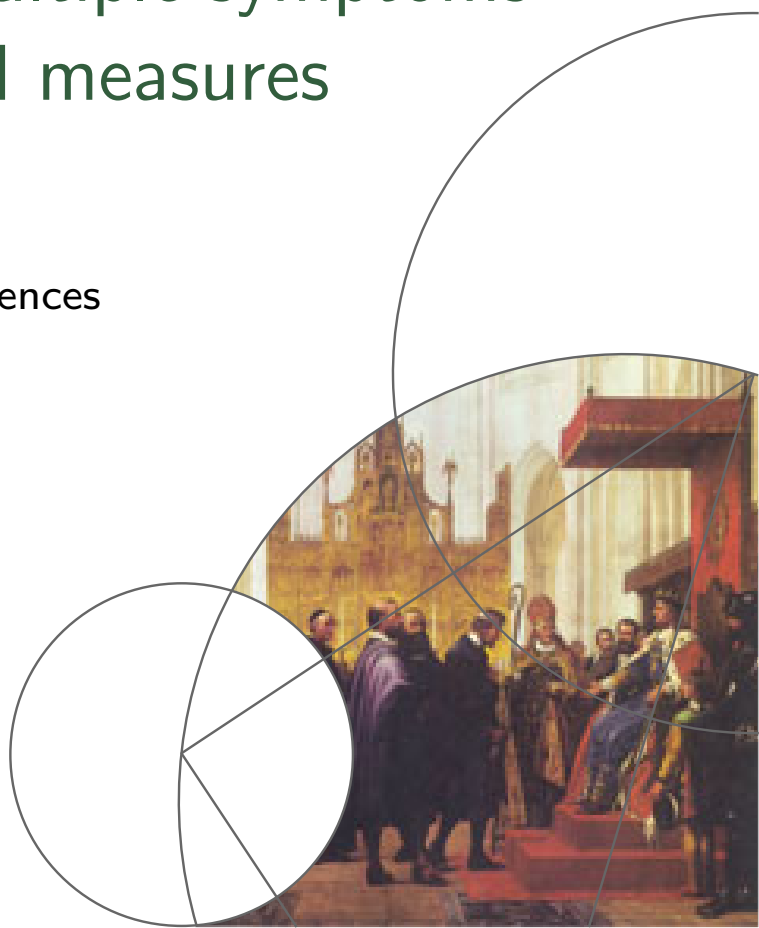
Statistical analyses: single & multiple symptoms and symptom clusters, repeated measures

Anders Tolver
Data Science Lab, Department of Mathematical Sciences

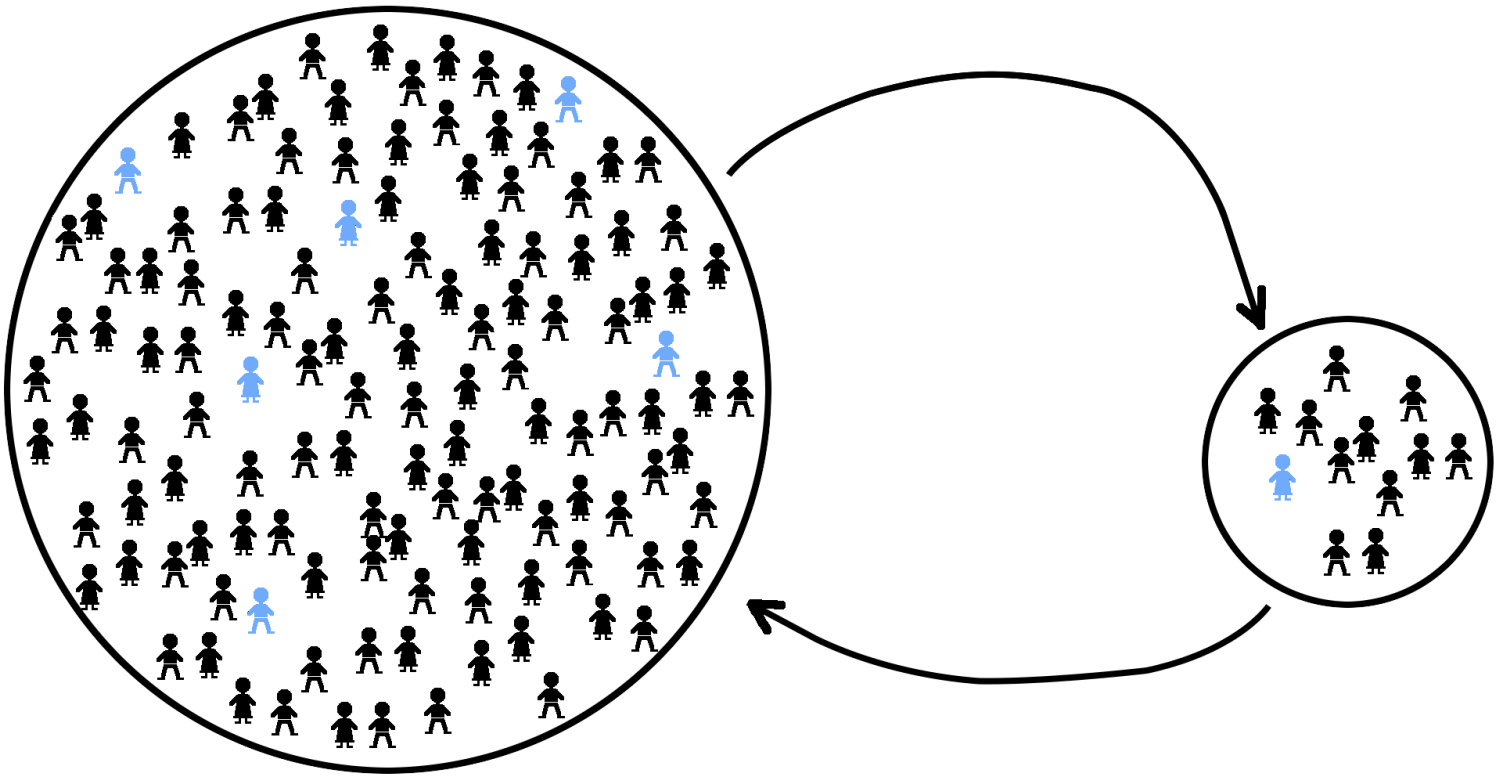


Outline

- What is statistics?
- What is your role?
- Statistical analysis of single symptoms
- Power calculations and sample size justification
- Repeated measurements of a symptom
- Statistical analysis of multiple symptoms
  - Adjustment for multiple testing
  - Symptom clusters
- Various other topics
  - Missing data, dropouts, compliance
  - Adjustment for other covariates, stratification

Population and sample

We wish to draw conclusions related to the entire population based on a finite sample



Picture from *Ekstrøm & Sørensen: Statistical Data Analysis for the Life Sciences (2nd Edition)*.

What is statistics?

How do we work?

Scientific question: related to one or more populations

Data: finite sample from relevant populations

Statistics: conclusions about the populations based on data

Example: Will prevalence of a symptom be different if we replace treatment A (standard) with treatment B (new)?

Let's collect data from populations receiving either treatment A or B. Compute the prevalence of symptom for each group and extrapolate the result to all future populations of patients receiving either treatment A or B.

Main problem:

Understand the variation of the results from small sample and how to relate it to the full population.

What is your role?

- You have the relevant scientific questions
- You are responsible for your data!
- You know your data
- You must be able to explain what has been done to data
- You must be able to discuss assumptions and limitations related to statistical data analysis

The path from data to results should not be a black box.

You can get far by simple statistical tools ...

You probably know more about statistics than a (random) statistician knows about your field of research.

Measurement of symptoms: data types

Symptom: subjective evidence of a disease

Measurement of symptoms will typicaly be either

binary	present / absent
ordinal	never, seldom, often, always
quantitative	(discrete) likert scale 1-5, number of occurrences
quantitative	(continuous) VAS 0-100, sum scale from questionnaire (PRO)

Data type determines what stat. analyses that can be used.

(Models may be useful without being correct in every respect)

Design of experiments: randomized clinical trial

Not only treatment affects prevalence of symptom.

Variation in age, gender, diagnosis, stage of the disease, personal frailty etc. may be associated with differences in prevalence of the symptom.

Willingness to try treatment B may be different for certain age group, stage of disease, gender etc.

To get comparable groups: **random allocation of treatments**

Possibly combined with

- **stratification:** separate randomization lists (for each diagnosis, center, ...)
- **adjustment**: group comparisons adjusted (for age, gender, ...)

Comparing prevalence of a symptom

Create a table of counts

treatment	present	absent	total
A	10	18	28
B	6	24	30

Estimate for prevalence in group A (and 95 %-conf.int):

$$\hat{p}_A = \frac{10}{28} \approx 0.357; \quad \hat{p}_A \pm 1.96 \cdot \sqrt{\frac{0.357 \cdot (1 - 0.357)}{28}} = [0.179, 0.535]$$

Test for same prevalence in groups A and B:

- χ^2-test: $p = 0.1809$
- Fisher's exact test: $p = 0.2432$

Online ressources: Chi-Square calculator - Fisher's exact test

Compare average perception of a symptom

Compute summary statistics of symptom measurements

treatment	mean	standard deviation (SD)	total
A	32.23	7.28	13
B	21.54	5.81	13

Make histogram for each group: bell shaped (Gaussian)?

If not: report instead median, range and/or quartiles!

Estimate and 95 %-CI for population mean: (Gaussian)

$$\hat{\mu}_A = 32.23; \quad \hat{\mu}_A \pm 1.96 \cdot \frac{7.28}{\sqrt{13}} = [28.27, 36.19]$$

Test to compare symptom burden between groups:

- Gaussian: t-test for two independent samples $p = 0.0002$
- not Gaussian: Mann-Whitney U-test $p = 0.0011$

Online ressources: Mann-Whitney U-test - t-test

Statistical hypothesis testing

Hypothesis, H_0:
Same prevalence/level of symptom in groups A and B.

Use data to compute: **test statistic** and **p-value**

If $p \geq 0.05$ accept H_0. If $p < 0.05$ reject H_0.

Four possibilties:

	H_0 accepted	H_0 rejected
H_0 is true	OK	error of **type I**
H_0 is false	error of **type II**	OK

- Probability of Type I error $\leq$ **0.05** $\leftarrow$ **significance level**
- 1 - (Probability of Type II error) is called **power**

Power calculations and sample size

Unfortunately: No guarantee that statistical test based on your data lead to the right conclusion!

- If H_0 is true then p-value computed from data should be high ($p \geq \alpha$).
- If H_0 is false then p-value computed from data should be low ($p < \alpha$).

Level of test (α): (type I error)
Is something that we decide/control!
Power of test: (type II error)
Probabilty of actually rejecting H_0 if there is a difference.

Waste of time/ressources to conduct a study with low power.

To get money/permission to do a study you need a power calculation (or a sample size justification).

Power calculations

Problem/goal:

Design intervention study that will reveal a difference between A and B (with high probability) if there is an effect!

We shall discuss two cases:

- Primary outcome (symptom) is a quantitative measure
- The primary outcome (symptom) is a binary measure

We here use a signficance level of $\alpha = 5\%$, but the power depends on α.

Lower level α (low type I error) will lead to a lower power (high type II error).

For fixed α a larger sample size will decrease type II error, but there is always a limit (ethics, time, economy, etc.)

Online ressources: about power calculations

Power: quantitative symptom measurement

Calculation often based on t-test and depends on

- **Significance level:** for example $\alpha = 0.05$
- **Minimal clinically important difference:** requires subject matter knowledge
- The (within group) **standard deviation** of outcome: find data from relevant study in literature
- The **sample size**: number of patients in each group

Exercises:

- Compute the power to detect a difference of 2 for an outcome with within-group SD $= 10$ for a study with 50 patients in each group.
- How many patients are required to get a power of 0.80? What if we expect 25 % to drop out?

Power: binary symptom measure

Calculation often based on two-sample test for equal proportions and depends on

- **Significance level:** for example $\alpha = 0.05$
- **Prevalence in group A:** for example $p_A = 0.5$
- **Prevalence in group B:** for example $p_B = 0.3$
- The **sample size**: number of patients in each group

Exercises:

- Compute the power to detect a reduction of the prevalence of a symptom from $p_A = 50$ % to $p_B = 30$ % for a study with 50 patients in each group.
- How large should the (true) reduction be to get a power of 0.8 in a study with 30 patients in each group?

Multiple assessment times: example

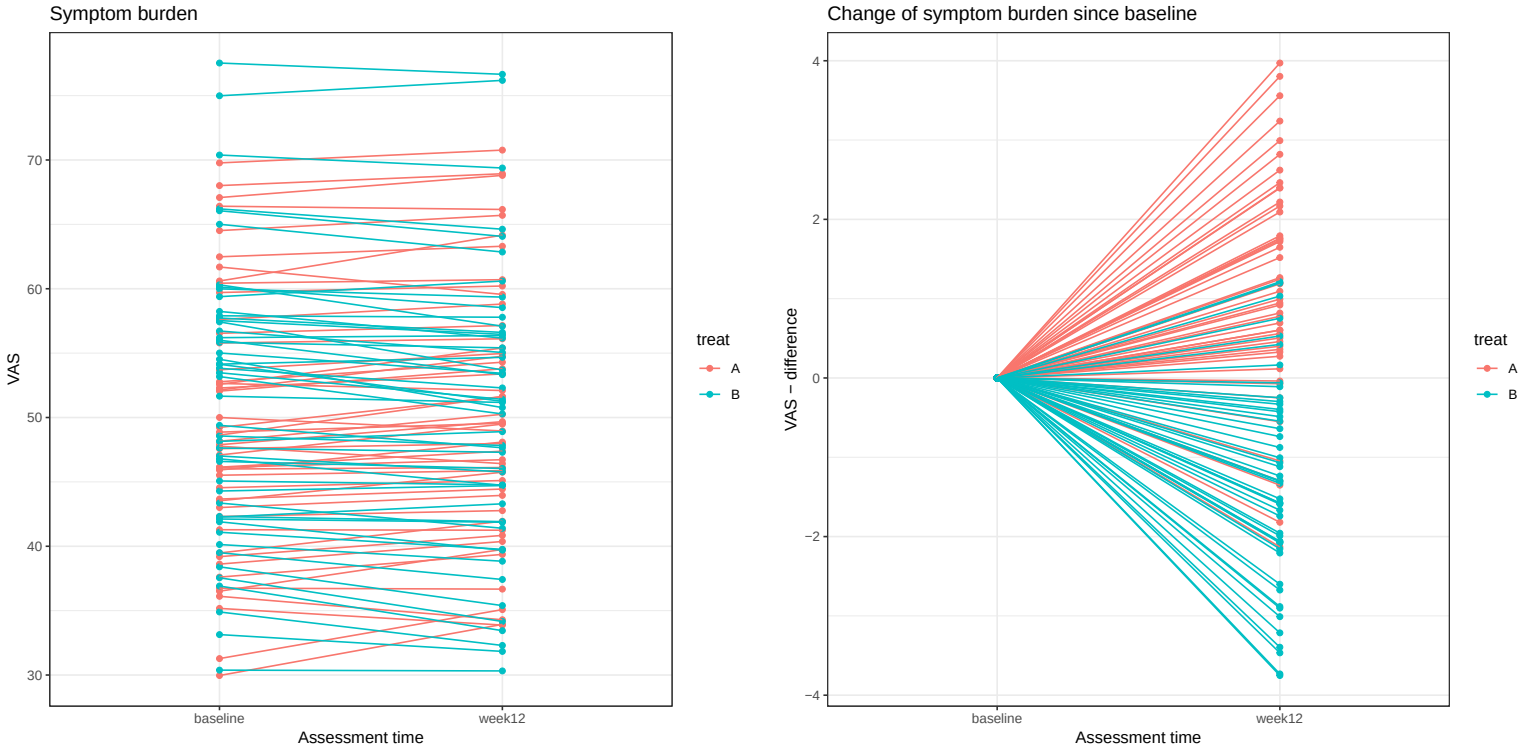
- **Primary outcome:**
  symptom measured on VAS after 12 weeks
- **Two treatment groups:** A and B
- **Power calculation:**
  SD $= 10$ on primary outcome and a clinical relevant difference of 2 a sample size of 50 yields a power of 0.17.

New idea: include also **baseline measurement** of symptom

Due to randomization there is no difference between symptom burden at baseline
(-if there is this just happened by chance!)

Differences between symptom burden after 12 weeks and baseline may be more relevant measure of treatment effect and have **lower variance**.

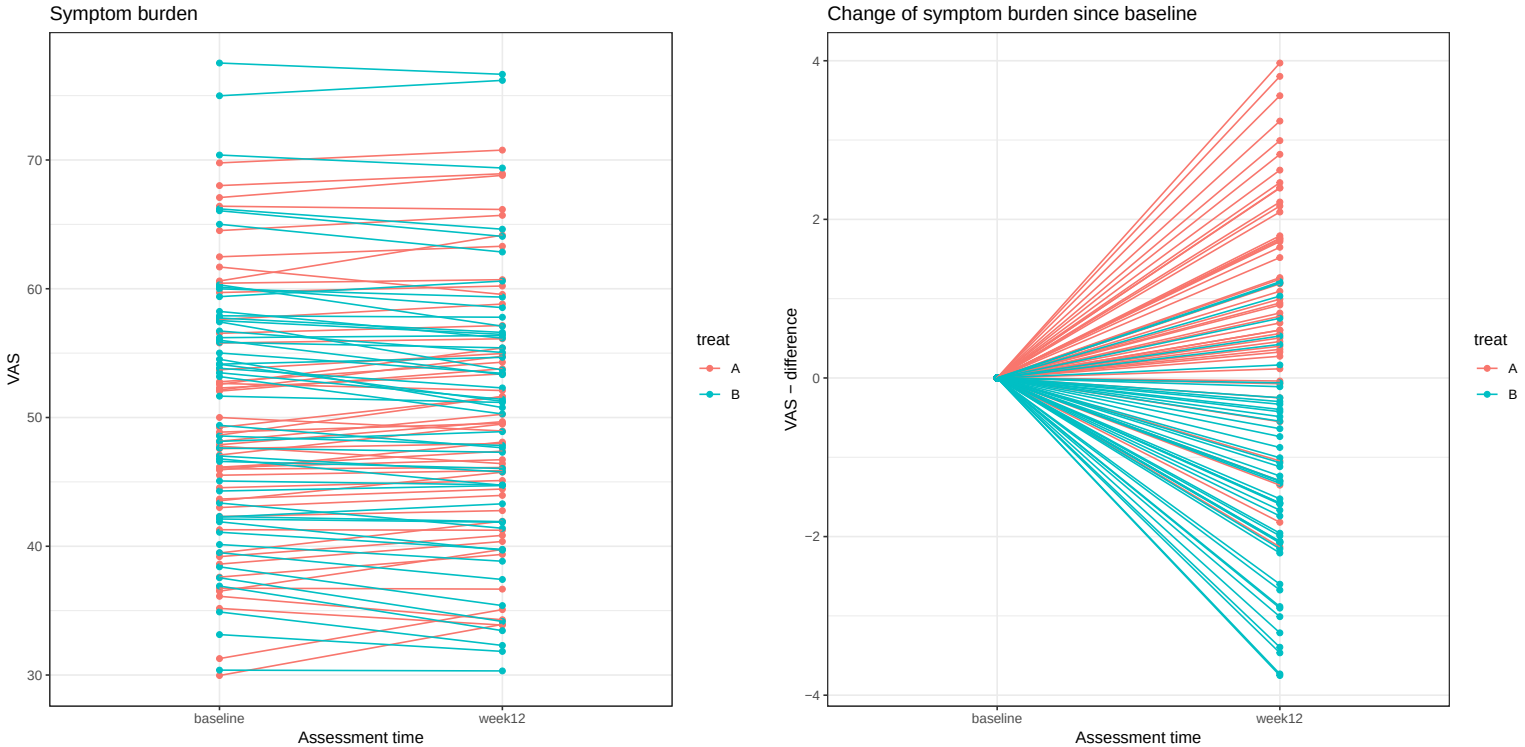
Multiple assessment times: use baseline



Left:

Range of VAS at week 12 is 30-83. No diff. between groups.

Multiple assessment times: use baseline



Left:

Range of VAS at week 12 is 30-83. No diff. between groups.

Right:

Range of Δ VAS is (-4)-4. Clear difference between groups.

Multiple assessment times: take home messages

Changes from baseline often the most relevant outcome

Remember to do power calculation for relevant outcome:

- Use clinically relevant differences for changes since baseline
- Use standard deviation (SD) for changes since baseline

Relevant SD can be harder to find in literature!

Statistical analysis: (pre- and post intervention data)
Compute changes and compare changes between groups using (unpaired) t-test or Mann-Whitney.

Repeated mesurements:
May refer to situation with more than two assessment times.
(Full) statistical analysis requires more complicated models.

Multiple symptoms

It is very common to have multiple measurements of symptom burden:

- **Repeated measurements** (over time) for each patient: baseline, week 12, etc.
- Several **different symptoms** for each patient: anxiety, depression, fatique, etc.

Symptom measurements may

- be temporally dependent
- have a common source

Repeating statistical analyses on each single symptom/measurement increase the risk of type I errors (false positive results).

No *right way* to adjust for this (personal opinion!).

Adjustment for multiple testing

Bonferroni correction (classic approach):

- Count the total no. of statistical tests you did (say: $k = 25$).
- Consider only $p < \frac{0.05}{25} = 0.002$ as a significant p-value.
- At most 0.05 (=5 %) risk of reporting a false positive.

Suggestions:

- **Robustness:** Report original and Bonferroni adjusted p-values and modify discussion of results accordingly.
- **Be honest:** For secondary outcomes there is a higher risk of false positive findings!
- **Get help:** Fit a more complex statistical model allowing a more advanced approach for adjustment of p-values.

(Symptom) clusters

Some researchers like to talk about symptom clusters.

The definition is not clear but may be something like

> *Symptom clusters are defined as two or more concurrent symptoms that are related and may or may not have a common cause.*

[Fan et al., Curr Oncol. 2007 Oct; 14(5): 173–179.]

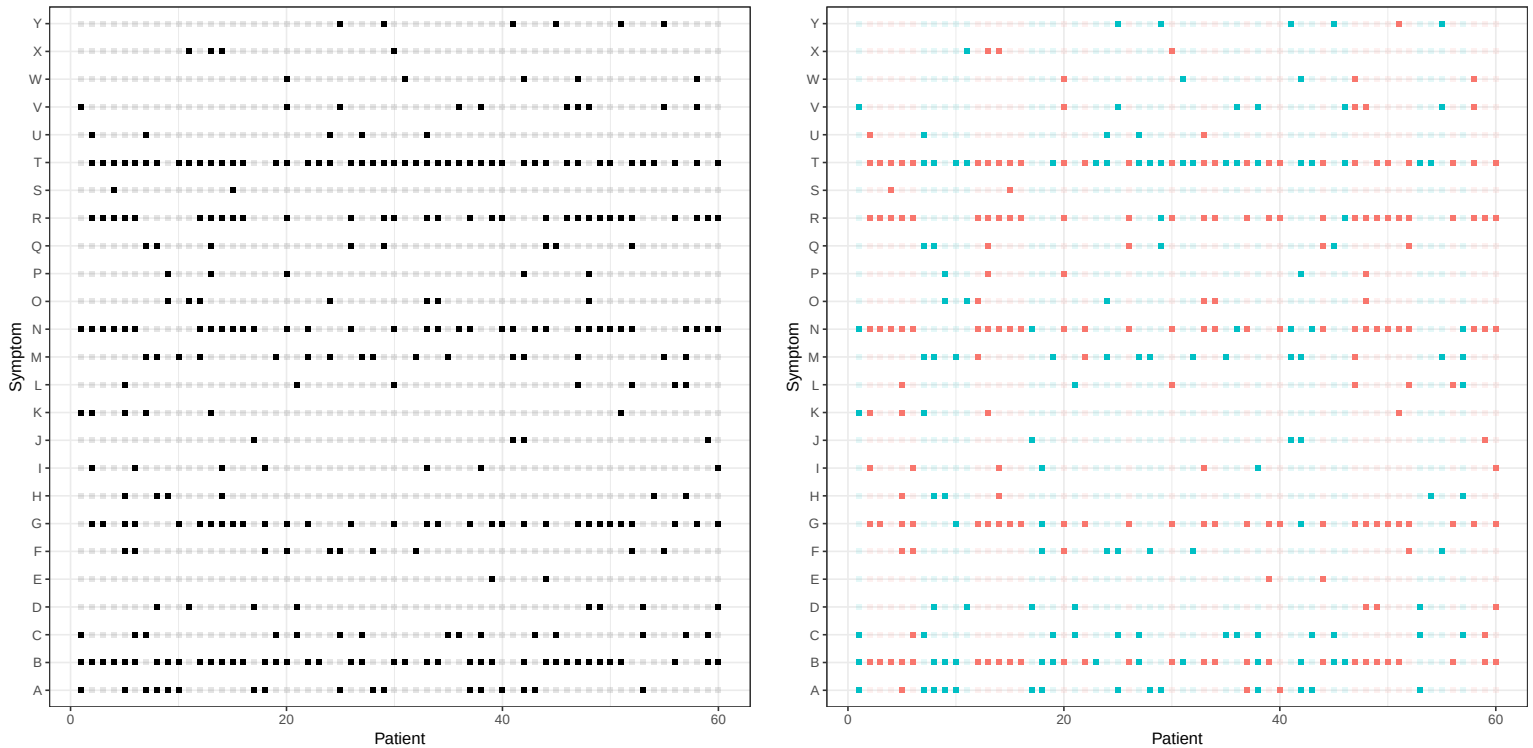
I prefer the part of the literature looking for

> *Groups of patients experiencing a similar pattern/fingerprint of symptoms.*

In particular:
Same two symptoms may be part of the symptom byrden for two groups of patients (but at different level).

Example: latent class analysis

25 binary symptoms (A-Y) registered for 50 patients.



Latent class analysis finds groups of patients with similar symptom burden:

B, G, N, R, T occur frequently together in one group.

A, C, M occur more frequently in the other group.

Missing data, dropouts, compliance

The randomized clinical trial allows us to get unbiased estimates of treatment effects.

Things may happen that could lead to biased sample:

- **Drop-outs:** some patients did not complete the study
- **Missing data:** incomplete observations for all patients
- **Compliance:**
  some patients did not receive the prescribed treatment

The statistician can't tell you how to adjust the analysis: Any method is based on assumptions that may not be verified internally from data.

You might have additional information that can justify a particular method for adjusting the analysis.

Advice: Keep information on why patients dropped out, why some data are missing, compliance to treatment, etc.

Missing data, dropouts, compliance: examples

Think about how the following may affect the results of the statistical analysis (i.e. comparison of groups A and B)?

- Patients drop out of the study if they get too sick/weak.
- Patients do not complete questionnaires on symptoms that they are strongly affected by.
- It was more embarrassing for patients in group B to show up for the final test to get a poor result than for patients in group A.
- Some patients receiving treatment B dropped out or returned to treament A if they did not see any improvement or due to side effects.
- Patients in group A who also heard about treatment B (exercise program) started to exercise more.

Stratification

Often primary outcome will be different for subgroups characterized by other factors (age, gender, diagnosis, etc.).

We have discussed that importance of randmomizing treatments A and B in a clinical study.

Stratified randomization:

Divide population into subgroups based on factors known to affect outcome. Use separate randomization lists for each subgroup.

Benefit:

Balanced treatment groups for each strata (=subgroup).

Stat. analyses should be adjusted for stratification variables.

Stratification necessary?

On average subgroups will be balanced w.r.t. treatment.
Maybe for smaller studies - hard to find strong arguments.

Adjustment for covariates: why?

For a randomized clinical trial: Difference between group means is an unbiased estimate of the treatment effect.

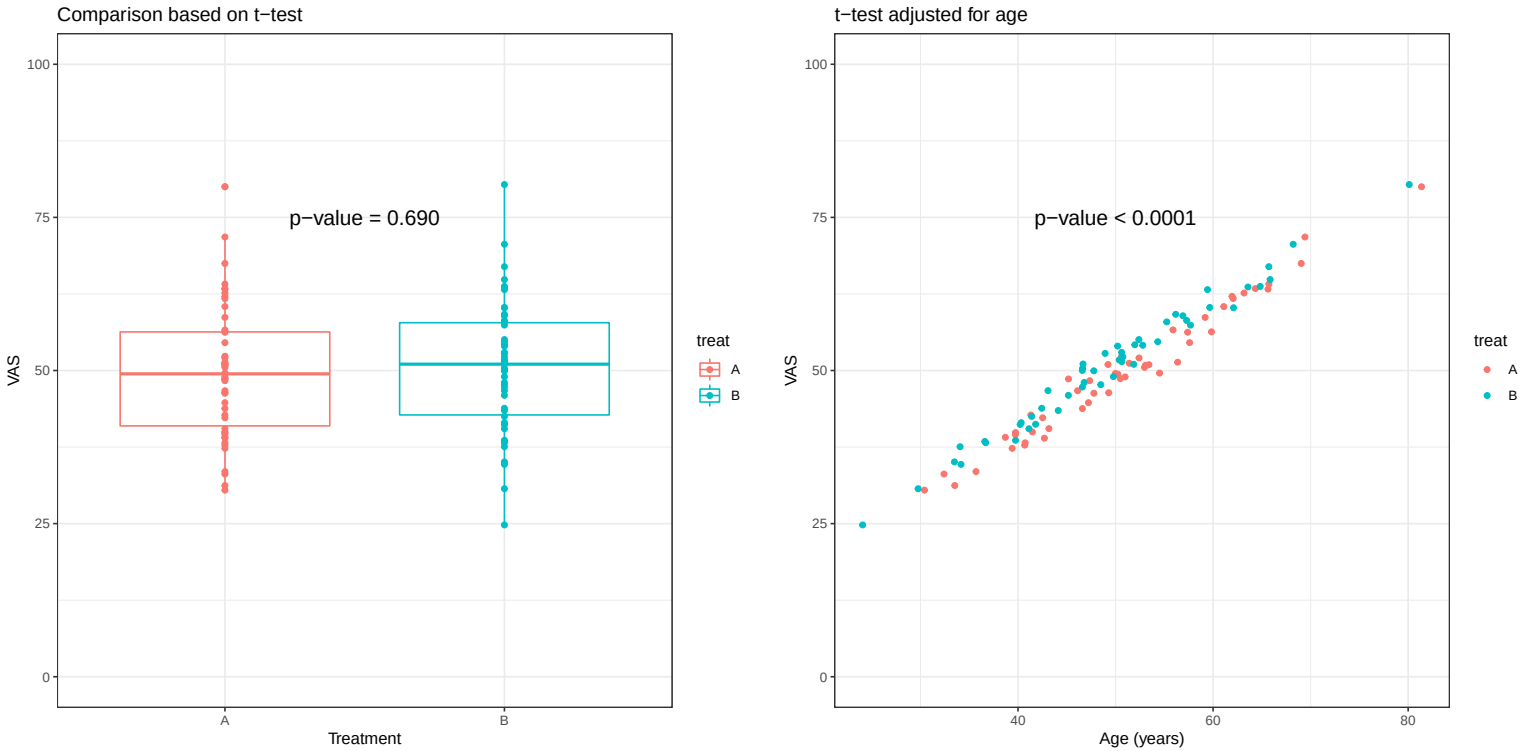
But if study population is heterogeneous ...

- ... it may be hard to demonstrate a treatment effect
- ... the treatment effect may be different for subgroups within treatment groups

Adjusted analyses: (purpose)
Use mathematical model to compute treament effect adjusted for other factors that may influence the outcome.

- Technical details/tools:
  take stat-course + learn some R/SAS/SPSS/STATA
- Today we illustrate the importance through an example

Adjustment for covariates: example



Left: Large within-group variance of VAS (mainly due to variation in Age).

Right: When adjusting analysis for Age there is a clear effet of treatment.

Adjustment for covariates: some piece of advice

- List observable variables/factors (apart from treatment) that may affect outcome
- (Stratify randomization according to 1-2 variables?)
- Quantify/test for treatment effect (unadjusted analysis)
- Quantify/test for treatment effect adjusted for relevant variables (=covariates/explanatory variables)
  - **Quantitative outcome:** analysis of (co)variance AN(C)OVA?
  - **Binary outcome:** logistic regression?

Effect modification:

To allow/model different treatment effects in subgroups you may need to include interactions
(ex: gender $\times$ treatment)

Concluding remarks

If you want to do research based on data you need to know something about statistics.

Statistics is not (only) about describing data, rather about understanding the uncertainty (or reproducibility) of the results from your study.

You can get far by:

- common sense
- knowing a few basic tools
- understanding and being able to talk about the main problems in non-technical terms
- using google, taking a stat-course, or engaging with a statistician to get access to more advanced tools