STATISTICAL LEARNING IN EPIDEMIOLOGY (An Introduction)

Outline

Block 1 • Epidemiology: intro & basic measures

Block 2 • Study designs

Block 3 • Regression models

Block 4 • Survival Analysis

gbarbati@units.it

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Members of the Biostatistics Unit:

Associate Professor, MED01

Lucio Torelli, Associate Professor, MED01

Ilaria Gandin, RTT MED01

Daniela Zugna, Ass prof @UNITO

Federico Turoldo, MD, Resident in Epidemiology

Paolo Dalena & **Giulia Zamagni**
PhD students ADSAI (BURLO)

Block 1.1

Slides, references, exercices... will be posted on the Moodle repository (https://moodle2.units.it)

Prerequisites:

- Descriptive statistics
- Random variables
- Sampling
- Confidence Intervals
- Hypothesis Testing
- Linear Regression Model

Statistics at Square One, 11th Edition [Chapters 1 to 8]
Michael J. Campbell, T. D. V. Swinscow

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute
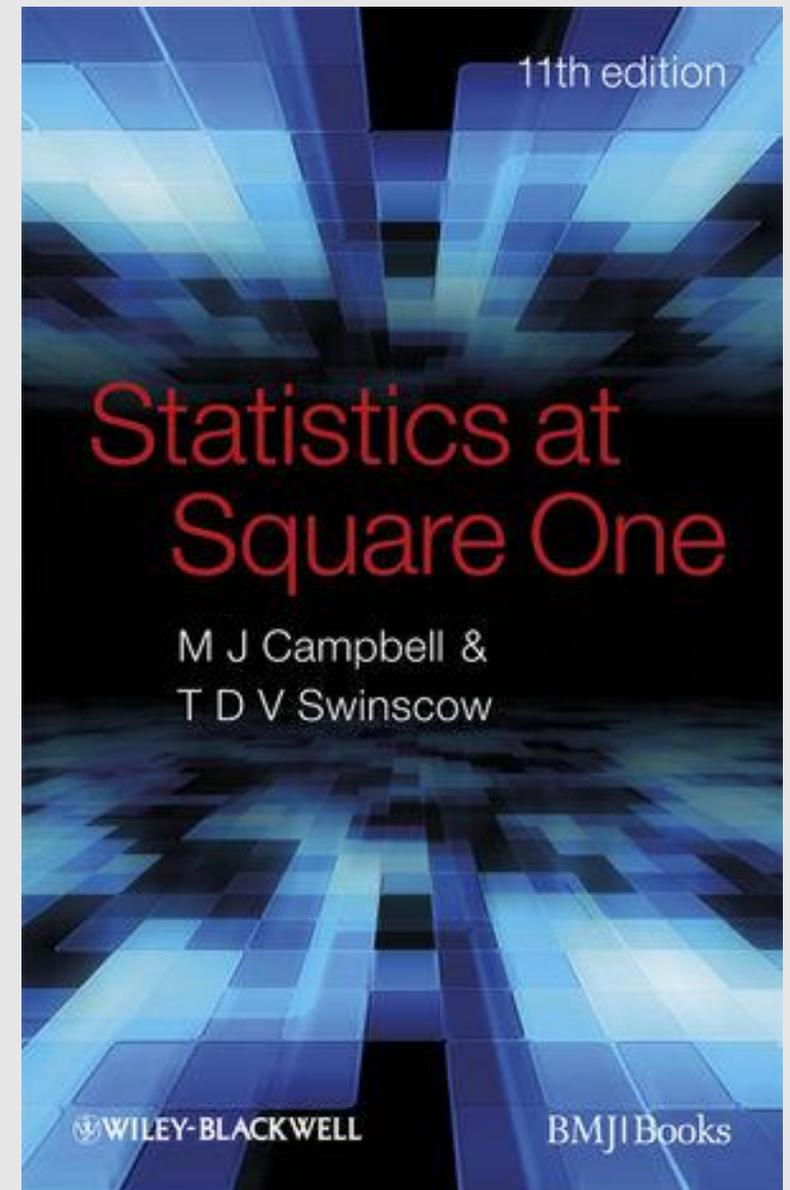
Bibliography of the course:
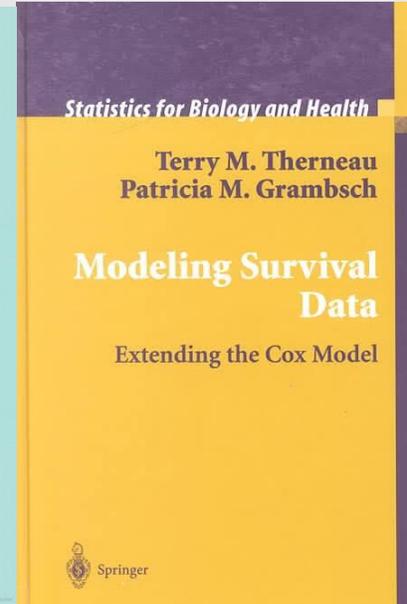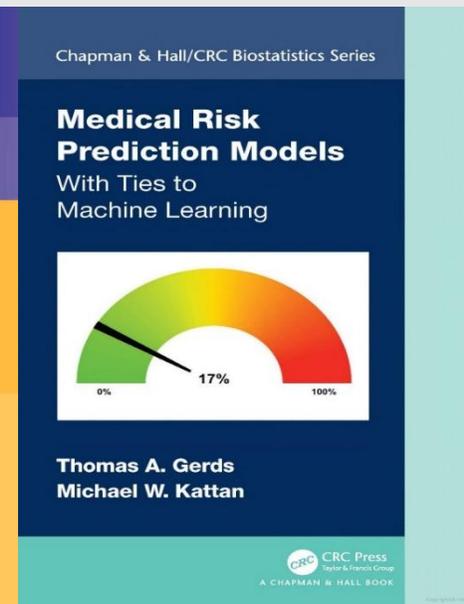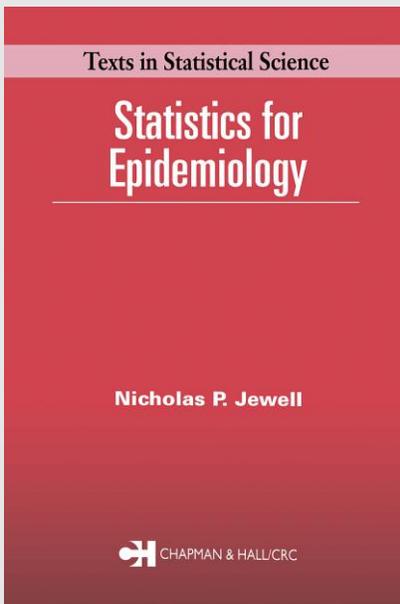
The central theme is an ***overview*** of statistical modelling of **health** data.

Focus more on ideas and principles rather than details of the statistical methodology.

In the first part (**blocks 1** and **2**) more emphasis will be given to **epidemiology concepts**, in the second part (**blocks 3** and **4**) more to **biostatistics methods**.
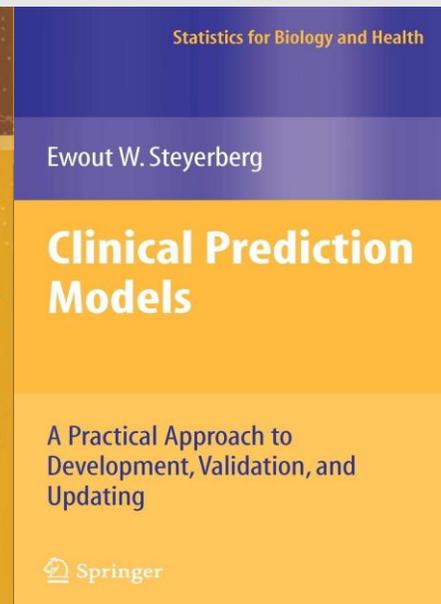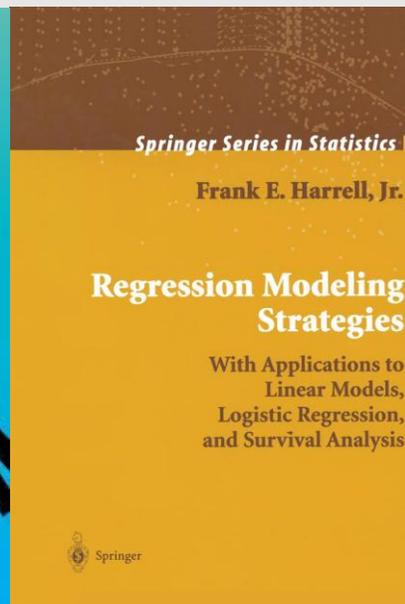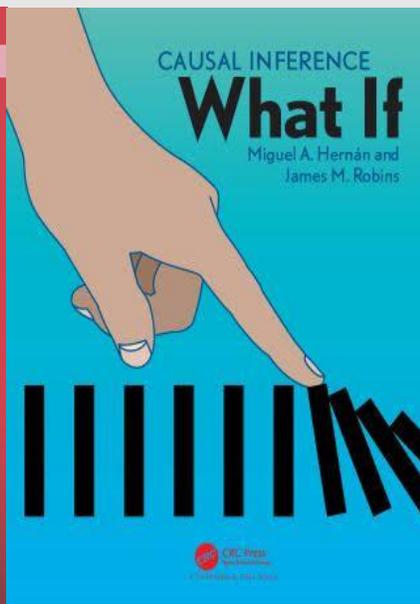
<u>Evaluation :</u>

- **<u>Project</u>**: dataset to be analyzed, presentation of the results. It should be done preferably in team (**2-4** students; **30** minutes of discussion max)

- **<u>Written exam</u>** (individual) : questions on the **theoretical** part of the course [quiz with open answers on Moodle]

- **Final mark** will be an **<u>average</u>** between project (team/individual) and (individual) written exam

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Project guidelines

Data and research questions

The structure of the data determines what you *can* do

Your research question determines what you *should* do!

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Project guidelines

1. Comprehension of the **scientific question** and of the **study design** that originated the data

2. Data **preprocessing**: *IDA* (initial data analysis [descriptive statistics])

3. *Model's* estimation to answer the scientific question

4. Report (**slides + code**) explaining **analyses** and **results**.

Please **finalize your teams as soon as possible** (no later than the beginning of **May**).

Once you are ready, I will send the dataset and the analysis questions to the **Team Captain**.

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Exam's dates:

Descrizione Appello	Data ora aula
Quarto Appello Sessione Estiva	15/09/2026 14:30 Aula Informatica 3A
Terzo Appello Sessione Estiva	21/07/2026 14:30 Aula Informatica 3A
Secondo Appello Sessione Estiva	07/07/2026 14:30 Aula Informatica 3A
Primo Appello Sessione Estiva	04/06/2026 14:30 Aula Informatica 3A

1 - Written exam (individual) : questions on the **theoretical** part of the course [quiz with **open answers** on Moodle] → 1hour (max)

2 - Project: dataset to be analyzed, presentation of the results. Preferably in team (**2-4** students; **30** minutes of discussion max)

📢 Let's Get Started: Understanding Your Background

Feedback   Impostazioni   Domande   Modelli   Risposte   Altro ⌄

Apertura: lunedì, 2 marzo 2026, 09:00

Some basic information about your background and skills

Introduction

- Epidemiology, Public Health, Clinical research

- Why a statistician (*data scientist*) is needed ?

1 Vai a **wooclap.com**

Codice evento
XKDMOE

2 Immettere il codice dell'evento nel banner superiore

Epidemiology

επί (epi)

δημος (démos)

Λόγος (logos)

Epi : upon, among

Demos : people

Ology : science, study of…

Epidemiology : the study of *what happens* to a population

It is the scientific method of *disease* investigation.

It involves the disciplines of medicine and **biostatistics**.

Hypothesis + data

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Formal definitions of *Epidemiology:*

1. The study of **distribution** and **determinants** of health, disease, or injury in human populations and the application of this study to the **control** of health problems.

2. The study of how the frequency of diseases varies in the **populations**, **places** and **times**.

3. The study of the relations between **diseases** and their potential **determinants**, controlling for the effects of *confounders* and *modifiers*.

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

The study of **distribution**

Outcomes of interest

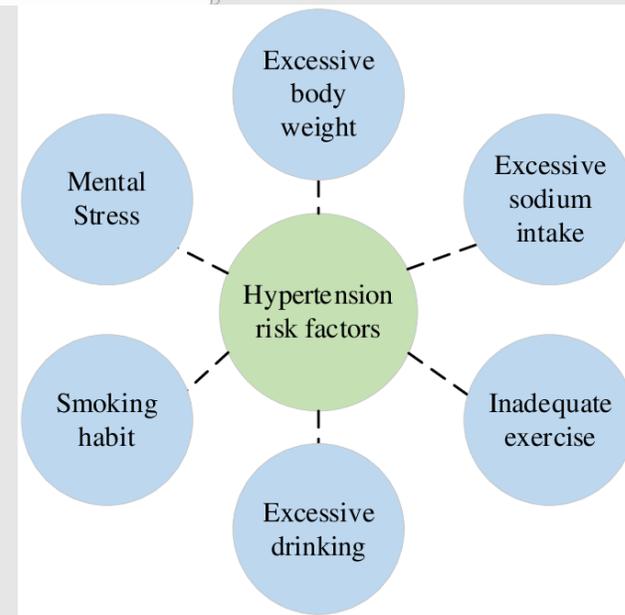**Mean** blood pressure, *prevalence* of hypertension, *incidence* of CHD (coronary artery disease....
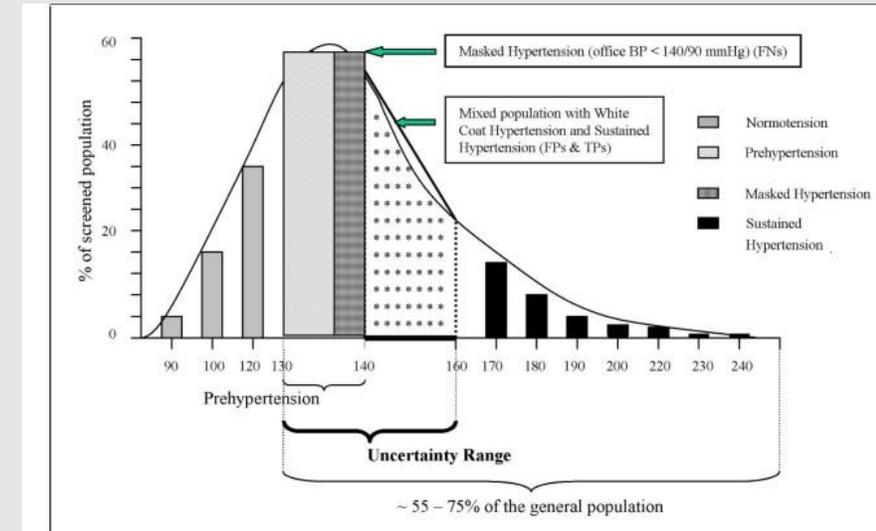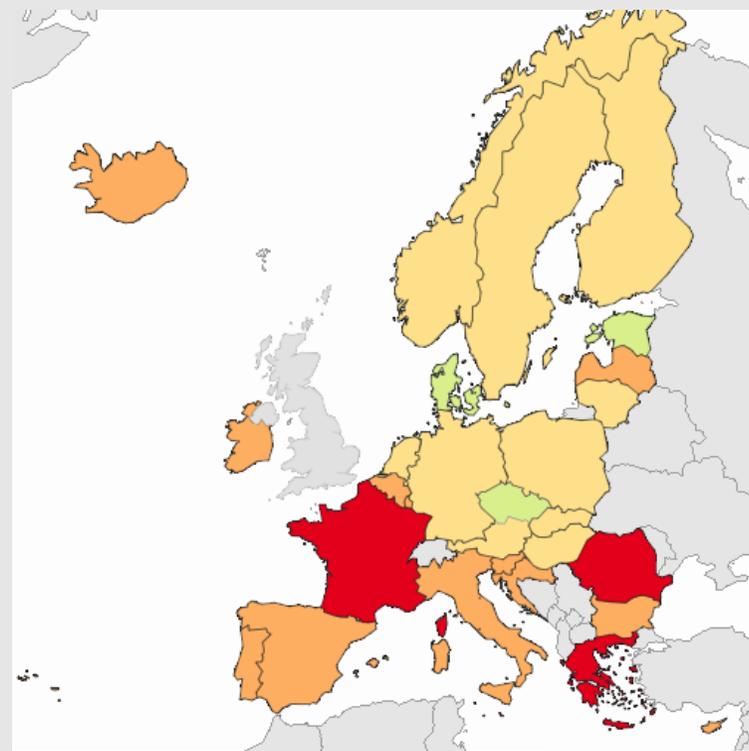
and **determinants** of health/disease

Associations between **Risk Factors** and **Outcomes** (causal ??)

[Relative Risk, Odds Ratio, Hazard Ratio, Regression Coefficients]

DATA SCIENCE & SCIENTIFIC COMPUTING

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Epidemiology & *Public Health*

So …what is epidemiology, and how does it contribute to the **health** of our society?

Consider the following statements:

- 10 years of hormone drugs therapy benefit *some women* with breast cancer

- Cellular telephone users who talk or text on the phone while driving *cause* 1 out of 4 car accidents

- Omega-3 pills, a popular alternative medicine, *may not help cure* depression

Epidemiology affects the way that individuals make **personal decisions** about their lives and the way that the government, public health agencies, and medical organizations make **policy decisions** that affect the way we live.

- It might prompt an oncologist to determine *which* of his patients would reap the benefits of hormone therapy

- It might prompt a road safety *campaign* *against the use of cellular telephone* while driving

- It might prompt a person to use a *traditional medication* for depression…

Epidemiology is the basis of **clinical** and **public health** actions

Public health is a multidisciplinary field whose goal is to promote the **health of the population** through organized community efforts.
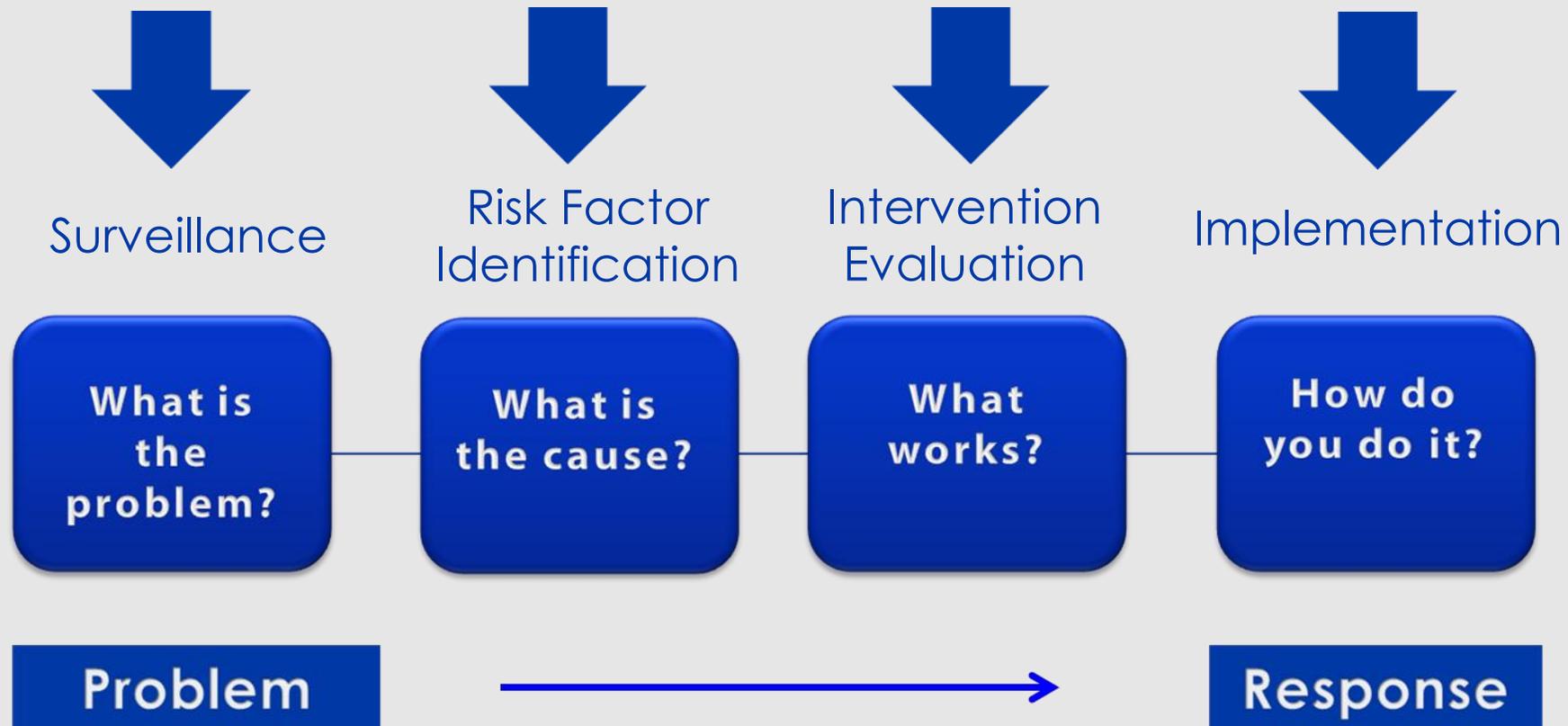
- **Public Health**: **monitor** health situation [**surveillance**].

- **Epidemiology**: **where** diseases originate, **how** or **why** move through populations, how we can **prevent** them *(communicable/non-communicable diseases)*.

- **Laboratories/Hospitals**: perform **tests** to confirm disease diagnoses. Drivers of **research** and training.

- **Informatics** + **Statistics**: collecting, compiling, interpreting & presenting **data**/study results.

- **Prevention/Action**: public health **policy/**clinical guidelines. Information for decision makers/doctors to help them choose the best option available.

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Different levels : Basic, Clinical, and Public Health

Characteristics	Basic	Clinical	Public Health
What/Who is studied	Cells, Tissues, Animals **Laboratory** Settings	Sick **patients** who come to **health care facilities**	**Populations** or community at large
Research goals	Understanding **disease mechanisms** and the effect of substances	Improving **diagnosis** and **treatment** of disease	**Prevention** of disease/Promotion of health/**Surveillance**
Examples	Toxicology, Immunology, Pharmacology…	Internal Medicine, Pediatrics, Cardiology…	Epidemiology, Environmental health sciences …
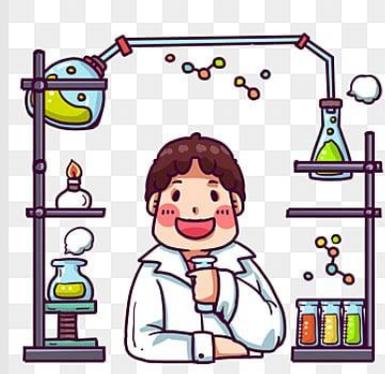
Epidemiology is … "*the study of the **distribution** and **determinants** of **disease frequency** in human **populations** and the application of this study to **control** health problems*"

- determine the **extent** of disease in a **population**
  [*prevalence*]

- identify **patterns** and **trends** in disease **occurrence**
  [*incidence*, *prognostic models*]

- identify the **causes/risk factors/exposures** related to the disease
  [*explanatory models*]

- study the **time course** of **disease** from onset to *resolution*
  [*survival analysis*]

- evaluate the **efficacy/effectiveness** of **measures** that prevent and treat disease
  [*cost-effectiveness models*]

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

POPULATION & DISEASES

Population will always refer to a group of people with a **common** characteristic, such as place of residence, gender, age, or use of certain medical services.

People who live in the city of Trieste are members of a geographically defined population.

Size of the population under study is the **denominator** for disease frequency measures (we refer often to *samples*..).

Disease **frequency** refers to quantifying **how often** a disease occurs in a population.

Estimation of disease frequency includes three steps:

(1) developing a **definition** of the disease

(2) instituting a mechanism for **counting** cases of the disease within a specified population

(3) determining the **size** of that population

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Disease **distribution**: patterns according to the characteristics of person, place, and time; **who** gets the disease, **where** it occurs, and **how** it changes over time.

Disease **determinants** : factors that bring change in a person's health or make difference in a person's health.

Individual determinants: a person's genetic makeup, gender, age, immunity level, diet, behavior, and existing diseases….
The risk of breast cancer is increased among women who carry specific genetic alterations, such as BRCA1 and BRCA2

Environmental/societal determinants: natural, social, and economic events and conditions.
Presence of infectious agents, poor and crowded housing conditions,…

Disease **control**: surveillance or active public health actions.
For every case of HIV, data are collected on the individual's demographic characteristics, transmission category, and diagnosis date.

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Epidemiology & **(bio) statistics [*data science*]**

Epidemiologists use **quantitative** methods.

Information (**data**) is collected to investigate a question.

The **methods** and **tools** of **biostatistics** are then used to **analyze** the data to aid decision making.

Biostatistics: statistical methods applied to the collection, analysis, and interpretation of *bio*logical *data,* especially data relating to human biology, health, and medicine.

The **goal** of biostatistics is to make valid **inferences** to answer questions in health research.

- Designing and conducting **experiments/observational studies** related to health problems

- **Collecting** and **analyzing** data to improve public health programs, answer to medical questions

- **Interpreting** the results of their findings

Why do you need a biostatistician?

Antonia Zapf[1*] [iD], Geraldine Rauch[2] and Meinhard Kieser[3]

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Roadmap to work with health data

1. Start from a public health/clinical **research question**:
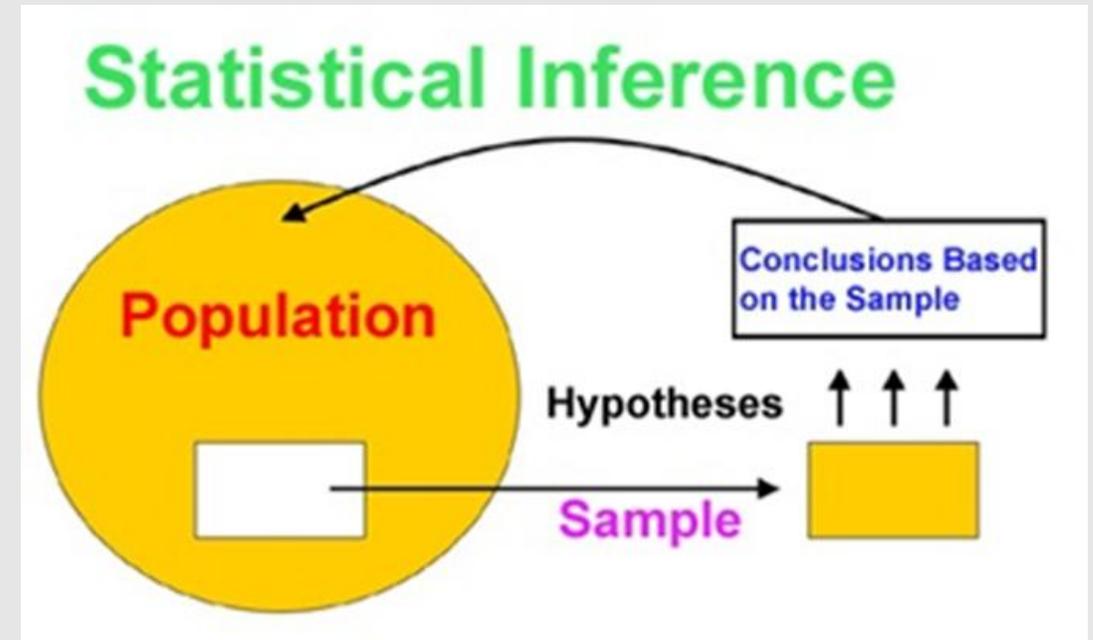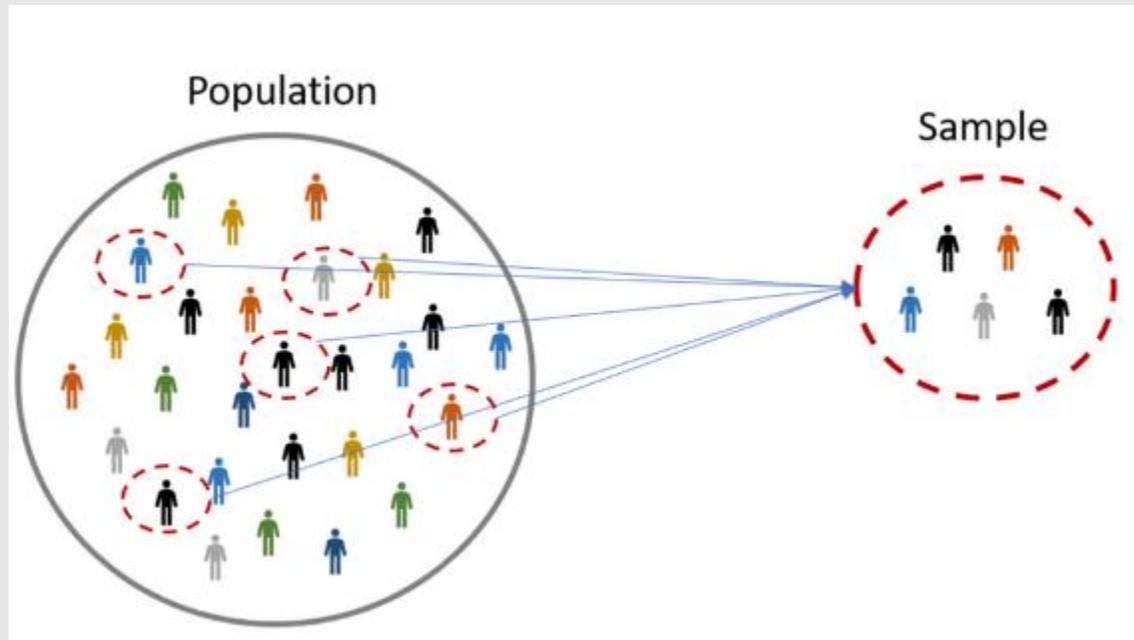
    Initial hypothesis

- The **risk** of **developing** lung cancer **remains constant** in the last five years in the U.S.

- The use of a cell phone **is associated with** developing brain tumor

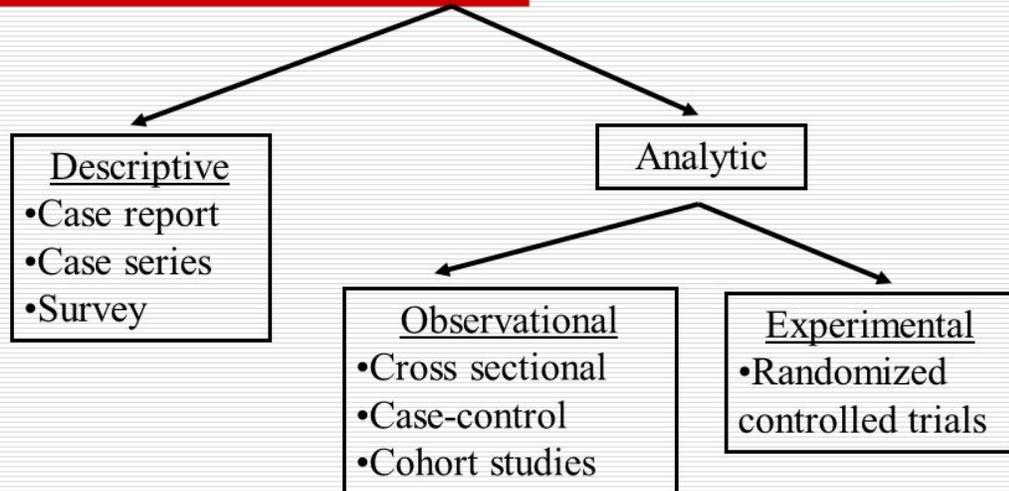- Vioxx (antinflammatory drug) **increases** the risk of heart disease

2. <u>**Study Design (I)**</u>:

rarely is individual information on disease status and possible risk factors available for an **entire** population.
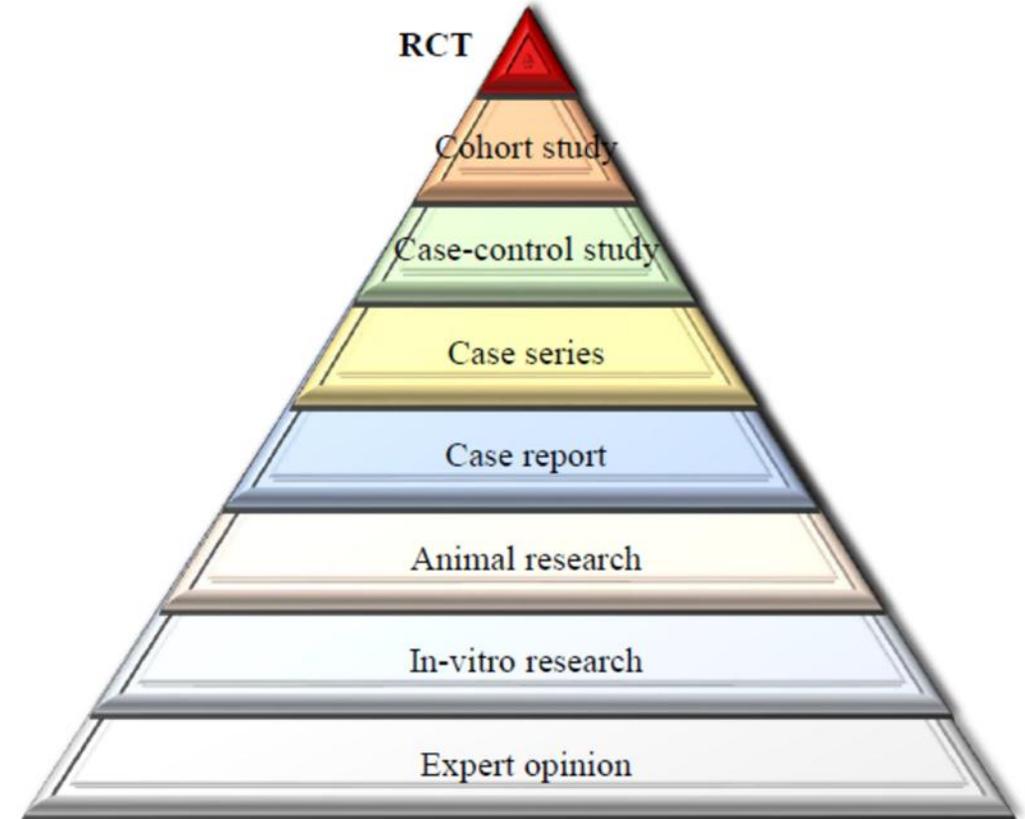
Study Design, Block 2!

2. Study design (II)

Hierarchy of Study Types

Descriptive
- Case report
- Case series
- Survey

Analytic

Observational
- Cross sectional
- Case-control
- Cohort studies

Experimental
- Randomized controlled trials

Strength of evidence for causality between a risk factor and outcome

RCT
Cohort study
Case-control study
Case series
Case report
Animal research
In-vitro research
Expert opinion

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute
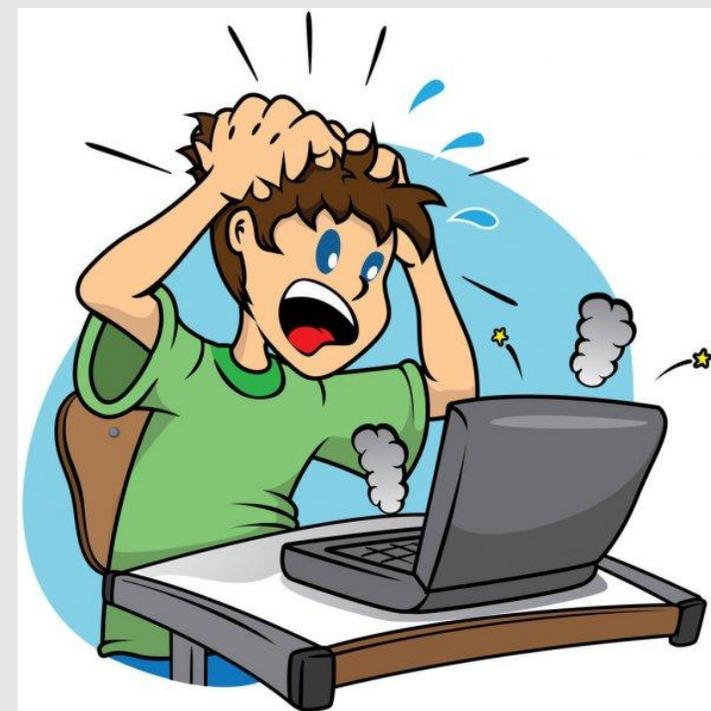
3. <u>Collect data:</u>

Define **what** to collect (and how)

- **#** of lung cancer cases from 2010 to 2022 in the United States
- **#** of people with heart attacks in a **sample** of individuals having used Vioxx and in a **sample** of non-users
- determinants…

4. <u>Describe data</u>

- Organization and summarization of data
- Tables
- Graphs
- Data cleaning
- Summary measures
- Missing data evaluations…

5. Assess the <u>**strength of evidence**</u> for/against a hypothesis**/estimate a risk**

Inferential/causal inference/predictive/prognostic methods:
- estimates from a **sample** to the whole group (**target** population)
- make **comparisons** between groups
- make **predictions**
- assess the **impact** of specific **exposures/treatments** on outcomes
- ask more questions… suggest future research

6. **Actions:** epidemiologists at the end recommend <u>interventions</u> or <u>preventive</u> programs:

- study results will prove or disprove the hypothesis, or sometimes fall into a grey area of "unsure"
- study results appear in a publication and/or are disseminated to the public by other means

The **policy** or **action** can range from developing specific *regulatory programs* to general personal behavioral changes, to modify treatments for specific diseases….

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Statistical approaches to health data
[*in this course…*]

With many diseases, we are still unable to *accurately* quantify the **amount** of disease beyond its *presence* or *absence*.

That is: we are often limited to a simple **binary indicator** of whether an individual is diseased or not.

For this reason, we will focus *mainly* on statistical techniques designed for a **binary outcome** variable.
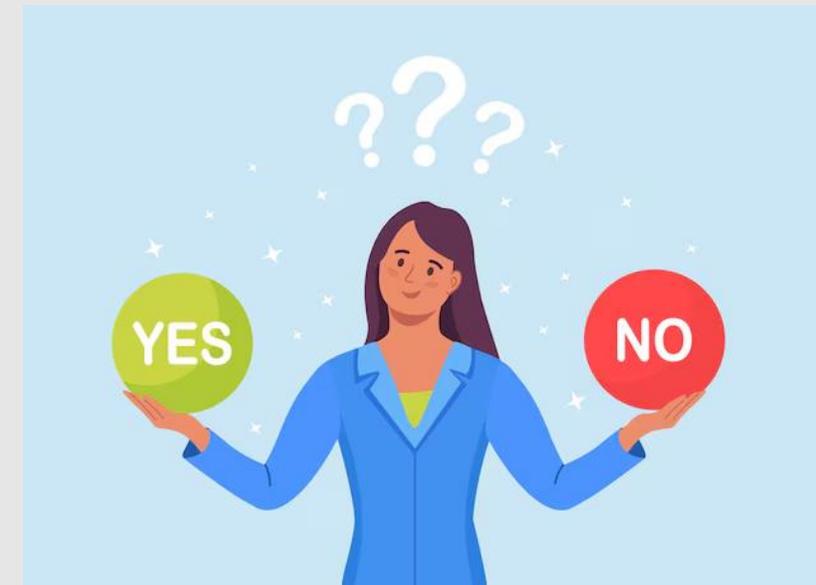
Statistical approaches to health data
[*in this course…*]

Risk factors/predictors/features (explanatory variables/exposures…) come in all possible forms, from binary (sex), to unordered discrete (ethnicity), to ordered discrete (coffee consumption in cups per day), to continuous (infant birthweight)*…

We will assume *mostly* that risk factors have a **fixed value**** and therefore do not vary over time…

*We will refer to **structured** data, even if active research is also on **unstructured** health information

Methods to accommodate exposures that **change over time, in the context of longitudinal studies, provide attractive extensions

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Different scientific aims, *different* methods

Descriptive modelling

Predictive modelling

Explanatory modelling

Describe the outcome of interest:
which factors *are associated with it* and how?

Estimate *prevalence of* CVD according to age and sex

Accurate **predictions** of future observations.
No concern about causality and confounding (association)

Estimate **risk** of developing CVD in the next x years

Testing and comparing existing **causal** theories.

Effect of LDL on CVD risk

Policy recommendations (statins) **to lower** CVD risk

Statistical approaches to health data
[*in this course…*]

Part of the course will be devoted to discuss statistical models used for ***explanatory*** purposes.

This is because health research, at public health or clinical level, mostly focus on the ***etiology*** of diseases

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330
© Institute of Mathematical Statistics, 2010

To Explain or to Predict?

Galit Shmueli

Explanatory models are focused on quantifying the ***causal*** effects of some *[pre-selected]* ***predictors*** of interest in ***causing*** a disease or its progression.

This does not mean that we are not interested in **predictions**/*prognosis* but this often is viewed more *as a consequence* of a (possibly good) explanatory model.

For these reasons, we will focus mostly on **classical** statistical tools

BUT…

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Block 1.1

IEA
International Epidemiological Association

Opinion

Intersections of machine learning and epidemiological methods for health services research

Sherri Rose

Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA, 02115, USA. E-mail: rose@hcp.med.harvard.edu

IEA
International Epidemiological Association

Education Corner

Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference

Tony Blakely,[1]* John Lynch,[2] Koen Simons,[1] Rebecca Bentley[1] and Sherri Rose[3]

CORSI ⌄   CONVEGNI ⌄   OPPORTUNITÀ ⌄   ACCREDITAMENTO ⌄   COLLEGIO ⌄   RIVISTA ⌄

EDITORIAL   Epidemiology Biostatistics and Public Health - 2019, Volume 16, Number 4

Machine learning in clinical and epidemiological research: isn't it time for biostatisticians to work on it?

https://www.sismec.info/

Machine Learning in Clinical Research Group [1]

Machine Learning nella ricerca clinica

f 🐦 📱

Coordinatori: Paola Berchialla (Università di Torino), Ileana Baldi (Università di Padova)
Gruppo di lavoro: Danila Azzolina (Università di Novara), Giulia Barbati (Università di Trieste), Daniele Bottigliengo (Università di Padova), Pasquale Dolce (Università di Napoli), Ilaria Gandin (Area Science Park, Trieste), Caterina Gregorio (Università di Padova), Dario Gregori (Università di Padova), Francesca Ieva (Politecnico di Milano), Corrado Lanera (Università di Padova), Giulia Lorenzoni (Università di Padova), Michele Marchioni (Università di Chieti), Alberto Milanese (Università La Sapienza), Andrea Ricotti (Università di Torino), Veronica Sciannameo (Università di Torino).

Obiettivi: (i) approfondire l'utilizzo delle tecniche di **Machine Learning (ML)** evidenziando i punti di **contatto e di integrazione** con le tecniche classiche di modellizzazione; (ii) dare ampia **diffusione alle conoscenze** alla base delle **tecniche ML** per rendere l'approccio all'analisi basato su tali strumenti più facilmente comprensibile e accessibile; (iii) promuovere l'utilizzo di strumenti appropriati che rendano **interpretabili** i modelli basati sul ML; (iv) censire le **risorse open source disponibili** (come software e modelli pre-addestrati che possono essere utilizzati).

Scheda di presentazione approfondita: leggi QUI

Interessato ad unirti al gruppo di lavoro? Scrivi a paola.berchialla@unito.it