Faculty of Health Sciences

Day 1: fundamental concepts

Biostatistics for the design and analysis of clinical trials

Paul Blanche

Section of Biostatistics, University of Copenhagen

November 18, 2025

"We have to trust the scientific judgement of the scientists who ran the study. Statistics should be their handmaiden, not their jailer."

David Salsburg [1]

[1] Stat Med, vol 9, 201-211, 1990.

Outline/Intended Learning Outcomes (ILOs)

About the course

ILO: to know what to expect about/from the course

Trial objectives, outcomes and endpoints

ILO: outline why it matters for the statistics

ILO: describe superiority, non-inferiority and equivalence studies

Why do we need a control group?

ILO: recognise and exemplify the need for a control group

ILO: exemplify regression to the mean

Why randomizing?

ILO: explain the benefit and limits of randomization

ILO: exemplify confounding, relate it to blinding and non-concurrent controls

Power and sample size calculations

ILO: identify why and how to make power and sample size calculations

ILO: analyse their strengths and limitations

ILO: describe the EZ principle and apply it using R

Overall ILOs for this course

A student who has met the objectives of the course will be able to:

1. Relate key concepts in clinical trials to statistical thinking.
2. Take advice from a statistician to design a clinical trial and write a relevant statistical analysis plan.
3. Describe consensual recommendations to design and analyze clinical trials and restate their rationale.
4. Exemplify the use of different statistical methods to design and analyze clinical trials.
5. Carry out commonly used basic computation using the R software.

Course plan & topics

Day 1:	Essential concepts, trial objectives, randomization, need for a control group, blinding, power and sampe size calculation, EZ principle.
Day 2:	Baseline table, covariate adjustment, robustness, ANCOVA, regression standardization, missing data, MMRM (i.e., Mixed Model for Repeated Measurements).
Day 3: (part 1)	Exact unconditional tests and confidence intervals, (simple) sensitivity analysis.
(part 2)	Adherence, intention to treat & per protocol analyses, estimand framework, Statistical Analysis Plan
Day 4:	Multiple testing, Group-sequential trials, miscellaneous topics (Heterogeneous Effect, Response Adaptive Randomization)

Formalities

1. We start at 8:30 and end at 15:00 (or earlier).
2. This course gives **2.8 ECTS** points, **if** you attend $\geq 7/8$ half days. [2]
3. We work with R and we planned the teaching assuming that you have the **prerequisites stated on the course description**.
4. Material is available at the course Homepage:

`http://paulblanche.com/files/RCTcourse.html`

[2] we count the signatures

Textbooks

This course is not based on a single textbook, but the following were used to prepare the course material (mostly the two first).

- *Statistical Issues in Drug Development*, by Stephen Senn (3rd Edition, 2021).
- *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022).
- *Fundamental Concepts for New Clinical Trialists*, Scott Evans & Naitee Ting (2015).

Outline/Intended Learning Outcomes (ILOs)

About the course
- ILO: to know what to expect about/from the course

Trial objectives, outcomes and endpoints
- ILO: outline why it matters for the statistics
- ILO: describe superiority, non-inferiority and equivalence studies

Why do we need a control group?
- ILO: recognise and exemplify the need for a control group
- ILO: exemplify regression to the mean

Why randomizing?
- ILO: explain the benefit and limits of randomization
- ILO: exemplify confounding, relate it to blinding and non-concurrent controls

Power and sample size calculations
- ILO: identify why and how to make power and sample size calculations
- ILO: analyse their strengths and limitations
- ILO: describe the EZ principle and apply it using R

Why are clinical trials needed?

"In God we trust. All others must bring data."[3]

- Clinical trials make it possible to collect high quality data to answer specific clinical research questions.

Why (often) chosing a single primary outcome?

*"we want to answer **many questions**, but we often choose one primary outcome or **endpoint**, the outcome upon which the benefit of treatment or intervention will be judged. [...] The reason for choosing only one primary endpoint is the the scourge of **multiplicity**."* [4]

[3] quote widely attributed to Edwards Deming 1900–93, American statistician and management theorist; source Oxford Essential Quotations (6 ed.), 2018

[4] *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022).

Usually a single primary outcome:

"The ICH E9 guideline on biostatistical principles in clinical trials recommends that generally clinical trials have one primary variable. A single primary variable is sufficient, if there is a general agreement that a treatment induced change in this variable demonstrates a clinical relevant treatment effect on its own."[5]

But exceptions exist, e.g.:

Concept	Description
Co-primary endpoints	Both primary endpoints need to be superior to the control.
Dual primary endpoints	At least one primary endpoint needs to be superior to the control.

(Großhennig et al. Pharm Stat 22.5 2023: 836-845)

"the use of co-primary outcomes ensures that both use and safety outcomes are considered jointly. Researchers in the field of antimicrobial resistance are encouraged to consider the use of co-primary outcomes"[6]

[5] EMA scientific guidelines: *Points to consider on multiplicity issues in clinical trials*, 2002, www.ema.europa.eu/en/documents/scientific-guideline/points-consider-multiplicity-issues-clinical-trials_en.pdf

[6] Gillespie et al. *"Use of co-primary outcomes for trials of antimicrobial stewardship interventions."* Lancet Infectious Diseases 18.6 (2018): 595-597

Digression: outcome vs endpoint

"The term outcome usually refers to the measured variable (e.g., peak volume of oxygen or PROMIS Fatigue score), whereas an endpoint refers to the analyzed parameter (e.g., change from baseline at 6 weeks in mean PROMIS Fatigue score)." [7]

But some use the two terms interchangeably...

[7] Karen Staman, https://rethinkingclinicaltrials.org/chapters/design/choosing-specifying-end-points-outcomes/choosing-and-specifying-endpoints-and-outcomes-introduction/ (accessed Oct 2025)

Digression: mind the competing risk of death!

Choosing an appropriate endpoint for a given outcome is not always completely straightforward.

Example: *"The* [mean] *number of days hospitalized after surgery was not directly compared between the two groups to avoid that early death contributed to favorable outcomes. Instead, we compared the number of days alive that were not spent in the hospital within 90 days after surgery, as recommended elsewhere [14,15]."* (Graeser et al., Acta Anaesthesiol Scand. 2023)

14. McCaw ZR, Tian L, Vassy JL, et al. How to quantify and interpret treatment effects in comparative clinical studies of covid-19. Ann Intern Med. 2020;173(8):632-637.

15. Beyersmann J, Friede T, Schmoor C. Design aspects of COVID-19 treatment trials: improving probability and time of favorable events. Biom J. 2022;64(3):440-460.

Trial Objectives

In this course, we focus on:

- Superiority trial: *"designed to detect a difference between treatments. The first step of the analysis is usually a test of statistical significance to evaluate whether the results of the trial are consistent with the assumption of there being no difference in the clinical effect of the two treatments"* [8]

- Non-inferiority trial: *"aims to demonstrate that the test product is not worse than the comparator by more than a pre-specified, small amount. This amount is known as the non-inferiority margin"* [9]

But others exist: equivalence trials, dose-finding trials,...

[8] EMA scientific guidelines: *Points to consider on switching between superiority and non-inferiority*, 2000, https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-switching-between-superiority-and-non-inferiority_en.pdf

[9] EMA scientific guidelines: *Guideline on the choice of the non-inferiority margin*, 2005, https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-choice-non-inferiority-margin_en.pdf

Which statistical results to conclude?

Superiority trial:

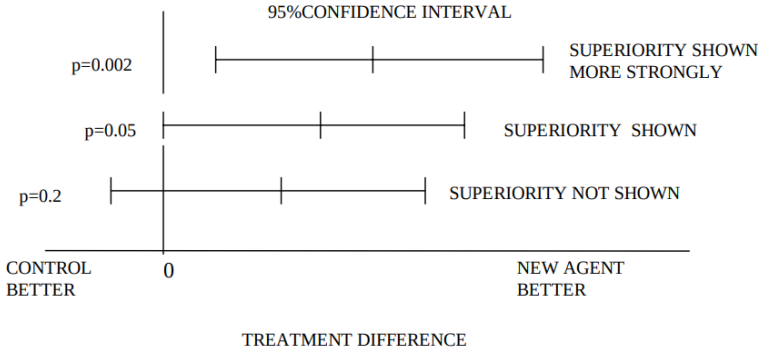


Figure 1: Relationship between significance tests and confidence intervals

"Whether the observed difference is indeed clinically relevant is a matter of judgement." (EMA scientific guidelines: *Points to consider on multiplicity issues in clinical trials*, 2002)

Non-inferiority trial:

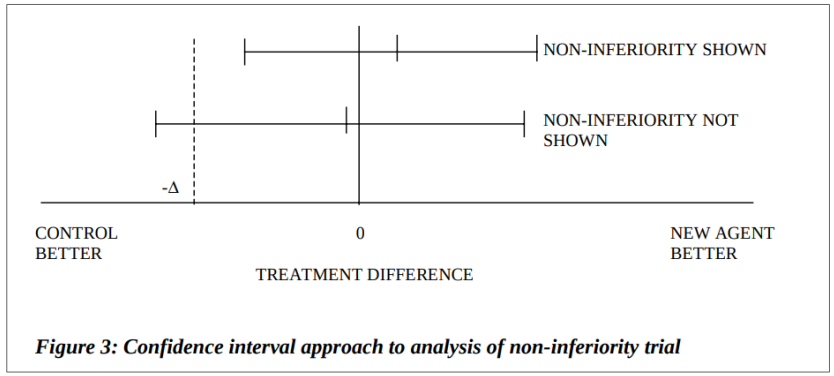


Figure 3: Confidence interval approach to analysis of non-inferiority trial

E.g.: $\Delta = 5$ mmHg in blood pressure

"The choice of delta must always be justified on both clinical and statistical grounds. It always needs to be tailored specifically to the particular clinical context and no rule can be provided that covers all clinical situations." (EMA scientific guidelines: *Guideline on the choice of the non-inferiority margin*, 2005)

Choice of non-inferiority margin

- Some guidelines are provided for trials for drug development in the pharmaceutical industry (M1, M2). [10] [11]
- They are unfortunately (often) of limited usefulness in pragmatic trial in acadamedia.
- Often, we should carefully consider:
  - Clinical relevance: benefit-risk ratio

    larger benefits (e.g., substnatially shorter duration of IV antibiotic treatment or hospitalization) may justify larger non-inferiority margins to evaluate safety (e.g., risk of re-infection).
  - Feasibility: power/sample size

    smaller margins call for larger sample size
  - Objectives: "alternative" or"replacement" trial. [12]

    smaller margins might be neede when we aim to replace current standard of care, but not when the aim is to provide an "alternative" options to clinicians and patients

[10] EMA scientific guidelines: *Guideline on the choice of the non-inferiority margin*, 2005, https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-choice-non-inferiority-margin_en.pdf

[11] FDA scientific guidelines: *Guidance for Industry Non-Inferiority Clinical Trials*, 2016, https://www.fda.gov/media/78504/download

[12] Tweed et al. "Exploring different objectives in non-inferiority trials." BMJ 385 (2024).

Choice of primary outcome: statistical considerations

In addition to clinical relevance, consider:

- Feasibility, especially power/sample size calculations
- Missing data: missingness might be informative and lead to bias

  (e.g.: long term outcome and risk of dropout; with patient reported outcomes not all patients reply to questionnaires)
- The measurement principle:

  "The process of measurement of the primary endpoint should not be influenced by treatment. Of course the value of the measurement will be influenced by an effective treatment, but how it is measured should not be." [13]

  Rationale: the two arms should be as similar as possible with respect to everthing except the treatment startegies being compared. If a difference is observed, we must be able to confidently conclude that it can only come from the difference in treatment startegies.

  Example 1: choose readmission within 28 days from treatment start instead of 21 after treatment stopped; two arms are "long" vs "short/individualized" antibiotic treatment duration (7 days vs 2-4 days).

  Example 2: the principle is violated when the amount of monitoring for events differs by arm (e.g., outcome retrieved from registries, diagnosis made during follow-up visits and clinical follow-up visits scheduled differently in the two arms).

[13] *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022).

Appendix: additional considerations

5.1.4.1 Desirable Characteristics of Endpoints

The motivation for every clinical trial begins with a scientific research question. The primary objective of the trial is to address the scientific question by collecting appropriate data. The selection of the primary endpoint is made to address the primary objective of the trial. The primary endpoint should be clinically relevant, interpretable, sensitive to the effects of the intervention, practical and affordable to measure, and ideally can be measured in an unbiased manner (Table 5.3).

TABLE 5.3

Considerations for Endpoint Selection

- Clinical relevance to address the objective
- Is it a measure of how a patient feels, functions, or survives?
- Interpretability
- Sensitivity to intervention effects (i.e., assay sensitivity)
- Practicality to obtain (e.g., invasiveness)
- Cost
- Accuracy of measurement (susceptibility to bias)
- Reproducibility of measurement
- Timeliness of result availability
- Susceptibility to manipulation
- Scale of measurement (e.g., continuous, binary, nominal, and event time)
- Impact on required sample size
- Susceptibility to missing data

Source: page 75 in *Fundamental Concepts for New Clinical Trialists*, Scott Evans & Naitee Ting (2015).

Why are clinical trials needed?

"In God we trust. All others must bring data."[14]

- Clinical trials make it possible to collect high quality data to answer specific clinical research questions.

As we shall see, getting valuable data (most often) requires:

- A control group
- Randomization
- Power and sample size calculation

[14] quote widely attributed to Edwards Deming 1900–93, American statistician and management theorist; source Oxford Essential Quotations (6 ed.), 2018

Clinical trials can deliver the "strongest" level of evidence to conclude that clinical practices can be improved via (almost) any possible intervention, e.g.:

- using a new drug
- a new usage of an existing drug
- a new surgery technique
- a new diagnostic procedures to guide a treatment decision

Although the studied interventions and clinical research questions differ widely from trial to trial, most trials are designed similarly with regards to control group, randomization and power calculation, for very sensible reasons!

The most common clinical trials are:

- paralell-group [15]
- 1:1 randomized.

And they are designed using:

- power and sample size calculations.

15 "paralell-group" refers to a specific choice of the control group.

The most common clinical trials are:

- paralell-group [15]
- 1:1 randomized.

And they are designed using:

- power and sample size calculations.

By the end today, you should be able to:

- explain
- exemplify

the rationale for each of these three concepts.

15 "paralell-group" refers to a specific choice of the control group.

Outline/Intended Learning Outcomes (ILOs)

About the course
- ILO: to know what to expect about/from the course

Trial objectives, outcomes and endpoints
- ILO: outline why it matters for the statistics
- ILO: describe superiority, non-inferiority and equivalence studies

Why do we need a control group?
- ILO: recognise and exemplify the need for a control group
- ILO: exemplify regression to the mean

Why randomizing?
- ILO: explain the benefit and limits of randomization
- ILO: exemplify confounding, relate it to blinding and non-concurrent controls

Power and sample size calculations
- ILO: identify why and how to make power and sample size calculations
- ILO: analyse their strengths and limitations
- ILO: describe the EZ principle and apply it using R

Why do we need a control group?

1.2 Purpose of control group

Control groups have one major purpose: to allow discrimination of patient outcomes (for example, changes in symptoms, signs, or other morbidity) caused by the test treatment from outcomes caused by other factors, such as the natural progression of the disease, observer or patient expectations, or other treatment. The control group experience tells us what would have happened to patients if they had not received the test treatment or if they had received a different treatment known to be effective.

Source: EMA scientific guidelines , ICH Topic E 10 Choice of Control Group in Clinical Trials, available at `www.ema.europa.eu/en/documents/scientific-guideline/ich-e-10-choice-control-group-clinical-trials-step-5_en.pdf`.

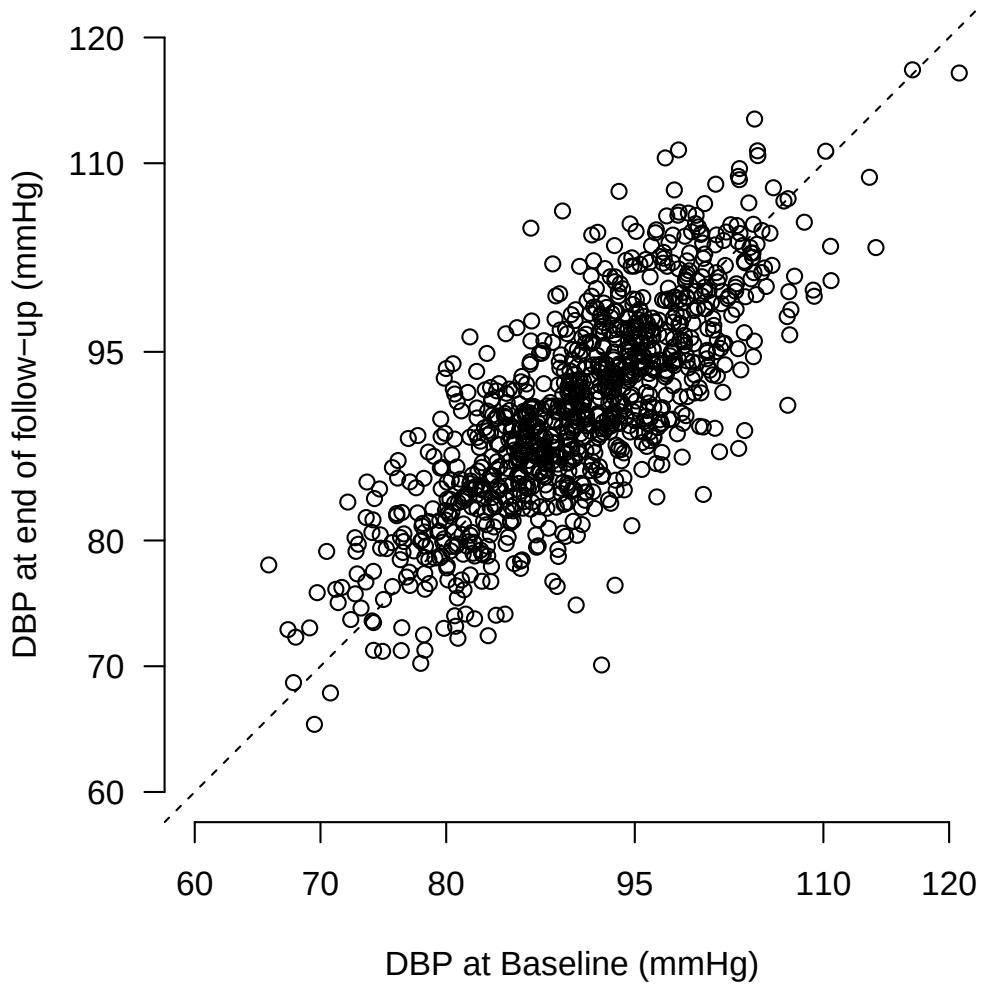
- "Natural progression" relates to regression to the mean, a concept which is:

  "so trivial that all should be capable of learning it and so deep that many scientists spend their whole career being fooled by it."[16]

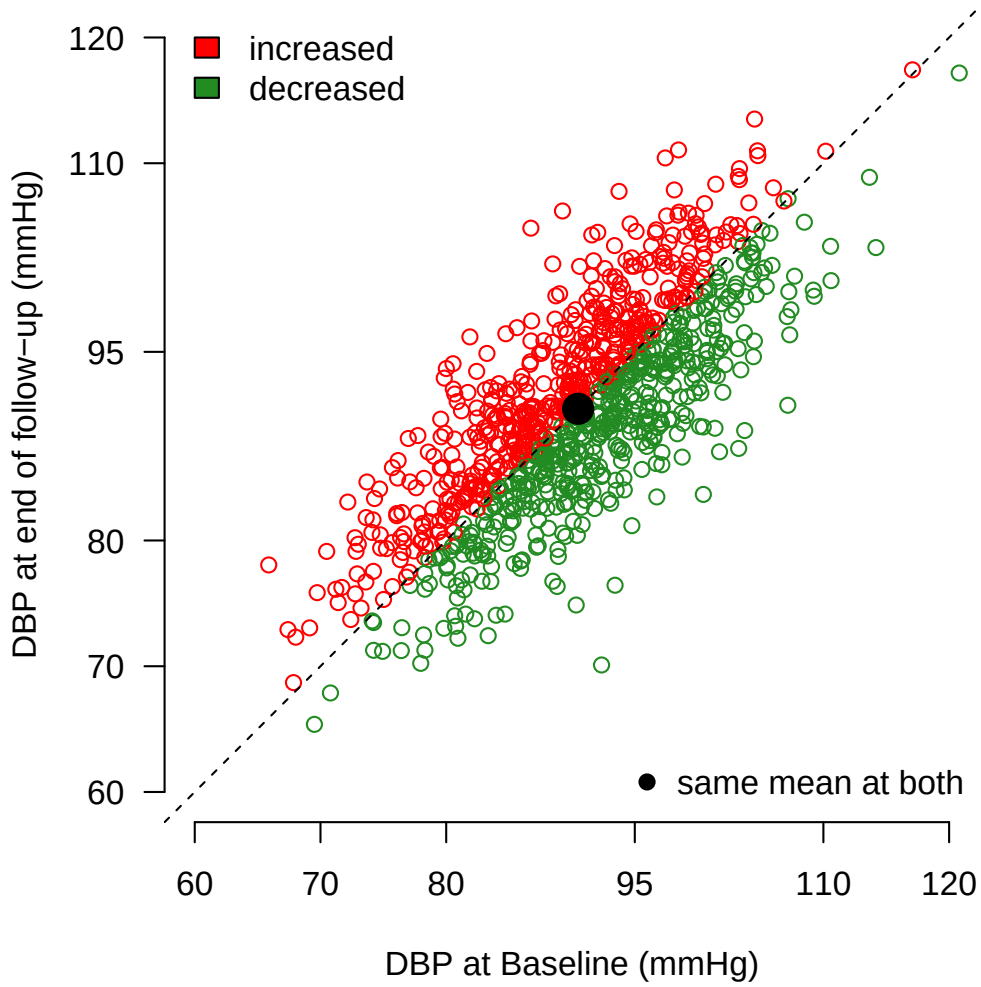
- In short, *"Regression to the mean is a consequence of the observation that, on average, extremes do not survive."*[16]

[16] Senn. "Francis Galton and regression to the mean." Significance 8.3 (2011): 124-126. The next slides closely follow this paper.

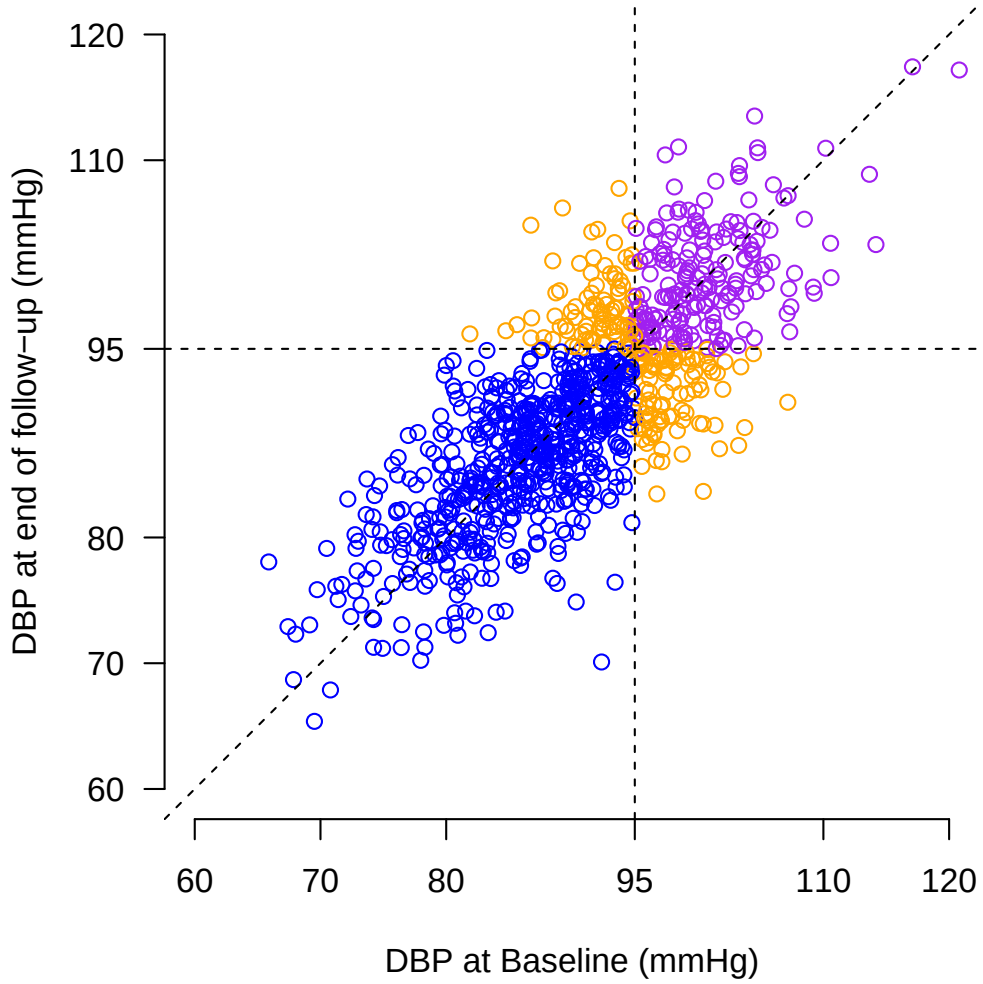
Diastolic blood pressure: "Random sample" (n=1000)



Diastolic blood pressure: "Random sample" (n=1000)

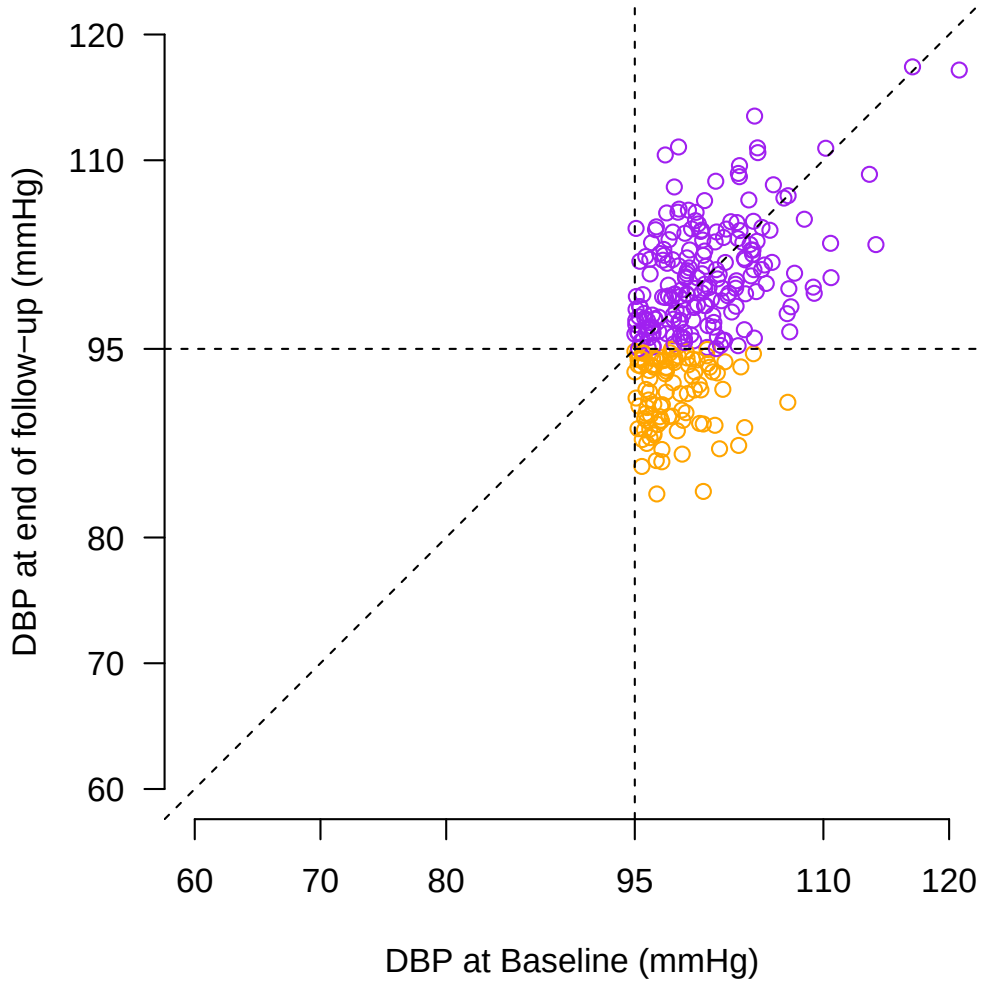


Diastolic blood pressure: "Random sample" (n=1000)



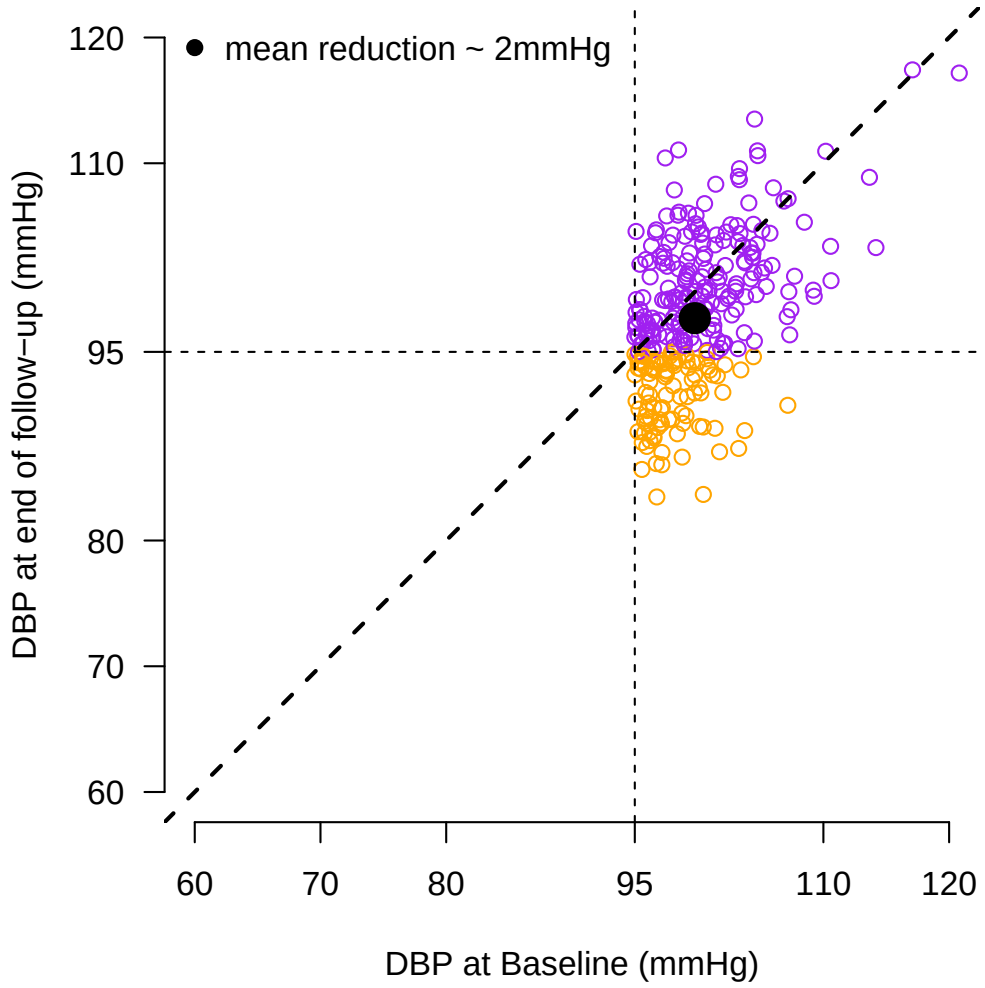
- "Hypertensive" if > 95mmg.

Diastolic blood pressure: "Clinical trial"



- Only "Hypertensive" patients are included in the trial.

Diastolic blood pressure: "Clinical trial"

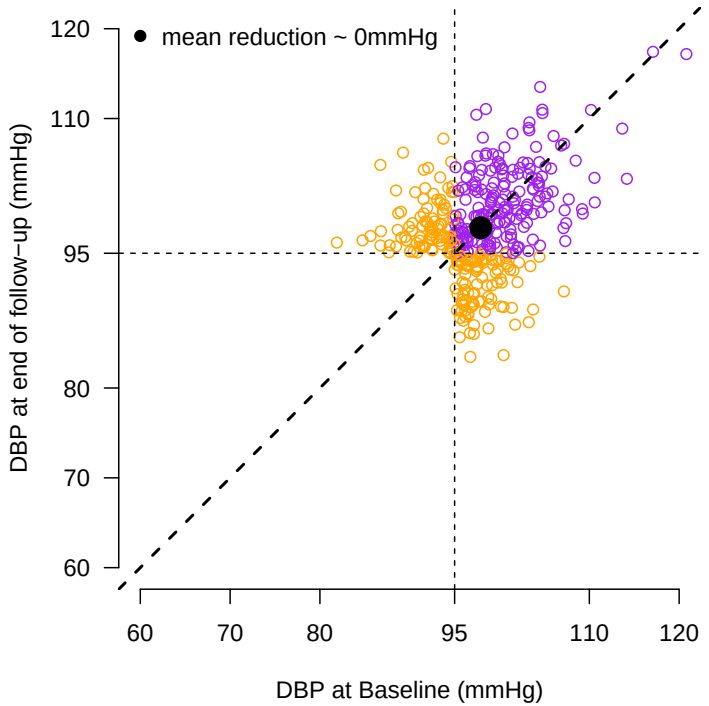


- Only "Hypertensive" patients are included in the trial.
- The subjects have improved!

Appendix: more details

[In this hypothetical trial,] *"We can only see patients who remain hypertensive or who become normotensive. We left out the patients who were normotensive but became hypertensive. They are shown in* [the right Figure]. *If we had their data they would correct the misleading picture in* [the previous Figure], *but the way we have gone about our study means that we will not see their outcome values."*

(Senn. "Francis Galton and regression to the mean." Significance 8.3 (2011): 124-126.)



Consequences on baseline follow-up studies

- We can (almost) always expect a spontaneous improvement from baseline when we include "symptomatic" patients.
- This is often observed in a placebo (control) group and it usually has nothing to do with a genuine placebo effect [17]. This is just the consequence of inclusion criteria and random variation.
- Lesson: The control group tells us what would have happened to patients if they had not received the treatment. By comparing the treated group to the placebo group, we can learn whether the improvement in a treated group is the results of more than just spontaneous improvements (i.e. regression to the mean) or standard care given to the patients irrespective of the studied treatment[18].

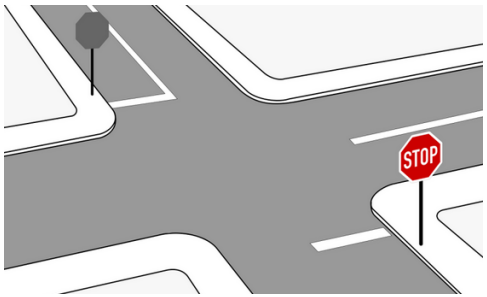
[17] Only in area of pain control does there seems to be reliable evidence of a placebo effect.

[18] Patients included in trials usually receive a lot more medical care than just the treatment studied and are often more closely followed up by doctors than they would outside of a trial.

Digression: thought-provoking joke

"Regression to the mean is not just limited to clinical trials. Did you choose dangerous road intersections in your region for corrective engineering work based on their record of traffic accidents? Did you fail to have a control group of similar black spots that went untreated? Are you going to judge efficacy of your intervention by comparing before and after? Then you should know that Francis Galton's regression to the mean predicts that sacrificing a chicken on such black spots can be shown to be effective by the methods you have chosen."

(Senn. "Francis Galton and regression to the mean." Significance 8.3 (2011): 124-126.)



Outline/Intended Learning Outcomes (ILOs)

About the course
- ILO: to know what to expect about/from the course

Trial objectives, outcomes and endpoints
- ILO: outline why it matters for the statistics
- ILO: describe superiority, non-inferiority and equivalence studies

Why do we need a control group?
- ILO: recognise and exemplify the need for a control group
- ILO: exemplify regression to the mean

Why randomizing?
- ILO: explain the benefit and limits of randomization
- ILO: exemplify confounding, relate it to blinding and non-concurrent controls

Power and sample size calculations
- ILO: identify why and how to make power and sample size calculations
- ILO: analyse their strengths and limitations
- ILO: describe the EZ principle and apply it using R

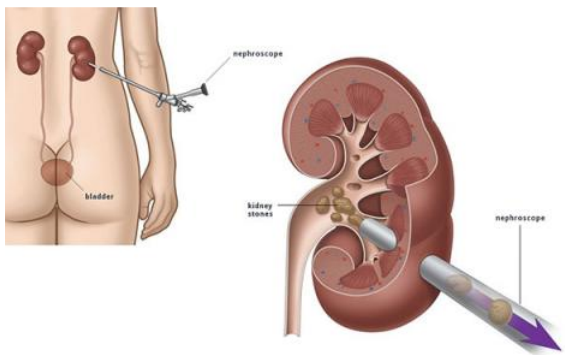
Example:

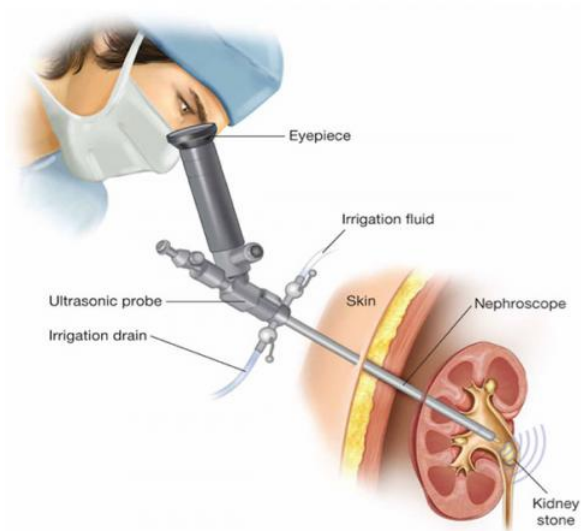
Comparing open surgery to percutaneous nephrolithotomy to remove kidney stones.

(like in "Grey's Anatomy" television series)

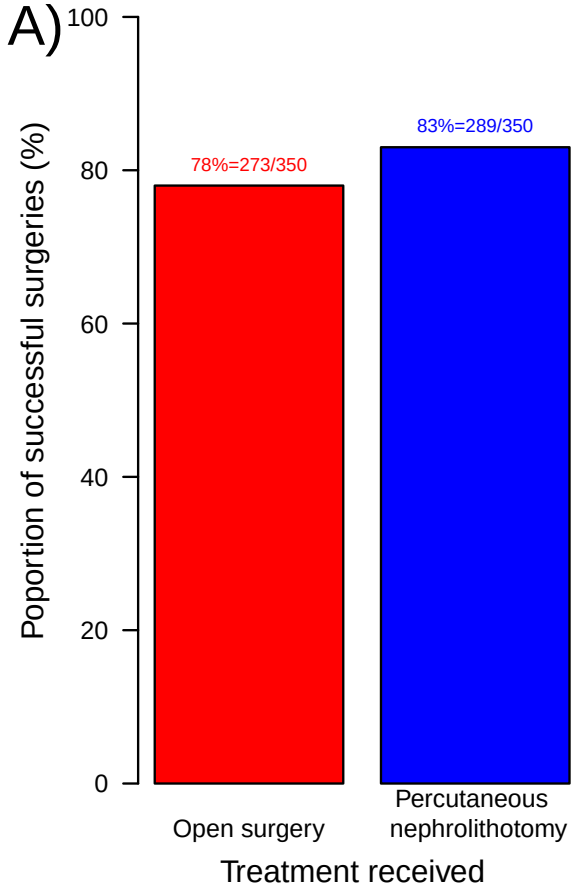
versus

A small tube is inserted into the kidney through the back. A nephroscope (a special telescope) is then inserted into the kidney, through which the stone can be disintegrated and removed.



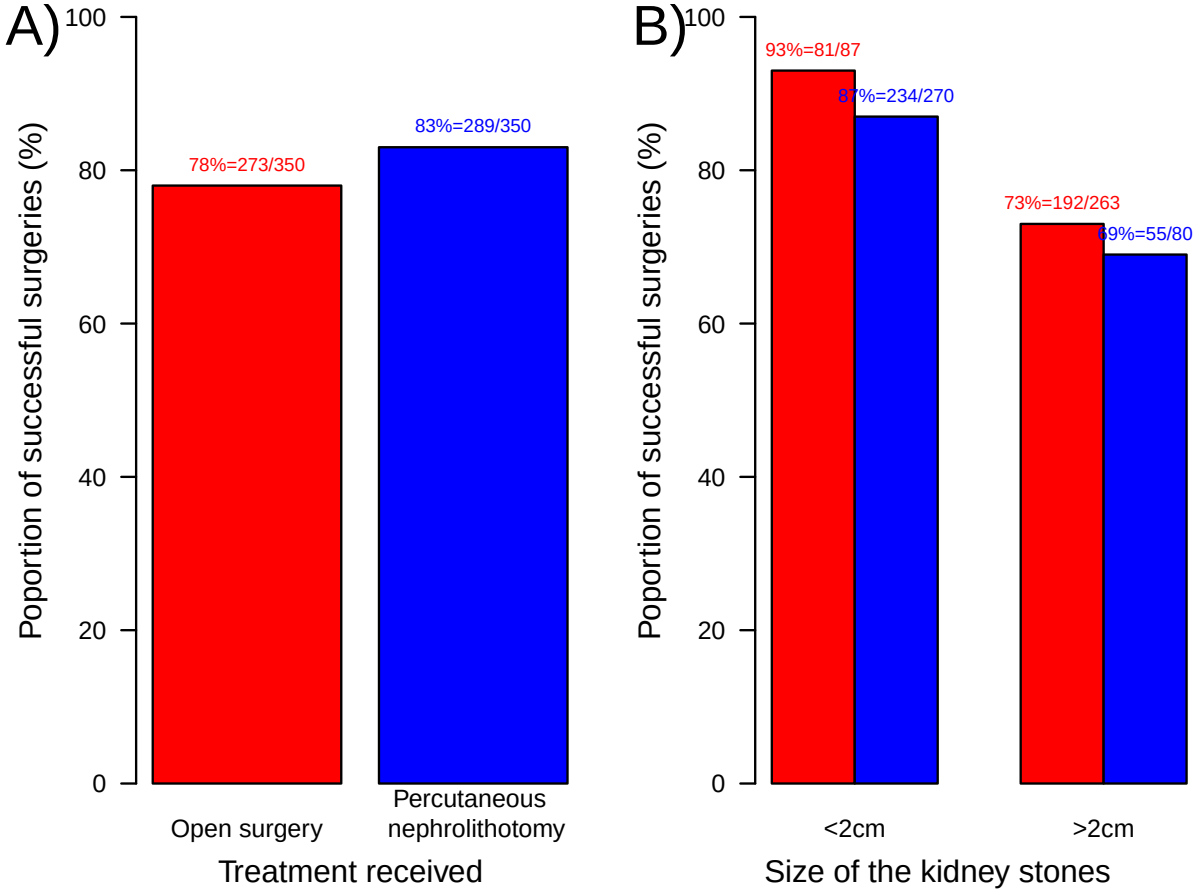


Imagine that you see these results (these are real data)



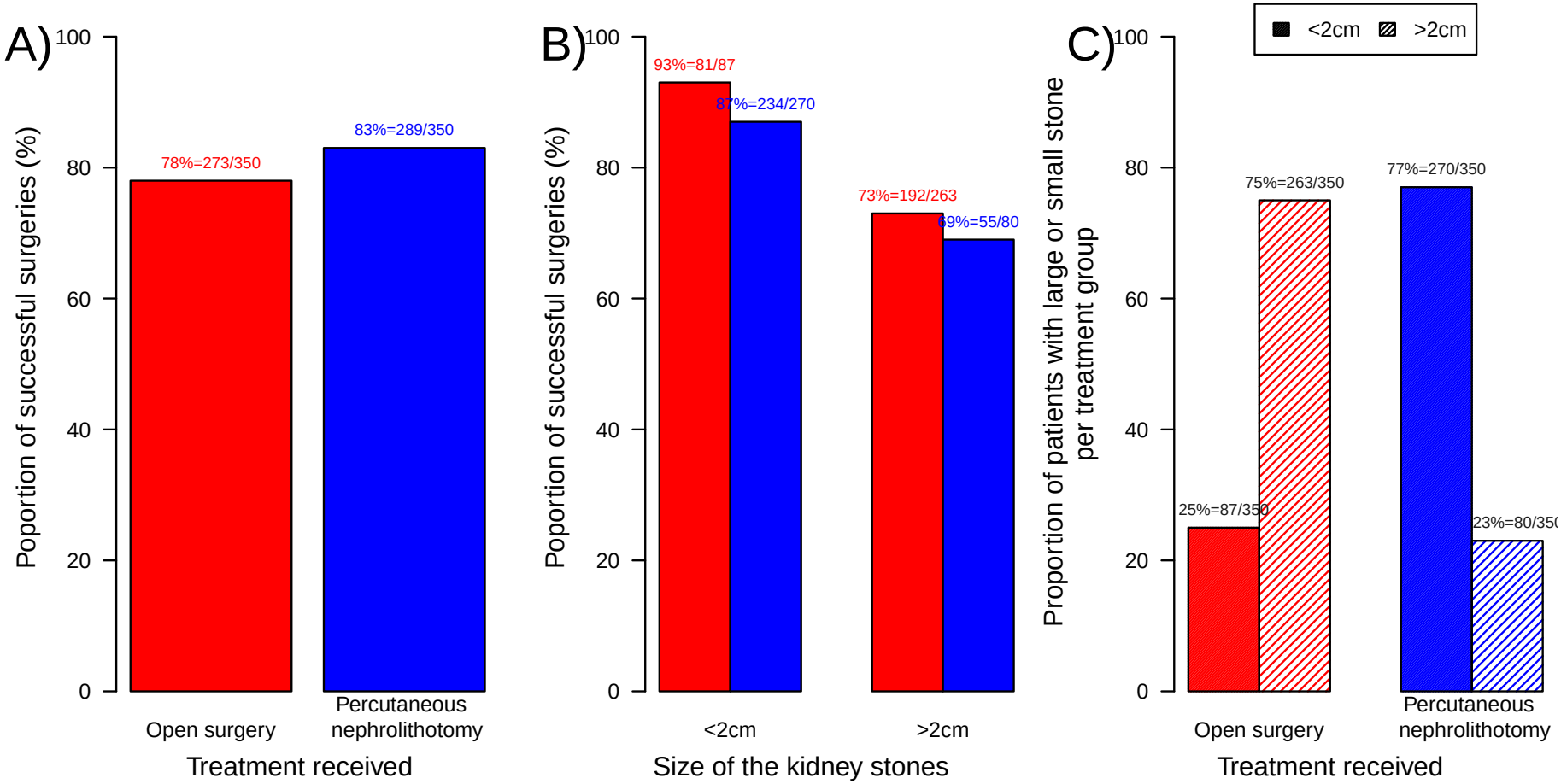
Which treatment looks best?

Imagine that you see these results (these are real data)



Which treatment looks best?

Imagine that you see these results (these are real data)



Which treatment looks best?

Source: Julious & Mullee. Confounding and Simpson's paradox. BMJ. 1994 Dec 3;309(6967):1480-1.

Wrapping-up

- Overall (i.e., not looking at subgroups), open surgery looks worst.
- "Paradoxically", open surgery looks best for both patients with small stones and patients with big stones.

Wrapping-up

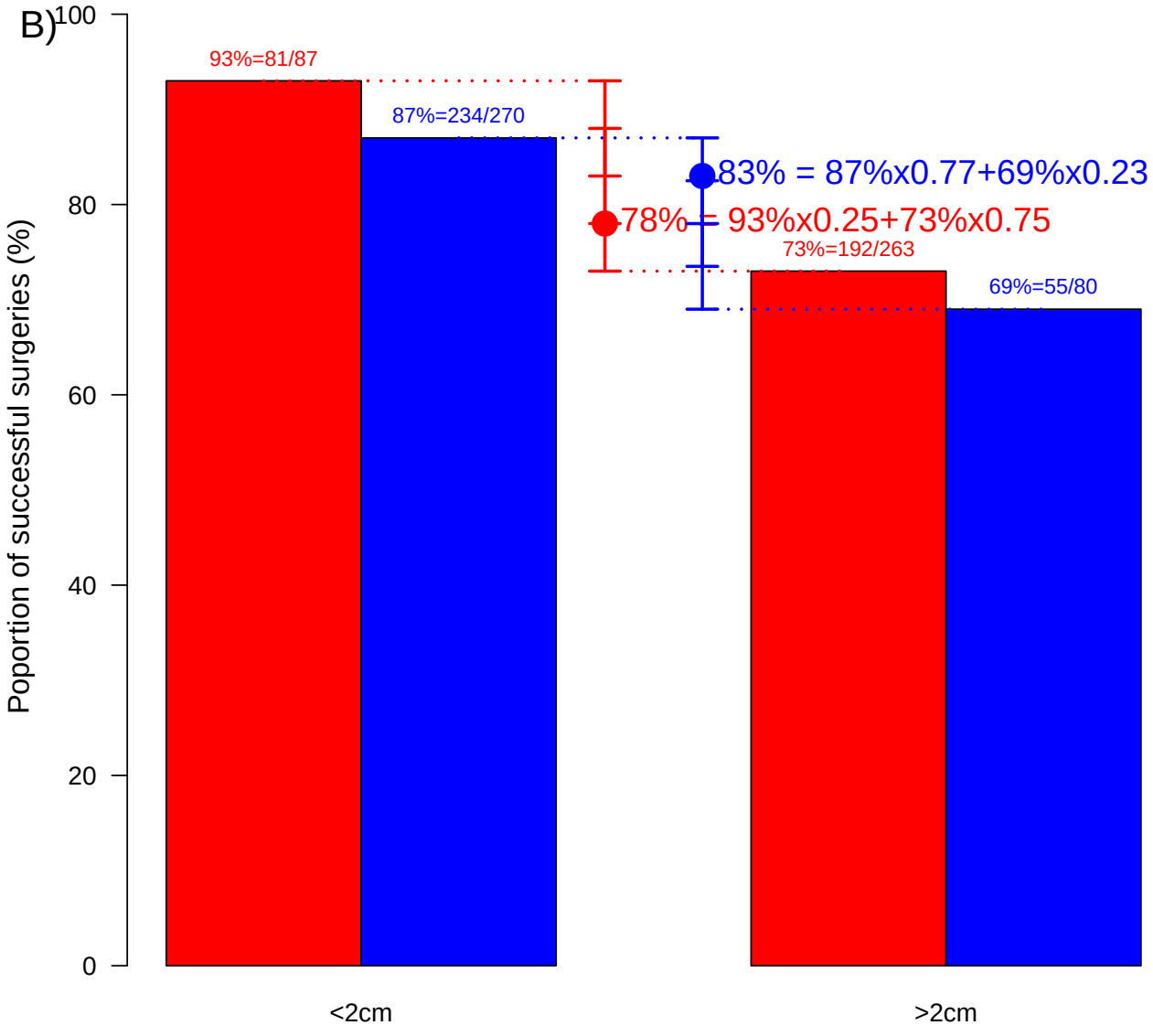
- Overall (i.e., not looking at subgroups), open surgery looks worst.
- "Paradoxically", open surgery looks best for both patients with small stones and patients with big stones.
- This happens because most patients receiving open surgery are "easy" patients (with small stones) while those receiving the other surgery are "difficult" patients (with big stones) (whatever the treatment, successful surgery is more difficult to achieve with big stones than with small stones).
  - $78\% = 93\% \times 0.25 + 73\% \times 0.75$
  - $83\% = 87\% \times 0.77 + 69\% \times 0.23$

Wrapping-up

- Overall (i.e., not looking at subgroups), open surgery looks worst.
- "Paradoxically", open surgery looks best for both patients with small stones and patients with big stones.
- This happens because most patients receiving open surgery are "easy" patients (with small stones) while those receiving the other surgery are "difficult" patients (with big stones) (whatever the treatment, successful surgery is more difficult to achieve with big stones than with small stones).
  - $78\% = 93\% \times 0.25 + 73\% \times 0.75$
  - $83\% = 87\% \times 0.77 + 69\% \times 0.23$

  Had the proportions of patients with small and big stones been the same in both treatment groups, the overall results (i.e., not looking at subgroups) would not have been misleading; they would have been in the same direction as those of the subgroups.

Weighted averages with different weights!



Why randomizing?

- Differences in patient characteristics between treatment arms can explain, at least to some extent, differences in outcomes between the treatment arms.

- Ideally, we would like the two arms to be identical with respect to everything except the treatment they receive. Hence, if a difference in outcome between the treatment arms is observed, we can safely conclude that this is due to the unique difference between the two arms: the treatment.

- How can we make sure that the two arms are identical, at least for all what matters (e.g., age, comorbidity, lifestyle, genotype...)?

19 Senn, *Statistical issues in drug development*, 2021, page 39.

Why randomizing?

- Differences in patient characteristics between treatment arms can explain, at least to some extent, differences in outcomes between the treatment arms.

- Ideally, we would like the two arms to be identical with respect to everything except the treatment they receive. Hence, if a difference in outcome between the treatment arms is observed, we can safely conclude that this is due to the unique difference between the two arms: the treatment.

- How can we make sure that the two arms are identical, at least for all what matters (e.g., age, comorbidity, lifestyle, genotype...)? We cannot: *"The two groups in a clinical trial will never be equal."* [19]

[19] Senn, *Statistical issues in drug development*, 2021, page 39.

Why randomizing?

- Differences in patient characteristics between treatment arms can explain, at least to some extent, differences in outcomes between the treatment arms.

- Ideally, we would like the two arms to be identical with respect to everything except the treatment they receive. Hence, if a difference in outcome between the treatment arms is observed, we can safely conclude that this is due to the unique difference between the two arms: the treatment.

- How can we make sure that the two arms are identical, at least for all what matters (e.g., age, comorbidity, lifestyle, genotype...)? We cannot: *"The two groups in a clinical trial will never be equal."* [19]

- How can we make sure that the two arms are "not too different" or, in other words, "similar enough"?

[19] Senn, *Statistical issues in drug development*, 2021, page 39.

Why randomizing?

- Differences in patient characteristics between treatment arms can explain, at least to some extent, differences in outcomes between the treatment arms.

- Ideally, we would like the two arms to be identical with respect to everything except the treatment they receive. Hence, if a difference in outcome between the treatment arms is observed, we can safely conclude that this is due to the unique difference between the two arms: the treatment.

- How can we make sure that the two arms are identical, at least for all what matters (e.g., age, comorbidity, lifestyle, genotype...)? We cannot: *"The two groups in a clinical trial will never be equal."* [19]

- How can we make sure that the two arms are "not too different" or, in other words, "similar enough"? We randomize!

[19] Senn, *Statistical issues in drug development*, 2021, page 39.

Why does randomization do the job?

Cornfield (1959, p. 245) summarized the importance of randomization:

> 1. It controls the probability that the treated and control groups differ more than a calculable amount in their exposure to disease, in immune history, or with respect to any other variable, known or unknown to the experimenter, that may have a bearing on the outcome of the trial. This calculable difference tends to zero as the size of the two groups increase.
> 2. It makes possible, at the end of the trial, the answer to the question "In how many experiments could a difference of this magnitude have arisen by chance alone if the treatment truly has no effect?" It may seem mysterious that a mathematician could actually predict the course of future experiments. All you have to do is compute what would happen if a given set of numbers were randomly allocated in all possible ways between the two groups. Randomization allows this.

Rosenberger & Lachin. *Randomization in clinical trials: theory and practice*. 2015, page 7.

Lessons

Theoretically and to a large extent in practice, randomization makes the two arms "similar enough" to derive reliable conclusions, because it:

1. prevents (confounding) bias. This means that, in average, the estimated treatment effects are not too small, not to large, just how they should be.

2. makes p-values and confidence interval reliable to quantify statistical uncertainty[20]. This means that about 95% of all 95% confidence intervals indeed contain the treatment effect that we seek to estimate and that less than 5% of all statistically significant results (p-value <5%) are false positive results.

[20] Importantly, this is achieved without the need of relying on modeling assumptions, unlike what we usually need in epidemiology, using observational data

Utmost good faith

Randomization is also *"a means of establishing utmost good faith between trialist and sponsor on the one hand and sponsor and regulator on the other."* [21].

"For example, even if a sponsor had balanced a clinical trial by known important prognostic factors (in practice this would be rather difficult, if not impossible) and these had been taken into account in the analysis, if he had allocated the patients according to a particular plan among many which might have satisfied the requirement for balance, the regulator might suspect his motives. He might suppose that some hidden characteristics of the patients were being used to give an advantage to the sponsor's drug. There is, in fact, evidence that there is a tendency for investigators to bias trials which are not randomized in favour of the experimental treatment."

Digression: we will see that writing a Statistical Analysis Plan also relate to the concept of utmost good faith.

[21] (*Statistical Issues in Drug Development*, Senn, 3rd Edition, 2021, page 40)

How do we randomize? (most common choices)

- Tossing a coin
- Permuted block randomization (balance not only at the end, but throughout a trial)

Table 6.1 Randomized blocks for a trial with two treatments and block sizes of four.

Patient	Block No.	Block sequence	Treatment
1	1	BAAB	B
2	1	BAAB	A
3	1	BAAB	A
4	1	BAAB	B
5	2	ABAB	A
6	2	ABAB	B
7	2	ABAB	A
8	2	ABAB	B
9	3	BAAB	B
10	3	BAAB	A
11	3	BAAB	A
12	3	BAAB	B

- Stratified permuted block randomization, consisting of using separate blocks within subgroups (e.g. stratified on disease severity, to force the balance between the two arms, with respect to disease severity at inclusion)

Table from page 78 from Senn, *Statistical Issues in Drug development*, 2021.

Digression: blinding

Randomization is necessary to make the patients in the two arms "similar enough", but it is usually insufficient. Blinding is often needed in addition to randomization.

1.2.2 Blinding

The groups should not only be similar at baseline, but should be treated and observed similarly during the trial, except for receiving the test and control drug. Clinical trials are often "double-blind" (or "double-masked"), meaning that both subjects and investigators, as well as sponsor or investigator staff involved in the treatment or clinical evaluation of subjects, are unaware of each subject's assigned treatment. Blinding is intended to minimize the potential biases resulting from differences in management, treatment, or assessment of patients, or interpretation of results that could arise as a result of subject or investigator knowledge of the assigned treatment. For example:

Remark: the purpose of using a placebo (instead of just no treatment) is blinding.

Source: EMA scientific guidelines, ICH Topic E 10 Choice of Control Group in Clinical Trials, available at www.ema.europa.eu/en/documents/scientific-guideline/ich-e-10-choice-control-group-clinical-trials-step-5_en.pdf

Digression: blinding & the double dummy technique

"When comparing two treatments in double blind trials, the treatments will almost never resemble each other. The most common way to hide what is being given is to use a placebo to each. Thus, in a trial comparing new to standard treatment each patient receives either new plus placebo to standard or standard plus placebo to new. This is called the double dummy technique." [22]

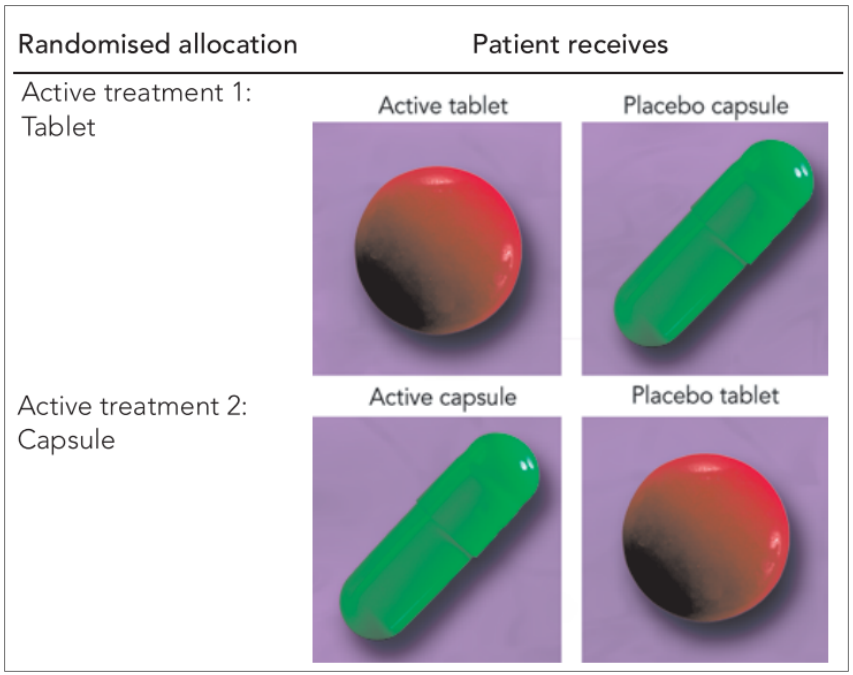


Figure from Forder et al. *"Allocation concealment and blinding: when ignorance is bliss."* Medical Journal of Australia 182.2 (2005): 8

[22] Senn, *Statistical issues in drug development*, 2021, page 41.

Digression: creativity might be needed for blinding!

"Acupuncture has gained increasing attention in the treatment of chronic pain. The lack of a satisfying placebo method has made it impossible to show whether needling is an important part of the method or whether the improvement felt by the patient is due to the therapeutic setting and psychological phenomena. Also, the effectiveness of acupuncture has not been demonstrated suficiently. We treated 52 sportsmen with rotator cuff tendinitis in a randomised single-blind clinical trial using a new placebo-needle as control. " [23]

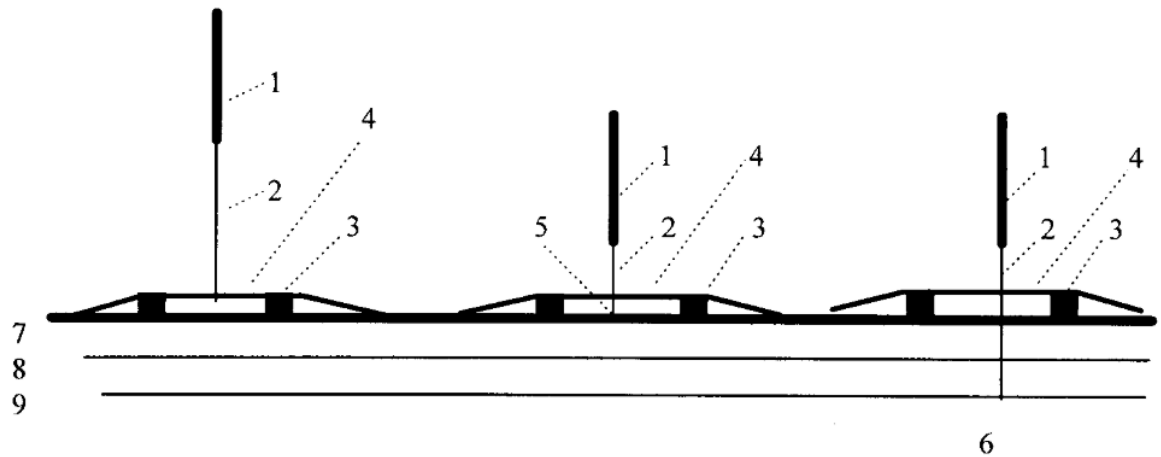


Fig. 1. Placebo-needle when touching the skin (left) and after retraction of the needle into the handle (middle), real acupuncture needle (right). Reproduction of this figure with permission of Streitberger and Kleinhenz, 1998. 1, Needle handle; 2, needle corpus; 3, plastic ring; 4, plaster; 5, blunt tip of the needle; 6, sharp tip of the needle; 7, cutis; 8, subcutis; 9, muscle.

[23] Kleinhenz et al. *"Randomised clinical trial comparing the effects of acupuncture and a newly designed placebo needle in rotator cuff tendinitis."* Pain 83.2 (1999): 235-241.

Blinding: simple, double, triple?

- Single blind: Either the patient or clinician (usually the patient) remains unaware of the treatment assignment
- Double blind: Both the patient and investigator are unaware of the allocated treatment.
- Open label: All parties are aware of treatment being received after randomisation.
- Triple blind: The patient, the investigator and either those who adjudicate the study outcomes (the outcomes assessment committee) or those who monitor the study safety (the safety and data monitoring committee) are unaware of the allocated treatment.

For more details, read e.g., Forder et al. "Allocation concealment and blinding: when ignorance is bliss." Medical Journal of Australia 182.2 (2005): 87-89.

Note of caution:

"A problem with this lexicon is that there is great variability in clinician interpretations and epidemiological textbook definitions of these terms.[...] *Authors should instead explicitly report the blinding status of the people involved for whom blinding may influence the validity of a trial."*[24].

[24] 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

Appendix: Blinded outcome assessment

Blinded assessment of outcomes is possible in most trials. It is usually suitable for subjective outcomes (e.g., cause of death, presence of sequelae).

"The primary outcome was sequelae after 6 months in patients with BJIs, defined as any atypical mobility or function of the affected bone or joint, assessed blindly" [25]

[25] 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.

Must blinded trials test for blindness?

Some think that patients should be asked to name their treatment to check whether the trial was blind.

But, many are skeptical about this idea.

- *"Testing for 'blindness' may not, and often can't, generate valid answers"* [...] *"End-of-trial tests for 'blindness' can't be done with validity, because they can't distinguish blindness from hunches about efficacy."* [...] *"We neither can nor need to test for blindness during and after trials, but we must bear in mind the bias-generating consequences that result from its loss"*[26]

- *"My view of blinding is that it is meant to make sure that identification of the treatment is not possible by the patient using irrelevant characteristics (such as for example colour or size of a pill). However, efficacy is not irrelevant, it is essential. If there is a difference in the effects of two treatments being compared, then this ought to provide information that would help patients to guess what treatment they are being given. Suppose for example, that all patients who consider their state of health at the end of the trial is satisfactory guess that they received the experimental treatment and all those who consider their state is less than satisfactory guess placebo. If the treatment is effective they will guess correctly more often than one would expect by chance."* [27]

[26] Sackett. "Commentary: Measuring the success of blinding in RCTs: don't, must, can't or needn't?." International journal of epidemiology 36.3 (2007): 664-665.

[27] page 97 in *Statistical Issues in Drug Development*, by Senn (3rd Edition, 2021)

Concurrent controls

> A concurrent control group is one chosen from the same population as the test group and treated in a defined way as part of the same trial that studies the test treatment, and over the same period of time. The test and control groups should be similar with regard to all baseline

Rationale for using concurrent controls:

- things might change over time. E.g., some intensive care units were organized very differently before, during and after the COVID-19 crisis.
- the two arms are not necessarily "similar enough" if patients are included at different times.

Source: EMA scientific guidelines, ICH Topic E 10 Choice of Control Group in Clinical Trials, available at `www.ema.europa.eu/en/documents/scientific-guideline/ich-e-10-choice-control-group-clinical-trials-step-5_en.pdf`.

Digression: platform trials

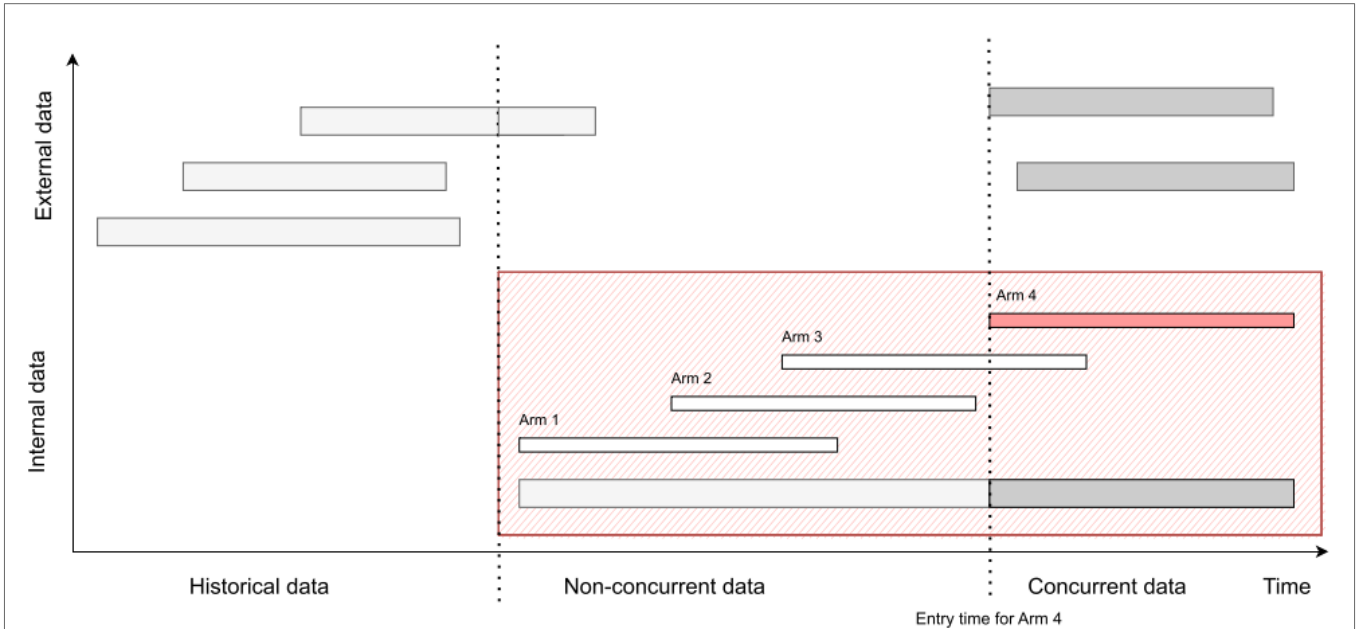


Fig. 1 Definition of controls depending on the source and time. The data within the red box represents the (internal) data from a platform trial, the data outside the red box represents the external data. Non-concurrent control data for arm 4 is represented in light grey boxes and concurrent controls are represented in dark grey

- Platform trials can lead to non-concurrent controls.
- Specific methods may be needed to mitigate concerns with non-comparability (e.g, methods similar to those for non-randomized studies).

Source: figure from Bofill Roig et al. *On the use of non-concurrent controls in platform trials: a scoping review.* Trials (2023) 24:408

Outline/Intended Learning Outcomes (ILOs)

About the course
- ILO: to know what to expect about/from the course

Trial objectives, outcomes and endpoints
- ILO: outline why it matters for the statistics
- ILO: describe superiority, non-inferiority and equivalence studies

Why do we need a control group?
- ILO: recognise and exemplify the need for a control group
- ILO: exemplify regression to the mean

Why randomizing?
- ILO: explain the benefit and limits of randomization
- ILO: exemplify confounding, relate it to blinding and non-concurrent controls

Power and sample size calculations
- ILO: identify why and how to make power and sample size calculations
- ILO: analyse their strengths and limitations
- ILO: describe the EZ principle and apply it using R

Case study

"All characters appearing in this story are fictitious. Any resemblance to real persons is purely coincidental."

Clinician: We are planning to start a new trial. We aim to show that an innovative way of treating some of our intensive care patients will reduce 30-day mortality. Can you help us to plan the trial, especially with regards to sample size calculation?

Statistician: Sure. What is the current 30-day mortality rate, approximately?

Clinician: Registry data show 17%, within the last 5 years.

Statistician: By how much do you think you can reduce this mortality, using the new treatment?

Clinician: By approximately one third, it's really promising you know! So, I expect approximately 11% mortality only, with the new intervention.

Statistician: Well, a standard (simple) calculation suggests that you need to include approximately 1048 patients, randomized 1:1, so 524 per arm, to have **80% power**, with a usual **type-I error** control at **5%**.

Reminders

- **Power:** the chance of obtaining a statistically significant result, when the scientific hypothesis that we aim to confirm is indeed correct (e.g. "this new treatment is better than standard of care"). In other words, the chance of a obtaining a "positive finding". It can be computed for a given treatment effect, e.g., the mortality changes from 17% (with current standard of care) to 11% (when using the new treatment).

- **Type-I error**: a false positive finding. In other words, concluding that the scientific hypothesis that we aim to confirm is correct although it is wrong (e.g., concluding that an ineffective drug is better than a placebo).

- **5% level**: *"Conventionally the probability of type I error is set at 5%"* [28]

Note: *"The probability of type II error* [i.e., one minus the power] *is conventionally set at 10% to 20%; it is in the sponsor's interest to keep this figure as low as feasible especially in the case of trials that are difficult or impossible to repeat."* [28]

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

Textbook formula ("large n" approximation)

$$n = \frac{\left\{ z_{\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} + z_{\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)} \right\}^2}{(p_1-p_2)^2}$$

- $z_{\alpha/2} = -1.96$ for $\alpha = 5\%$.[29]
- $z_{1-\beta} = 0.84$ and 1.28 for $1-\beta = 80\%$ and 90%.
- $\bar{p} = (p_1+p_2)/2$.
- n: number of observations in **each** group.

Useful for computing:

- Sample size: n for given "guesses" of p_1 and p_2 and desired $1-\beta$ and α.
- Power for a given budget/sample size: $1-\beta$ for "guesses" of p_1 and p_2 and desired n and α.
- Least detectable difference (or ratio): $\delta = p_1 - p_2$ (or $r = p_1/p_2$) for given n, "guess" of p_1 and desired α and minimal power $1-\beta$.

[29] where z_γ is the γ-quantile of a standard normal distribution.

Why randomizing 1:1?

- The sample size of the two arms do not necessarily need to be equal. We can e.g. randomize 2:1 (formulas also exist for that case)
  - Randomizing 2:1 is sometimes better, e.g. to show that the risk of side effects of a new drug is rare enough (a larger sample size in the experimental arm can help)
- However, roughly, for a given total sample size, the randomization ratio that leads to the highest power is 1:1.

Clinician: Sh**!... it is a rare disease you know. We see only about 50 such patients per year in Denmark and that is the number we can include at Rigshospitalet, per year. I think that they can include approximately 50 per year in Karolinska in Stockholm too and 30 in Charité in Berlin. So, with our current consortium, we could include about 130 per year and it will take approximately 8 years to include 1048 patients...
Would it be so bad to include patients for only 5 years, hence aiming for a total sample size of only approximately $130 \times 5 = 650$ patients?

Statistician: Well, the standard (simple) calculation suggests that this would provide 60% power only. Would it be worth it and **ethical** to run the trial with this power?

Clinician: Sh**!... it is a rare disease you know. We see only about 50 such patients per year in Denmark and that is the number we can include at Rigshospitalet, per year. I think that they can include approximately 50 per year in Karolinska in Stockholm too and 30 in Charité in Berlin. So, with our current consortium, we could include about 130 per year and it will take approximately 8 years to include 1048 patients...
Would it be so bad to include patients for only 5 years, hence aiming for a total sample size of only approximately $130 \times 5 = 650$ patients?

Statistician: Well, the standard (simple) calculation suggests that this would provide 60% power only. Would it be worth it and **ethical** to run the trial with this power?

Clinician: No, definitely not. For the trial, we will need to transfer severely ill patients from all over Denmark to Rigshospitalet. We cannot guarantee that those receiving standard of care after their transfer will be as well off as those receiving standard of care without being transferred. Transferring patients is never completely safe! The trial is ethical to run only if this potential increase in mortality risk is largely compensated by the fact that lots of patients will benefit from the new treatment, in the future, once we have demonstrated that it is better than standard of care. Future patients will not benefit from the new treatment if we cannot show it works... and we will not likely enough be able to show it works, if the power is low.

Clinician: Sh**!... it is a rare disease you know. We see only about 50 such patients per year in Denmark and that is the number we can include at Rigshospitalet, per year. I think that they can include approximately 50 per year in Karolinska in Stockholm too and 30 in Charité in Berlin. So, with our current consortium, we could include about 130 per year and it will take approximately 8 years to include 1048 patients...
Would it be so bad to include patients for only 5 years, hence aiming for a total sample size of only approximately $130 \times 5 = 650$ patients?

Statistician: Well, the standard (simple) calculation suggests that this would provide 60% power only. Would it be worth it and **ethical** to run the trial with this power?

Clinician: No, definitely not. For the trial, we will need to transfer severely ill patients from all over Denmark to Rigshospitalet. We cannot guarantee that those receiving standard of care after their transfer will be as well off as those receiving standard of care without being transferred. Transferring patients is never completely safe! The trial is ethical to run only if this potential increase in mortality risk is largely compensated by the fact that lots of patients will benefit from the new treatment, in the future, once we have demonstrated that it is better than standard of care. Future patients will not benefit from the new treatment if we cannot show it works... and we will not likely enough be able to show it works, if the power is low.

Statistician: Could you get more hospitals, from other countries, to participate?

Clinician: Maybe our Australian colleagues, they could proably include about 40 patients per year. But I am not sure we want them to join... the regulation about randomizing unconscious patients are different over there. This might complicate the design of the study we had in mind... By how much the power could increase, if they join and we plan for 5 year of inclusion?

Clinician: Maybe our Australian colleagues, they could proably include about 40 patients per year. But I am not sure we want them to join... the regulation about randomizing unconscious patients are different over there. This might complicate the design of the study we had in mind... By how much the power could increase, if they join and we plan for 5 year of inclusion?

Statistician: Well, with 650+ 5×40=850 patients, the power is approx 71%.

Clinician: Hum... that seems still too low. But waiting 8 years to enroll 1048 is really long.... By the way, am I really sure to have a decent power in that case? Maybe we are a bit optimistic about the new treatment. What about if the treatment reduces mortality from 17% to 13% only (instead of 11%) would we still have a decent power?

Clinician: Maybe our Australian colleagues, they could proably include about 40 patients per year. But I am not sure we want them to join... the regulation about randomizing unconscious patients are different over there. This might complicate the design of the study we had in mind... By how much the power could increase, if they join and we plan for 5 year of inclusion?

Statistician: Well, with 650+ 5×40=850 patients, the power is approx 71%.

Clinician: Hum... that seems still too low. But waiting 8 years to enroll 1048 is really long.... By the way, am I really sure to have a decent power in that case? Maybe we are a bit optimistic about the new treatment. What about if the treatment reduces mortality from 17% to 13% only (instead of 11%) would we still have a decent power?

Statistician: Well, with 1048 patients, risks 17% vs 13%, the power is approx 44%....

Clinician: So low, really? How good should the new treatment be to have at least 75% power, if we wait for the 8 years to include 1048 patients?

Clinician: Maybe our Australian colleagues, they could proably include about 40 patients per year. But I am not sure we want them to join... the regulation about randomizing unconscious patients are different over there. This might complicate the design of the study we had in mind... By how much the power could increase, if they join and we plan for 5 year of inclusion?

Statistician: Well, with 650+ 5×40=850 patients, the power is approx 71%.

Clinician: Hum... that seems still too low. But waiting 8 years to enroll 1048 is really long.... By the way, am I really sure to have a decent power in that case? Maybe we are a bit optimistic about the new treatment. What about if the treatment reduces mortality from 17% to 13% only (instead of 11%) would we still have a decent power?

Statistician: Well, with 1048 patients, risks 17% vs 13%, the power is approx 44%....

Clinician: So low, really? How good should the new treatment be to have at least 75% power, if we wait for the 8 years to include 1048 patients?

Statistician: Well, with 1048 patients, to have at least 75% power, you need to reduce the risk from 17% to 11.5% or lower. Smaller risk reductions are not enough, sorry.

Clinician: Maybe our Australian colleagues, they could proably include about 40 patients per year. But I am not sure we want them to join... the regulation about randomizing unconscious patients are different over there. This might complicate the design of the study we had in mind... By how much the power could increase, if they join and we plan for 5 year of inclusion?

Statistician: Well, with 650+ 5×40=850 patients, the power is approx 71%.

Clinician: Hum... that seems still too low. But waiting 8 years to enroll 1048 is really long.... By the way, am I really sure to have a decent power in that case? Maybe we are a bit optimistic about the new treatment. What about if the treatment reduces mortality from 17% to 13% only (instead of 11%) would we still have a decent power?

Statistician: Well, with 1048 patients, risks 17% vs 13%, the power is approx 44%....

Clinician: So low, really? How good should the new treatment be to have at least 75% power, if we wait for the 8 years to include 1048 patients?

Statistician: Well, with 1048 patients, to have at least 75% power, you need to reduce the risk from 17% to 11.5% or lower. Smaller risk reductions are not enough, sorry.

Could you be too pessimistic? If it is not unlikely that the treatment is actually better than you expect, we can plan an interim analysis. This would provide you with a decent chance of stopping the trial early for efficacy, say after 6 years instead of 8, without the need to include the full sample size. Should we make few calculations about how likely you could stop early?

(more complicated calculations are required in that case....)

Wrapping-up

- Several sample size and power calculations are often interesting to perform, to best understand the consequences of the choice of a specific sample size.
- Ethical, financial and logistical constraints are also important to take into account, such that there is no simple recipe to chose the optimal sample size for a clinical trial.
- Planning a study is not all about statistics, but statistical thinking has an important role to play!

Time for computation ! Use R!

Exercise.1.1: reproduce all the numbers provided by the statistician in the previous discussion:

1. $n = 1048$
2. 60% power if $n = 650$
3. 71% power if $n = 850$
4. 44% power if mortality is reduced from 17% to 13% only.
5. treatment efficacy needs to reduce the risk to 11.5% or below, to have 75% power.

Help: you can use the `power.prop.test()` function of R. Just use 3 out of the 4 relevant input parameters[30] and provide the relevant 3 values for each computation.

[30] 'p1', 'p2', 'power' and 'n'

Additional background details:

Assume that the expected risks of 17% and 11%, for the risk under standard of care and using the new treatment, respectively, came from the following expectations.

- Approximately 40% of the patients will be in septic shock; 60% will not.
- Among patients in septic shock, we expect a risk of approximately 30% under standard of care versus 20% using the new treatment.
- Among patients **not** in septic shock, we expect a risk of approximately 7.5% under standard of care versus 5% using the new treatment.

Additional questions:

6. Check that 17% and 11% are indeed (approximately) consistent with what we expect in the two subgroups ("septic shock" and "not in septic shock").
7. What could be the advantages and disadvantages of including only patients in septic shock in the trial?

EZ Principle

Many test statistics used in clinical trials are of the form

$$Z = \frac{\widehat{\delta}}{\widehat{\text{se}}(\widehat{\delta})}$$

where $\widehat{\delta}$ is an estimator of the treatment effect δ and $\widehat{\text{se}}(\widehat{\delta})$ is an estimator of the standard error of $\widehat{\delta}$. Often,

- Z is is approximately normal with mean δ and variance 1.
- $\delta = 0$ under the null hypothesis and $\delta > 0$ if treatment is effective
- The ratio of estimated to true standard errors is close to 1. More specifically, $\widehat{\text{se}}(\widehat{\delta})/\text{se}(\widehat{\delta}) \approx 1$ (for large sample size n).
- Luckily, we know a "relatively simple" formula for $\text{se}(\widehat{\delta})$, which is usually is proportional to $1/\sqrt{n}$. Hence, we can compute $\boxed{E(Z) = \delta/\text{se}(\widehat{\delta})}$. E.g., $\delta = p_1 - p_2$ and $\text{se}(\widehat{\delta}) = \sqrt{2\bar{p}(1-\bar{p})/n}$ when comparing two proportions by computing their difference.

Principle 8.1. The EZ Principle (it's easy!) *Power* $1-\beta$ *for a one-tailed test at level* $\alpha/2$ *or a two-tailed test at level* α *requires the expected z-statistic to satisfy:*

$$\boxed{\mathrm{E}(Z) = z_{\alpha/2} + z_{\beta},} \tag{8.2}$$

We often use a two-tailed test at level 0.05, so $z_{\alpha/2}$ is 1.96. The most common values of power are between 0.80 and 0.90 (i.e., type II error rates β between 0.10 and 0.20). For these power values, z_β is given in Table 8.1.

Table 8.1: Values of z_β for power 0.80, 0.85, or 0.90.

Power	0.80	0.85	0.90
z_β	0.84	1.04	1.28

From the EZ principle flows everything of interest about power and sample size. Equation (8.2) emphasizes that the only thing that matters is the expected z-score. Table 8.1 shows that if the expected z-score is $1.96 + 0.84 = 2.80$, $1.96 + 1.04 = 3$, or $1.96 + 1.28 = 3.24$, then power for a two-tailed t-test at $\alpha = 0.05$ or a one-tailed t-test at $\alpha = 0.025$ will be approximately 80%, 85%, or 90%. If the expected z-score is substantially smaller, there is

page 129 in *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022)

Appendix

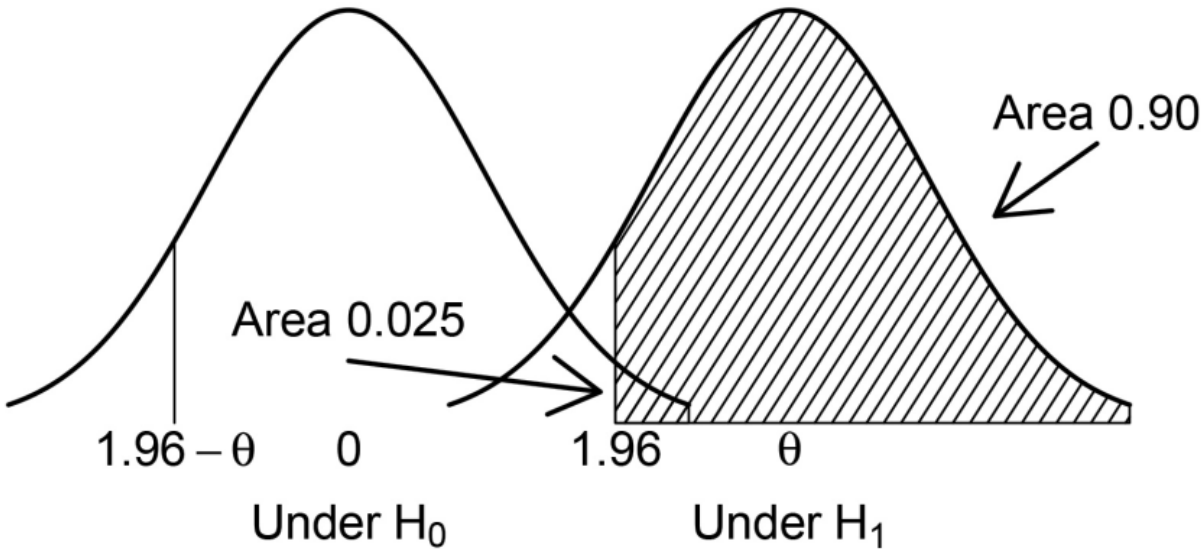


Figure 8.1: The null (left curve) and alternative (right curve) normal densities for the z-statistic. Power is the area to the right of 1.96 under the right curve, which equals the area to the right of $1.96 - \theta$ under the left curve. This area is $1 - \Phi(1.96 - \theta)$.

- Here θ denotes what δ in the previous slides.
- Source: page 128 in *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022)

8.3.2 Test of Proportions

The EZ principle can be used to compute sample size and power for a test of proportions. Substitute the expected proportions for $\hat{p}_C$ and $\hat{p}_T$ in the usual z-statistic formula, equate to $z_{\alpha/2} + z_\beta$, and solve for the per-arm sample size n:

$$\begin{aligned} \frac{p_C - p_T}{\sqrt{2\bar{p}(1-\bar{p})/n}} &= z_{\alpha/2} + z_\beta \\ \sqrt{n}(p_C - p_T) &= (z_{\alpha/2} + z_\beta)\sqrt{2\bar{p}(1-\bar{p})} \\ n &= \frac{2(z_{\alpha/2} + z_\beta)^2 \bar{p}(1-\bar{p})}{(p_C - p_T)^2}, \end{aligned} \tag{8.12}$$

where $\bar{p} = (\hat{p}_C + \hat{p}_T)/2$.

Source: page 133 in *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022)

Power and the EZ principle

$$\text{Power} = \Phi\big\{E(Z) - z_{\alpha/2}\big\} \quad \text{and} \quad E(Z) = \frac{\delta}{\text{se}(\widehat{\delta})}$$

with $\Phi(z) = P(Z \leq z)$ when Z is a standard normal random variable.

Example & Exercise.1.2, use R!: use this formula to re-compute the results that the power is approximately 71% and 60% for n=850/2 and n=650/2 per arm.

Help: to compute $\Phi(z)$ for any value z with R, use `pnorm(z)`. E.g., check that `pnorm(1.96)` is 97.5% and `pnorm(0)` is 50%.

8.8 Appendix: Other Sample Size Formulas for Two Proportions

Sample size formula (8.12) was based on the *score statistic*, with variance of $\hat{p}_C - \hat{p}_T$ estimated under the null hypothesis. The *Wald statistic*,

$$Z_W = \frac{\hat{p}_C - \hat{p}_T}{\sqrt{\hat{p}_C(1-\hat{p}_C)/n_C + \hat{p}_T(1-\hat{p}_T)/n_T}},$$

uses a variance that does not assume the null hypothesis. Substituting the expected values of $\hat{p}_T$ and $\hat{p}_C$ under the alternative hypothesis yields the sample size formula

$$n = \frac{(z_{\alpha/2} + z_\beta)^2\{p_C(1-p_C) + p_T(1-p_T)\}}{(p_C - p_T)^2}. \tag{8.18}$$

The most common per-arm sample size formula for the test of proportions merges features of the score and Wald statistic:

$$n = \frac{\left\{z_{\alpha/2}\sqrt{2p(1-p)} + z_\beta\sqrt{p_C(1-p_C) + p_T(1-p_T)}\right\}^2}{(p_C - p_T)^2}. \tag{8.19}$$

The three formulas are asymptotically equivalent, even though they are always ordered in the following way:

$$(8.18) \leq (8.19) \leq (8.12). \tag{8.20}$$

- Source: page 145 in *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022)
- Eq. (8.18) uses: $\mathsf{se}(\hat{\delta})^2 = \{p_C(1-p_C) + p_T(1-p_T)\}/n$ instead of $2\bar{p}(1-\bar{p})/n$; this is often what we use by default for non-inferiority trials, when we expect $p_C \neq p_T$.

A non-inferiority case study (TRACTION)

The TRACTION trial is a multicenter, investigator-initiated, parallel group, 1:1 randomized, assessor-blinded, non-inferiority trial.[31] Patients admitted with NSTEACS[32] and an indication for Invasive Coronary Angiography (ICA) are randomized to either:

- Coronary Computed Tomography Angiography (CCTA) and team-based interventional triage (intervention group)
- or standard-of-care with conventional ICA (control group)

Primary outcome is MACE[33] within one year.

Primary clinical question: *"Is the risk of MACE within one year from admission not more than 5% higher when using CCTA instead of ICA for initial diagnostic testing?"* [34].

[31] More details in SAP available online: `https://cdn.clinicaltrials.gov/large-docs/62/NCT06101862/SAP_000.pdf`

[32] NSTEACS: Non-ST-segment Elevation Acute Coronary Syndrome

[33] Major Adverse Cardiac Event: death, myocardial infarction, hospitalization with either refractory angina or heart failure.

[34] shorter version of that written in the SAP

Null hypothesis and alternative with non-inferiority

Non-inferiority hypothesis test:

- Null hypothesis: the risk of MACE within one-year after admission is **at least** 5% higher when using CCTA than when using ICA. Formally, $\mathcal{H}_0 : \pi_1 - \pi_0 \geq 5\%$, where π_1 and π_0 are the one-year risk after CCTA or ICA, respectively.

versus

- Alternative hypothesis: the risk of MACE within one-year after admission is **at most** 5% higher when using CCTA than when using ICA. Formally, $\mathcal{H}_1 : \pi_1 - \pi_0 < 5\%$.

- Here $\delta = \pi_0 - \pi_1$, and we do not have $\delta = 0$ under the null hypothesis and $\delta > 0$ if treatment is effective, as in superiority testing/studies (previous slides)
  Remarak: we use $\boxed{\delta = \pi_0 - \pi_1}$ instead of $\delta = \pi_1 - \pi_0$ to have the inequality sign in the alternative hypothesis as in the previous slides.
- We can easily convert a nonzero null hypothesis $\mathcal{H}_0 : \delta = -5\%$ to the zero null hypothesis $\boxed{\mathcal{H}_0 : \delta - (-5\%) = 0}$. We just substract the non-inferiority margin $\delta_0 = -5\%$. Therefore, sample size and power can be computed using the expected z-score method, but now the z-score is:

$$Z = \frac{\widehat{\delta} - \delta_0}{\widehat{\text{se}}(\widehat{\delta})}$$

Remark: we can use $\mathcal{H}_0 : \delta = \delta_0$ instead of $\mathcal{H}_0 : \delta \leq \delta_0$ for all practical puposes, as this is the "worst" case for type-I error control.

Exercise.1.3, use R !

1. The desired power was 90% and the expected proportion of primary endpoint in both groups was 15%. Using the non-inferiority margin of 5%, check that the EZ principle leads to a sample size calculation of approximatively n=1070 per arm.

2. To ensure a power of power 90%, is is better to overestimate or underestimate the 1-year risk assumed to be equal in both arm (15%)? For instance, using $n = 1070$ per arm, what is the power if risk in both arms are larger than 15% (e.g., 17% and 20%) or lower (e.g., 13% or 10%)?

3. Assume that CCTA is non-inferior to ICA at the non-inferiority margin of 5%, but still less efficacious with π_1=17% versus π_0=15%. Using $n = 1070$ per arm, what is the power of the trial? Hint: use $\text{se}(\hat{\delta})^2 = \{p_C(1-p_C) + p_T(1-p_T)\}/n$ as in Eq. (8.18) in previous slides, as $p_C \neq p_T$.

Because the EZ principle is based on a large sample approximation, the resulting sample size n cannot be trusted if it is small. Similarly, power calculations are not accurate if n is small.

Additionally, with a binary outcome, we cannot trust the results if the expected number of events x (or $n - x$) is very small, say ≤ 5. Indeed, in that case the large sample approximation behind the EZ principle will also often perform poorly.

7 Sample size determination and power calculation

As stated in Nielsen et al. [11], the sample size of 180 children with confirmed BJI (90 in each group) was computed to provide 90% power assuming: a non-inferiority margin of 5%, a risk of sequelae of 1% in both group and 10% dropout rate. The usual asymptotic normal approximations was used, i.e.,

$$\text{Power} \approx \Phi\left(\frac{(0.99-0.99)+0.05}{\sqrt{0.99\times 0.01/81+0.99\times 0.01/81}}-1.96\right)\approx 90\%$$

where $n = 90 \times (1 - 0.1) = 81$ subjects are expected fully observed (with no dropout) in each group and Φ is the cumulative standard normal distribution function, see e.g. Chow et al. [1, Sec. 4.2.2].

Because the expected number of events in each group are very low (0 or 1) and because we will use an exact test (see Sec 6.1 above), the initial (above) asymptotic power calculation has been suspected to not be sufficiently precise. Hence, an alternative exact power computation was performed and showed that with $n = 81$ subjects in each group, the power is actually approximately 51%, 68% or 82% if we expect 1%, 0.5% or 0.25% risk of sequelae in both group[5]. This suggests that the study is substantially underpowered if there is indeed 1% risk of sequelae in both group, but decently powered if the risk is slightly lower (equal to 0.25%), which is not unrealistic. These additional results complement those presented in the protocol paper [11]. They are the consequence of additional, more recent, thinking.

[5] The computation was performed by computing the likelihood of observing $(x_1, x_0) = (0,0), (0,1), (1,0), (1,1), \ldots$ and summing-up the likelihood of each case leading to a significant result, where (x_1, x_0) denote the number of sequelae in the OT and IVT groups.

- Source: SAP available at `https://cdn.clinicaltrials.gov/large-docs/25/NCT04563325/SAP_000.pdf`
- Main paper: Nielsen et al. "Oral versus intravenous empirical antibiotics in children and adolescents with uncomplicated bone and joint infections: a nationwide, randomised, controlled, non-inferiority trial in Denmark." The Lancet Child & Adolescent Health 8.9 (2024): 625-635.

EZ principle and quantitative outcomes

When comparing the mean of a quantitative outcome (e.g., number of days with antibiotic treatment) in two arms via the computation of a mean difference $\widehat{\delta} = \widehat{\mu}_1 - \widehat{\mu}_0$, then we can use $\text{se}(\widehat{\delta}) = \sigma\sqrt{2/n}$, where σ is the (expected) standard deviation of the outcome (when it assumed to be the same in the two groups).

Interpretation: the smaller the variability of the data (i.e., the smaller σ) (or the larger the sample size) and the more presicely we can estimate the mean difference $\delta = \mu_1 - \mu_0$.

Hence, $E(Z) = z_{\alpha/2} + z_\beta$ becomes $\delta/\{\sigma\sqrt{2/n}\} = z_{\alpha/2} + z_\beta$ and

$$n = \frac{2\sigma^2(z_{\alpha/2} + z_\beta)^2}{\delta^2} \quad \text{and} \quad \text{Power} = \Phi\left\{\frac{\delta\sqrt{n/2}}{\sigma} - z_{\alpha/2}\right\}$$

Exercise.1.4, use R !

Consider a recent trial about early termination of empirical antibiotics[35] in febrile neutropenia in children with cancer. The primary outcome was number of days with antibiotic treatment within the first 28 days after treatment initiation. [36] A mean of 12 days was expected in the experimental group, versus 15 days in the control group. A standard deviation of 6.5 days in each arm was expected. These expected values were, among other things, based on previous results from a similar study in adults.

1. When expecting these values, which sample size ensures a power of 90%?
2. The study had a substantially lower accrual rate than anticipated (changes in cancer treatment protocols resulted in shorter neutropenia periods, hence less patients fulfilling includion criteria). Hence, early termination of the trial was decided. To anticipate the consequence of different choices for the date of end of inclusion (and resulting sample size), it was useful to compute the expected power of the trial with samller sample sizes n than initially planned. What is the expected power with $n = 80$, 100, 120 and 150?
3. It was decided to plan the end of inclusion at a date when $n = 100$ was expected. What difference in mean should exist, instead of $15 - 12 = 3$ days, to have a power of 90% in that case? (again assuming SD=6.5 days)

Ideally: use both the `power.t.test()` function of R (just use 3 out of the 4 relevant input parameters `'delta'`, `'sd'`, `'power'` and `'n'`) and computation "by hand" using the EZ principle (for double checking).

[35] early termination means end after 48 hours of apyrexia and clinical stability despite neutrophil count $< 0.5 \times 10^9$ cells/L; standard duration is until either neutrophil count is $\geq 0.5 \times 10^9$ cells/L, and the child is afebrile and clinical stable or the child has received 10 days of antibiotics and has been afebrile and clinically stable for 7 days.

[36] See e.g., SAP online: `https://cdn.clinicaltrials.gov/large-docs/64/NCT04637464/SAP_000.pdf`

Appendix: which difference δ and σ to use? (in sample size calculation)

- Principled choices for δ:
  - expected/hypothesized difference.
  - minimum (clinically) relevant difference.
- Pragmatic choice for δ: smallest difference "disappointing" to overlook.
- Principled choices for σ:
  - Estimate from previous studies from your research group or published in the literature (be aware of statistical uncertainty).
  - Expert guess obtained from ideally **several** senior collaborators (reaching consensus). In that case, asking them the expected normal range (and divide its width by 4) can help. Eg., ask *"In which interval do you expect that approximately 95% of the outcome values will fall?"* [37]

Recommended practice:

- use several likely values to do several calculations. Well informed decisions and final choices may require the results of several iterations (guesses $\longrightarrow$ calcualtion $\longrightarrow$ updated guesses $\longrightarrow$ new claculation etc). This might help to realize how sensitive are the results to our best guesses, and provide incentive to give more "thought throught" guesses and/or to be more or less conservative.
- consider ethical issues (whenever relevant).

[37] Personal experience suggests that this provides substantially more accurate guesses than asking the expected σ directly. Rationale is that the interval "mean $\pm$ 2σ" includes 95% of the observations, if the data are normally distributed, and more than 75% in any case.

Important repetitions

- Several sample size and power calculations are often interesting to perform, to best understand the consequences of the choice of a specific sample size.
- Ethical, financial and logistical constraints are also important to take into account, such that there is no simple recipe to chose the optimal sample size for a clinical trial.
- Planning a study is not all about statistics, but statistical thinking has an important role to play!