Statistical learning in epidemiology

Lecture III

Giulia Zamagni

Clinical Epidemiology and Public Health Research Unit, IRCCS «Burlo Garofolo»

University of Trieste

The study design

- Determines the **validity of the results**
- Helps distinguishing between **correlation and causation**
- Guides **data collection and** the subsequent **analysis**

“If we want to know whether smoking causes lung cancer, how we design the study changes the answer?”

families:

Observational studies:

the researcher does not intervene

descriptive → *«what's happening?»*

analytical → *«why is this happening?»*

Descriptive studies

- Cross-sectional studies
- E.g.: Prevalence studies: *«how many cases of diabetes did we have in Italy in 2025?»*

Analytical studies

- Case-control studies, cohort studies
- Compare groups, look for associations
- E.g.: «is smoking associated with lung cancer?»

Experimental studies:

- the researcher intervenes directly
- He/she assigns the exposure or controls the study conditions
- *«what happens if I do something?»*

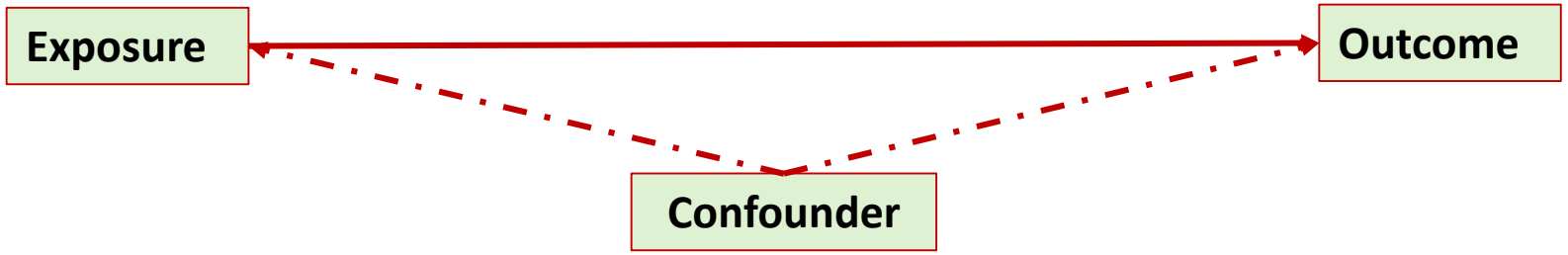
Randomized controlled trials (RCTs)

- the Gold Standard for assessing **causality**
- participants are randomly assigned to intervention and control group
- blind or double blind

Confounding

When the goal is to estimate a causal effect, confounding is a key challenge

What is a confounding factor?



A confounder is associated <u>both</u> with the exposure and the outcome

People drinking coffee have an increased risk of developing lung cancer

Drinking coffee is linked to smoking, as people who consume more coffee tend to smoke more.

Smoke is associated with lung cancer→ **Smoke is a confounder!**

Sampling

Confounding is an **analytical problem** that can create false associations if data are not analyzed appropriately.

In addition, **sampling** introduces another potential bias: **individuals included in a study must be representative of the target population.**

Problem	**Origin**	**Effect**
Counfounding	Variables	False association
Sampling	Participants	Non-representative sample

- *Analyzing data correctly is not sufficient*: we must also carefully control *who* is included in the study.
- An improper sampling approach can *introduce* or *amplify* existing confounding.

Sampling

Random sampling

Every individual in the population has the same probability of being selected

- Selection based on **chance**
- Minimized selection bias
- Improves representativeness

- **Simple**
- **Stratified**: population is divided into subgroups and random samples are taken from each group to ensure proper *representation*.
- **Cluster**: population divided into groups (clusters), and a random selection of entire clusters is included in the study.
- **Systematic**: individuals are selected at regular intervals from an ordered list after a random starting point.

Non-random sampling

- Selection based on non-random criteria

- Higher risk of bias
- Higher risk of non-representative sample

- **Convenience sampling:** people easiest to reach (students who are present in class in a given day)
- **Volunteer sampling**: people that accept to participate

Population vs Sample

Population: the set of all subjects of interest.

Sample: the *subset* of the population in which we measure data.

Often, collecting data only on a subset of the population is ***easier*** and ***quicker***

If correctly designed, we can use a sample to **infer** or **predict** something about the population

Example: if the goal is to predict the population mean using data from the sample, we have to consider that **the accuracy of our *sample mean* relies upon how well our *sample* represents the *population*** at large.

After defining how to select a representative sample, the next key question is: **how large should this sample be to obtain reliable results?**

The (minimum) sample size

- A good study requires both: the **right participants *and* a sufficient number of them**
- The sample size can be estimated based on **two** main **approaches**:

Precision:

we want to estimate a parameter accurately (e.g., prevalence of a disease)

→ The goal is to have a precise estimate

→ higher sample, narrower confidence intervals

Effect size

we want to detect a difference or an association (e.g., whether a drug reduces BP levels by at least 5mmHg)

→ the goal is to detect an effect

Sample size based on precision

Calculated to ensure that an estimate is obtained with a desired level of **accuracy**

Focus: **estimation**, *not hypothesis testing*

We want the estimate to be close to the true population value

Precision is reflected by:

1. Confidence interval width
2. Margin of error

Key determinants:

- Desired precision (margin of error)
- Variability of the outcome
- Confidence level (e.g., 95%)

Typical relationships:

- **Higher precision → larger sample size**
- **More variability → larger sample size**

A small sample may give an estimate, but with a wide confidence interval, making it unreliable.

Sample size based on effect size

Calculated to ensure sufficient statistical power to detect a specified **difference** or **association** between groups.

Focus: hypothesis testing

We want to detect whether:

- a treatment works
- an exposure is associated with an outcome

Key determinants:

- Effect size (minimum meaningful difference)
- Variability of the data
- Power (commonly 80% or 90%)
- Significance level (α) (commonly 0.05)

Common pitfalls:

Too small samples → risk of missing a real effect

Overpowered studies → detect trivial, clinically irrelevant differences

The sample size must be large enough to:

1. **detect the real effect**
2. **avoid false negatives**

Type I and Type II errors

So far, we have seen how to choose the sample size based on precision and effect size.

But there is another key question: ***how confident do we want to be in our results?***

Sample size is a balance between two types of errors

Type I error (alpha)
Probability of rejecting the null when it is true
→ Probability of a False Positive

Type II error (beta)
Probability of failing to reject the null when it is false
→ Probability of a False Negative

- **Smaller alpha → Larger n**
- **Smaller beta → Larger n**