Day 8: Repeated measurements and clustered data

Paul Blanche

Section of Biostatistics, University of Copenhagen

Mars 19, 2025

Outline/ILOS

Scientific & Statistical reasoning:

ILO: to recognize repeated measurements and clustered data and list contexts in which we meet them.
ILO: to distinguish between situations in which simple analyses are sufficient and situations where they are not.

Statistical methods

ILO: to summarize these data to provide appropriate descriptive statistics/plots and to make simple inference.
ILO: to examplify why missing data is a common challenge with repeated measurements and deal with it.
ILO: to fit a random effect model and interpret the results.
ILO: to fit a Mixed Model for Repeated Measurements (MMRM) and interpret the results.

What and why clustered data?

We say we have clustered data when the same outcome is measured on groups of subjects/animals that are somehow related. They share a heritage, an environment, or a more or less random experimental condition.

Examples:

- Rats from the same litter/cage
- Children from the same family/school
- Patients from the same hospital
- Follicles from the same woman
- Febrile episodes from the patient
- Knees from the patient

Why are they common? Often easier to collect the data.

Why does clustering matter?

Example: when studying episodes of febrile neutropenia in high-risk children with leukemia, we can collect data from several episodes of the same children. In that case, we might expect that the outcomes from two febrile episodes of the same child are more likely similar than the outcomes of two episodes from two different children. That is, we cannot rule out within-patient correlation between episodes.

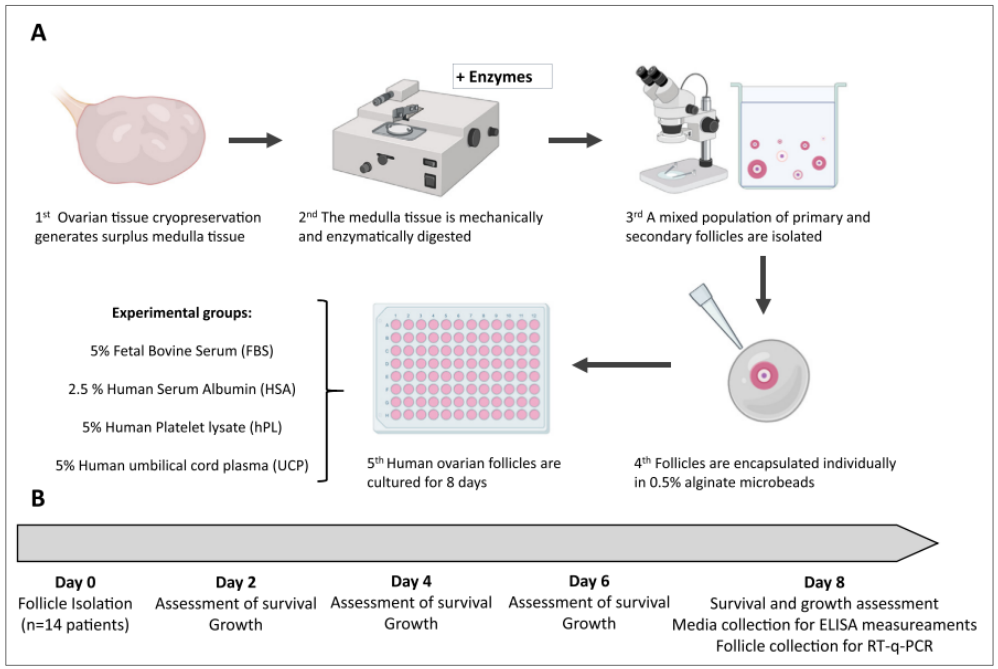
General concept: we cannot rule out that the outcome from two patients/animals from the same cluster are systematically more (or less) similar than the outcome of two patients/animals from two different clusters. This implies that within-cluster correlation might exist and that the usual "independence" assumption assumed by most "standard" statistical methods might be wrong (and potentially seriously wrong).

Consequently, estimates, standard errors, 95% confidence intervals and p-values produced by "standard" methods ignoring the clustering might be unreliable.

Case: human follicle data

Data: n=724 follicles, from J=14 patients (diameter in μm; NA if follicle is dead).

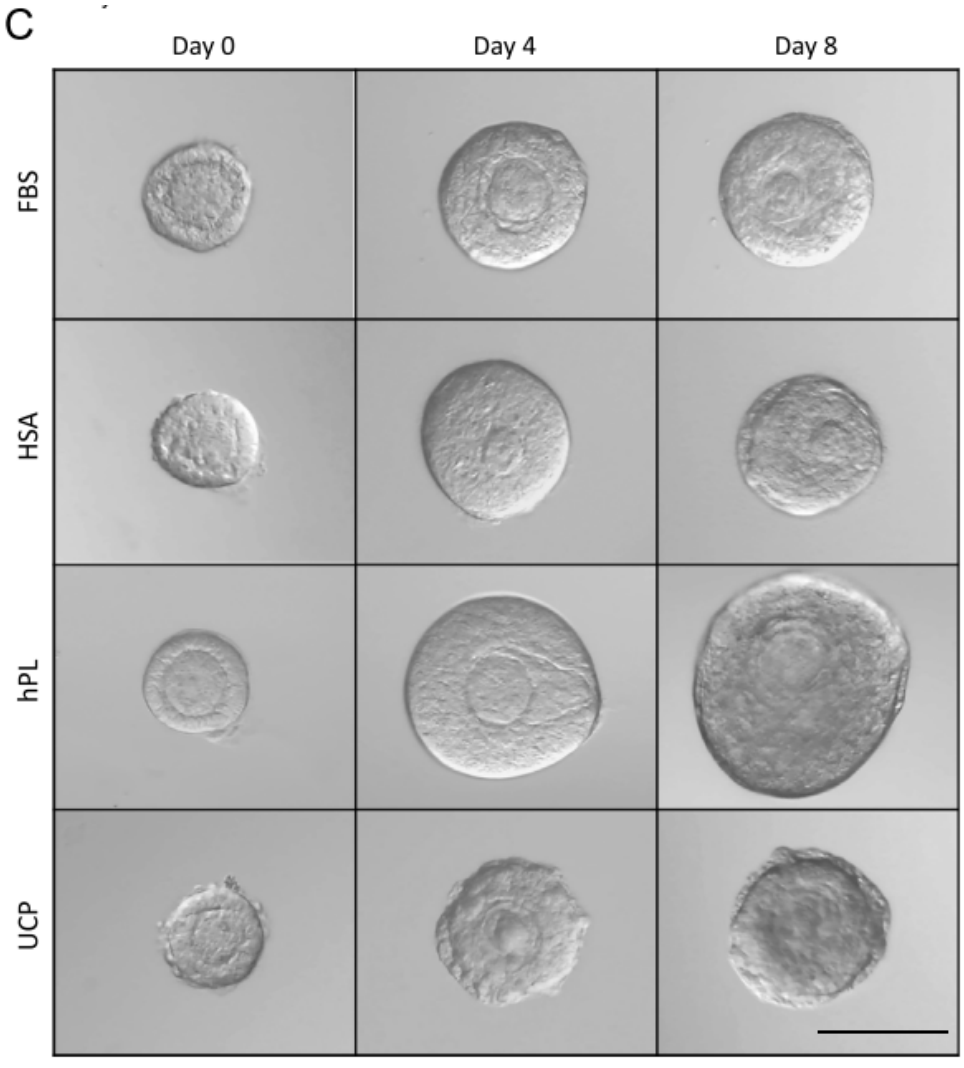
	Patient	Treatment	Day0	Day8
1	1	FBS	101.4590	161.500
2	1	FBS	89.8315	NA
3	1	FBS	90.2835	129.447
4	1	FBS	120.3145	170.740
5	1	FBS	93.0085	120.940
6	1	FBS	73.0530	109.947



Research question: How do platelet-rich plasma products like human platelet lysate (HPL), human serum albumin (HSA), fetal bovine serum (FBS) and umbilical cord plasma (UCP) affect the growth of isolated human pre-antral follicles in vitro?

Ref: Adrados et al. Reproductive BioMedicine Online 47.5 (2023): 103256.

Growth/size is defined via the diameter



The follicles were cultured at 37°C and 5% CO_2 for 8 days. Every second day, one-half of the culture media (50 μl) was exchanged with fresh media and survival and growth were assessed. The diameter was defined as the mean of two perpendicular lines drawn through the middle of the follicle, using the basal membrane as a measure. A

Ref: Adrados et al. Reproductive BioMedicine Online 47.5 (2023): 103256.

Random effect model, aka "Mixed model"

Model for the outcome "$\log_2$-growth at day 8" of the j-th follicle of the i-th woman:

$$Y_{ij} = \mu + \mathtt{x}_{ij}\beta_1 + \mathtt{z}_{ij}\beta_2 + \mathtt{u}_{ij}\beta_3 + \mathtt{v}_{ij}\beta_4 + a_i + \varepsilon_{ij}$$

There are "fixed" effects and "random" effects (hence the term "mixed"). The random variation is modeled using two random terms,

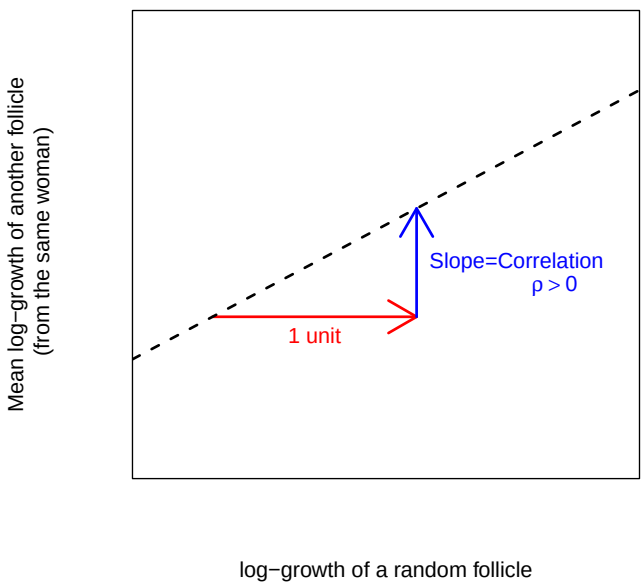
$$a_i \sim N(0, \omega_B^2) \quad \text{and} \quad \varepsilon_{ij} \sim N(0, \tau_W^2) \ .$$

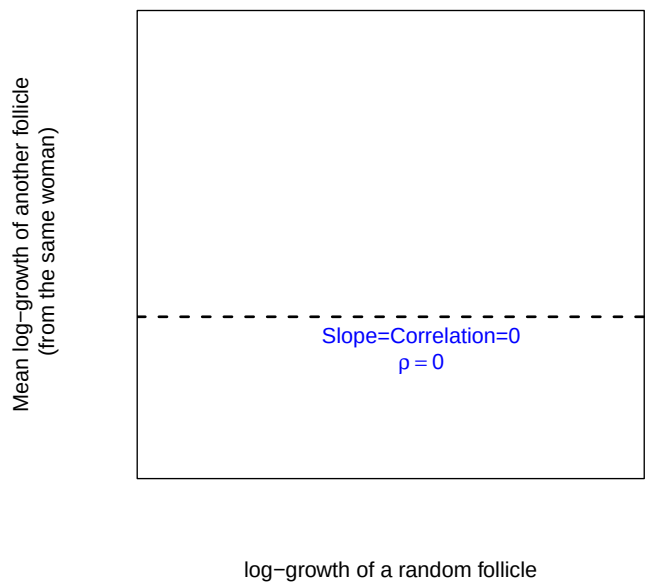
- The variances ω_B^2 and τ_W^2 are called variance components
- ω_B^2 is the variance **B**etween clusters (here, between women) and τ_W^2 the variance **W**ithin clusters. We say that the total variance is $\sigma_T^2 = \omega_B^2 + \tau_W^2$ (sum of variance "explained by the woman" and "unexplained" variance).

- Data from two different women are assumed to be independent
- Data from two follicles of the same woman are **NOT** assumed to be independent, but correlated. The correlation is modelled by

$$\rho = \frac{\omega_B^2}{\sigma_T^2} = \frac{\omega_B^2}{\omega_B^2 + \tau_W^2}$$

and it is called the intra-class correlation. Here, it represents the correlation between two follicles of the same woman.





Mixed model assumption versus "standard" linear model assumption

Interpretation/details

- $\mathrm{x}_{ij} = 1$ (resp. $\mathrm{z}_{ij} = 1$, $\mathrm{u}_{ij} = 1$) if the j-th follicle of the i-th woman is grown using HPL (resp. HSA, UCP), 0 otherwise (i.e., reference is FBS).
- v_{ij} is the ($\log_2$ of the) baseline size (at day 0), of the j-th follicle of the i-th woman, minus $\log_2(75)$ (i.e., minus the $\log_2$ of the average baseline size of 75μm).
- β_k for $k = 1, \ldots, 3$: the mean difference in the $\log_2$-growth at day 8 (outcome) between two follicles, either of the same woman or of two different women, having the same baseline size, one is grown using HPL (when $k = 1$, HSA if $k = 2$, UCP if $k = 3$) and the other is grown using FBS (the reference plasma product).
- β_4: the mean difference in outcome between two follicles, either of the same woman or of two different women, one having a $\log_2$ baseline size of $v + 1$, the other of v, when both follicles are grow using the same plasma product (i.e., same group).
- μ (intercept): the mean outcome in the reference group FBS, for a follicle of the average baseline size (75μm), for an average woman.
- a_i: the woman's random effect, i.e., the (average) deviation in $\log_2$-growth of a random follicle from the i-th woman to that of the average woman, when comparing two follicles grown with the same (plasma product) group and baseline size (i.e. random variation "explained" by the woman).
- ε_{ij}: "error" term, i.e., random variation neither explained by the woman nor by the covariates. Random deviation in $\log_2$-growth of the j-th follicle to the average follicle of the same woman, when comparing two follicles grown with the same (plasma product) group and of the same baseline size.

Note: data should be in the "long" format, see slide for other case study.

R code:

```
library(lmerTest)
fitlmer <- lmer(loggrowth ~ Treat + logDay0 + (1|PatientID), data=d)
summary(fitlmer)
```

Output (partial):

```
Random effects:
 Groups    Name        Variance Std.Dev.
 PatientID (Intercept) 0.02922  0.1709
 Residual              0.16535  0.4066
Number of obs: 401, groups:  PatientID, 14

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  5.33230    0.06444 31.40511  82.743  < 2e-16 ***
TreathPL     0.86209    0.05570 391.56039 15.476  < 2e-16 ***
TreatHSA     0.10280    0.06550 395.99353  1.570    0.117
TreatUCP     0.42651    0.07524 389.59172  5.668 2.81e-08 ***
logDay0      0.77064    0.04868 395.61392 15.832  < 2e-16 ***
```

Note: 401 out of 724 follicles are still alive by day 8; the dataset `d` here only contains data from the survivors.

`Fixed effects:`

- `(intercept)`: 5.332 is the estimated value of μ
- `TreathPL`: 0.862 is the estimated value of β_1
- `TreatHSA`: 0.103 is the estimated value of β_2
- `TreatUCP`: 0.427 is the estimated value of β_3
- `logDay0`: 0.771 is the estimated value of β_4

`Random effects:`

- `PatientID (Intercept)`: 0.1709 is the estimated value of ω_B
- `Residual`: 0.4066 is the estimated value of τ_W

Note: we can further deduce $\sigma_T = \sqrt{0.1709^2 + 0.4066^2} = 0.441$ and (with the σ_T^2 being the total variance) and the intra-class correlation $\rho = 0.1709^2/0.441^2 = 0.15$.

Another illustration of intra-class correlation (simulated data)

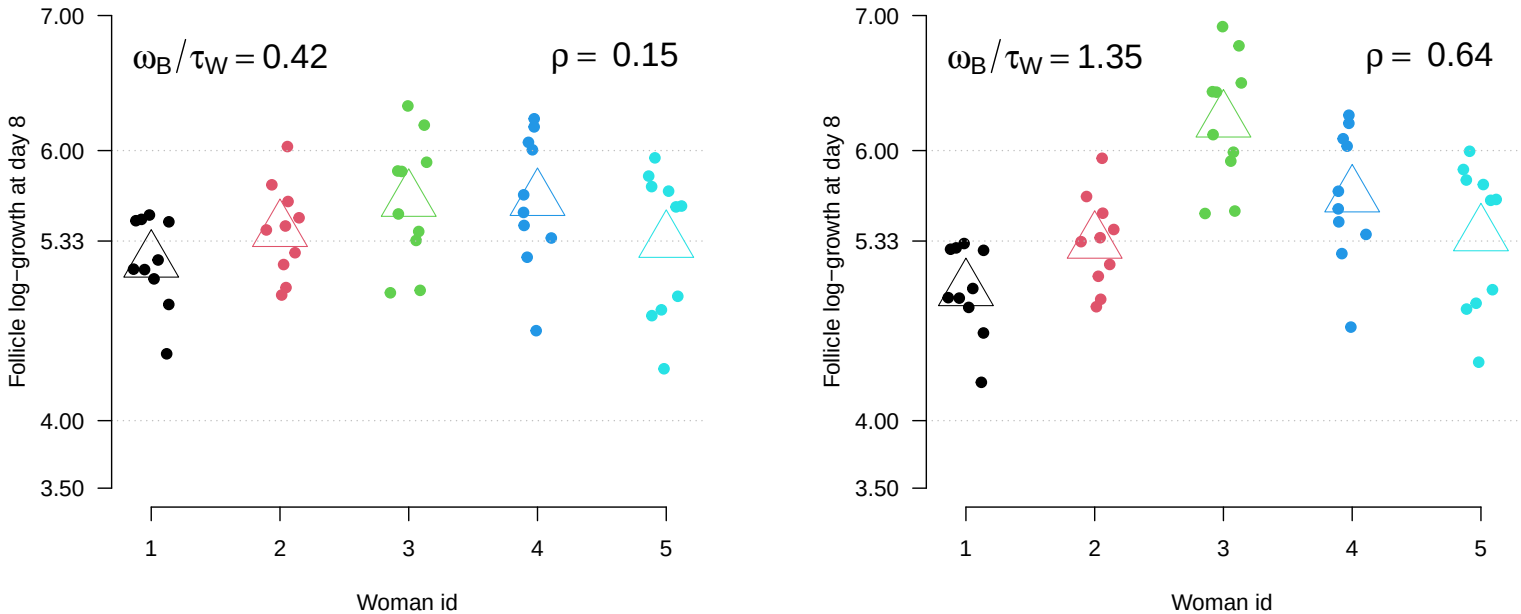


Illustration. Left panel corresponds to the fitted model; right is for comparison, with larger intra-class correlation ρ. Shown are 10 follicles for 5 women, assuming all follicles have the same treatment (FBS) and size at day 0 ("average" log baseline size). The larger the "between" variance ω_B^2 relative to the "within" variance τ_W^2, i.e., the higher the intra-class correlation ρ, the larger the difference between the means per woman (triangles) relative to the distances between the differences between two follicles of the same woman (the dots of the same color).

Results: `lm` ("naive") versus `lmer` (recommended)

R code:

("standard" linear model, lm)

```
fit1 <- lm(loggrowth ~ Treat + logDay0 ,
                data=d)
summary(fit1)
```

(linear "mixed" model, lmer)

```
fitlmer <- lmer(loggrowth ~ Treat + logDay0
            + (1|PatientID), data=d)
summary(fitlmer)
```

Output (partial):

```
            Estimate Std. Error Pr(>|t|)
(Intercept)  5.34283    0.04734 < 2e-16 ***
TreathPL     0.85701    0.05806 < 2e-16 ***
TreatHSA     0.17032    0.06515 0.00928 **
TreatUCP     0.41879    0.07916 2.03e-07 ***
logDay0      0.72752    0.04980 < 2e-16 ***
---

Residual standard error: 0.4387
```

```
            Estimate Std. Error Pr(>|t|)
(Intercept)  5.33230    0.06444 < 2e-16 ***
TreathPL     0.86209    0.05570 < 2e-16 ***
TreatHSA     0.10280    0.06550 0.117
TreatUCP     0.42651    0.07524 2.81e-08 ***
logDay0      0.77064    0.04868 < 2e-16 ***
```

$\sqrt{\text{total variance}}$ (i.e., σ_T)=0.441.

Note: not all women have follicles grown in all products. Especially, three women have no follicles grown in HSA and additional results (unshown) suggest that follicles of these women were estimated to grow less than those of an average woman. The mixed model somehow accounts for this random unbalance, resulting from random variation.

Digression: further interpretation of variance components

Typical difference between $\log_2$-growth of two follicles of **the same woman**, when comparing two follicles grown with the same plasma product and of the same baseline size:

- within: $\pm 1.96\sqrt{2 \times \omega_B^2} = \pm 1.96\sqrt{2 \times 0.17^2} = \pm 0.48$ (in log μm)

Typical difference between $\log_2$-growth of two follicles of **different women**, when comparing two follicles grown with the same plasma product and of the same baseline size:

- within: $\pm 1.96\sqrt{2 \times (\underbrace{\tau_W^2 + \omega_B^2}_{=\sigma_T^2})} = \pm 1.96\sqrt{2 \times (0.17^2 + 0.41^2)} = \pm 1.23$

Note: we multiply the variances by 2 because there are 2 women (error terms, hence variances, add up). Here, "within" means the width of the 95% prediction interval or, equivalently, of the normal range.

Results back transformed (original scale)

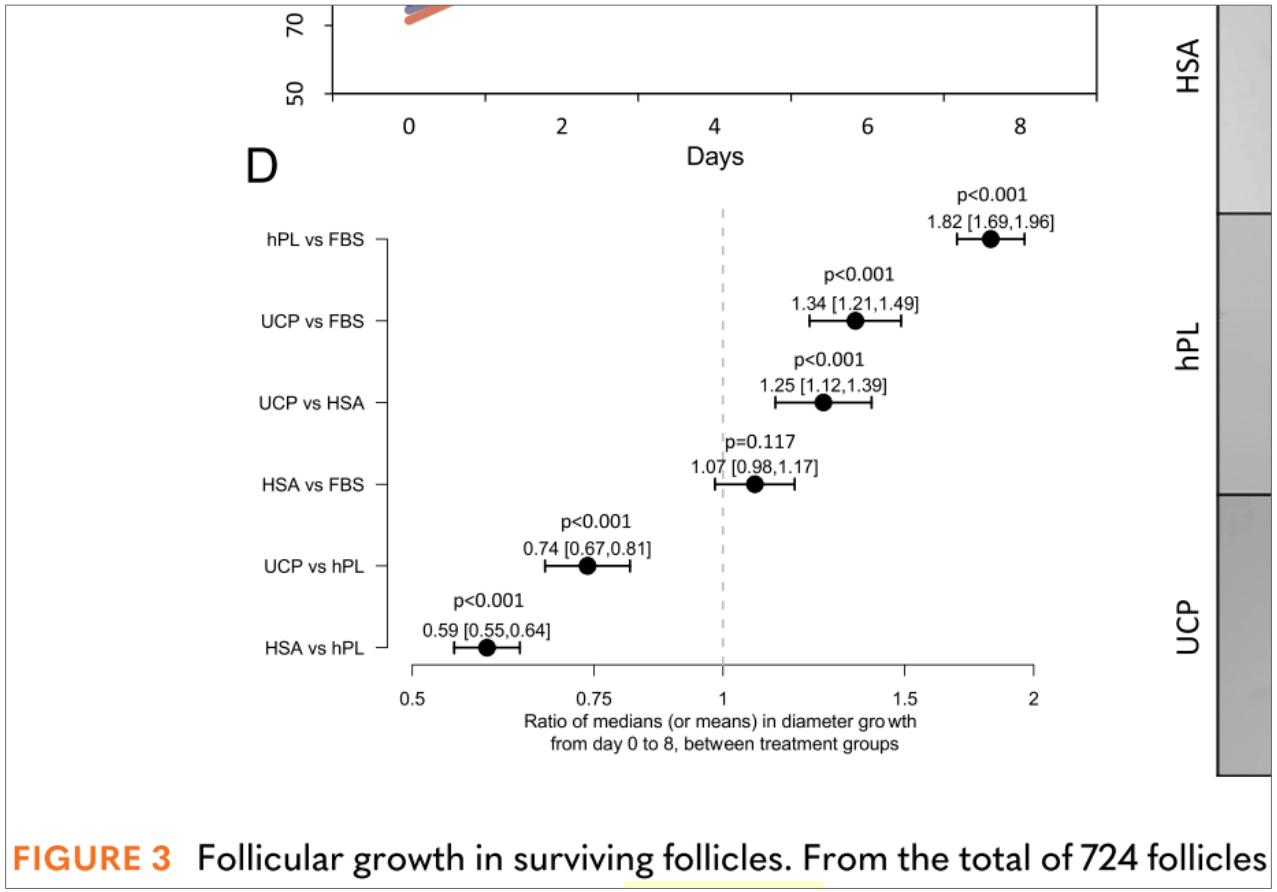


FIGURE 3 Follicular growth in surviving follicles. From the total of 724 follicles

- For example, $2^{\hat{\beta}_1} = 2^{0.86209} = 1.82$ (hPL vs FBS)

Ref: Adrados et al. Reproductive BioMedicine Online 47.5 (2023): 103256.

Does clustering always matter?

- No! Sometimes we can reasonably expect that the variance between clusters (ω_B^2) is negligible as compared to the variance within cluster (τ_W^2), resulting in an intra-class correlation (ρ) so small that it can be considered as zero for all practical purposes.
- But, making this assumption requires that we are ready to carefully defend its rationale. Skeptical/thorough reviewers will require the rationale! And sometimes the random effect model analysis as a sensitivity analysis too...
- In case of doubt, it is usually recommended to *"let the data speak freely"*, hence use a mixed model that allows for intra-class correlation $\rho > 0$.

Example: A similar study was conducted with follicles coming from mice and the intra-class correlation was considered negligible in the statistical analysis. The rationale was that there was not much phenotypical variation from mouse to mouse because both the genotype and the environment were similar for all included mice. The mice were all fed similarly, kept in the same cage, of the same age and coming from the same litter. This is unlike with human data, because individuals included in typical human studies (such as our case study) are coming from different genotypical backgrounds, are exposed to different environments, have different lifestyles and have different ages and comorbidities.

Digression: truncation by death (1/2)

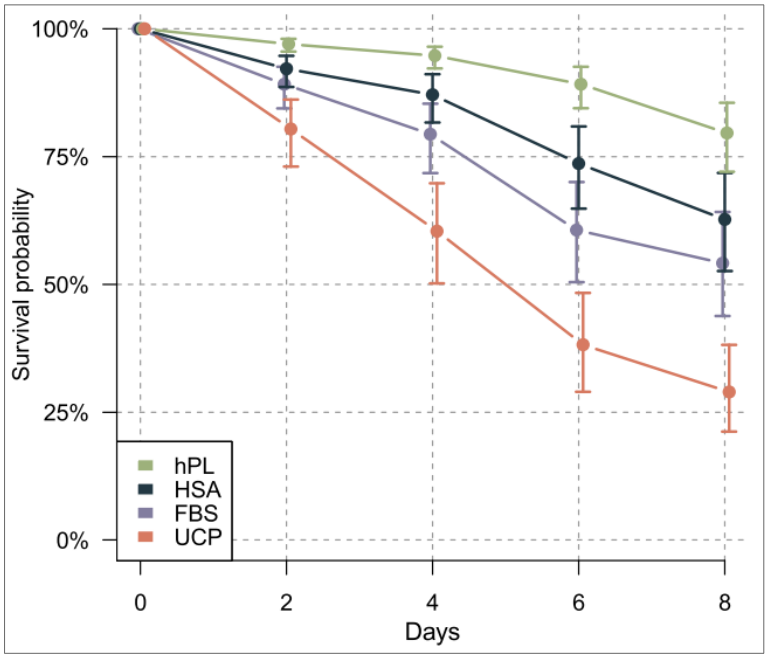
Here we studied and compare follicle growth among follicles alive after 8 days. Here it is not important that many follicles die, as long as many survive and grow well. This is because (as I understood it) the ultimate goal of this research is to improve the way we can obtain a few good follicles to transfer to a woman in needs of infertility treatment.

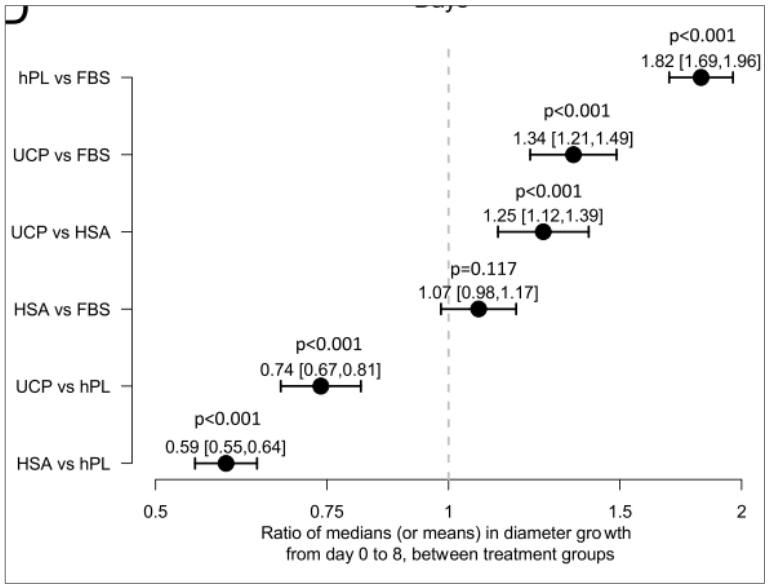
In some other contexts, looking at a change in outcome among the survivors only might be misleading. For instance, comparing improvements in Quality of Life (QoL) scores might be misleading if survival rates differ between the two groups that we compare. Often, comparing outcomes among the survivors makes sense only if the survival rates are similar in both groups[1].

Statistical jargon: we often say that an outcome is truncated by death if this outcome does not exist (i.e., "is not defined") for patients who die (e.g., QoL).

[1] or if the differences is in the same direction for survival rate and for the change in outcome (e.g., QoL) among the survivors

Digression: truncation by death (2/2)





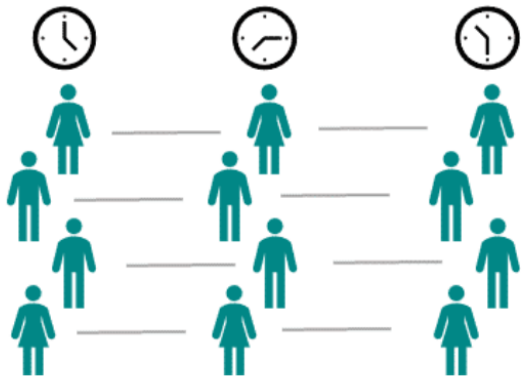
In the case study, we can conclude that UCP "works better" than "FBS", because of better growth (in average), even though the survival rate is less good. This is because sufficiently many follicles survive.

If we were comparing "QoL in patients" instead of "growth in follicles", the conclusion would be rather different!

Note & ref: *''Follicular survival was analysed by fitting a logistic mixed model with follicle survival as the outcome, patient as a random effect and experimental group as a fixed effect.''* (Adrados et al, 2023) . This is a "similar" mixed model, but the presentation of this model is beyond the objective of this course.

What are repeated measurements data?

Repeated measurements usually refer to data where the same outcome has been measured several times on the same subject.



Examples:

- follicle growth after 2, 4, 6 and 8 days (previous case study).
- GLP-2 concentration [2] in the blood 10, 20, 30,. . . and 240 mins after food intake (next case study).
- score at 6 and 12 weeks after treatment initiation (last case study).

[2] GLP-2: Glucagon-like peptide 2

Why are they common?

Often, either...

They are needed: having repeated measurements allow health science researchers to study changes over time within the same subjects and factors that influence them such as treatment.

or...

They are "easy"/"cheap" to obtain: often, the main outcome is measured after some time, at end of follow-up (e.g., 8 days in the follicle growth case study, 12 weeks in the next case study). It is often relatively easy and not much additional effort to collect additional data for the same outcome at a couple of earlier timepoints.

The stat analysis does not always need to advanced!

- A relevant outcome to analyze (and sometimes the most relevant) might be a single value that summarizes all the repeated measurements of the same patient.
  - e.g., GLP-2 next case study.

- Sometimes we have collected repeated measurements mostly because we could, not because we needed. E.g., when the main outcome is the repeated measurement at end of follow-up.
  - e.g., follicle growth in the previous case study.

IMPORTANT: in these two cases, the main statistical analysis can often be performed using"simple" statistical methods that are not specific to repeated measurements data (e.g., linear model, ANOVA, ANCOVA).

Case: GLP-2 stimulation data

Data: GLP-2 concentrations (pmol/L) of n=10 patients at 10, 20, 30,. . . and 240 mins after food intake and 15 mins before, when eating three different meals (first columns shown).

ID	meal	GLP2.minus.15	GLP2.10	GLP2.20	GLP2.30
1	High carbohydrate	1	18	38	44
1	High fat	1	30	27	24
1	High protein	8	7	4	13
2	High carbohydrate	4	21	25	23
2	High fat	1	1	13	8
2	High protein	2	16	23	31

Research question: Does the stimulation of GLP-2 secretion differ after eating carbohydrate, fat or protein enriched (iso-energetic) meals?

Ref: Prahm et al. Peptides. 2023 Nov 1;169:171091.

Table 2
Dietary components of the three isoenergetic meals.

High carbohydrate meal	High fat meal	High protein meal
Wholemeal rye bread, 25 g	Wholemeal rye bread, 20 g	Cod's roe, 200 g
Whole grain bread, 50 g	Butter (salted), 15 g	Shrimps, 300 g
Cheese (30+), 40 g	Cheese (45+), 60 g	Crispbread (wholegrain), 22 g
Raspberry jam, 20 g	Crispbread (wheat), 10 g	Cheese (30+), 40 g
Hazelnut-spread/nutella, 20 g	Hazelnuts, 45 g	Ylette (lowcalorie dairy product), 270 g
Ensini (nutrition supplement with 89 % carbohydrates), 200 g	Oatmeal, 25 g	Whey protein powder, 36 g
Fruit-youghurt, 200 g	Ymer/junket (dairy product - 44 % energy from fat), 200 g	Lemon juice, 20 g

Ten healthy subjects were admitted on three occasions, at least a week apart, after a night of fasting. In an open-label, crossover design, they were randomized to receive a high carbohydrate (HC), high fat (HF) or high protein (HP) meal. The meals were approximately ~3.9 MJ. Venous blood was collected for 240 min, and plasma concentrations of GLP-2, GIP and PYY were measured with specific radioimmunoassays.

Ref: Prahm et al. Peptides. 2023 Nov 1;169:171091.

How to define simple yet relevant outcomes?

- The main interest was in the stimulation, hence this is the change in concentration as compared to before food intake, that defines the repeated measurements outcome of primary interest.
- The main comparison was based on the area under the curve (AUC) of the change from baseline. This is a simple summary of change in concentration over time, which is sufficient to answer the research question.
- Secondary interests were in the stimulation within the first 60 mins and after 60 mins (and up to 240 mins). AUCs using the corresponding time intervals was therefore used too.

Common summary outcomes (sometimes relevant, not always)

- outcome at end of follow-up (last measurement occasion)
- change at end of follow-up (i.e., difference between outcomes at end of follow-up and before intervention; common in e.g., RCT for pain/functioning scores or weight loss outcomes)
- area under the curve (AUC, common in e.g., pharmacokinetics; it is a "weighted average".)
- average of all responses ($\approx$ same as AUC if equal time intervals between measurement occasions)
- minimum/maximum (e.g. PSA nadir)
- time to reach a specific value (e.g., time to max or time to 30% decrease)

Appendix I

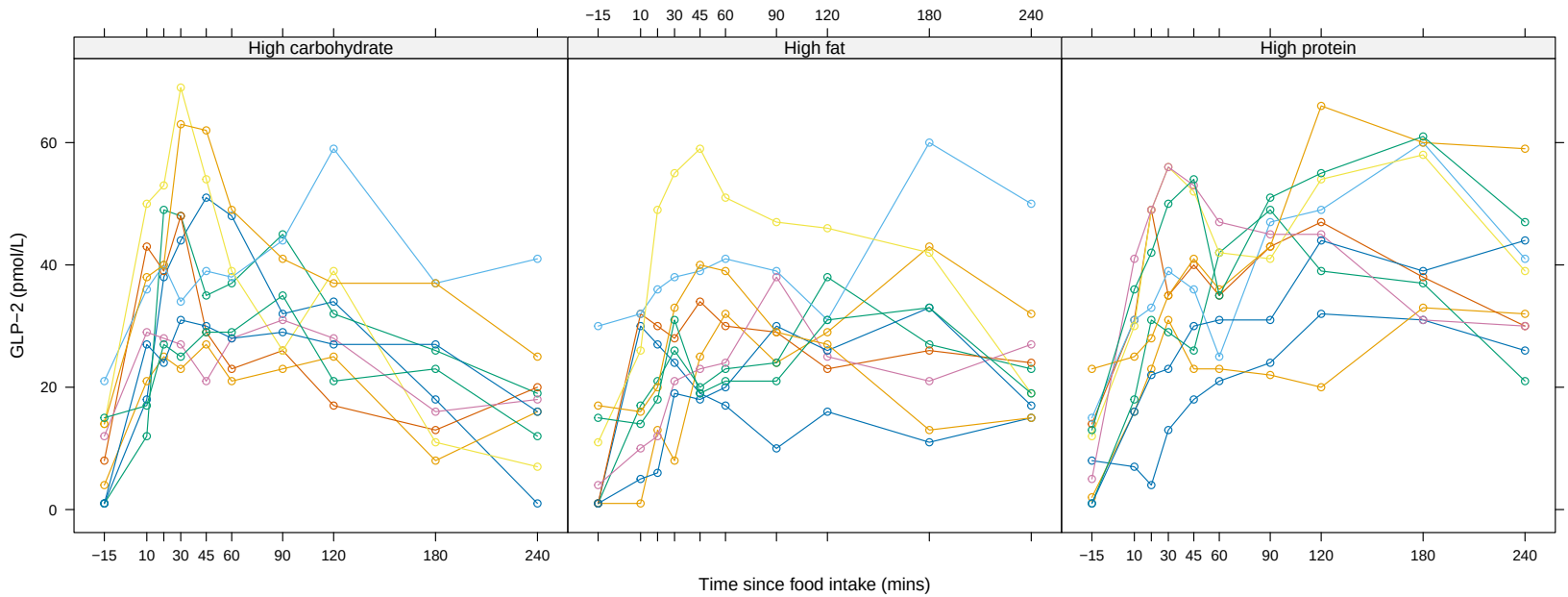
Some summary measures

Type of data	Question to be answered	Summary measure
Peaked	Is the overall value of the outcome variable the same in different groups?	Overall mean (equal time intervals) Area under curve (unequal time intervals)
Peaked	Is the maximum (minimum) response different between groups?	Maximum (minimum) value
Peaked	Is the time to maximum (minimum) response different between groups?	Time to maximum (minimum) response
Growth	Is the rate of change of the outcome variable different between groups?	Regression coefficient
Growth	Is the eventual value of the outcome variable the same between groups?	Final value of outcome measure or difference between last and first values, or percentage change between first and last
Growth	Is the response in one group delayed relative to the other?	Time to reach a particular value (for example, a fixed percentage of baseline)

Ref: Matthews et al. "Analysis of serial measurements in medical research." British Medical Journal 300.6719 (1990): 230-235.

Spaghetti plot

An often useful descriptive plot of repeated measurements data, especially with small sample sizes, is the Spaghetti plot. Observations from the same subjects are linked, to emphasize the correlation structure of the data.



Note: it is a crossover study and here the same colors are used for the same subjects (that eat all three meals at different days, at least one week apart).

Digression: "wide" vs "long" data format

The original data we import from e.g. Excel is often in wide format:

ID	meal	GLP2.minus.15	GLP2.10	GLP2.20	GLP2.30
1	High carbohydrate	1	18	38	44
1	High fat	1	30	27	24
1	High protein	8	7	4	13
2	High carbohydrate	4	21	25	23
2	High fat	1	1	13	8
2	High protein	2	16	23	31

To fit mixed models or perform similar analyses (e.g., to produce Spaghetti plots), we often need the data in the long format:

ID	meal	time	GLP2
1	High carbohydrate	-15	1
1	High carbohydrate	10	18
1	High carbohydrate	20	38
1	High carbohydrate	30	44
1	High fat	-15	1
1	High fat	10	30
1	High fat	20	27
1	High fat	30	24
1	High protein	-15	8
1	High protein	10	7
1	High protein	20	4
1	High protein	30	13
2	High carbohydrate	-15	4

Note: we can go from one format to the other using the `reshape()` function of `R` (see R-demo).

From "complex" to "simple" data

In this case study, the correlation structure between the observations of GLP-2 concentrations is complex. The data are correlated because of:

- repeated measurements: data from the same patient at different times after food intake.
  - represented by the lines on the Spaghetti plot.
- cross-over design: three series of repeated measurements for each patient, one after each of the three meals.
  - represented by colors in the Spaghetti plot.

Defining the AUC as the primary outcome of interest reduces the complexity of the data to "simpler" paired data.

Case (con't): GLP-2 stimulation "paired" data

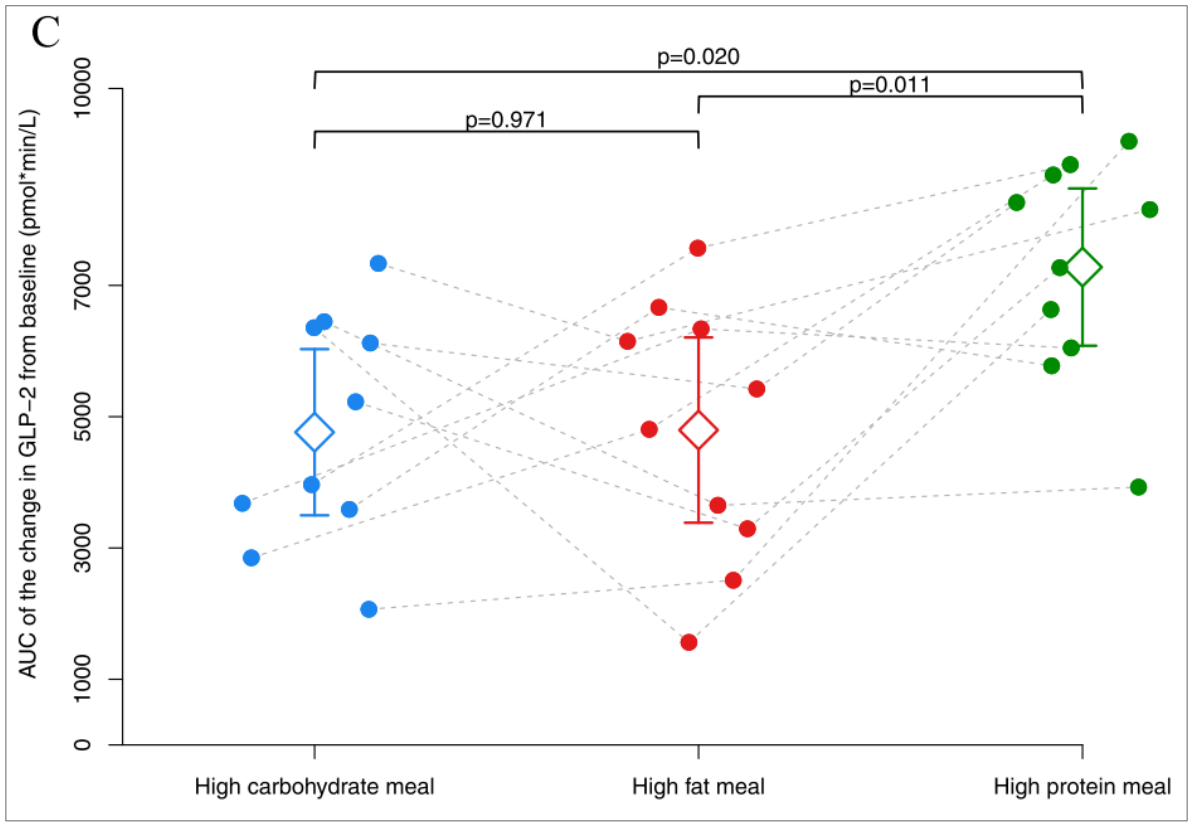
Data: AUC of the curve of the change of GLP-2 concentration within 240 mins (pmol.min/L) of n=10 patients, after eating three different meals at three occasions. Change refers to change from baseline (15 mins before food intake).

	High carbohydrate	High fat	High protein
1	6447.5	3650.0	3927.5
2	3587.5	6665.0	5775.0
3	7335.0	6147.5	8155.0
4	3680.0	6337.5	6047.5
5	5227.5	3292.5	7270.0
6	3965.0	7567.5	8840.0
7	2850.0	4807.5	8682.5
8	6122.5	5422.5	8262.5
9	6355.0	1562.5	6632.5
10	2065.0	2507.5	9195.0

Research question: Does the stimulation of GLP-2 secretion differ after eating carbohydrate, fat or protein enriched (iso-energetic) meals?

Ref: Prahm et al. Peptides. 2023 Nov 1;169:171091.

Shown are individual observations (AUCs), estimated mean AUC and 95%-CI for each meal (see lecture 1) and corrresponding p-values from a paired t-test (see next slide). Observations from the same subjects are linked, to emphasize the correlation structure of the data (same subjects).



Note: reporting differences in mean (between meals) with 95%-CI is also common and usually good practice (in addition or instead of these results).

Ref: Prahm et al. Peptides. 2023 Nov 1;169:171091.

Digression: common mistake!

In lecture 1, we have seen that we can sometimes deduce whether a difference in mean is statistically significant from a visual comparison of the confidence intervals and estimates in the two groups that we compare.

This works for two "independent" samples, but this does not work for paired data as in the previous plot.

R code:

```
t.test(dauc[,"High protein"],dauc[,"High fat"],paired=TRUE)
```

Output (partial):

```
        Paired t-test

data:  dauc[, "High protein"] and dauc[, "High fat"]
t = 3.1966, df = 9, p-value = 0.01089
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
  725.7702 4239.7298
sample estimates:
mean difference
        2482.75
```

Reminder: paired t-tests are one-sample t-tests (see lecture 2)

R code:

```
t.test(dauc[,"High protein"]-dauc[,"High fat"])
```

Output (partial):

```
        One Sample t-test

data:  dauc[, "High protein"] - dauc[, "High fat"]
t = 3.1966, df = 9, p-value = 0.01089
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  725.7702 4239.7298
sample estimates:
mean of x
  2482.75
```

Digression: why are crossover designs popular?

- In short, patients included in the two arms of a usual randomized trials will never be exactly similar for all of what could matter (e.g., age, disease severity, comorbidity, lifestyle, genotype...). Randomization will only make the differences "small" and non-existent "in average".
- This random unbalance at baseline might cause some false positive findings.
- Usual statistical analysis accounts for that, but the price to pay is to compute sufficiently large confidence intervals and p-values.
- In crossover trials, the design substantially reduces random unbalance at baseline. Only time-varying factors, e.g., blood pressure, might differ. Properly accounting for that in the statistical analysis, using appropriate methods for paired data, makes it possible to compute narrower confidence intervals and p-values.

Note: a similar reasoning applies in contexts in wchich stratified randomization is used.

Case: baseline follow-up study (of a neurodegenerative disease, RCT)

Data: scores of n=166 patients randomized 1:1, some missing values at follow-up visits (missed visit). The higher the score the better. Shown are changes from baseline.

	id	trt	baseline	week6	week12
1	1	SoC	1.4149135	0.5570839	0.4874180
2	2	SoC	0.5392197	0.6747093	0.2686820
3	3	Exp	0.6554562	-0.7778319	0.6444571
4	4	Exp	1.7226614	2.2641281	0.7723273
5	5	SoC	-2.8416278	0.9717710	2.1931570
6	6	SoC	2.7684744	-2.5536338	NA

Research question: Does the experimental treatment (Exp) improve the score of the patients at week 12, as compared to standard of care (SoC)?

Missing data are challenging!

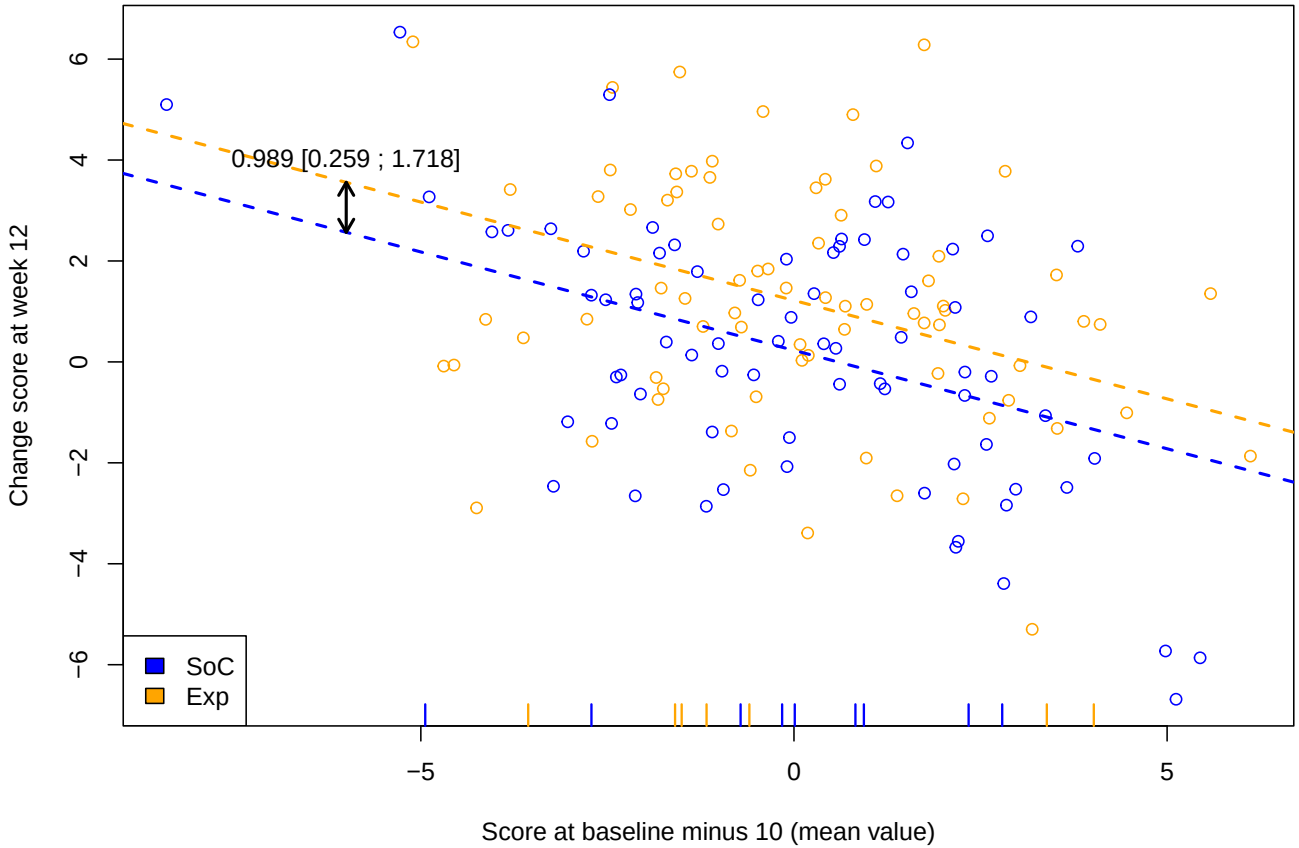
PRINCIPLES

There is no universal method for handling incomplete data in a clinical trial. Each trial has its own set of design and measurement characteristics. There is, however, a set of six principles that can be applied in a wide variety of settings.

- *"First, it needs to be determined whether missingness of a particular value hides a true underlying value that is meaningful for analysis."* E.g., Quality of Life, pain or functioning scores or CD4 counts do not exist after death!
- *"Third, reasons for missing data must be documented as much as possible."*
- *"Fourth, the trial designers should decide on a primary set of assumptions about the missing data mechanism."*
- *"Fifth, the trial sponsors should conduct a statistically valid analysis under the primary missing data assumptions."*

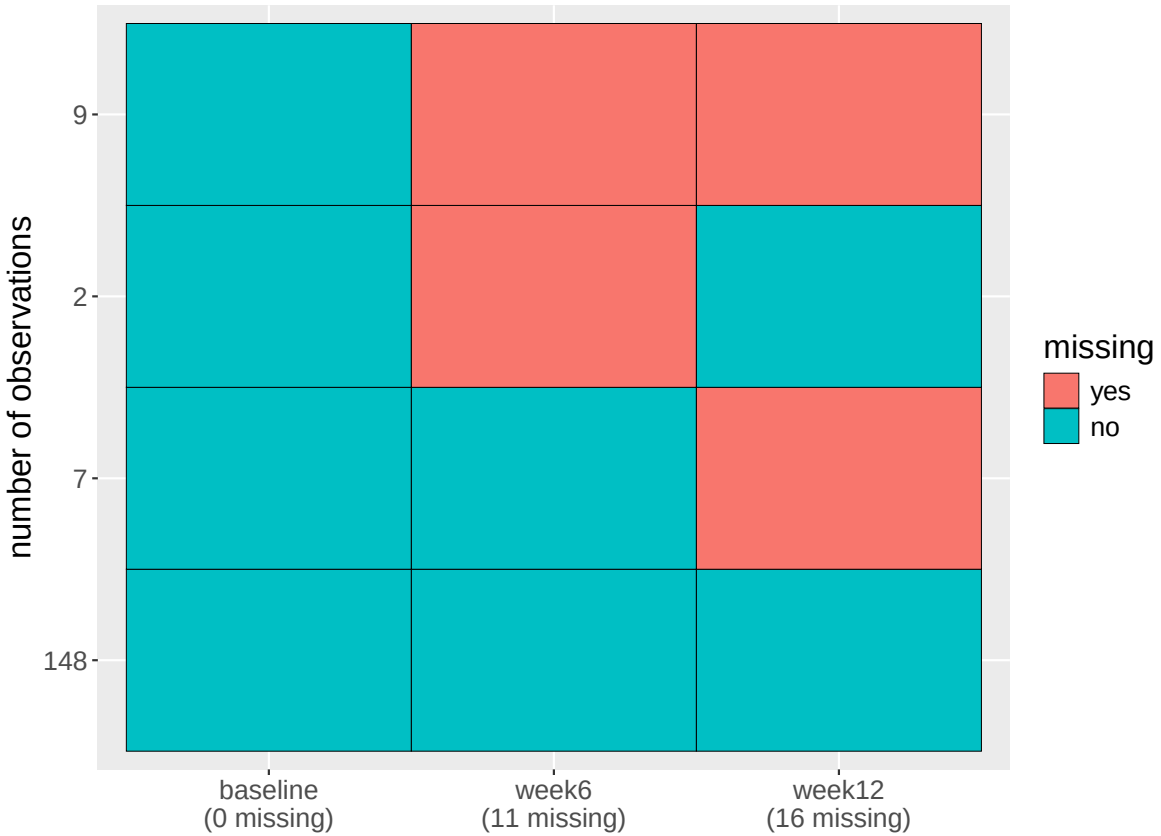
Note: the second principle is about well-defined estimands, the sixth about sensitivity analysis.

Ref: National Research Council. 2010. *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12955 (pages 48-49)



- Without missing data, the "standard" approach is to fit a usual ANCOVA model (lecture 7). It would provide "model robust" conclusions thanks to randomization.
- Results shown are those obtained from the "complete case analysis" (i.e., excluding patients with a missing change score at 12 weeks).
- Baseline scores of the 16 patients with missing values for the change score at week 12 are shown by "ticks" on the x-axis.

Missing data pattern



It is usually useful to transparently report the missingness pattern and explain what could be the most likely causes of each case. Recommendation: always produce this plot, at least to check that there is nothing unexpected or implausible (e.g., with the follicles data the second line, showing intermittent missing data, would have indicated a typo in the data collection, as NA encoded a follicle "dead" at that day).

Missing data: issues and a solution (often useful, but not always)

Challenges with missing data include:

- if the missing data are not "Missing Completely At Random" (MCAR), the complete case analysis is usually biased. Informally, we say that the missing data are MCAR if missingness is unrelated to outcome and covariates
- even with MCAR, using some available information about the excluded patients usually increases the power (e.g., change score at week 6). Idea: some kind of information is better than none!

Solution using a Mixed-effect Model for Repeated Measures (MMRM):

- prevents bias if the data are "Missing At Random" (MAR), meaning that the missingness may depend on covariates and previous measures of the outcome (e.g., change score at week 6) , but is otherwise completely random.
- more powerful in case of MCAR.

The MMRM model

Model for the outcome "change score" at the j-th visit (j=1 for visit at week 6; j=2 for visit at week 12) of the i-th patient:

$$Y_{ij} = \mu + \mathbf{x}_i\beta_1 + \mathbf{z}_j\beta_2 + \mathbf{u}_i\beta_3 + \mathbf{x}_i \cdot \mathbf{z}_j\beta_4 + \mathbf{u}_i \cdot \mathbf{z}_j\beta_5 + \varepsilon_{ij}$$

with

$$\varepsilon_{ij} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} , \begin{bmatrix} \sigma_1^2 & \rho\sigma_2\sigma_1 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix} \right)$$

- $\mathbf{x}_i$: baseline score of the i-th patient minus 10 (the "average" baseline score).
- $\mathbf{z}_j$: indicates the week, equal 1 if j=1 (week 6), 0 if j=2 (week 12)
- $\mathbf{u}_i$: indicates the arm, equal 1 if i-th patient randomized to "Experimental" arm, 0 if randomized to "Standard of Care".

- Observations from different patients are assumed to be independent.
- The name "MMRM" is because this model can is an extension of the random effect/mixed model which assumes $\sigma_1 = \sigma_2 = \sigma_T$.

R code:

```
library(LMMstar)
lmmfit <- lmm(score~baseline*visit + trt*visit, repetition = ~visit|id,
              structure = "UN", data = long)
summary(lmmfit)
```

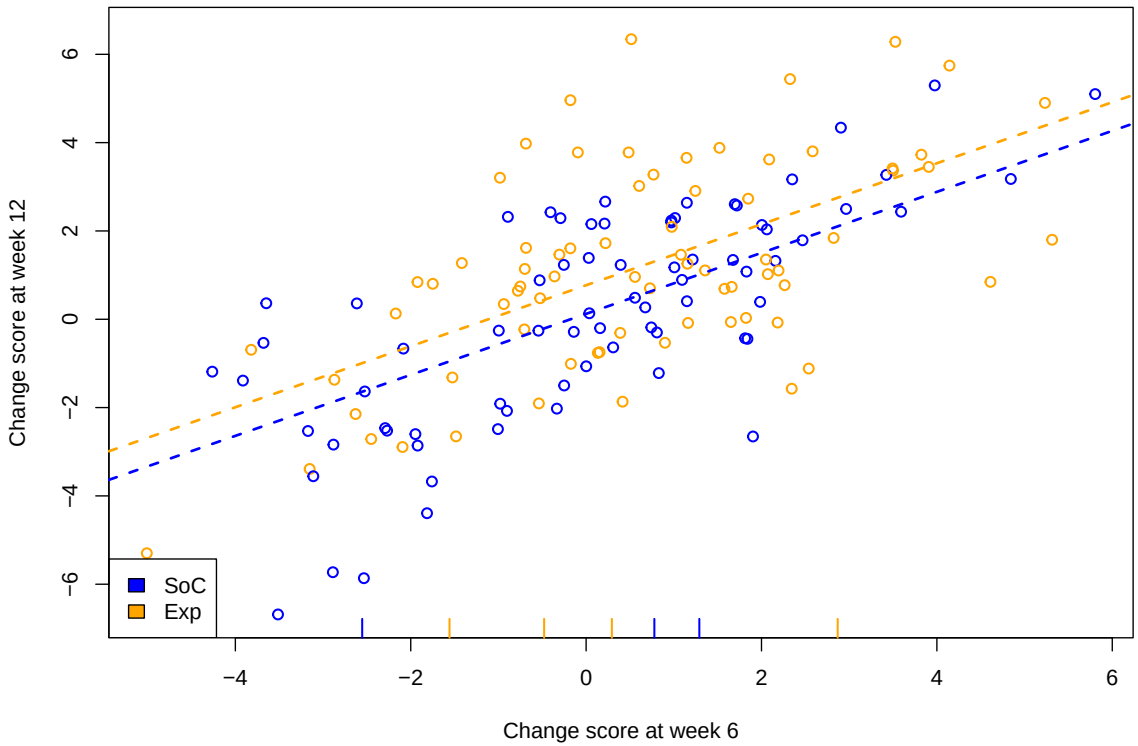
Output (partial):

```
Residual variance-covariance: unstructured

  - correlation structure: ~0 + visit
        2     1
  2 1.000 0.627
  1 0.627 1.000

  - variance structure: ~visit
          standard.deviation ratio
  sigma.2               2.25 1.000
  sigma.1               2.04 0.908

Fixed effects: score ~ baseline * visit + trt * visit

                  estimate    se    df  lower  upper p.value
  (Intercept)        0.226 0.258 152.5 -0.283  0.735 0.38236
  baseline          -0.395 0.072 154.2 -0.537 -0.252 < 1e-04 ***
  visit1            -0.083 0.216   147  -0.51  0.343 0.69972
  trtExp             0.984 0.364   153  0.265  1.703 0.00766  **
  baseline:visit1    0.157 0.061 148.9  0.036  0.277 0.01099   *
  visit1:trtExp     -0.497 0.305 147.3 -1.099  0.105 0.10513
```

Parameters interpretation (reference group: trt=SoC, visit=2 (12 weeks) and baseline score - 10=0)

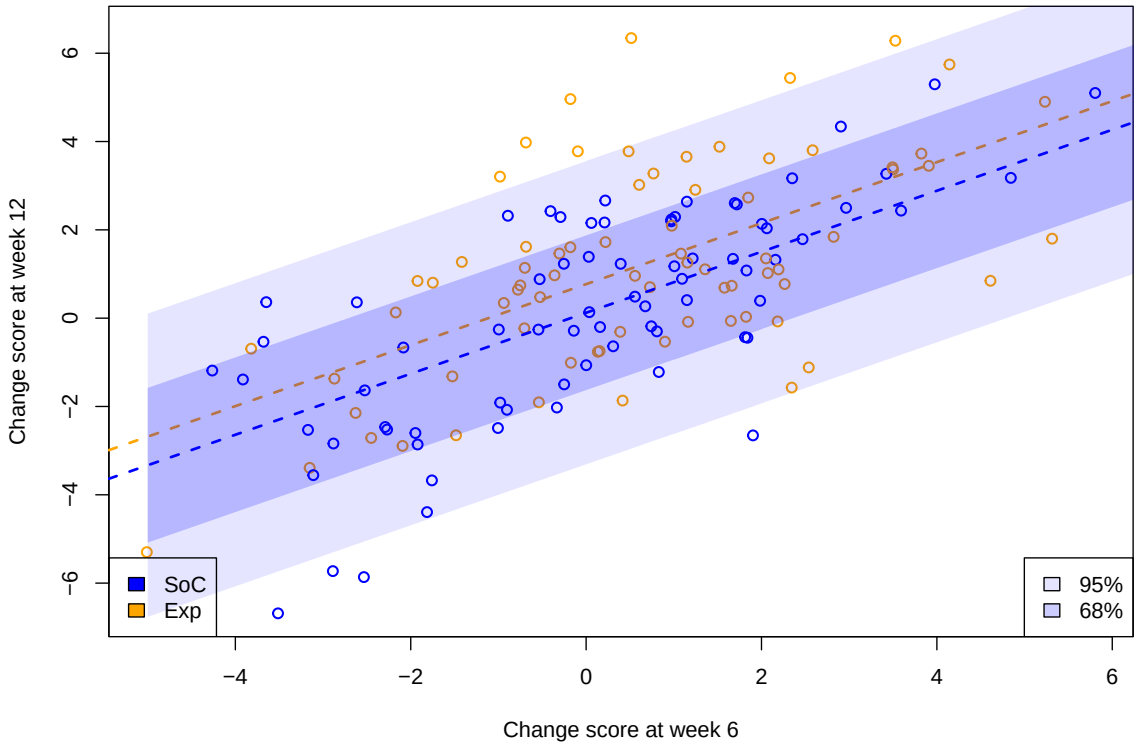
- μ: estimated as 0.226, is the mean change score at 12 weeks for a patient randomized to "SoC", with baseline score=10.
- β_1: estimated as -0.395, is the mean change in outcome (change score at 12 weeks) when comparing two patients of the same arm, one with a baseline score one unit larger than the other.
- β_2: estimated as -0.083, is the mean difference between the change score at week 6 and at week 12, for a patient of baseline score=10 and randomized to "SoC".
- β_3: estimated as 0.984, is the mean difference in outcome (change score at 12 weeks) for a patient randomized to "Exp" as compared to a patient randomized to "SoC", when both patient have the same baseline score (or "in average", due to randomization).
- β_4: estimated as 0.157, is a difference of differences in mean. In short, the difference β_1 becomes $\beta_1 + \beta_4$ when comparing change scores at week 6 instead of change scores at week 12. It is just to let the data speak freely and model a possibly different association between the baseline score and the change scores at 6 and 12 weeks.
- β_5: estimated as -0.497, is a difference of differences in mean. In short, the difference β_2 becomes $\beta_2 + \beta_5$ for a patient randomized to "Exp". It is just to let the data speak freely and model possibly different treatment effects on the change scores at 6 and 12 weeks.
- σ_1: estimated as 2.04, is the standard deviation of the "unexplained" variability of the change score at 6 weeks (i.e., standard deviation of error term ε_{i1}; unexplained because neither explained by the treatment nor by the baseline score; prediction interval is "estimated mean" $\pm 1.96 \cdot \sigma_1$; see plot on next slides).
- σ_2: estimated as 2.25, same interpretation as for σ_1, but for the change score at 12 weeks.
- ρ: estimated as 0.627, is the correlation between the change score of the same patient at 6 and 12 weeks (see plot on next slide).

How does the MMRM handle missing data? (1/2)



- The lines show the estimated average change score at 12 weeks for a patient of baseline score=10 (the average value), for both arms (lines would be shifted up or down for other baseline scores).
- The slope (assumed to be the same in the two groups) is the estimated value of $\rho \cdot \sigma_2/\sigma_1$
  (if $\sigma_2 \approx \sigma_1$, then slope $\approx \rho$, i.e., the correlation between the change score of the same patient at 6 and 12 weeks)
- The change score at week 6 of the 7 patients with missing values for the change score at week 12 (but no missing value at week 6) are shown by "ticks" on the x-axis.

How does the MMRM handle missing data? (1/2)



- Shown are 68% and 95% prediction intervals for the change score at week 12, given an observed value of the change score at week 6 and baseline score=10, in the standard of care arm (intervals would be shifted up or down for other baseline values and/or treatment arm).
- Implicitly, the MMRM "guesses" the likely values of the missing scores at 12 weeks, given the available information at baseline and week 6. Because the "guesses" use this information, the results will be robust to missingness mechanisms that depend on baseline score and change score at week 6, which is more realistic than assuming that it depends only on what is observed at baseline (which is what the complete case analysis using ANCOVA assumes). This will also typically lead to power gains, when data are MCAR. This is intuitive. E.g., if correlation $\rho \approx 1$, then knowing the change score at week 6 is almost as good as knowing it at week 12, so not much loss of information hence not much loss of power.

Why are MMRM commonly used?

- Most modeling assumptions are not so important for the analysis of (sufficiently large) randomized data.
  - E.g., the assumption that the error terms are normally distributed isn't important unless sample sizes are small. Linear mixed models are highly robust due to the central limit theorem.
- The MAR assumption is often more realistic than the MCAR assumption. The MAR assumption is sufficient[3] to avoid bias and it is difficult to use another modeling approach without assuming MAR (at least a "simple" alternative; complementary sensitivity analyses can be useful).
- User-friendly software exist.
- Many (reliable) guidelines and textbooks recommend MMRM as a good "default choice" in many contexts. (E.g., Mallinckrod et (2008), *"Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials,"* Drug Information Journal, 42, 303–319.)

[3] Precisely, it is sufficient under "correct" model specification.

Would you like to learn more?

Statistical analysis of repeated measurements and clustered data

- Ph.D. course running each year in May. Room for 78 students.
- Open webpage at `https://absalon.ku.dk/courses/47665`.
- Offers DIY guidance for most commonly used study designs.
- We plan to make new video of the lectures in 2025.

Content:

- Linear mixed models for randomized and observational follow-up studies, crossover studies, reproducibility studies, and cluster RCTs.
- Generalized linear mixed models and generalized estimating equations (GEE) for non-normal data with focus on binary data.
- Handling of missing data with emphasis on randomized trials.