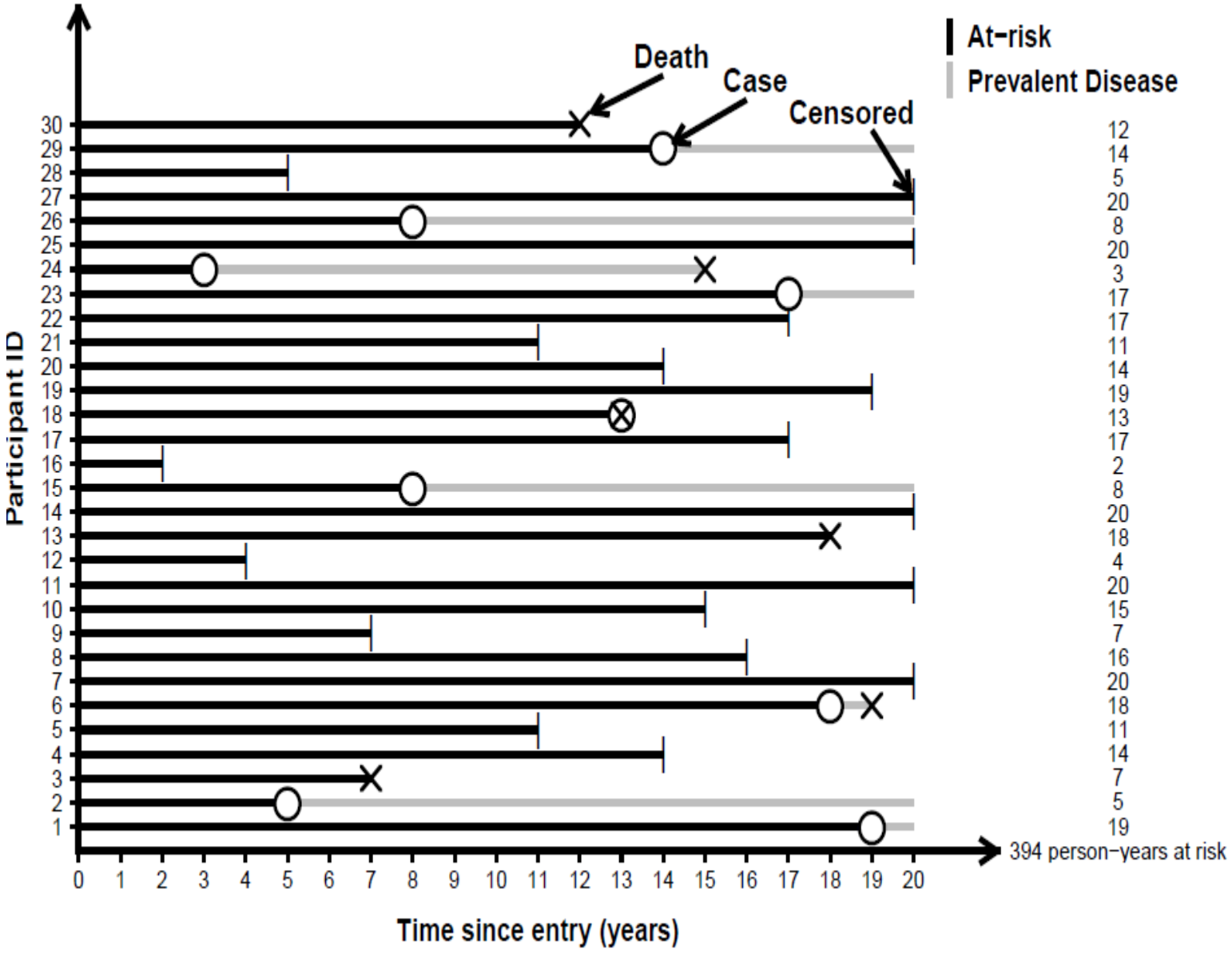
12 women with MI during the year (8 users, 4 non-users)

$$\text{IRR} = \frac{8}{40000\ py} \div \frac{4}{80000\ py} = 4$$

Person-years can also be calculated individually from "entry time" to "exit time"

**Vandenbroucke and Pearce, IJE, 2012*

Calculating Incidence Density (closed cohort , exact individual follow-up)



Incidence rate
=9/394
=2.28/100py

Difficulties with cohort studies

1. Loss-to-follow up:
   Subjects may lose interest, die, move to another area…
   Reduced study size weakens analysis
   More importantly may cause ***bias***

2. Follow-up may be expensive (time and money)

3. Changes in habits (time trends) and hence exposure.

4. Not suitable for studying rare diseases, except very large samples used.

1. and 2. of less concern if outcome is available electronically (e.g. electronic registers)

Case-control Study (Retrospective Study)

Start by identifying a group of people with the disease (cases) and a suitable reference group of people without the disease (controls).

Cases and controls compared for prevalence of the risk factor

Controls may be matched to the cases on certain important variables (e.g. age, sex): called a matched design (more in Lecture 2.2)

Advantage of Case-control studies

1. Suited to study of ***rare diseases***

2. Requires comparatively ***few subjects***

3. **Existing records** can be used

2. and **3.** ⇒ ***Speed:*** no waiting for outcome

Economy especially for rare diseases

4. Allows study of **multiple potential causes** of disease

Difficulties with Case-control studies

1. Cannot calculate **incidence** (at least not directly!)

2. Choosing controls

3. Availability of previously recorded data

4. ***Bias***

   especially recall bias (e.g. cases may “remember” more adverse exposures)

5. Maybe difficult to establish **sequence of events**

6. Unsuitable for study of **rare exposure**

Electronic data and cohort studies

A “real-life” longitudinal study identifies individuals now,
and follows them into the future.

Such studies can take a long time to gather enough data to address the research question.

Availability of continuously recorded data from the past, enables us to identify individuals “back then” and “*follow*” them in time

Hence we can study changes and trends over time
without waiting for data to accumulate!

Studies that use previously-recorded data in this way are called
“Retrospective Longitudinal” Studies

Electronic data and case-control studies

For case-control studies, electronic registers give us a convenient way of identifying cases of a disease

and perhaps also appropriate controls

(depending on the registers available)

Which design is best?

A cohort/incidence study often considered the gold standard

…. Sometimes easier to define research question in case-control setting

…. it helps to ask yourself “what would I do if I had the whole cohort”?

Traditional “pyramid of evidence” too simplistic

Tugwell P, Knottnerus JA. Is the 'Evidence-Pyramid' now dead? J Clin Epi. 2015
Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. EBM. Aug 2016.

Exercise 1.1

Description of your own work in terms of design and sampling scheme