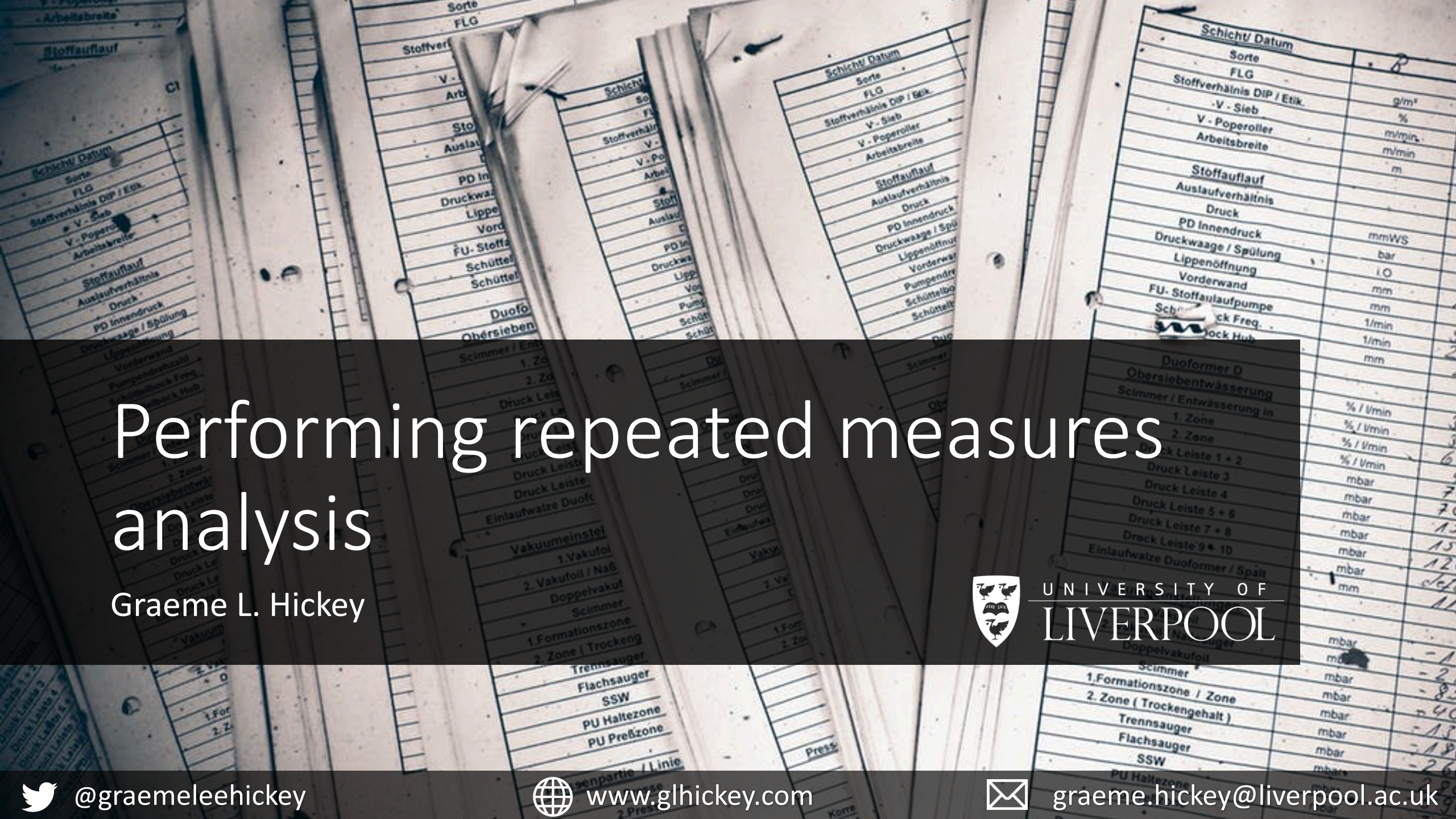


Performing repeated measures analysis

Graeme L. Hickey

UNIVERSITY OF LIVERPOOL

@graemeleehickey

www.glhickey.com

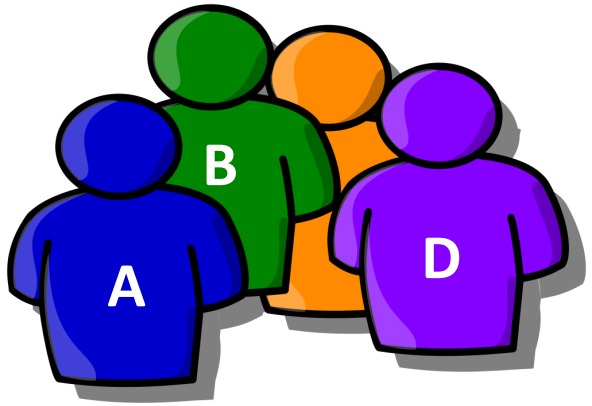
graeme.hickey@liverpool.ac.uk

Conflicts of interest

- None
- Assistant Editor (Statistical Consultant) for EJCTS and ICVTS

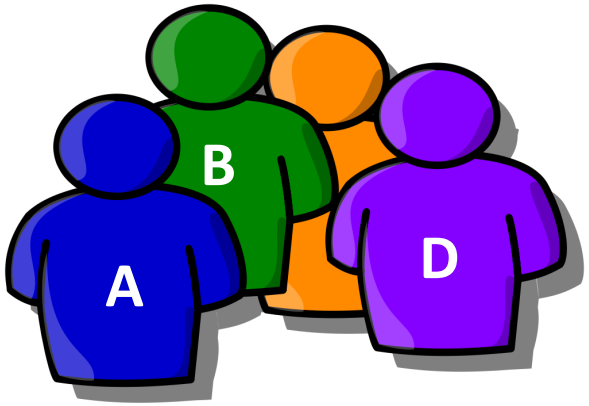
What are “repeated measures” data

“Condition”: chocolate cake



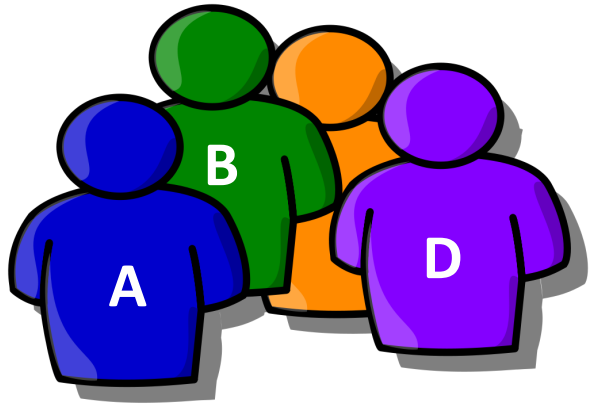
Measurement: taste score

“Condition”: lemon cake



Measurement: taste score

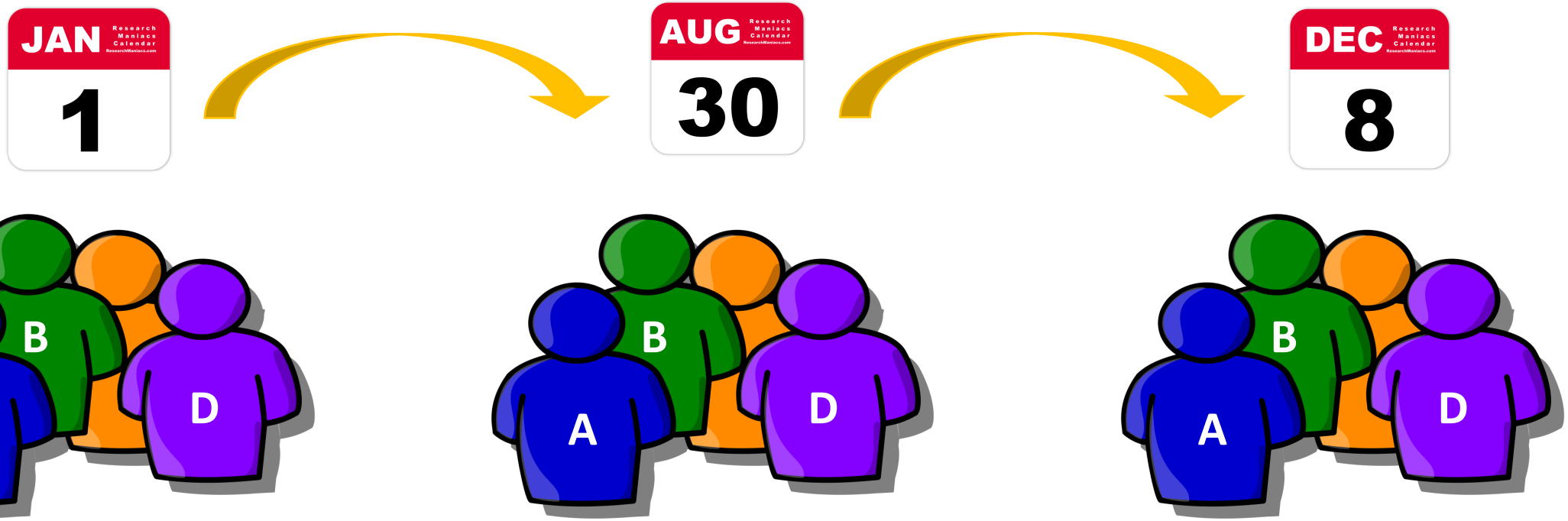
“Condition”: cheesecake



Measurement: taste score

Same people score each condition

What are “repeated measures” data



Measurement: systolic BP

Measurement: systolic BP

Measurement: systolic BP

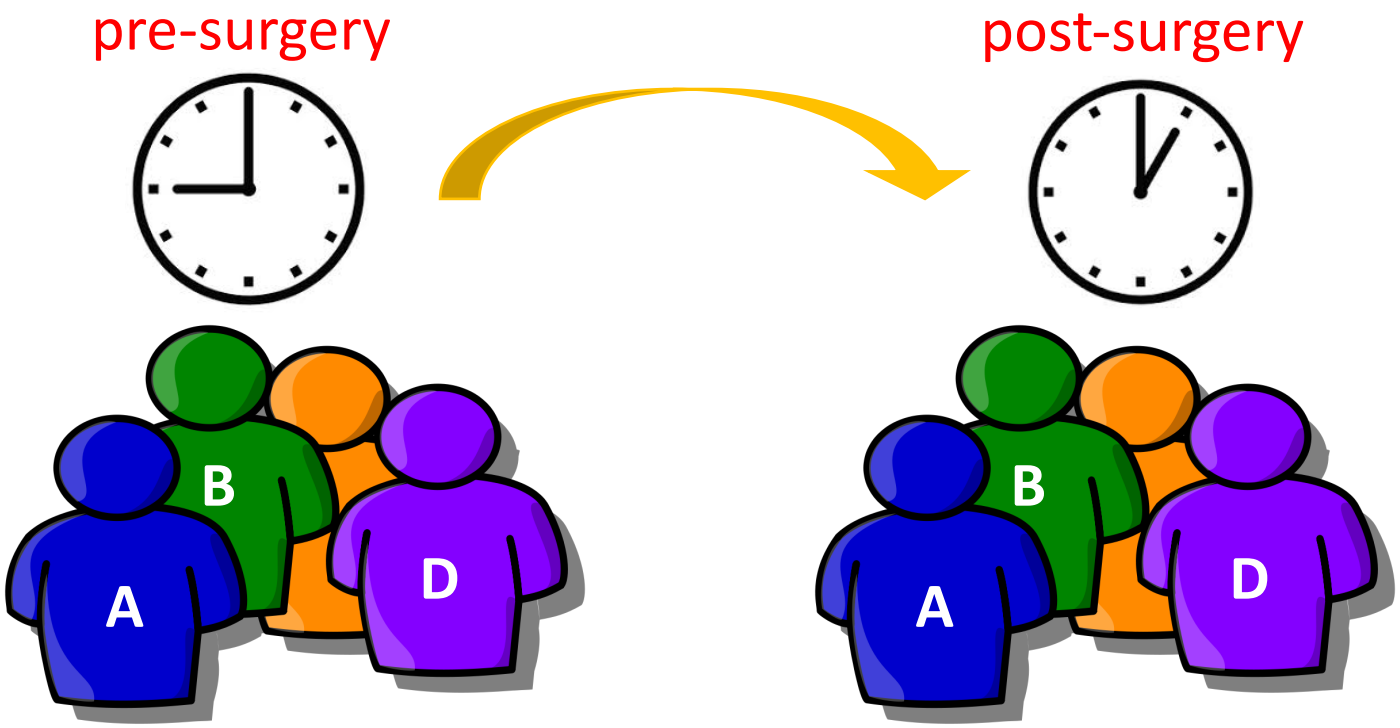
Same people provide BP at every follow-up appointment

Why do we need special methodology?

- Data are not independent: repeated observations on the same individual will be more similar to each other than to observations on other individuals

- Guidelines for reporting mortality and morbidity after cardiac valve interventions also propose the use of longitudinal data analysis for repeated measurement data

Simplest case: 2 measurement times

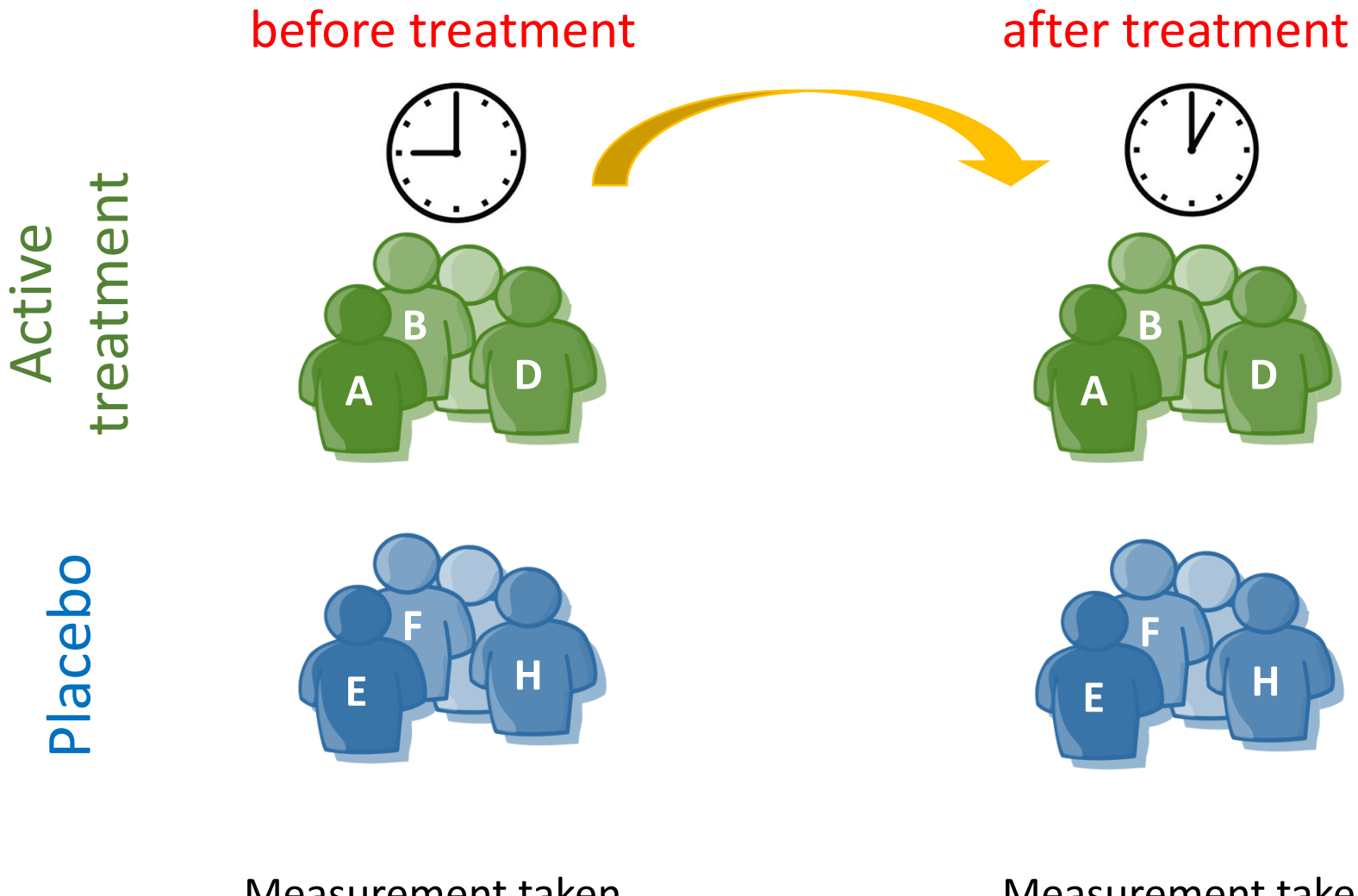


Measurement: AV gradient

Measurement: AV gradient

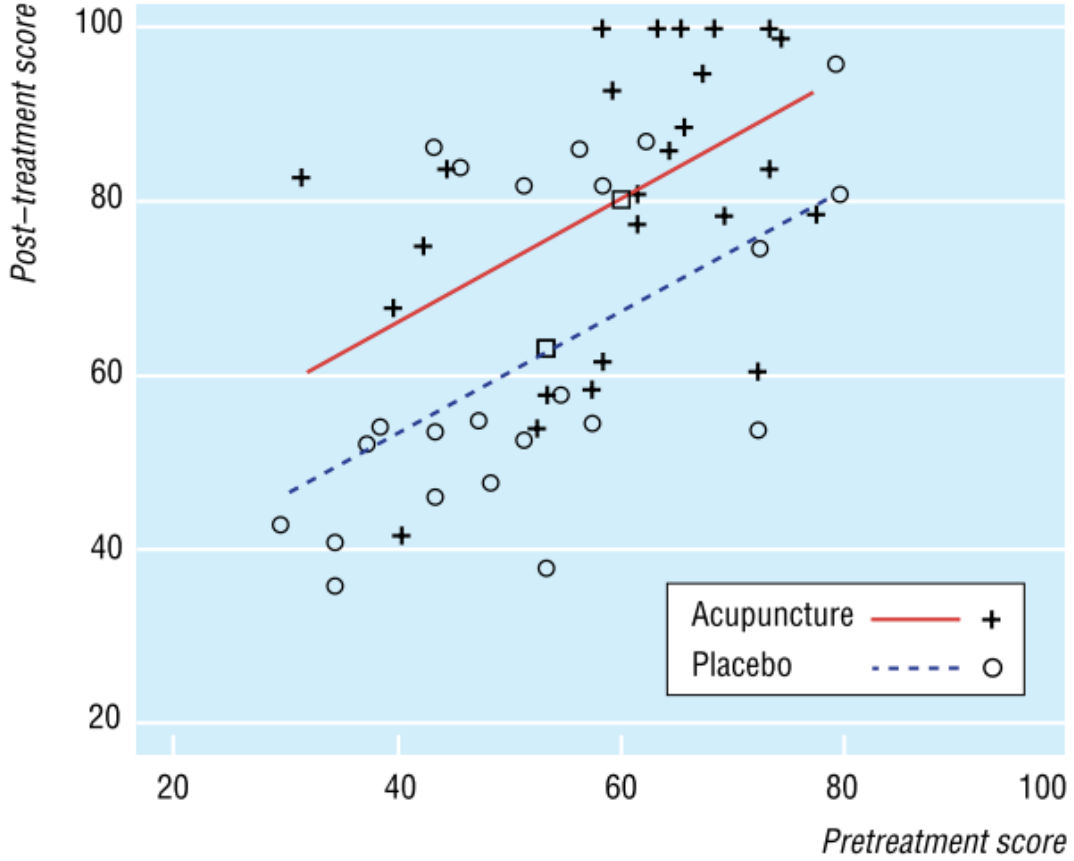
Suitable methods: paired *t*-test or Wilcoxon signed-rank test

What if we have treatment groups?



Question: if patients are randomised to treatment arms, how can we test whether active treatment is more effective than placebo?

Methods: shoulder pain example



	Placebo (*n* = 27)	Acupuncture (*n* = 25)	Difference between means (95% CI)	*P*
Follow-up	62.3 (17.9)	79.6 (17.1)	17.3 (7.5 to 27.1)	<0.001
Change score	8.4 (14.6)	19.2 (16.1)	10.8 (.3 to 19.4)	0.014
ANCOVA			12.7 (4.1 to 21.3)	0.005

General rule-of-thumb: analysis of covariance (ANCOVA) has the highest statistical power

Note: never use percentage change scores!

Source: Vickers & Altman. *BMJ*. 2001; 323: 1123–4.

More general scenario

- We record measurements of each patient >2 times
- Two (or more treatment groups)

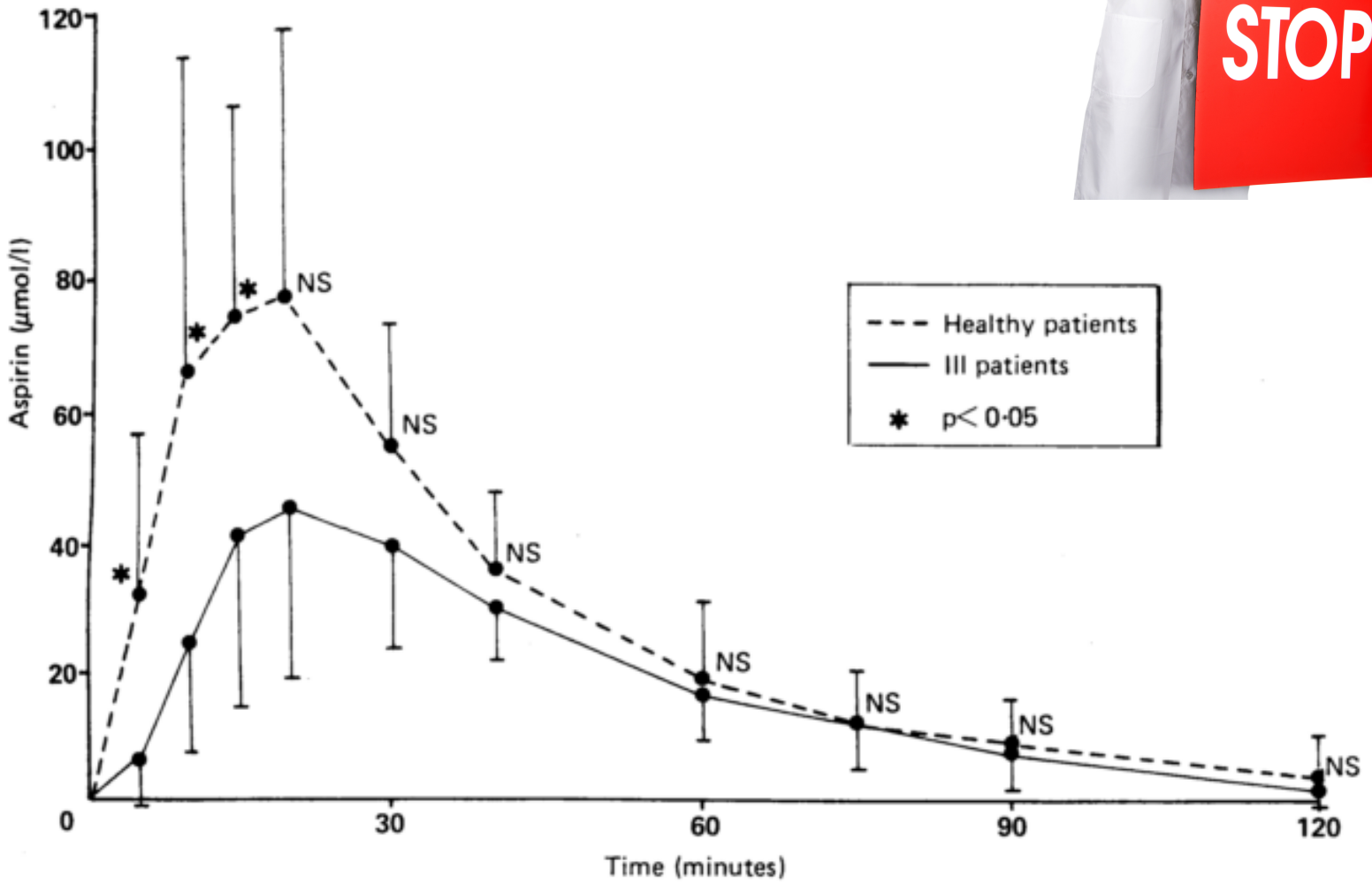
Design considerations

- **Balanced *versus* unbalanced**
  - Balanced follow-up (e.g. baseline, 1-hr, 2-hr, 8-hr, 16-hr, 24-hr)
  - Unbalanced (e.g. patient A visits their physician on days 1, 4, 6, 9, 12, and patient B visits only on days 5, 9, and 15)

- **Missing data**
  - E.g. patient fails to attend *scheduled* follow-up appointment

How *not* to proceed

- Multiple testing issues
- No account of same patients being measured ⇒ successive observations likely correlated
- Visualization + reporting issues



Source: Matthews et al. *BMJ*. 1990; 300: 230–5.

Data format / collection

Wide format

Subject	Jan 01	Aug 30	Dec 08
A	120	113	115
B	94	94	110
C	140	145	160
D	100	101	100

Good for balanced datasets

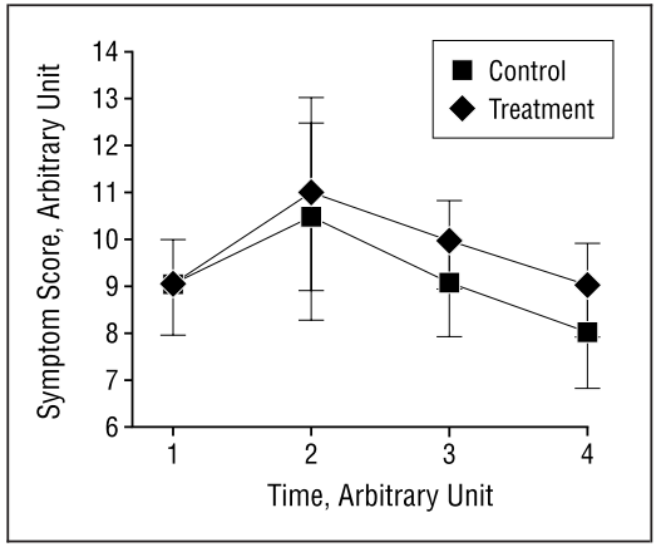
Good for unbalanced datasets

Long format

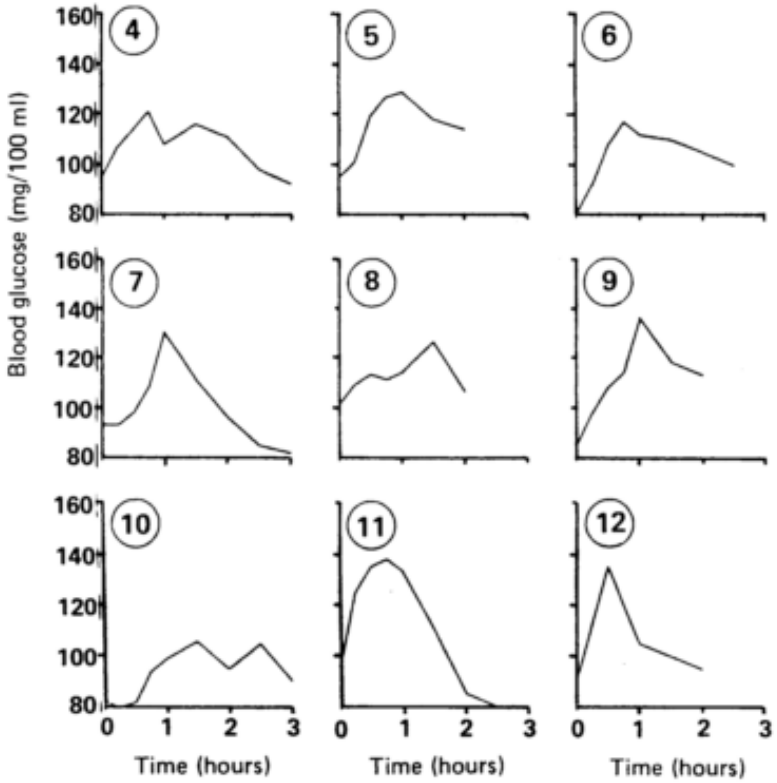
Subject	Date	BP (mmHg)
A	Jan 01	120
A	Aug 30	113
A	Dec 08	115
B	Jan 01	94
B	Aug 30	94
B	Dec 08	110
⋮	⋮	⋮
D	Aug 30	101
D	Dec 08	100

First step (always!): visualize the data

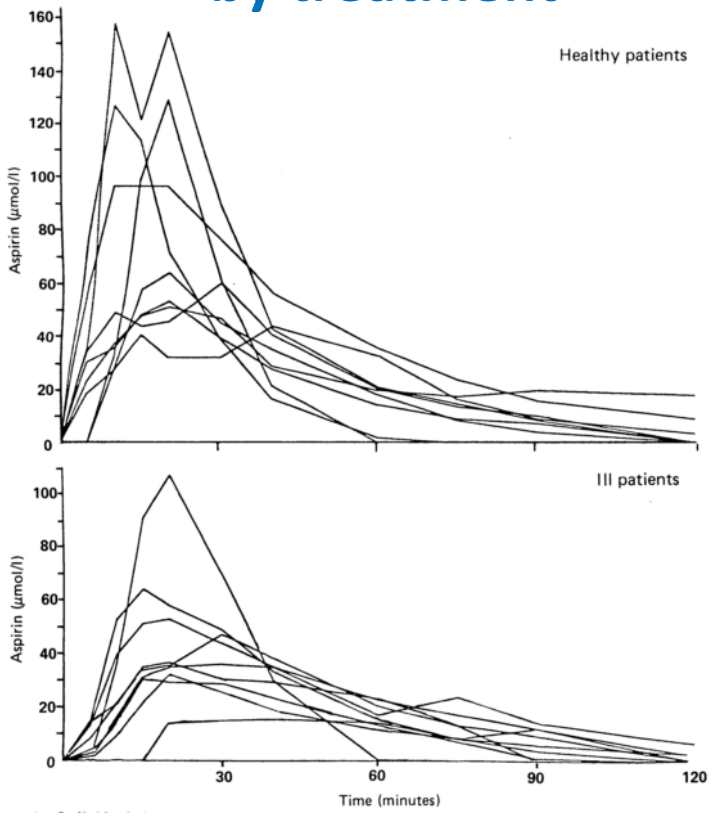
Mean profile plot



Individual panel plots



Individual plots grouped by treatment



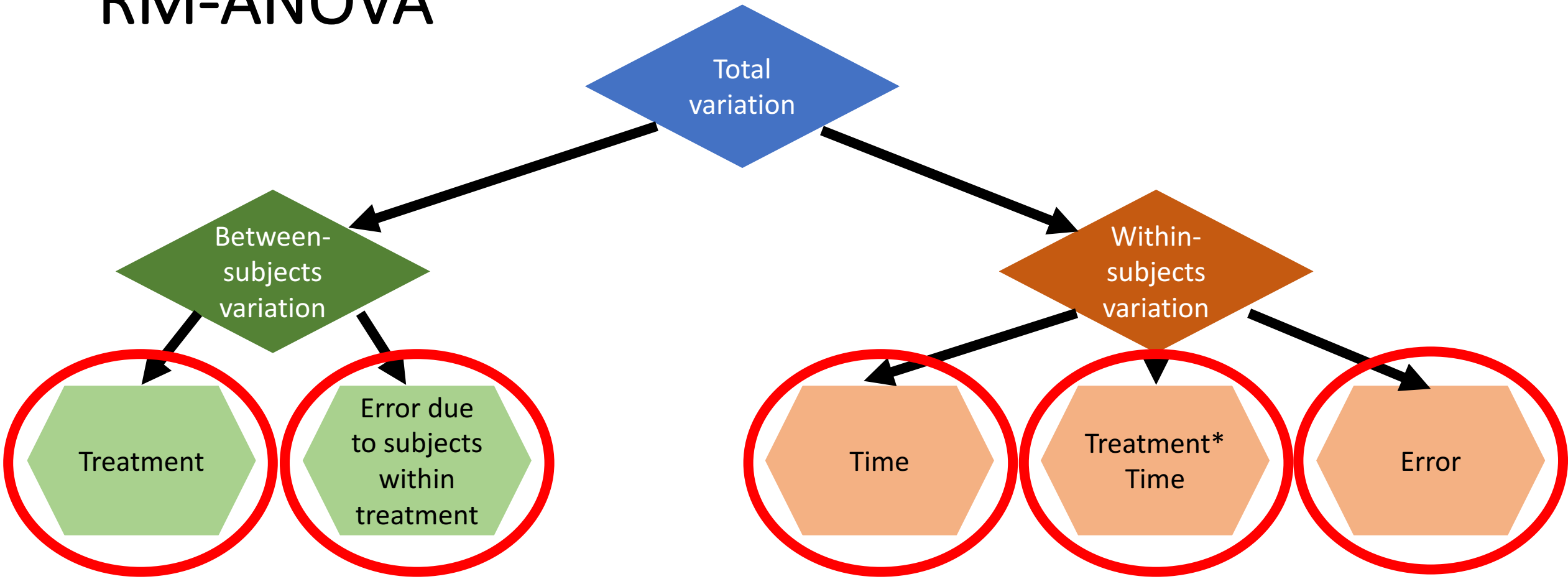
Source: Gueorguieva & Krystal. *Arch Gen Psychiatry*. 2004; 61: 310–317.

Source: Matthews et al. *BMJ*. 1990; 300: 230–5.

Analysis options

- Repeated measures analysis of variance (RM-ANOVA)
- Linear mixed models (LMMs)
- Summary statistics / data-reduction techniques
- Multivariate analysis of variance (MANOVA)
- Generalized least squares (GLS)
- Generalized estimating equations
- Non-linear mixed effects models
- Empirical Bayes methods
- …

RM-ANOVA



Test for: interaction effect

Sphericity

Tomorrow (14:15 – 15:45): *Checking model assumptions with regression diagnostics*

- RM-ANOVA depends on the usual assumptions for ANOVA…
- … and the assumption of sphericity

$$SD_{T2-T1} \cong SD_{T3-T1} \cong SD_{T3-T2} \cong \ldots$$

- Restrictive for longitudinal data $\Rightarrow$ measurements taken closely together are often more correlated than those taken at larger time intervals

- Test for sphericity using Mauchly's test

When sphericity is violated

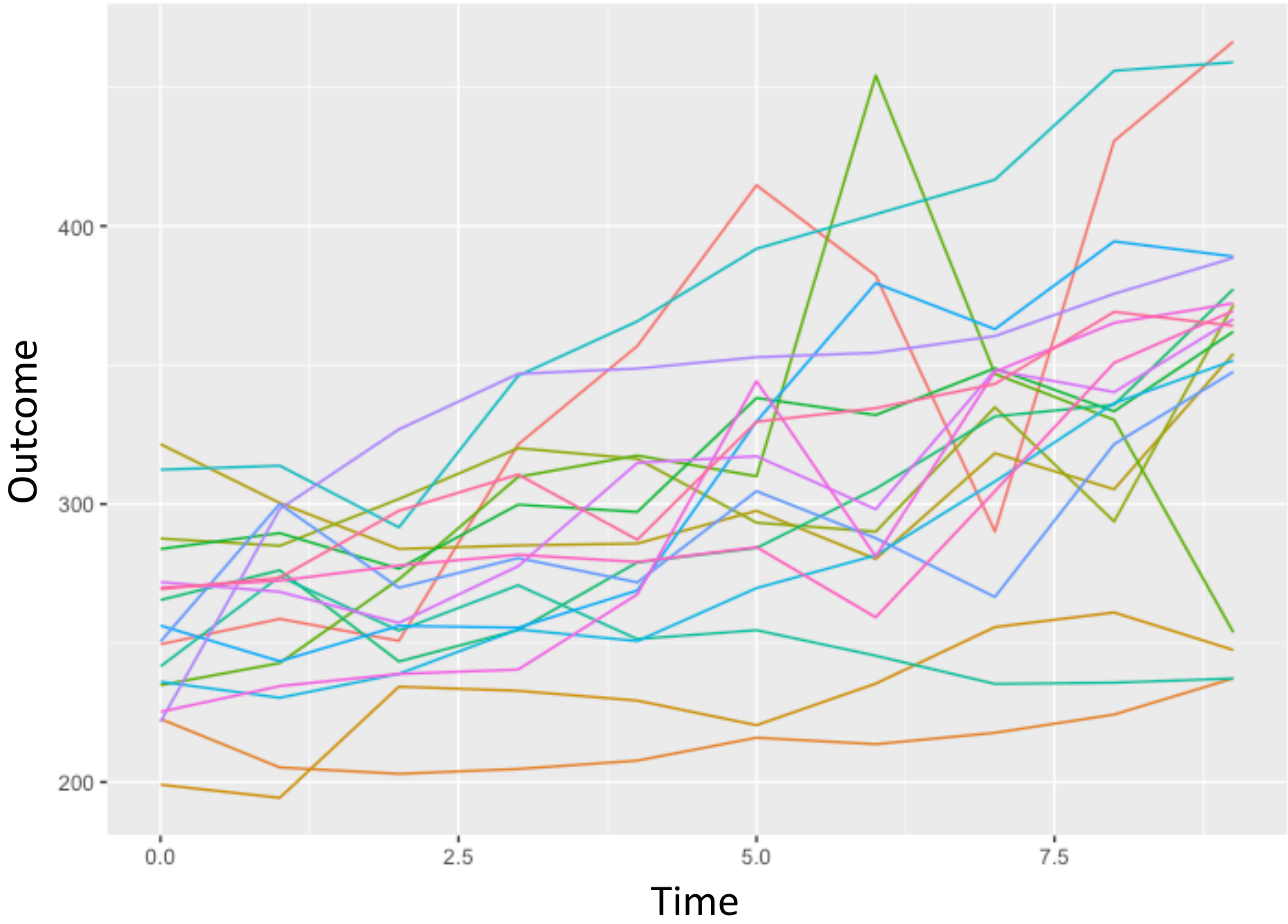
- If sphericity is violated, then type I errors are inflated and interaction term effects biased **– that is serious**
- ⚠ Mauchly's test may not reject sphericity if the sample size is small, even if the variances are vastly different

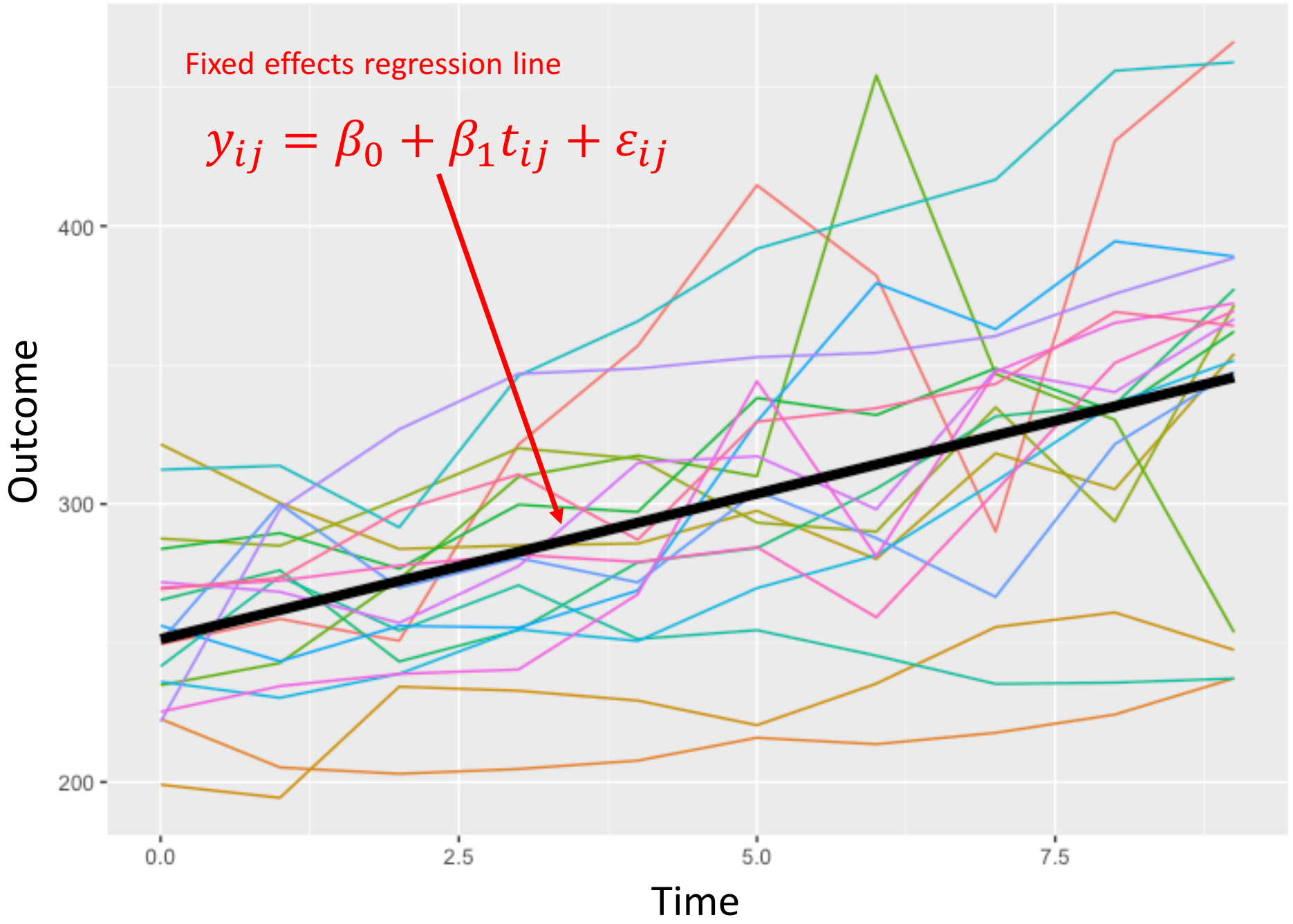
Correction proposal:

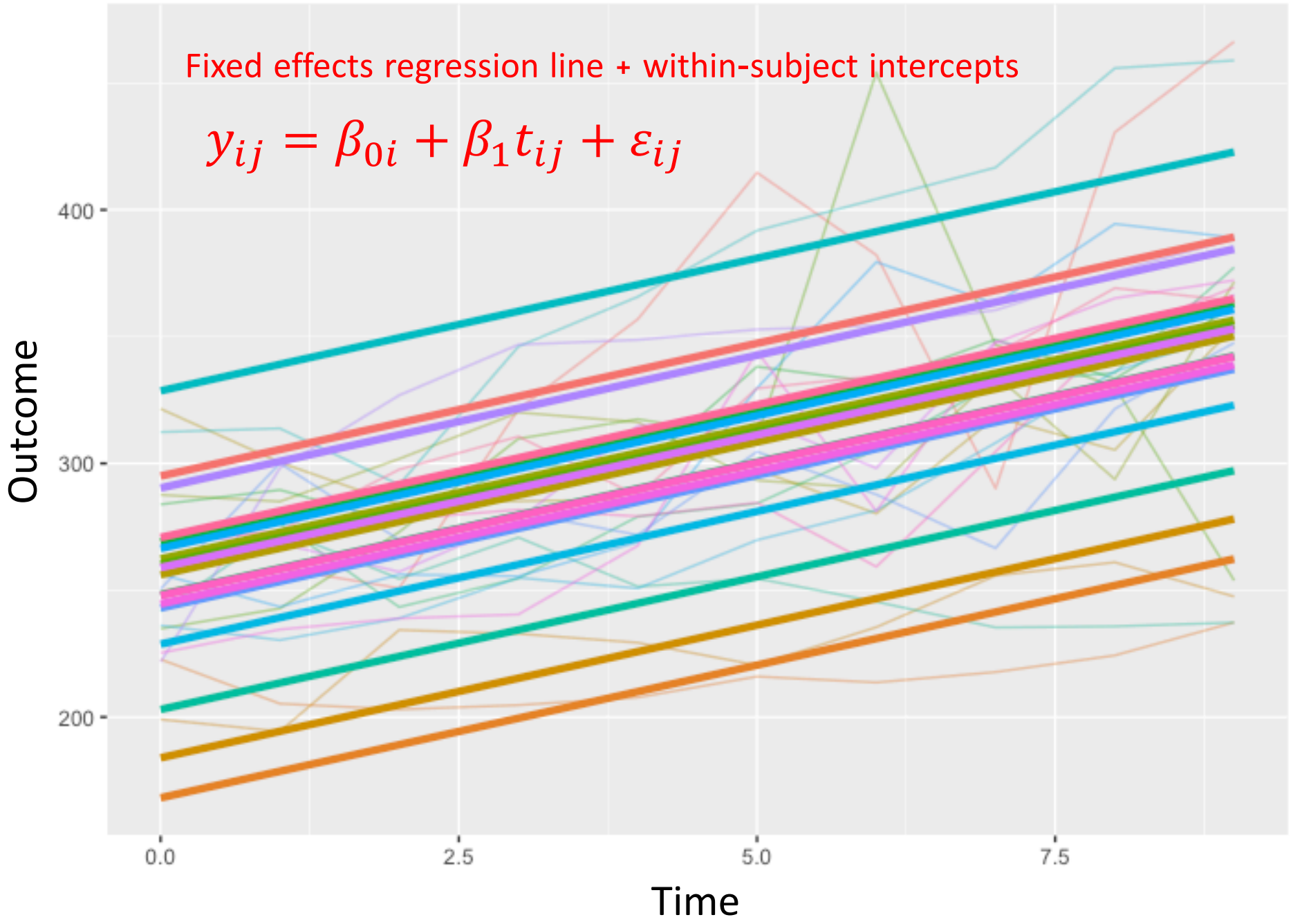
1. Calculate the epsilon statistic
   i. Greenhouse-Geisser
   ii. Huynh-Feldt
2. Multiply the *F*-statistic degrees of freedom by epsilon

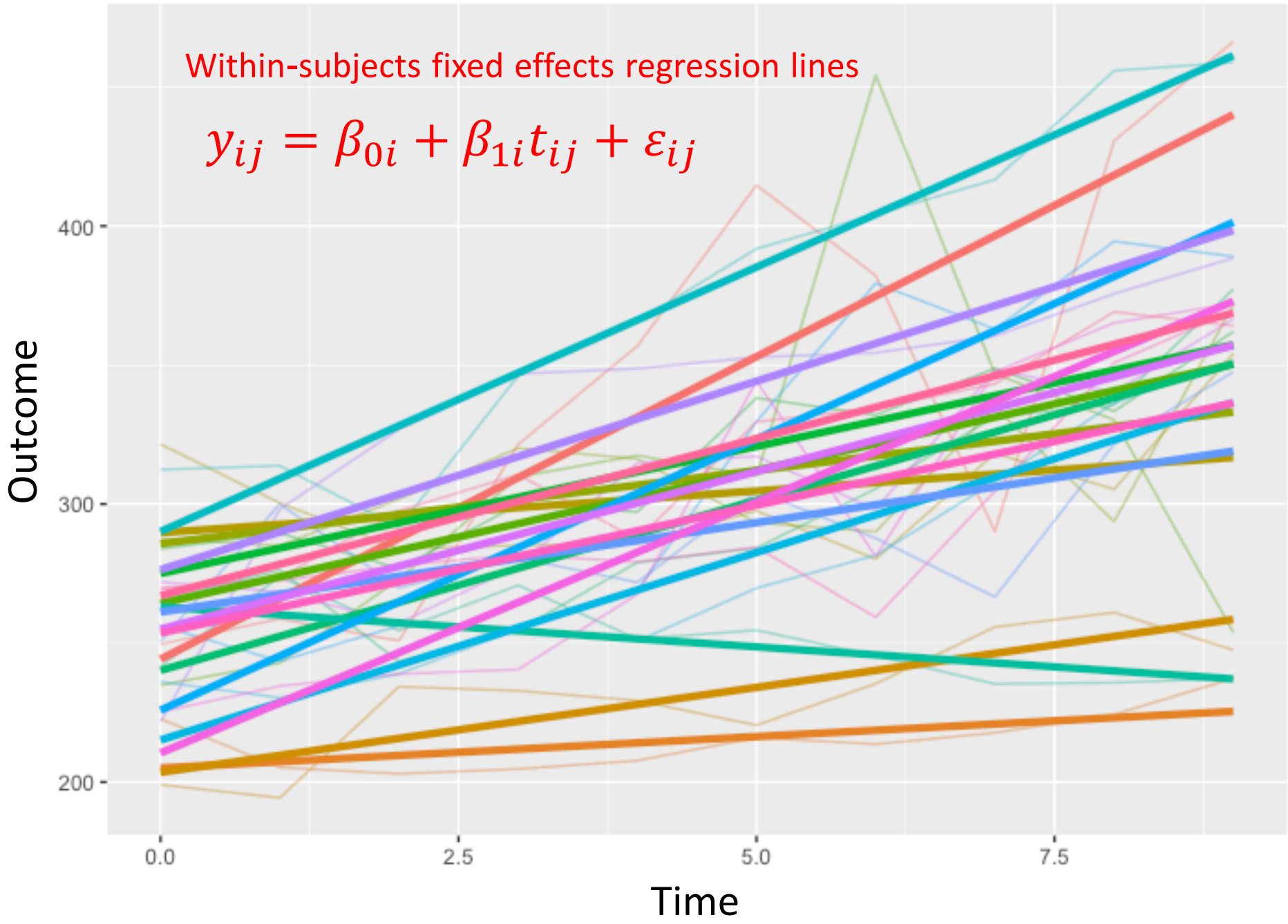
Linear mixed models

- Generalizes linear regression to account for correlation in repeated measures within subjects
- Also described as random effects models, mixed effects models, random growth models, multi-level models, hierarchical models, …


Outcome
400
300
200
0.0
2.5
5.0
7.5
Time


Fixed effects regression line
$y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}$
Outcome
400
300
200
0.0
2.5
5.0
7.5
Time


Fixed effects regression line + within-subject intercepts
$y_{ij} = \beta_{0i} + \beta_1 t_{ij} + \varepsilon_{ij}$
Outcome
400
300
200
0.0
2.5
5.0
7.5
Time


Within-subjects fixed effects regression lines
$y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \varepsilon_{ij}$
Outcome
400
300
200
0.0
2.5
5.0
7.5
Time

Linear mixed models

- A compromise is the model

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{ij}$$

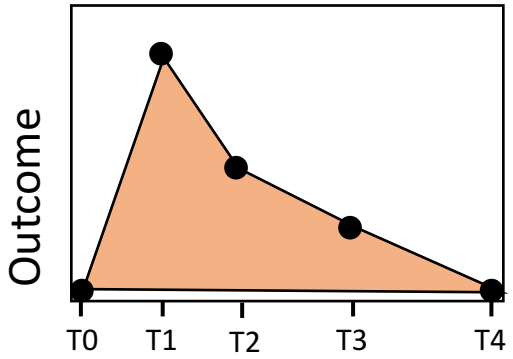
- (b_{0i}, b_{1i}) are called subject-specific random intercepts: intercept and slope respectively, distributed $N_2(0, \Sigma)$
- Observations *within*-subjects are more correlated than observations *between*-subjects
- Can be adjusted for other (possibly time-varying) covariates and baseline measurements

Summary statistics

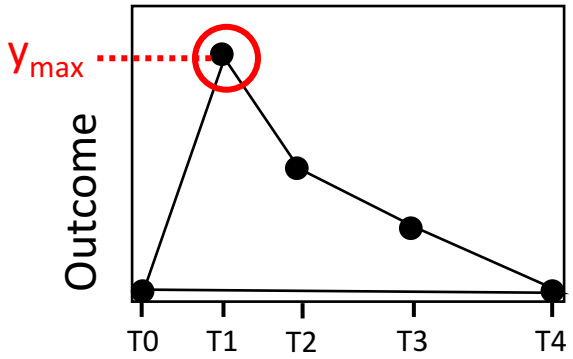
- A two-stage approach:
  1. Reduce the repeated measurements for each subject to a single value
  2. Apply routine statistical methods on these summary values to compare treatments, e.g. using independent samples *t*-test, ANOVA, Mann-Whitney *U*-test, …
- Benefits
  - Easy to do, and conceptually easy to understand
  - Can be used to contrast different features of the data
  - Encourages researchers to think about the features of the data most important to them in advance
- Choice of summary statistic depends on the data

If the data display a ‘peaked curve’ trend…

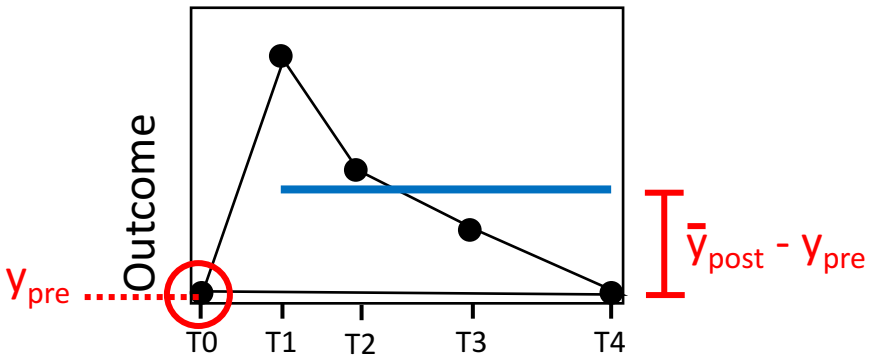
Area under the curve



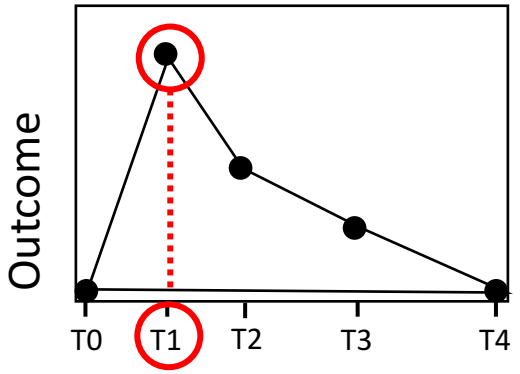
Maximum measurement



Mean follow-up – baseline

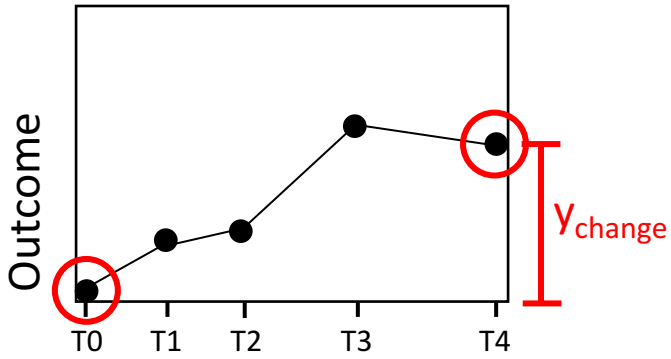


Time to reach maximum

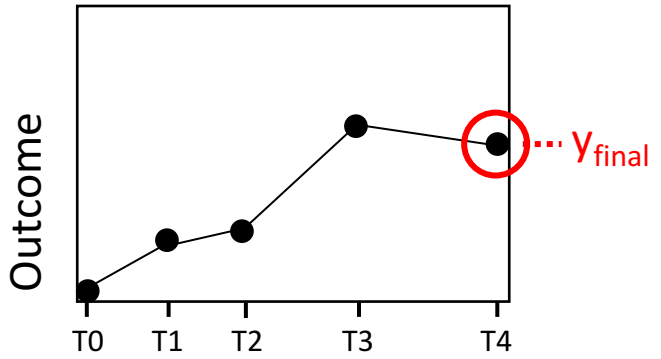


If the data display a ‘growth curve’ trend…

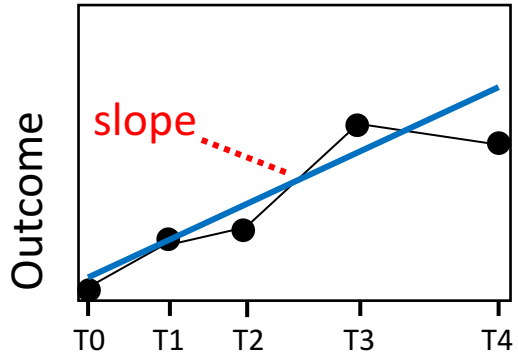
Change score



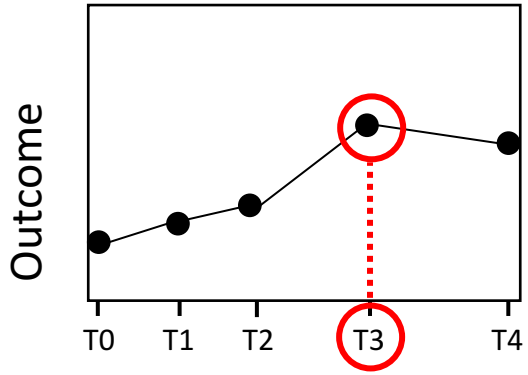
Final value



Slope



Time to a certain % increase/decrease



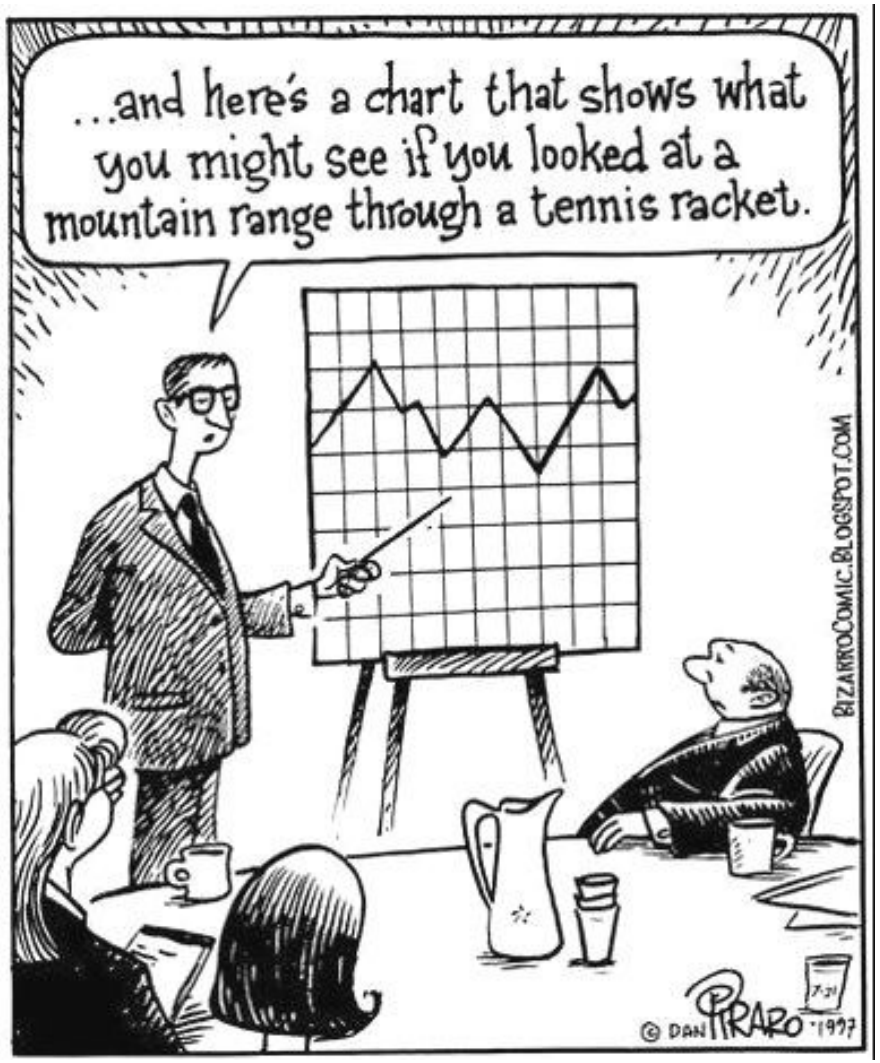
Missing data

Method	Can it handle missing data?	Can it handle unbalanced data?
RM-ANOVA	No – typically exclude patients with 1 or missing value	No
LMM	Yes – for data that is missing (completely) at random	Yes
Summary statistics	Depends on the choice of summary statistic	Depends on the choice of summary statistic

Software

- All methods implemented in standard statistical software

- Summary statistics usually require ‘manual’ calculation, but can be done easily in Microsoft Excel or programmed in a statistics software package



Thank you for listening…

any questions?

Statistical Primer article to be published soon!

Slides available (shortly) from: www.glhickey.com