Biostatistics & Epidemiological Data Analysis using R

4
Epidemiological study designs

Stefan Konigorski

Health Intervention Analytics Group, HPI

November 9, 2022

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Poll

Please go to

`https://pollev.com/skonigorski070`

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2022.10.19
	2	First steps in data analysis using R	2022.10.26
	3	Second steps in data analysis using R	2022.11.02
Epidemiology & Statistics: concepts	4	Epidemiological study designs	2022.11.09
	5	Estimation	2022.11.16
	6	Hypothesis testing and study planning	2022.11.23
	7	Missing data	2022.11.30
Data analysis w/ regression models	8	Linear regression I	2022.12.07
	9	Linear regression II	2022.12.14
	10	Regression models for binary and count data	2023.01.11
	11	Analysis of variance & Linear mixed models I	2023.01.18
	12	Linear mixed models II & Meta analysis	2023.01.25
	13	Survival analysis	2023.02.01
	14	Causal inference & Data analysis challenge	2023.02.08

(see full schedule online)

Overview

Review of classes 1-3

- Overview of data analysis steps.
- Get to know R and R Markdown.
- Do all steps of a data analysis (up until the main analysis with estimation and statistical testing).
- Questions?

Overview of classes 4-6

- Today: Why do we do this actually (from a population health perspective)?
- Class 5-6: Theory and practice of statistical estimation and hypothesis testing.

Learning objectives

- Get an overview and discuss important epidemiological concepts and terminology.
- Get an overview and discuss different epidemiological study designs.

Epidemiological concepts

Epidemiology

Overall Aim

- Study the distribution of health & disease on the population level, understand relevant factors (risk factors = exposures) that can be acted upon, in order to improve health/disease outcomes (public health).
- Find out which factors are relevant, for which disease, for whom; (understand disease etiology), build prediction models, derive risk groups.

Approach

- Study a sample from the population.
- Get insights about the population by using knowledge about the population, sampling, measurements, statistical models.
- Think causal!

Causal reasoning

Say we observe a statistical association between two variables X and Y in a study, e.g. inflammation of teeth (X), lung cancer (Y).

- How do you infer that X is a causal risk factor for Y?

It could be that

- X causes Y
- Y causes X
- X and Y have a common cause, e.g. smoking (confounding)
- X and Y both affect a third variable, which determined the sampling (selection bias, e.g. only include those that survived)
- the statistical test just gave a false positive alarm, in reality there isn't an association.

$\longrightarrow$ Visualize in a directed acyclic graph (DAG).

Leading questions

- How can you characterize the population you are studying?
- How are the people sampled?
- How can you assess the disease/outcome of interest?
- What factors affect the disease?
- How can you evaluate whether these factors really affect the disease (see previous slide)?

How can you characterize the population you are studying?

- Population = the group of individuals we want to investigate
- Examples: general population, children, healthcare workers, animal models, cell line
- Defined by e.g. time & place or characteristics of the individuals in the population
- Can be dynamic (change over time) or stationary (specific set of people that doesn't change)

- Examples?

How can you sample people from the population?

- Convenience sample
- (Representative) random sample
- Targeted sample, e.g. selected by exposure or disease

Which sampling should you (not) choose to:

- Determine if aspirin or ibuprofen works better to treat a certain kind of headache.
- Develop a new screening procedure to detect those at high risk of a tuberculosis infection.
- Get an estimate of the life-time prevalence of having a panic attack.
- Get an estimate about the 5-year survival rate after having brain surgery for brain cancer.

What do you study (i.e. which outcome)?

... and how do you measure it? $\longrightarrow$ Through a proportion.

Which proportion?

- How many people have the disease (prevalence) vs. how many people develop the disease (incidence)
- Prevalence: at a give time or in a given time frame

- Which population and sample should you study, respectively?
- What are advantages, disadvantages, challenges of these measures?

How can you study the people in your sample?

See study designs later.

How do you investigate if/which factors affect disease?

- Which setting would be ideal? $\longrightarrow$ e.g. counterfactual model
- Which real setting can you use? $\longrightarrow$ randomize, RCT, N-of-1 trial
- Which insights can you get from just observing people?
- What challenges are there?
- What methods can you use to consider these challenges? $\longrightarrow$ standardize, adjust, stratify, match

PICO

Formulate your study question e.g. using the "Participant-Intervention-Comparator-Outcome" (PICO) framework[1]:

- Participant: Who? $\longrightarrow$ population/sample
- Intervention: Effect of what? $\longrightarrow$ exposure, intervention
- Comparator: Compared to what? $\longrightarrow$ control group
- Outcome: Effect on what?

- Example?

$\longrightarrow$ Hypothesis-based vs. explorative analyses.

[1] Schardt et al. (2007). BMC Med Inform Decis Mak 7(1):16.

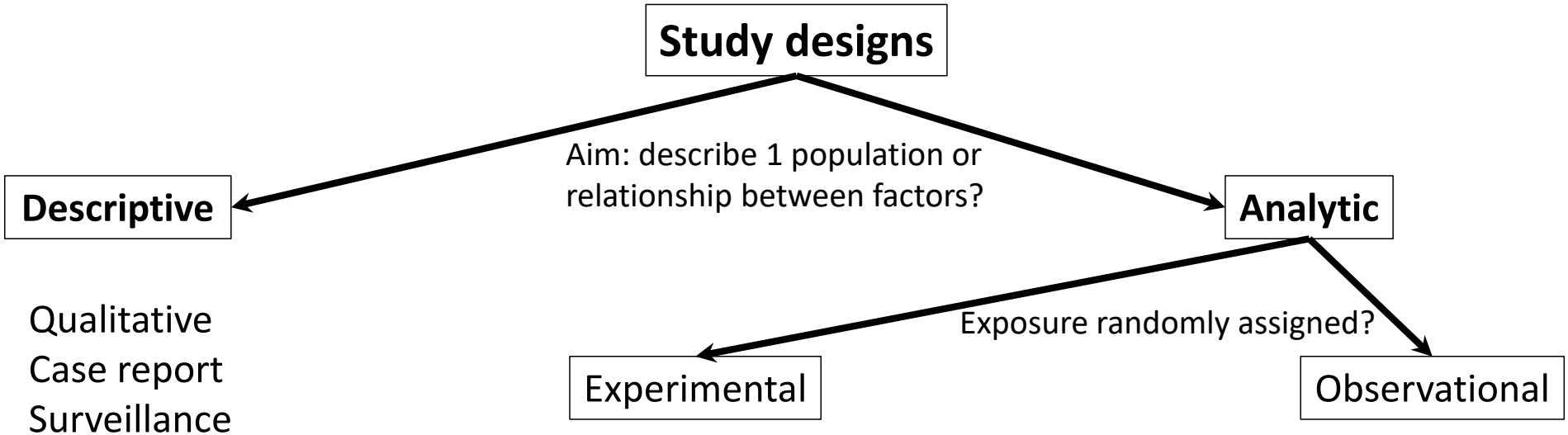
Exercise 1

Plan your own study

- Get together in breakout rooms.
- Think of a (specific!) study question that you want do investigate.
- Discuss what your underlying target population is, how you want to sample, what outcome measure you are measuring and how, and what you have to consider to get meaningful results.

Study designs

Overview



- Qualitative
- Case report
- Surveillance
- Cross-sectional
- Ecological

- RCT
- N-of-1 trial

- Cross-sectional
- Cohort
- Case-Control
- Nested case-control

cf. `http://www.cebm.net/index.aspx?o=1043`

Descriptive studies

Aim

Describe one population.

Types

- Qualitative study
- (Clinical) case report
- Surveillance study
- Cross-sectional study
- Ecological study

Descriptive studies - Qualitative study

- Study that aims to understand people's beliefs, attitudes, experiences, behaviors, interactions.
- Often use interviews, surveys, observations, focus groups etc.
- (Primary data is usually non-numeric).

Descriptive studies - Case report

- Detailed report about a single clinical case (patient).
- Guidelines: `https://www.aerzteblatt.de/archiv/145657/Die-Case-Reporting-(CARE)-Guideline`

Descriptive studies - Ecological study

- = aggregate study.
- Study where the unit of observation is a group of people rather than an individual.
- Examples: school classes, schools, factories, cities, countries.
- Important for interpretation of results: associations between exposure and disease on such a higher level (e.g. school-level) don't generally reflect individual-level associations.

Analytic studies

Aim

Compare multiple populations, explore relationships between exposures and outcomes

Differentiation

Was exposure (= intervention) assigned?

- Yes $\longrightarrow$ experimental study
- No $\longrightarrow$ observational study

Analytic studies

Experimental study

- Randomized controlled trial
- Non-randomized controlled trial
- N-of-1 trial

Observational study

- Cross-sectional study
- Cohort study
- Case-control study
- Nested case-control study

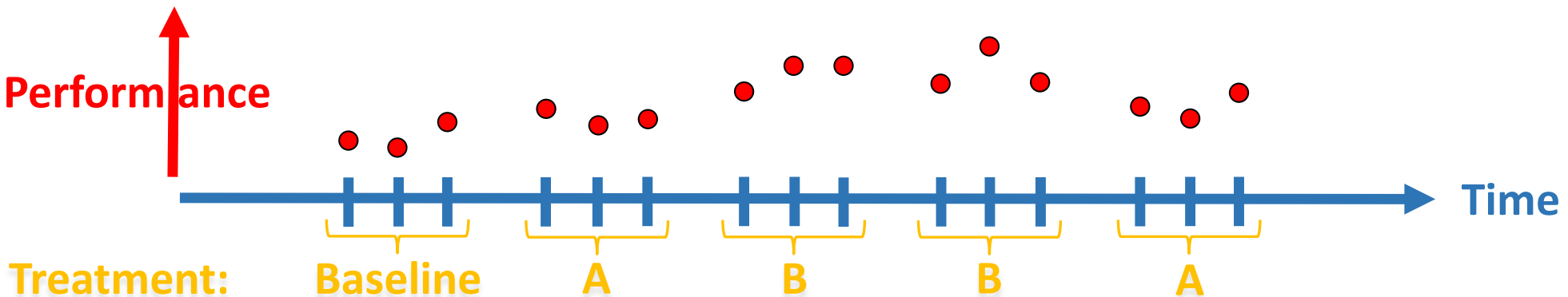
Analytic studies - Randomized controlled trial

- Experimental study in which the study participants are assigned to exposure/intervention groups randomly, and their outcome is observed prospectively.
- Often recommended as best study design to understand the effect of an intervention.
- Randomization can prevent different biases[1] and ensure that potential confounders don't systematically bias the results.
- Can be non-, single- or double-blinded.
- Can be cross-over: each participant has both interventions.
- Expensive, not always possible.
- (There exist other forms of trials not covered here.)

[1] bias = systematic error

Analytic studies - N-of-1 trial

- Perform a separate cross-over trial for each person to get individual-level effect estimates (compared to population-level effect estimates).
- Helpful for personalized medicine approaches.



Analytic studies - N-of-1 trial

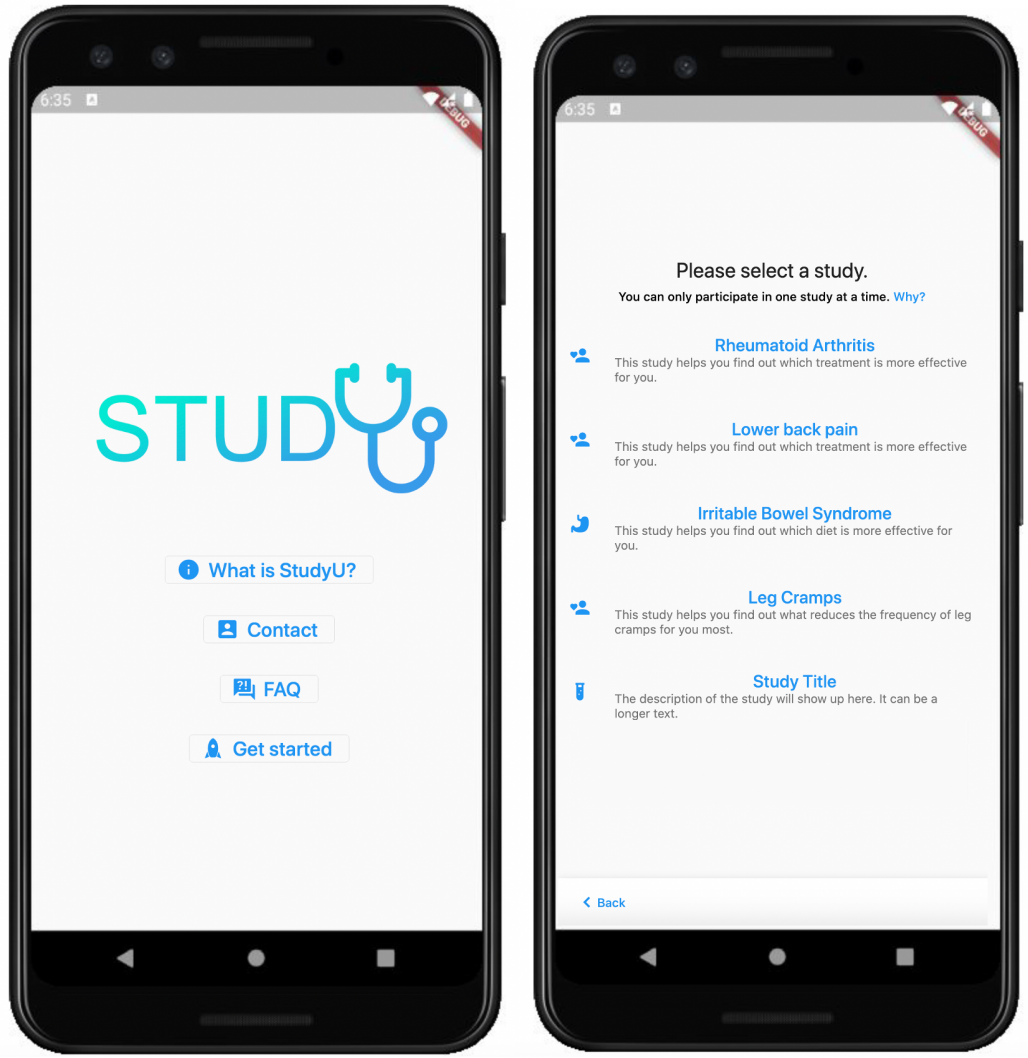
Example study:

Compare

Digital physical exercise interventions

in their effect on

chronic non-specific low back pain.



Analytic studies - Cross-sectional study

- Descriptive/observational study design in which exposures and diseases are assessed at the same time (frame) in the population.
- Can be used to estimate the prevalence of a disease or risk factor, or to determine the properties of a diagnostic test.
- Easy and cheap.
- Not randomized, can have hidden confounders and different biases.
- Can only get prevalence estimates.

Analytic studies - Cohort study

- Longitudinal observational study (i.e. study over time)
- Define a group of people that are free of the disease, assess all exposures of interest at baseline, then observe over time who develops the disease.
- I.e. assess exposures first, observe disease later.
- Can be prospective or retrospective (respective to the time when exposure and outcome were assessed).
- Can calculate incidence rates.
- Not randomized, can have hidden confounders.
- Usually expensive, hard for rare diseases or exposures, might need large sample sizes and/or long follow-up time.
- Important: response rate, drop-out rate.

Analytic studies - Case-control study

- Observational study, in which two groups (case and control group) are defined based on their outcome (e.g. sick/not sick), and then compared with respect to their attributes (i.e. who had a give exposure).
- I.e. compared to a cohort study, there is an extra step of sampling from the population (with respect to disease status).
- More efficient than cohort study, especially for rare diseases.
- Have to account for group selection (sampling) in the analysis, e.g. can do matching/re-weighting.

Analytic studies - Nested case-control study

- = case-control study nested in a cohort study.
- I.e. take all cases in the cohort study, and define a control group also within the cohort study (e.g. of a specific size, matched according to some criteria to the case group).

Exercise 2

Plan your own study

- Get together in the same groups as in exercise 1.
- Discuss the study that you have planned in exercise 1 again, if you want to do anything different now, and which specific study design you want to use for your study.
- Do a quick literature search (e.g. on PubMed) for studies that have investigated your study question. Look at one paper, and discuss how they have approached it.

Questions?

References

- Rothman KJ, Greenland S, Lash TL (2008). Modern epidemiology. Lippincott Williams & Wilkins.
- Ahrens W, Pigeot I (2014). Handbook of epidemiology. Springer.
- Rosner B (2010). Fundamentals of biostatistics. Brooks/Cole, Cengage Learning.
- Wasserman L (2010). All of statistics. A concise course in statistical inference. Springer.
- Knight K (1999). Mathematical statistics. CRC Press.

Homework

Homework

See file `R_4_homework.Rmd`.