Faculty of Health Sciences

Day 2: Covariate adjustment, robustness, missing data

Paul Blanche

Section of Biostatistics, University of Copenhagen

November 19, 2025

Outline/Intended Learning Outcomes (ILOs)

Baseline table

- ILO: repeat the main recommendations
- ILO: explain their rationale to skeptical colleagues and reviewers
- ILO: describe the relationship between randomization, baseline imbalance and the validity of statistical analyses

Covariate adjustment

- ILO: repeat the main recommendations
- ILO: explain how this can lead to power gains
- ILO: restate key robustness properties
- ILO: perform conventional analyses for binary and quantitative outcomes
- ILO: recall examples

Missing data

- ILO: list a few principles and type of missing data
- ILO: fit a Mixed Model for Repeated Measurements (MMRM) and interpret the results
- ILO: restate the rationale for using an MMRM analysis

Baseline Tables

- Baseline tables are usually presented in articles presenting the results of a clinical trial.
- Guidelines (e.g., CONSORT) encourage us to present them.[1]

CONSORT 2010 checklist of information to include when reporting a randomised trial*

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts [45,65])	
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	
	2b	Specific objectives or hypotheses	
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	

Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	
	13b	For each group, losses and exclusions after randomisation, together with reasons	
Recruitment	14a	Dates defining the periods of recruitment and follow-up	
	14b	Why the trial ended or was stopped	
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	

[1] Moher et al (2010). CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. Journal of Clinical Epidemiology, 63(8), e1–e37

Why presenting a baseline table?

5.4.4. Item 15. A table showing baseline demographic and clinical characteristics for each group

Example—See Table 4

Explanation—Although the eligibility criteria (see item 4a) indicate who was eligible for the trial, it is also important to know the characteristics of the participants who were actually included. This information allows readers, especially clinicians, to judge how relevant the results of a trial might be to an individual patient.

What to put in a baseline table? Simple descriptive statistics (only!).

ported, along with average values. Continuous variables can be summarised for each group by the mean and standard deviation. When continuous data have an asymmetrical distribution, a preferable approach may be to quote the median and a centile range (such as the 25th and 75th centiles) [177]. Standard errors and confidence intervals are not appropriate for describing variability—they are inferential rather than descriptive statistics. Variables with

Source: Moher et al (2010). CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. Journal of Clinical Epidemiology, 63(8), e1–e37.

Example

	Initial oral antibiotics (n=101)	Initial intravenous antibiotics (n=91)
Sex		
Female	37 (37%)	31 (34%)
Male	64 (63%)	60 (66%)
Age, years	2·4 (1·3–6·7)	2·3 (1·4–6·7)
0–4	66 (65%)	63 (69%)
5–9	20 (20%)	13 (14%)
10–17	15 (15%)	15 (17%)
Localisation		
Bone infection	41 (41%)	41 (45%)
Joint infection	43 (43%)	36 (40%)
Spondylodiscitis or sacroiliitis	17 (17%)	14 (15%)
Clinical presentation		
Symptom duration before treatment, days	5·0 (3·0–12·0)	5·0 (2·5–10·5)
Pain or immobilisation	101 (100%)	90 (99%)
Temperature ≥38·0 °C	66 (65%)	60 (66%)
Local signs of infection	41 (41%)	42 (46%)
C-reactive protein, mg/L		
At study enrolment	34 (10–65)	34 (14–58)
Diagnostic evaluations		
Blood culture	94 (93%)	88 (97%)
Joint puncture	35 (35%)	25 (27%)
Bone biopsy	4 (4%)	4 (4%)
MRI	77 (76%)	63 (69%)
Ultrasound	77 (76%)	72 (79%)
X-ray	83 (82%)	69 (76%)
Pathogen identified (sterile site)		
All pathogens†	25 (25%)	23 (25%)
S aureus‡	16 (16%)	13 (14%)
K kingae§	5 (5%)	8 (9%)
S pyogenes	4 (4%)	1 (1%)
S pneumoniae	0	1 (1%)

"Data are n (%) or median (IQR)."[2]

[2] Nielsen et al. (2024). Oral versus intravenous empirical antibiotics in children and adolescents with uncomplicated bone and joint infections: a nationwide, randomised, controlled, non-inferiority trial in Denmark. Lancet Child Adolesc Health, 4642(24), 1–11.

Don't add p-values to baseline tables!

- Although it is commonly done and also commonly asked by reviewers, unfortunately. But, usually not by statistical reviewers!
- CONSORT guidelines do not support adding such p-values:

  "Such hypothesis testing is superfluous and can mislead investigators and their readers." [3]

  (most medical journals encourage authors to follow the CONSORT guidelines; check their website)
- In the website of the *New England Journal of Medicine*, section "Statistical Reporting Guidelines", it is even stated:[4]

> b. P values should not be included in the traditional Table 1 of a randomized trial manuscript showing the distribution of baseline variables by treatment group. However, authors should note imbalances in potential confounders that could be due to chance or inconsistencies in randomization.

[3] Moher et al (2010). CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. Journal of Clinical Epidemiology, 63(8), e1–e37

[4] accessed Oct 29, 2025: https://www.nejm.org/author-center/new-manuscripts#statistical-reporting-guidelines

Why is it irrational to add p-values to baseline tables?

"Hardly a statistician of repute can be found to defend the practice common among physicians of comparing the treatment groups in a randomized clinical trial at baseline using hypothesis/significance tests on covariates. The reason for the statistician's dislike is that such a test appears to be used to say something about the adequacy of the given allocation whereas it could only be a test of the allocation procedure: the randomization process itself."

Why is it irrational to add p-values to baseline tables?

"Hardly a statistician of repute can be found to defend the practice common among physicians of comparing the treatment groups in a randomized clinical trial at baseline using hypothesis/significance tests on covariates. The reason for the statistician's dislike is that such a test appears to be used to say something about the adequacy of the given allocation whereas it could only be a test of the allocation procedure: the randomization process itself."

Can't I do like many others and add them anyway?

"In short, the test of baseline balance is a misuse of the significance test. The fact that it is frequently performed does not constitute a defence any more than the fact that antibiotics are commonly employed to 'treat' viral infections proves that they are effective antivirals. And the fact that baseline tests are commonly performed without much apparent harm is no more of a defence than saying of the policy of treating colds with antibiotics that most patients recover."

Ref: quotes from Section 7.2.1, pages 112-113, in *Statistical Issues in Drug Development* (3^{rd} Edition), by Stephen Senn.

Does random imbalance invalidate the results?

This seems to be a common misunderstanding!

"It is not necessary for the groups to be balanced. In fact, the probability calculation applied to a clinical trial automatically makes an allowance for the fact that groups will almost certainly be unbalanced, and if one knew that they were balanced, then the calculation that is usually performed would not be correct." [5]

In other words, standard errors, confidence intervals and p-values are computed accounting for the fact that we expect some random imbalance.

[5] Senn S. Seven myths of randomisation in clinical trials. Stat Med 2013;32:1439e50.

Random imbalance and trial sample size

Question: Some think that larger trials are generally less vulnerable to covariate imbalance than smaller ones. Is that really true?

[6] Senn S. Seven myths of randomisation in clinical trials. Stat Med 2013;32:1439e50.

Random imbalance and trial sample size

Question: Some think that larger trials are generally less vulnerable to covariate imbalance than smaller ones. Is that really true?

Answer: No, not if look at confidence intervals and p-values, i.e., if we look at results that rigorously account for randomness /uncertainty.

"When sample sizes increase, it is certainly the case that the expected random difference between two groups will reduce, and this reflects, amongst other things, the greater expected balance in proportionate terms between groups. In this sense, the belief that larger trials are more balanced than smaller ones is not a myth. However, by the same token, the standard error of the treatment effect will be smaller and the confidence interval will be narrower, and for any given observed difference at outcome, the p-value will be smaller. Thus, the effect of increasing sample size is consumed by conventional analyses in terms of increased precision."[6]

[6] Senn S. Seven myths of randomisation in clinical trials. Stat Med 2013;32:1439e50.

Outline/Intended Learning Outcomes (ILOs)

Baseline table
- ILO: repeat the main recommendations
- ILO: explain their rationale to skeptical colleagues and reviewers
- ILO: describe the relationship between randomization, baseline imbalance and the validity of statistical analyses

Covariate adjustment
- ILO: repeat the main recommendations
- ILO: explain how this can lead to power gains
- ILO: restate key robustness properties
- ILO: perform conventional analyses for binary and quantitative outcomes
- ILO: recall examples

Missing data
- ILO: list a few principles and type of missing data
- ILO: fit a Mixed Model for Repeated Measurements (MMRM) and interpret the results
- ILO: restate the rationale for using an MMRM analysis

Rationale for baseline covariate adjustment

Should baseline covariates be ignored because of randomization?

"one should not use the fact that one has randomized as an excuse for ignoring baseline prognostic information in analyzing a clinical trial. Such prognostic information provides the means of distinguishing the particular clinical trial from the much larger set of trials one might have run in which the baseline prognostic distribution might have been different." [7]

Baseline covariates provides important information that can lead to power gains, when this information is used well in the statistical analysis, using so-called "adjusted analyses".

"incorporating prognostic baseline covariates in the design and analysis of clinical trial data can result in a more efficient use of data to demonstrate and quantify the effects of treatment"[8]

[7] Senn. (2005). Baseline balance and valid statistical analyses: common misunderstandings. Applied Clinical Trials 14 (3): 24–27

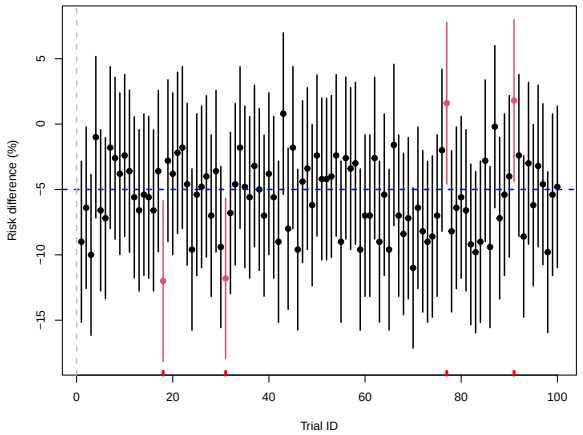
[8] FDA scientific guidelines: "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry", 2023, accessed 29/10/2025, https://www.fda.gov/media/148910/download

Why are adjusted analyses more powerful? (1/4)

- Randomization balances covariates in average
- But in average only! For any trial, there will be some degree of imbalance.
- If some covariates are prognostic of the outcome, random baseline covariates imbalance creates random variation, when we estimate a treatment effect, e.g., a risk difference with binary outcomes.
- When we adjust for baseline covariates, we (implicitely) compare subjects between arms within subgroups of subjects who have the same covariates profile. Within these subgroups, covariates balance necessarily holds. This is similar to the situation where we stratify the randomization.
- When covariates balance is enforced, there is less random variation, which results in smaller SEs (when computed appropriately) and more power.

Conclusion: when performing an unadjusted analysis, we make an allowance for the fact that there will be random covariates imbalance. But when we adjust we don't, as we do not need, and that is how we gain power.

Why are adjusted analyses more powerful? (2/4)



Setting:

- 100 trials simulated
- shown are 95%-CI; unadjusted analysis
- 95/100 contain the "true" risk difference (blue line), as expected
- Those who do not are shown in red
- mean(SE)=3.15%

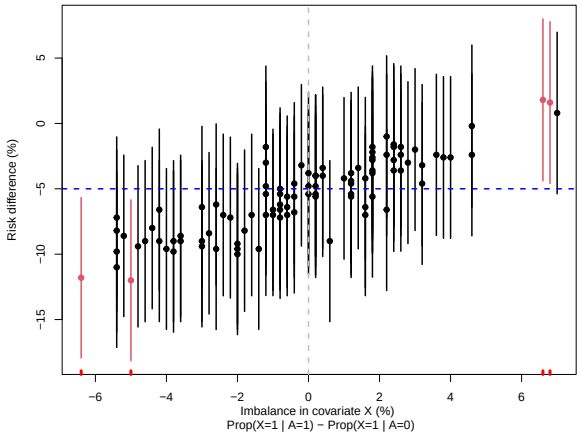
Details:

- $n = 500$ per group, randomized 1:1
- One binary covariate X ($p_X = 0.5$)
- Marginal (average) risks: $\pi_1 = 50\%$ vs $\pi_0 = 55\%$, from 4 subgroups, 25% prevalence for each.

A (intervention)	X (severity)	risk of death (90-day)
0	0	11%
1	0	10%
0	1	99%
1	1	90%

- $RR_{X|A} = 9$ and $RR_{A|X} = 0.91$

Why are adjusted analyses more powerful? (3/4)

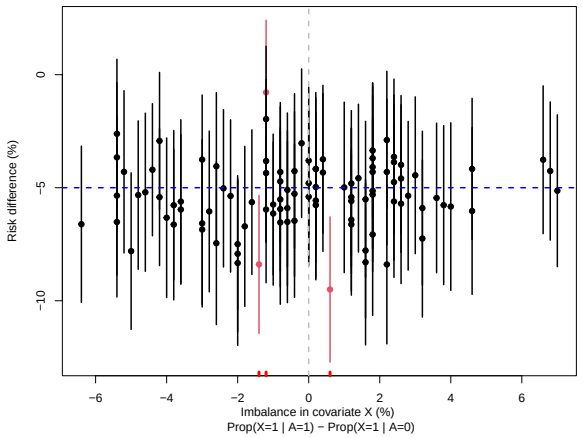


- shown are 95%-CI; unadjusted analysis
- imbalance ↑ ⇒ estimated risk difference ↑
- it reflects that X is prognostic of the outcome (X ↑ ⇒ risk ↑)
- red CIs that do not include the truth corresponds to large imbalance.
- Random large imbalance corresponds to a situation randomly far from the average, as in average there is balance.
- The CIs are on average centered around the truth
- But there are NOT centered in average, within the subgroup of "atypical" trials with large positive (or large negative) unbalance.

"To use some statistical jargon, the baselines provide a means of identifying a recognizable subset, a group for which the average no longer applies. An analogy from life insurance may be helpful here. A life-table may provide a reasonable assessment of the expectation of life of a male aged 50. However, if it is based on an "average" population, then it will overstate the expectation, other things being equal, for a 50-year-old male who is a smoker and underestimate it if he is not." [9]

[9] Senn (2019). The well-adjusted statistician: Analysis of covariance explained, Applied Clinical Trials.

Why are adjusted analyses more powerful? (4/4)



- shown are 95%-CI; adjusted analysis (logistic regression followed by standardization; non-parametric Bootstrap to compute SEs, n.boot=400)
- No association between imbalance and estimated risk difference
- 97/100 contain the "true" risk difference (blue line), approximately as the 95 expected (random, because only 100 simulated datasets)
- Those who do not are shown in red; random location.
- Note: mean(SE)=1.70% instead of mean(SE)=3.15% when unadjusted analysis. Much narrower CIs and more power !
- Corresponds to a redution in sample size needed of $\approx$ 70%, in this *extreme* example ($RR_{X|A} = 9$); because 1- ($\text{Var}_a/\text{Var}_u$) ≈ 0.70.

How to adjust? (binary outcome)

- Different approach exist to analyze the data while adjusting for covariates.
- One which is increasingly promoted (and which was used in the previous example with the simulations) is "Regression standardization". This approach is commonly used in epidemiology to account for counfounding[10]. This approach is also termed the 'standardized', 'plug-in', or 'g- computation' estimator.[11]
- This approach targets an unconditional (also called marginal, or average) treatment effect, often quantified as a risk difference. That is, it targets the same as an unadjusted analysis which simply compare the proportions in the two arms.
- By contrast, usual multiple (aka adjusted) logistic regression targets a conditional treatment effect, quantified via odds ratios. That is, the same as odds ratio computed within strata of baseline variables.

[10] Sjölander. "Regression standardization with the R package stdReg." European Journal of Epidemiology 31.6 (2016): 563-574

[11] FDA scientific guidelines: "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry", 2023, accessed 29/10/2025, https://www.fda.gov/media/148910/download

Conditional vs unconditional effect

"In general, treatment effects may differ across subgroups. However, with some parameters such as odds ratios, even when all subgroup treatment effects are identical, this subgroup-specific conditional treatment effect can differ from the unconditional treatment effect (i.e., the effect at the population level from moving the target population from untreated to treated) (Gail et al. 1984). This is termed non-collapsibility (Agresti 2002), which is distinct from confounding and can occur despite randomization and large sample sizes" [12]

Table 1: Non-collapsibility of the Odds Ratio in a Hypothetical Target Population

	Percentage of target population	Success rate		Odds ratio
		New drug	Placebo	
Biomarker-positive	50%	80.0%	33.3%	8.0
Biomarker-negative	50%	25.0%	4.0%	8.0
Combined	100%	52.5%	18.7%	4.8

Reminder: odds = risk/(1-risk), Odds ratio= ratio of odds (in the two arms)

[12] FDA scientific guidelines: "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry", 2023, accessed 29/10/2025, https://www.fda.gov/media/148910/download

Regression standardization (in a nutshell)

- As an example, the following are steps for one reliable method for covariate adjustment for unconditional treatment effects with binary outcomes that produces a resulting estimator (Steingrimsson et al. 2017; Freedman 2008) termed the "standardized," "plug-in," or "g-computation" estimator:

  (1) Fit a logistic model with maximum likelihood that regresses the outcome on treatment assignments and prespecified baseline covariates. The model should include an intercept term.

  (2) For each subject, regardless of treatment group assignment, compute the model-based prediction of the probability of response under treatment using the subject's specific baseline covariates.

  (3) Estimate the average response under treatment by averaging (across all subjects in the trial) the probabilities estimated in Step 2.

  (4) For each subject, regardless of treatment group assignment, compute the model-based prediction of the probability of response under control using the subject's specific baseline covariates.

  (5) Estimate the average response under control by averaging (across all subjects in the trial) the probabilities estimated in Step 4.

  (6) The estimates of average responses rates in the two treatment groups from Steps 3 and 5 can be used to estimate an unconditional treatment effect, such as the risk difference, relative risk, or odds ratio.

- Recommendation: compute standard errors by bootstrapping.

Source: FDA scientific guidelines: "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry", 2023, accessed 29/10/2025, https://www.fda.gov/media/148910/download

Reminder: logistic regression

Multiple logistic model is a widely used model for the risk of observing a binary outcome $(D = 1)$ given treatment A and baseline covariates $\boldsymbol{L}$:

$$P\{D = 1|\boldsymbol{L}, A\} = \frac{\exp(\boldsymbol{X}^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}^T\boldsymbol{\beta})} ,$$

where $\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{L}, A)$ is a pre-specified vector of variables constructed from $(\boldsymbol{L}, A)$, possibly including interaction terms. Often, $\boldsymbol{X} = (1, A, \boldsymbol{L})$ (i.e., the simplest possible model with no interaction terms, no categorized variables, no splines.)

Reminder: odds ratio and logistic regression

Odds: defined as "risk of event divided by risk of no event".

$$\boxed{\text{odds} = p/(1-p)} ,$$

Odds ratio (OR): defined as the ratio of the odds,

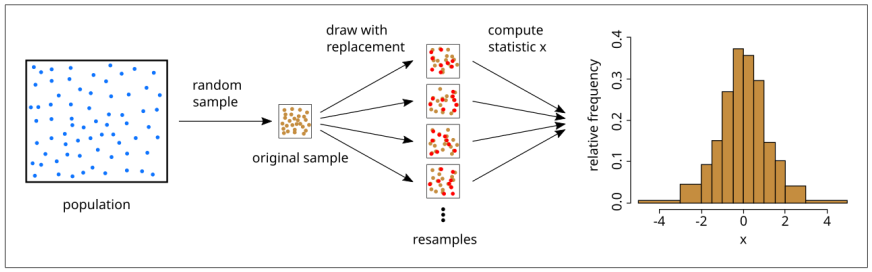
$$\boxed{OR = \frac{\text{odds}_1}{\text{odds}_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}} .$$

With logistic regression (without interactions)

$$e^{\beta_j} = OR_{(X_j = x+1 \text{ vs } X_j = x)}$$

is the odds ratio that compares the risks of two subjects, when both subjects are similar for all covariates adjusted for, except X_j, whith one subject having a value one unit larger than the other subject (i.e. $X_j = x + 1$ vs $X_j = x$).

Digression: bootstrapping (to compute standard errors)



"A sample is drawn from a population. From this sample, resamples are generated by drawing with replacement (orange). Data points that were drawn more than once (which happens for approx. 26.4% of data points) are shown in red and slightly offsetted. From the resamples, the statistic x [e.g., risk difference] *is calculated and, therefore, a histogram can be calculated to estimate the distribution of x [e.g., risk difference]."*

The standard deviation corresponding to that distribution is then computed to estimated the standard error of x [e.g., risk difference].

Source: https://en.wikipedia.org/wiki/Bootstrapping_(statistics), accessed 29/10/2025.

Robustness to logistic model misspecification

- The better the fit of the logistic model the larger the power gain.
- *"All models are wrong, but some are useful"* (George Box). The model does not need to be perfect to provide power gains. Sensible models will often do the job almost as well as perfect models.

[13] FDA scientific guidelines: "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry", 2023, accessed 29/10/2025, https://www.fda.gov/media/148910/download

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Robustness to logistic model misspecification

- The better the fit of the logistic model the larger the power gain.
- *"All models are wrong, but some are useful"* (George Box). The model does not need to be perfect to provide power gains. Sensible models will often do the job almost as well as perfect models.
- Interestingly, the resulting unconditional treatment effect estimates are unbiased and confidence intervals and p-values are valid (i.e., correct coverage and type-I error) even when the model is misspecified; with large sample sizes. That is, even when the relationship between the outcome and the treatment and baseline variables is it wrongly modeled (e.g., interaction exist but are not included in the model). This is a consequence of randomization. Hence, adjusting for baseline covariates is now widely considered safe:

  "incorporating prognostic baseline covariates in the design and analysis of clinical trial data can result in a more efficient use of data to demonstrate and quantify the effects of treatment. Moreover, this can be done with minimal impact on bias or the Type I error rate." [...] *"Covariate-adjusted estimators of unconditional treatment effects that are robust to misspecification of regression models have been proposed for randomized clinical trials with binary outcomes (e.g., Steingrimsson et al. 2017)"*[13]

[13] FDA scientific guidelines: "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry", 2023, accessed 29/10/2025, https://www.fda.gov/media/148910/download

R implementation:

- It is easy![14]
- FDA scientific guidelines[15] cite Steingrimsson et al (2017)[16], which contains easy to use R code in appendix.

Robustness with conditional treatment effect estimation:

- when simply using logistic regression without standardization, we estimate conditional treatment effects, with 95%-CI and p-values.
- Interestingly, a consequence of randomization is that type-I error will also be controlled in that case, even in case of model misspecification; with large sample sizes.[17]
- Howerver, misspecification might introduce bias and 95%-CI might have the wrong coverage in case of model misspecification. This is unfortunate, but "sensible" modelling choices should prevent serious problems.

[14] It is very easy when using the "simplest" possible logistic regression model, without interaction etc, which is relatively often good enough. A few extra lines of code might be needed for more complicated modeling approaches.

[15] FDA scientific guidelines: "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry", 2023, accessed 29/10/2025, https://www.fda.gov/media/148910/download

[16] Steingrimsson et al, *Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes without regression model assumptions*, Contemporary Clinical Trials 54 (2017) 18–24

[17] Rosenblum & van der Laan, 2009, Using Regression Models to Analyze Randomized Trials: Asymptotically Valid Hypothesis Tests Despite Incorrectly Specified Models, Biometrics, 65(3):937-945

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Practice with R!

Practice with the "Exercise.2.1" from the webpage.
It is about thrombolytic strategies for acute myocardial infarction.

Covariate adjustment: quantitative outcomes (1/2)

- Instead of a multiple logistic model, one can use a multiple linear model (aka ANCOVA[18]).
- *"The resulting estimated regression coefficient for the treatment indicator is the estimate of the average treatment effect."* [19] (when the is no baseline × treatment interaction terms; otherwise centering or other tricks might be needed.)
- Similar robustness and power gain properties exist:

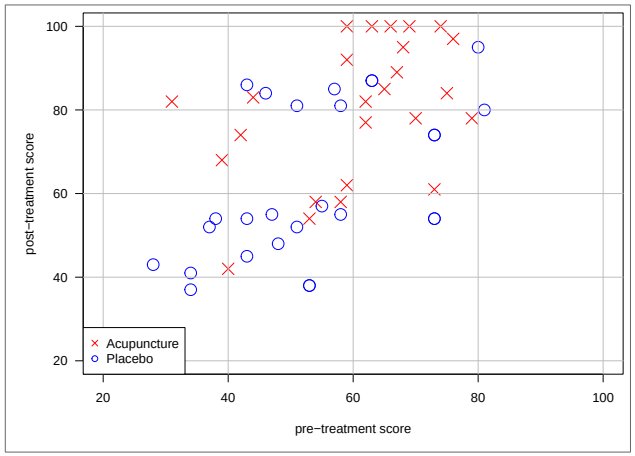
  "Even when the linear regression model is misspecified and does not accurately capture the relationships between the outcome, covariates, and treatment, covariate adjustment through a linear model is a valid method for estimating and performing inference for the average treatment effect."

[18] ANCOVA stands for ANalysis of COVAriance

[19] FDA scientific guidelines: "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry", 2023, accessed 29/10/2025, https://www.fda.gov/media/148910/download

Intuition & Example: baseline-follow up studies

Clear association between baseline and follow-up scores, within each arm: the larger the pre-treatment score the larger post-treatment score.

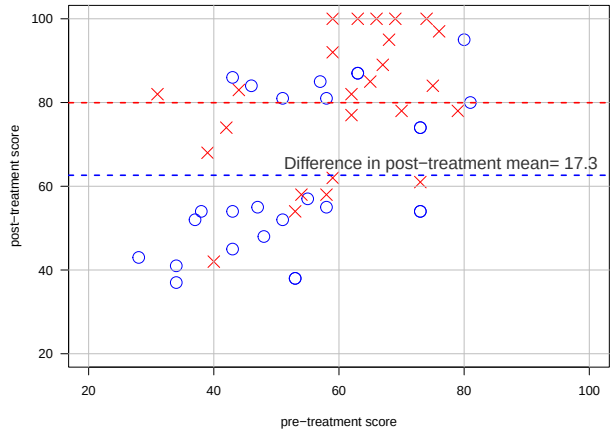


Data: reconstructed patient level data nearly perfectly matching those graphically presented in Vickers & Altman. BMJ. 2001 Nov 10;323(7321):1123-4.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Intuition & Example: baseline-follow up studies

Clear association between baseline and follow-up scores, within each arm: the larger the pre-treatment score the larger post-treatment score.

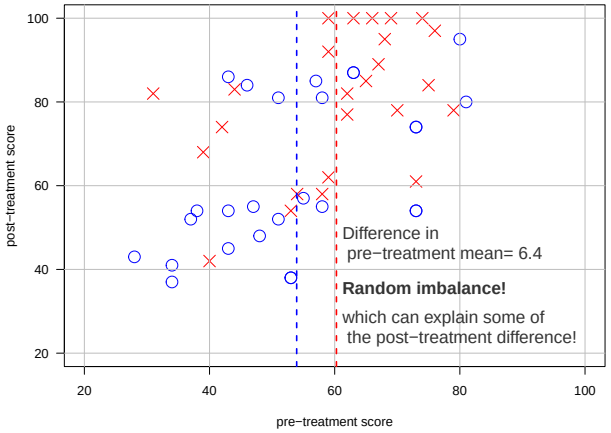


Data: reconstructed patient level data nearly perfectly matching those graphically presented in Vickers & Altman. BMJ. 2001 Nov 10;323(7321):1123-4.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Intuition & Example: baseline-follow up studies

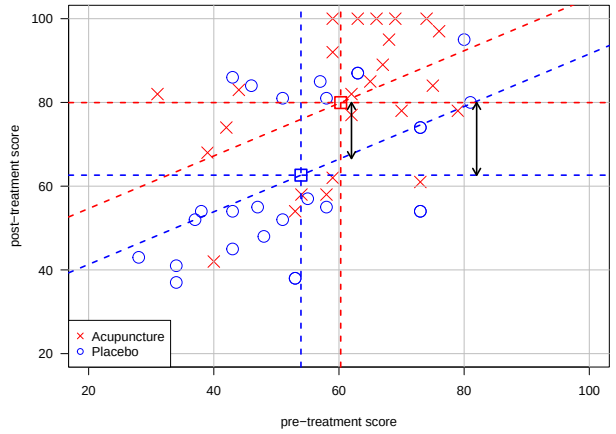
Clear association between baseline and follow-up scores, within each arm: the larger the pre-treatment score the larger post-treatment score.



Data: reconstructed patient level data nearly perfectly matching those graphically presented in Vickers & Altman. BMJ. 2001 Nov 10;323(7321):1123-4.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Intuition & Example: baseline-follow up studies (2/2)



The multiple linear model, adjusted for pre-treatment score, will always correct/adjust for random imbalance in baseline scores. Implicitly it estimates the mean difference in post-treatment score as if the mean pre-treatment score were the same in the two arms. This reduces the variability of the estimated difference, since random imbalance makes the unadjusted estimate randomly too small or randomly too large. Hence we can compute smaller standard errors for the adjusted differences. This leads to narrower confidence intervals and smaller p-values, in average (but in average only!).

Common questions

- How much power can we gain when using an adjusted analysis (i.e., ANCOVA) instead of a simple t-test?

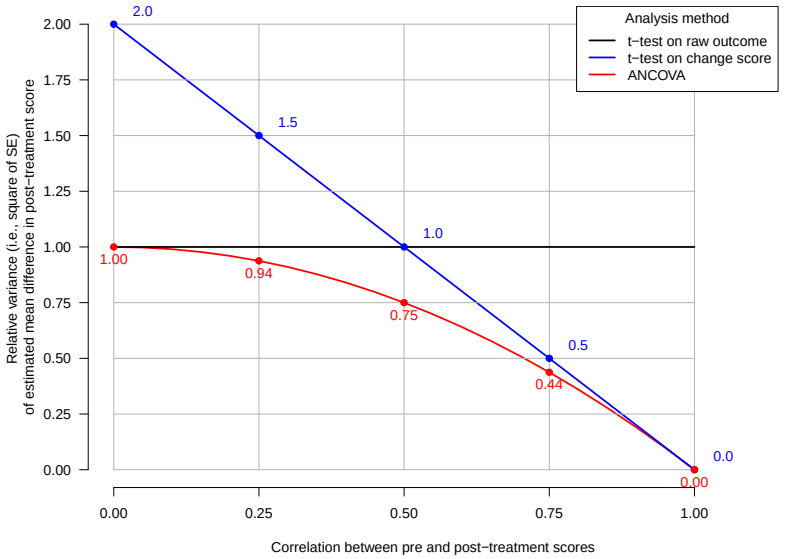
  Potentially a lot! The larger the correlation between the pre and post scores and (i.e., the steeper the slope) and the larger the power we gain. In terms of sample size needed, one might actually use $\sigma' = \sigma\sqrt{1-\rho^2}$ instead of σ in sample size calculation formulas, where σ is the SD of the post-treatment score and ρ the correlation.

- Isn't it sufficient to adjust for baseline by comparing the change scores (i.e., difference between pre and post treatment score)? Isn't it even better than ANCOVA?

  No! It is not as good as using an ANCOVA, and sometimes worse than using a t-test to compare post-treatment scores (in case of weak correlation, when $\rho < 0.5$).

Details: see e.g., Section 7.2.3, in *Statistical Issues in Drug Development* (3^{rd} Edition), by Stephen Senn.

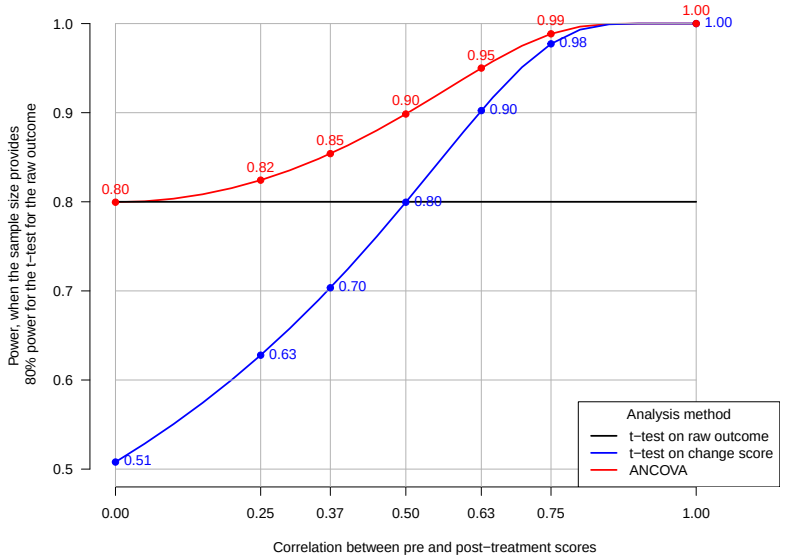
More detailed answers (1/2)



Note: "large" sample, but usually very accurate approximation.

Remark: relative variance equals relative sample size needed to have the same power.

More detailed answers (2/2)



Note: "large" sample, but usually very accurate approximation.

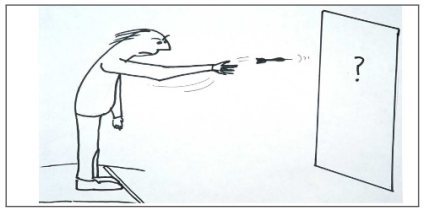
Remarks

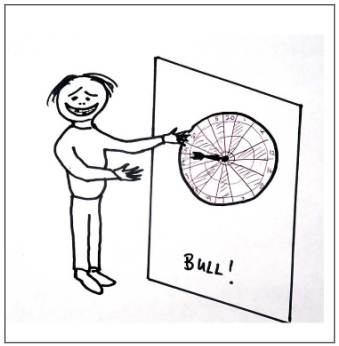
- The same "intuition" and "mathematics" work when adjusting for more than one baseline covariate. Usually, we use 5 ± 2, that we expect are (altogether) the most predictive of the outcome. But it really depends on the sample size and the background knowledge. Each additional covariate need to add "non-negligible" prognostic information on top of the others, to be useful. There is no simple rule!
- Power gains are often substantially larger for quantitative outcomes (when using ANCOVA) than whith binary outcomes (when using logistic regression, with or without standardization).
- When adjusting for more than one covariate, formulas such as $\sigma' = \sigma\sqrt{1-\rho^2}$ are becoming more difficult to use. Hence, conservative choices are usually made (e.g., sample size and power are computed as if only one covariate only will be adjusted for, although more covariates will be adjusted for).
- If the prognostic ability of baseline covariates is unclear / unknown, it may be wise to be conservative; this is a common and well accepted choice:

  "In a trial that uses covariate adjustment, the sample size and power calculations can be based on adjusted or unadjusted methods. The latter will often lead to a more conservati[ve] sample size."[20]

[20] FDA scientific guidelines: "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry", 2023, accessed 29/10/2025, https://www.fda.gov/media/148910/download

Digression: prespecification matters!





Concluding significance without prespecification is like drawing a dart-board around where the dart lands.

Of course, pre-specification of whether we plan to use an ANCOVA, a t-test or a t-test on change score is essential. We cannot perform the three analyses and report the best looking results!

Digression: three results with the acupuncture trial

Analysis Method	Est. ($\hat{\delta}$)	95%-CI	SE	p-value ($\mathcal{H}_0 : \delta = 0$)
ANOVA	13.3	[4.4;22.3]	4.4	0.42%
t-test (on post-treatment score)	17.3	[7.5;27.2]	4.9	0.09%
t-test on change score	11.0	[1.9;20.0]]	4.5	1.86%

Note: here $\hat{\rho} = 0.53$, baseline imbalance (difference in pre-treatment mean)$= 6.4$.

Remark: The recommended approach (ANCOVA) had the smallest SE (as expected) but not the smallest p-value (unlike what is expected "on average" or "the most often") because of random imbalance at baseline.

Remark: comparing change scores or post-treatment scores are two approaches that both target the same treatment effect, as in average there is no difference pre-treatment thanks to randomization. In other words, the mean difference post-treatment (at the population level) that we aim to estimate is equal to the mean difference in change score, as a consequence of randomization.

Exercise 2.2

1. Practice with R! Run the code of the script "`acunpunctureRDemo.R`" and recognize the main results of the previous slides.

2. Many researchers wonder whether they should define their primary outcome as the score at end of follow-up (i.e., post-treatment score) or as the change score at end of follow-up (i.e., post-treatment score minus pre-treatment score). Do you think it matters, if the primary analysis is made using an ANCOVA to adjust for the baseline score? Think...

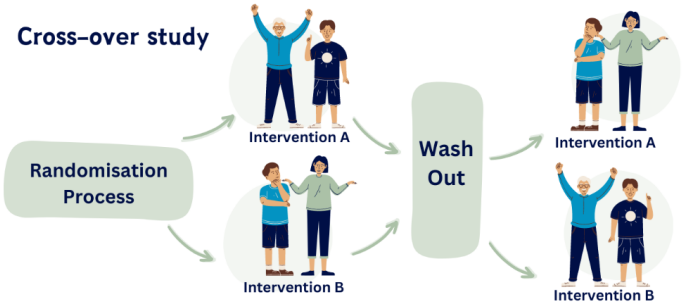
3. Perform the ANCOVA using the change from baseline as the outcome. What do you observe?

Remark: what you have just observed on this specific dataset always holds. It is an old results[21], which is relatively well-known.

[21] See e.g., Section 7.2.4, in *Statistical Issues in Drug Development* (3^{rd} Edition), by Stephen Senn, in which there is a reference to the famous paper of Laird (1983, *The American Statistician* 37: 329–330).

Digression: why are crossover designs useful? (1/2)

In a crossover study, each participant receives more than one treatment at different times. Participants are first given one treatment and then after a washout period they change to the other treatment. This study design sees every person acting as their own comparison, which in turn helps the researchers better understand the effects of each treatment.



Source: `https://unimelb.libguides.com/whichstudytype/Cross-over`, accessed Nov 4, 2025.

Digression: why are crossover designs useful? (2/2)

A short answer:

"The within-patient aspect of the design eliminates a staggering number of potential confounders:[22] *perhaps upward of 20,000 genes, any epigenetic factors, and the whole environmental history of the subject up until the start of the trial. The analysis could not possibly account for these factors individually but does not have to. Their joint effect as confounders is eliminated by subtracting outcomes under placebo from those under ISF240* [i.e., experimental treatment] *to form paired differences, which are the basis for analysis."* [23]

[22] Here 'confounders' refers to prognostic baseline variables that will inevitably be randomly imbalance, to some extent, despite randomization.

[23] Senn. Journal of Clinical Epidemiology 148 (2022): 184-188.

Stratified randomization and adjustment

- *"Randomization is often stratified by baseline covariates. A covariate adjustment model should generally include strata variables and can also include covariates not used for stratifying randomization."* [24]
- Not including strata variables usually lead to confidence intervals that are too wide, type-I error rates that are too low and a reduction in power.[25]
- Although it certainly reduces random imbalance, stratified randomization does not provide more power than simple randomization, when adjusting on strata variables in any case (in "large" trials; asymptotic results). In other words, to maximize the power, covariate adjustment matters, but stratified randomization does not.[26]

[24] FDA scientific guidelines: "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry", 2023, accessed 29/10/2025, https://www.fda.gov/media/148910/download

[25] See e.g., Kahan and Morris (2012). Statistics in Medicine, 31(4):328–340.

[26] See e.g., Wang et al. (2021). Journal of the American Statistical Association, 116(542), 1152–1163.

Further remarks

- Do not adjust on post-randomization variables. Only use baseline variables (i.e., those known at time of randomization). Otherwise, this might introduce 'collider' bias.

  "It is not advisable to adjust the main analyses for covariates measured after randomisation because they may be affected by the treatments." [27]

- Pre-specification of which baseline covariates should be adjusted for, and how, is strongly recommended (i.e., using or not using interaction terms, categorization, splines, etc...).

  "Sponsors should prospectively specify the detailed procedures for executing covariate-adjusted analysis before any unblinding of comparative data. FDA review will emphasize the prespecified primary analysis rather than post-hoc analyses using different models or covariates."[28]

[27] EMA scientific guidelines: *ICH Topic E 9 Statistical Principles for Clinical Trials*, 1998, https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf

[28] FDA scientific guidelines: "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry", 2023, accessed 29/10/2025, https://www.fda.gov/media/148910/download

Outline/Intended Learning Outcomes (ILOs)

Baseline table
- ILO: repeat the main recommendations
- ILO: explain their rationale to skeptical colleagues and reviewers
- ILO: describe the relationship between randomization, baseline imbalance and the validity of statistical analyses

Covariate adjustment
- ILO: repeat the main recommendations
- ILO: explain how this can lead to power gains
- ILO: restate key robustness properties
- ILO: perform conventional analyses for binary and quantitative outcomes
- ILO: recall examples

Missing data
- ILO: list a few principles and type of missing data
- ILO: fit a Mixed Model for Repeated Measurements (MMRM) and interpret the results
- ILO: restate the rationale for using an MMRM analysis

Missing values in baseline follow-up study (Case in neurodegenerative disease)

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!

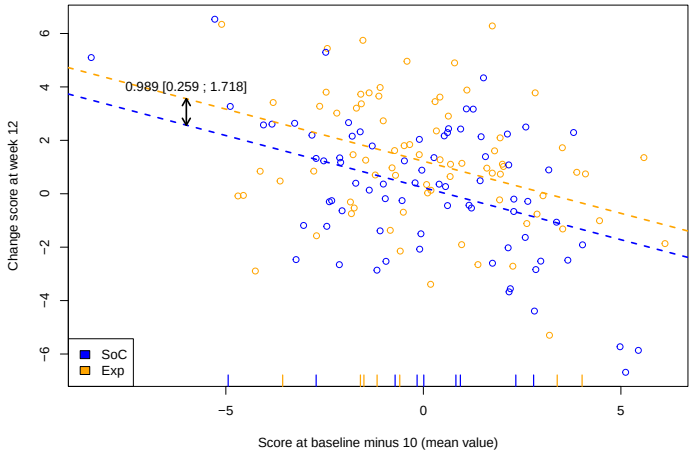
PRINCIPLES

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."*

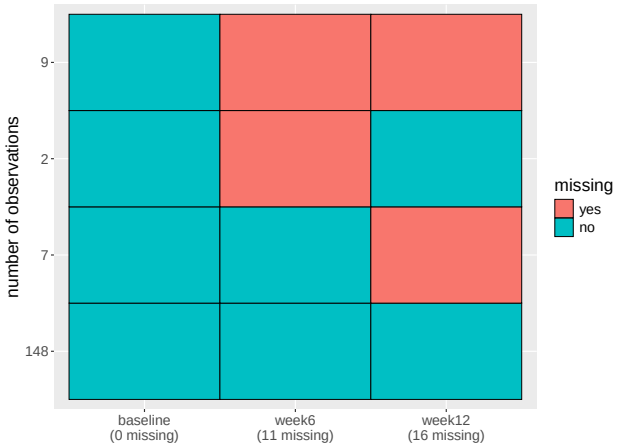
Note: the second principle is about well-defined estimands, the sixth about sensitivity analysis.

Source: 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.

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:

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

with

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

- x_i: baseline score of the i-th patient minus 10 (the "average" baseline score).
- z_j: indicates the week, equal 1 if j=1 (week 6), 0 if j=2 (week 12)
- 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 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)
```

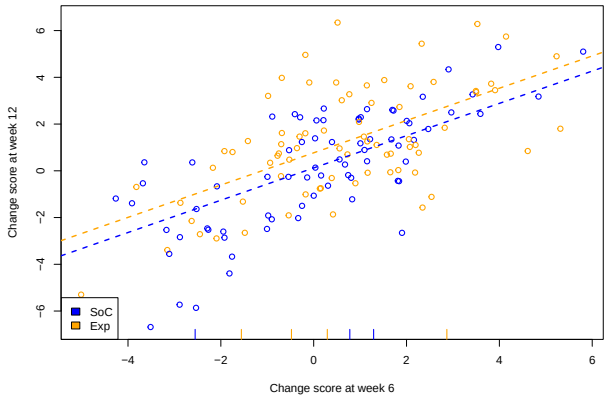
Output (partial):

```
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)

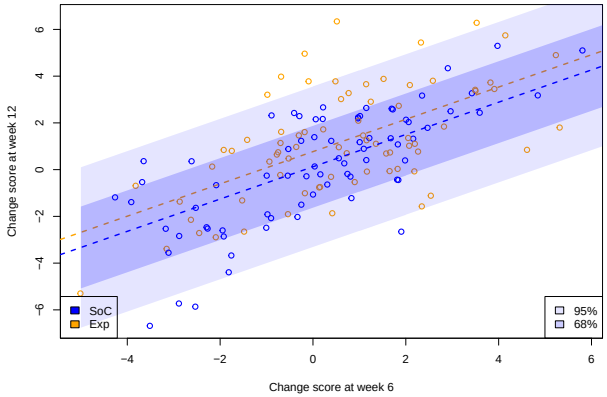
- μ: estimated as 0.226, is the mean change score at 12 weeks for a patient randomized to "SoC", with baseline score=10.
- β_1: estimated as -0.395, is the mean change in outcome (change score at 12 weeks) when comparing two patients of the same arm, one with a baseline score one unit larger than the 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.[29]
  - 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[30] 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.)

[29] See e.g., Wang et al. (2021). Journal of the American Statistical Association, 118(542), 1152–1163.

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

Exercise 2.3

1. Practice with R! Run the code of the script "`neuroRDemo.R`" and recognize the main results of the previous slides.
2. Do the complete case analysis, using an ANCOVA to adjust for baseline. How different are the results? Hint: to create the complete case data, you can use: `dCCA <- d[!is.na(d$week12),]`.
3. Fit the MMRM on the same complete data. What do you observe? Is it reassuring?

Exercise 2.4

Practice with the "Exercise.2.4" from the webpage.
It is about a knee surgery trial.