STATISTICAL LEARNING IN EPIDEMIOLOGY (An Introduction)

gbarbati@units.it

Outline

Block 1
- Epidemiology: introduction & basic measures

Block 2
- Study designs

Block 3
- Regression models

Block 4
- Survival Analysis

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

Block 1.1

Composition of the Biostatistics Unit:

https://dsm.units.it/

Associate Professor, MED01

Lucio Torelli, Associate Professor, MED01

Ilaria Gandin, RTD-A MED01

Daniela Zugna, Associate Professor, MED01 **UNITO**

Giovanni Baj, PhD student

Paolo Dalena, PhD student & biostat (**BURLO**)

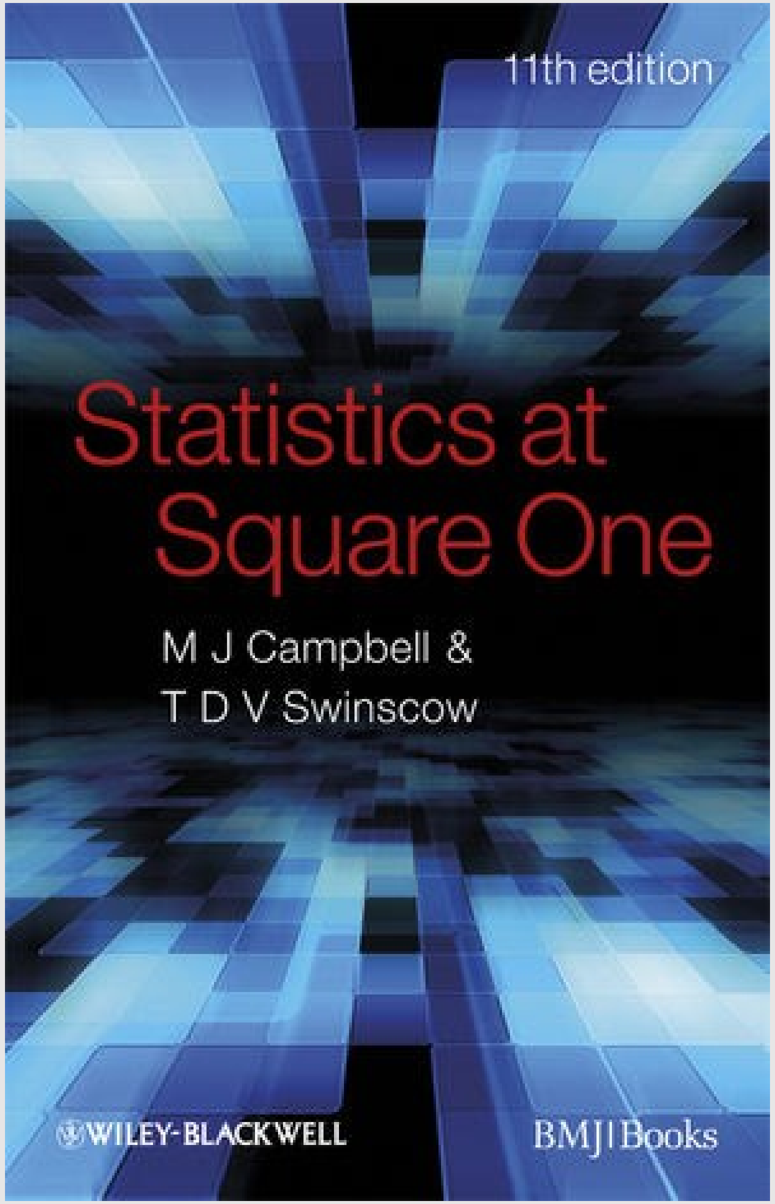
NITÀ DI BIOSTATISTICA
ipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

STATISTICAL LEARNING IN EPIDEMIOLOGY

Prerequisites:

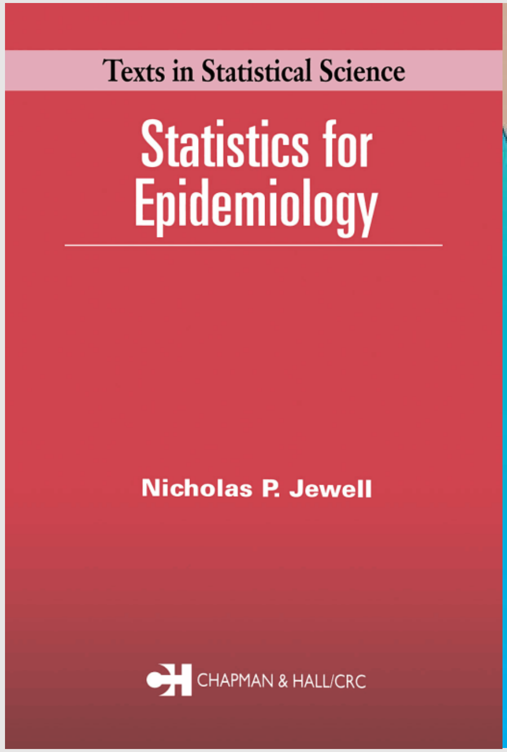
- Descriptive statistics
- Random variables
- Sampling
- Population parameters and estimation [*mean and proportion*]
- Confidence intervals and hypothesis testing [*Type I / Type II errors, p-values, power…*]
- Standard hypothesis tests for means and proportions [*t-tests, Chi-square test, non parametric tests…*]

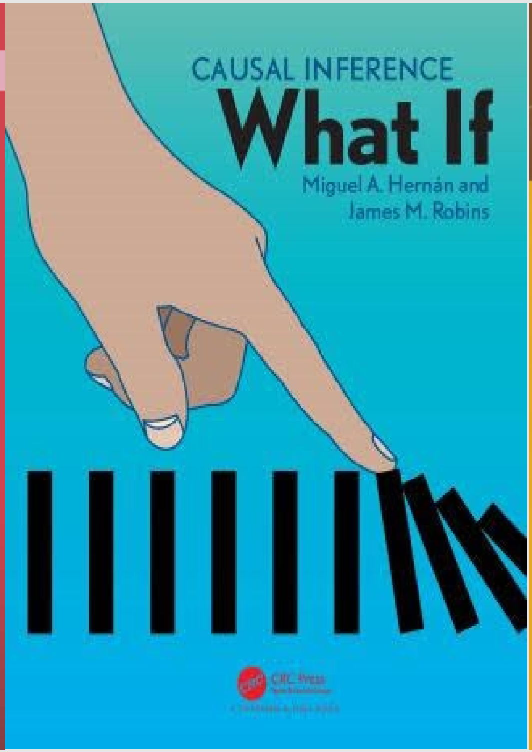
Statistics at Square One, 11th Edition [Chapters 1 to 8]
Michael J. Campbell, T. D. V. Swinscow

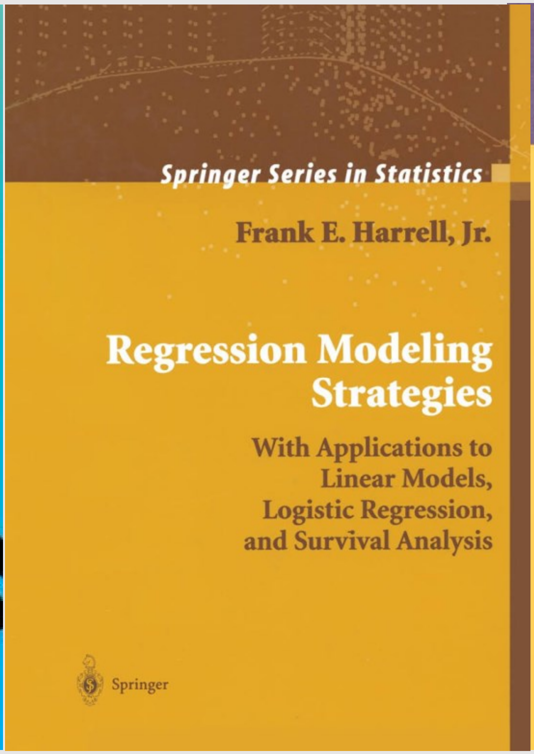


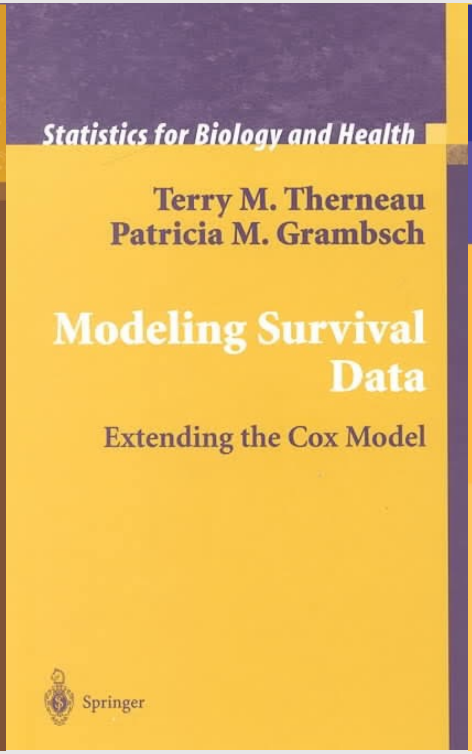
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

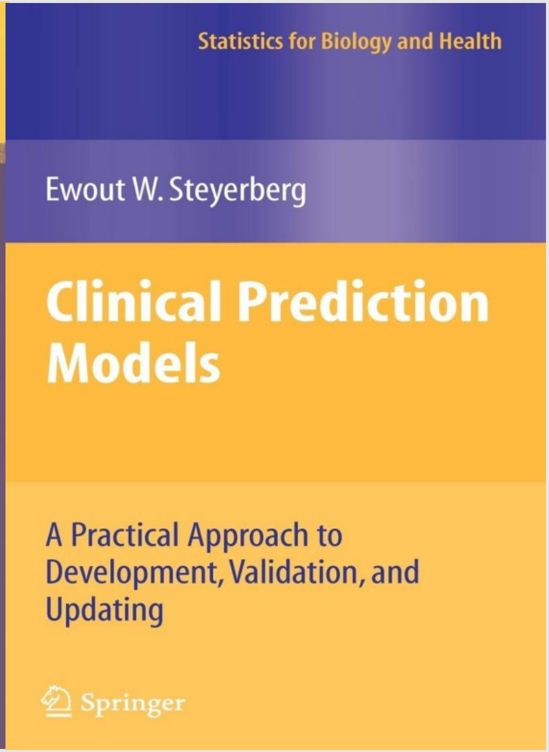
Suggested Books:











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

The central theme of the course will be an ***overview*** of statistical modelling of health data, yet the focus will be more on ideas and principles rather than details of the statistical methodology.

Methodological details will be at a minimum, more time will be devoted to discuss **examples**.

Slides, references, codes… + any announcement will be posted on the Moodle repository (https://moodle2.units.it) and in Teams class.

In the first part (**blocks 1** and **2**) more emphasis will be given to epidemiology, in the second part (**blocks 3** and **4**) more to biostatistics.

Evaluation :

- **Project**: dataset to be analyzed, presentation of the results. You can choose the statistical software you prefer (R, Python…); it could be done individually or in team (**max 3** students; **20-30** minutes)

- **Oral questions** at the end of the presentation (individual…)

- **Final mark** will be an **average** between project (team/individual) and (individual) answers

Project guidelines

1. **Choose a dataset** suggested websites are the following:

- https://cran.r-project.org/web/packages/medicaldata/index.html
- https://www.kaggle.com/datasets [search with some **keywords** as «health»…]
- https://hbiostat.org/data/
- https://archive-beta.ics.uci.edu/datasets
- https://cran.r-project.org/web/packages/NHSRdatasets/index.html
- https://www.causeweb.org/tshs/category/dataset/
- https://datarepository.stat.unipd.it/
- https://www.causeweb.org/tshs/category/dataset/
- https://aimidatasetindex.stanford.edu/
- ……

See dates in ESSE 3 for the exams

Project guidelines

2. Identification of the **scientific question** and (possibly) of the **study design** that originated the data (blocks 1-2)

3. Data **preprocessing**: **IDA** (initial data analysis / univariable analyses)

4. *Model's* estimation to answer the scientific question [blocks 3-4]

5. Report (**R markdown** or similar) explaining step by step **analyses** and **results**.

End of the course (3 June) each student/team should prepare a **5 minute** oral presentation in which:

- the selected **dataset** and **scientific question** are briefly presented
- **goals** and **roadmap** of the project should be *approximately* defined…

Introduction

Epidemiology & Public Health/Clinical research

Statistical approaches to epi/clinical data

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

Epidemiology

επί (epi) → **Epi** : upon, among

δημος (démos) → **Demos** : people

Λόγος (logos) → **Ology** : science, study of…

Epidemiology : the science or the study of *epidemic (diseases)*

It is the scientific method of disease investigation.

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

Hypothesis + data

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**… :

- Measures **outcomes** (usually presence/absence of diseases)

- Example: **mean** blood pressure, **prevalence** of hypertension, **incidence** of CHD (coronary artery disease), **survival** probability, **cumulative incidence** of some events ….

[Block 1]

…and **determinants** of health, disease, or injury in human populations :

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

- Relative Risk, Odds Ratio, Hazard Ratio, Regression Coefficients…

[Blocks 1-3]

Epidemiology & Public Health

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

Most people don't know the answer to this question. This is somewhat paradoxical because **epidemiology**, one of the basic sciences of **public health**, affects nearly everyone.

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

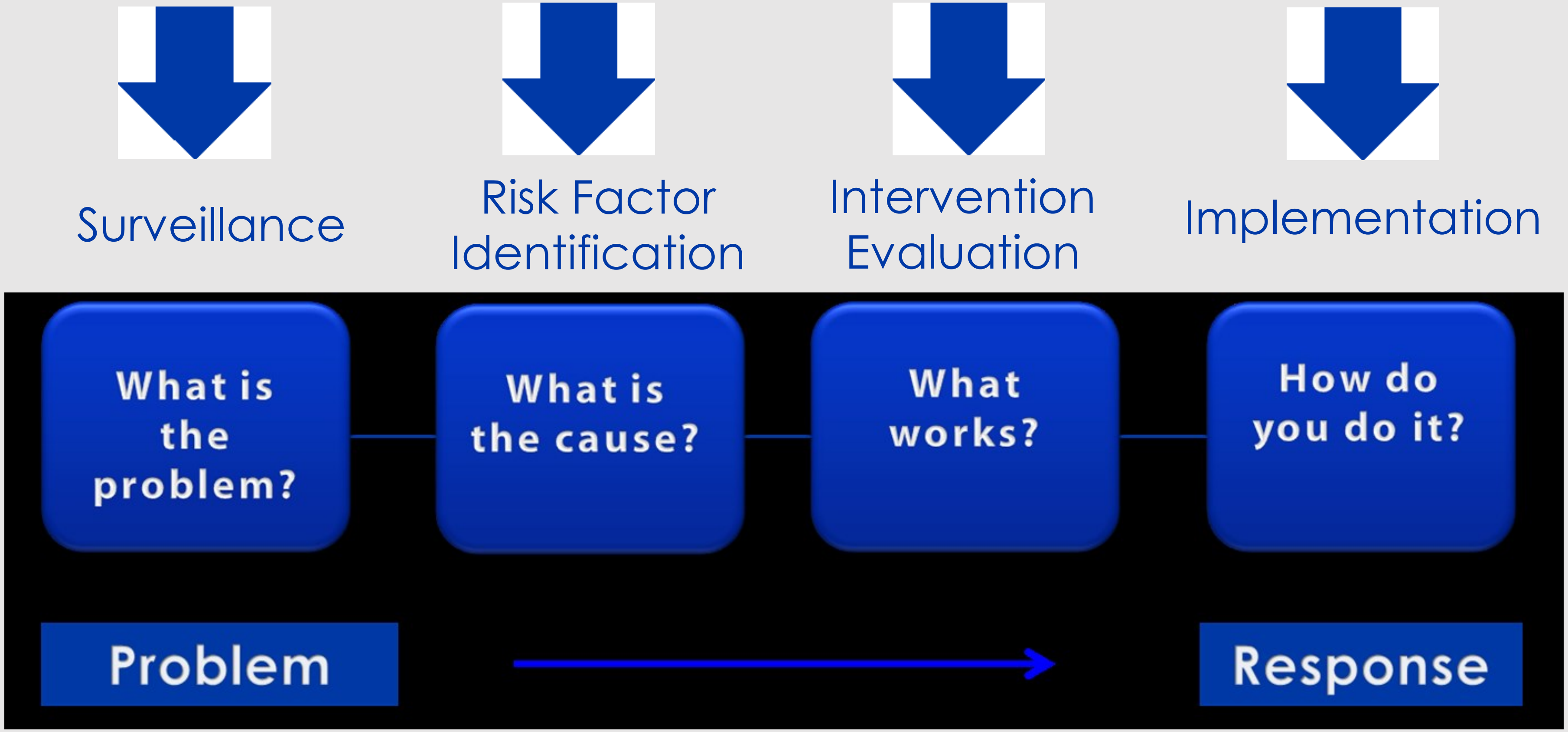
Epidemiology directly affects the daily lives of most people!

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 breast cancer 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 **Public Health** actions (or should be..)

Public health is a multidisciplinary field whose goal is to promote the health of the population through organized community efforts.



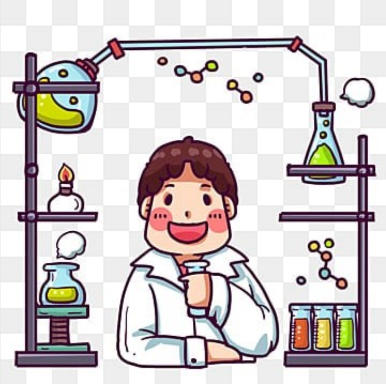
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

- **Public Health** Surveillance: **monitor** a public 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/Effectiveness**: 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

Basic, Clinical, and Public Health Science Research:

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****, block 1*]
- identify **patterns** and **trends** in disease **occurrence**
  [***incidence****/regression models for prediction, blocks1-3*]
- identify the **causes/risk factors/exposures** related to the disease
  [*regression modelling in* ***causal inference*** *framework, blocks 3 and 4*]
- study the **time course** of **disease** from onset to *resolution*
  [***survival*** *analysis, block 4*]
- evaluate the **efficacy/effectiveness** of **measures** that prevent and treat disease
  [causal inference tools, blocks 2-3-4]

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.

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

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.

Epidemiology and **biostatistics** are the basic sciences of public health.

Public health investigations use **quantitative** methods, which combine the two disciplines.

Epidemiology is about the **understanding** of disease development and the methods used to uncover the etiology, progression, and **treatment** of the disease.

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 *biological data* and especially data relating to human biology, health, and medicine.

The **goal** of biostatistics is to make valid **inferences** that can be used to solve problems in public health (turning data into knowledge).

- 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

DEBATE

Zapf *et al. BMC Medical Research Methodology* (2020) 20:23
https://doi.org/10.1186/s12874-020-0916-4

Why do you need a biostatistician?

Antonia Zapf[1*], 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 [scientific rationale, observations or anecdotal evidence]

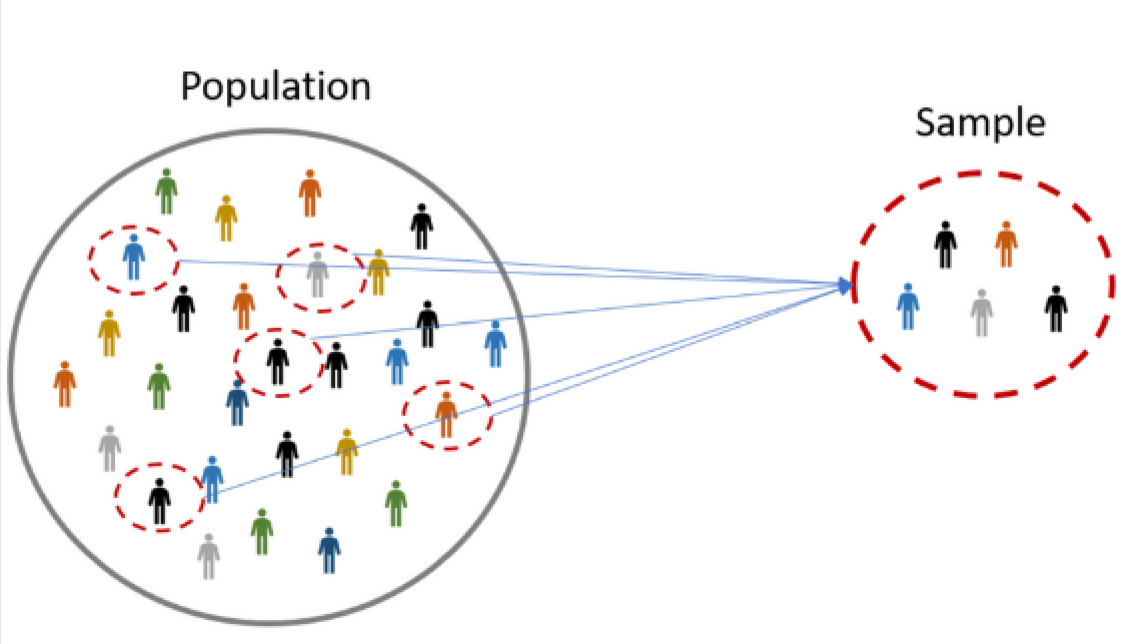
- The risk of developing lung cancer **remains constant** in the last five years in the U.S.
- The use of a cell phone **is associated with** developing brain tumor
- Vioxx (antinflammatory drug) **increases** the risk of heart disease

2. **Study Design (I)**: rarely is individual information on disease status and possible risk factors available for an **entire** population.

We work with some **fraction** of our population of interest, and we use **statistical tools** to select individuals (**sampling**) and to analyze data collected through a particular **study design**.

2.1 We wish to use **sample data** to most effectively make applicable statements about the larger population from which a sample is drawn (**inference**)

2.2 Since accurate data collection is often expensive and time-consuming, we want to ensure that we make **the best use** of available resources.



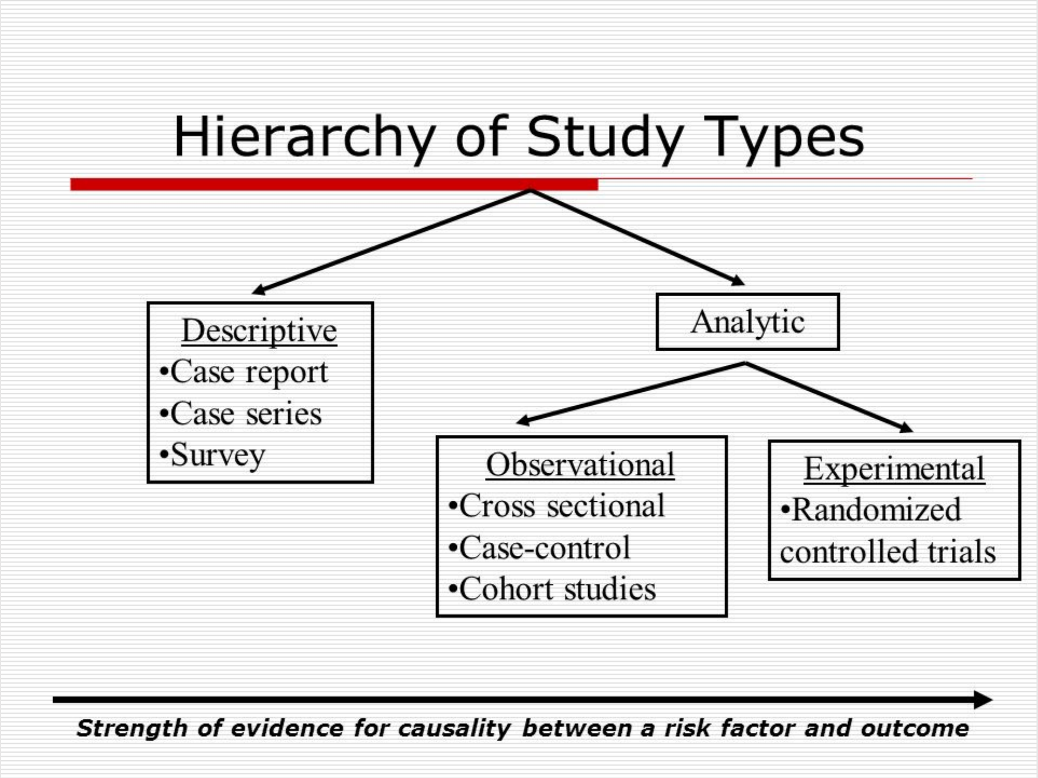
Study Design, Block 2!

2. **Study design (II)**

- *Survey/Cross-sectional [descriptive]*
  Estimate the **extent** of the disease in the population

- *Observational* [*analytical*]
  Association [causal?] between an exposure and a disease. ***Natural*** allocation of individuals to *exposed* or *non-exposed* groups

- *Experimental/Randomized Controlled Trials*
  Causal relationship between an exposure, often therapeutic treatment, and disease. Individuals are ***intentionally*** (but **randomly**…) placed into the treatment groups by the investigators.



3. <u>Collect data:</u>

Define **what** to collect. Numerical facts, measurements, or observations obtained from an investigation to answer a question.

- # 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

4. <u>Describe data</u> : exploratory assessment of the data from a study (initial data analysis, IDA, techniques)

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

5. Assess the **strength of evidence** 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 predictors on outcomes
- ask more questions… suggest future research

6. **Actions:** epidemiologists at the end recommend interventions or preventive 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….

Statistical approaches to health data

[in this course…]

In studying the relationship between two (or more…) variables, it is most effective to have *refined measures* of both the **explanatory** and the **outcome** variables.

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.

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

Statistical approaches to health data

[in this course…]

On the other hand, **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

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 epidemiological or clinical level, mostly focus on the ***etiology*** of diseases

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

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 on **classical** statistical tools instead of *black-box-type (ML)* approaches.

BUT…

…the door is open for contributions from the machine learning/AI community!!

Block 1.1

International Journal of Epidemiology, 2020, 1763–1770
doi: 10.1093/ije/dyaa035
Advance Access Publication Date: 1 April 2020
Opinion

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

International Journal of Epidemiology, 2020, 2058–2064
doi: 10.1093/ije/dyz132
Advance Access Publication Date: 11 July 2019
Education Corner

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]

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?

Machine Learning in Clinical Research Group [1]

https://www.sismec.info/

CORSI CONVEGNI OPPORTUNITÀ ACCREDITAMENTO COLLEGIO RIVISTA

zazione
mento
nza
sione Didattica e Formazione
menti
sione Ricerca e Sviluppo
menti
sione Normativa &
a
di Lavoro
ec
Intersocietari

Machine Learning nella ricerca clinica

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 comment

Steps to avoid overuse and misuse of machine learning in clinical research

Machine learning algorithms are a powerful tool in healthcare, but sometimes perform no better than traditional statistical techniques. Steps should be taken to ensure that algorithms are not overused or misused, in order to provide genuine benefit for patients.

SPECIAL COMMUNICATION

A Clinician's Guide to Artificial Intelligence (AI): Why and How Primary Care Should Lead the Health Care AI Revolution

Steven Lin, MD

Open access Protocol

BMJ Open

Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence

npj | Digital Medicine www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review

Anne A. H. de Hond[1,2,3,8], Artuur M. Leeuwenberg[4,8], Lotty Hooft[4,5], Ilse M. J. Kant[1,2,3], Steven W. J. Nijman[4], Hendrikus J. A. van Os[2,6], Jiska J. Aardoom[6,7], Thomas P. A. Debray[4], Ewoud Schuit[4], Maarten van Smeden[4], Johannes B. Reitsma[4], Ewout W. Steyerberg[2,3], Niels H. Chavannes[6,7] and Karel G. M. Moons[4]

RESEARCH METHODS AND REPORTING

Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension

Xiaoxuan Liu,[1,2,3,4,5] Samantha Cruz Rivera,[5,6] David Moher,[7,8] Melanie J Calvert,[4,5,6,9,10,11] Alastair K Denniston,[1,2,4,5,6,12] On behalf of the SPIRIT-AI and CONSORT-AI Working Group

JAMA Network | Open

Original Investigation | Health Informatics

Randomized Clinical Trials of Machine Learning Interventions in Health Care A Systematic Review

Deborah Plana, BS; Dennis L. Shung, MD, PhD; Alyssa A. Grimshaw, MSLIS; Anurag Saraf, MD; Joseph J. Y. Sung, MBBS, PhD; Benjamin H. Kann, MD