Day 3 (part 1): Exact unconditional tests and confidence intervals: what, why, how?

Paul Blanche

Section of Biostatistics, University of Copenhagen

November 20, 2025

Outline/Intended Learning Outcomes (ILOs)

What?

ILO: recall unconditional tests and confidence intervals

Why?

ILO: contrast them to simpler, more popular, alternatives
ILO: recognize their theoretical advantages and remember convincing examples

How?

ILO: perform the computation in R
ILO: describe their use in a SAP

Missing data

ILO: perform and exemplify a basic sensitivity analysis

Reminder

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.

Additional SAP details

6.1.1 Main Analytical Approach

For this analysis we will use the "Main analysis set" detailed in Section 5. We will estimate the primary estimand, that is, the difference in risk $\pi_1 - \pi_0$, as the empirical (i.e. observed) proportion of participants with sequelae after 6 months in the high-dose OT intervention group minus the same proportion in the IVT intervention group. An appropriate one-sided 97.5% confidence interval and a matching p-value for the null hypothesis $\mathcal{H}_0 : \pi_1 - \pi_0 \geq 5\%$ will be computed using an exact

(page break)

unconditional test. An exact approach will be used instead of alternatives based on (asymptotic) normal approximations, because the expected number of events in each group are very low (say, 0 or 1) and thus usual (asymptotic) approximations do not seem reliable. Specifically, we will use an exact unconditional test using the score statistic ordering, as implemented in the function `uncondExact2x2` of the `R` package `exact2x2`; see [4] for the mathematical details and Appendix 8.1 for the specific `R` code that we will use. The score statistic ordering was chosen for its good power properties, as compared to alternatives [4].

If the upper limit of the 97.5% confidence interval of the risk difference $\pi_1 - \pi_0$ does not include the non-inferiority margin 5% or, equivalently, if the matching p-value is less than 2.5%, we will reject the null hypothesis[3].

- 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.

Digression: one-sided vs two-sided hypothesis tests

- Although debates and controversies exist [1] [2], it is often recommended that two-sided tests should be used by default [3].
- But, it is often considered acceptable and equivalent for all practical purposes to use either a one-sided test at 2.5% or a two-sided test at 5%. [2] [4] [5]
- Of course *"It is important to clarify whether one- or two-sided tests of statistical significance will be used"*. [4]
- This is a bit different in group-sequential or adaptive trials (i.e., in trial with interim analyses).

---

[1] Freedman, Laurence S. "An analysis of the controversy over classical one-sided tests." Clinical Trials 5.6 (2008): 635-640.

[2] Senn, Stephen S. Statistical Issues in Drug Development. 3^{rd} Edition, 2021 (see chapter 12).

[3] Bland JM, Altman DG. One and two sided tests of significance. BMJ 1994; 309: 208

[4] 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

[5] 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

What do "exact" and "unconditional" mean?

- *"Unconditional tests are not commonly used in clinical trials, but provide an attractive alternative to Fisher's exact test when sample sizes are small."* [6]

Exact:

- *"'Exact test' does not mean that the type I error rate is exactly 0.05."*[7]. Often, it is impossible to construct such tests.
- Instead, 'Exact' means that the type-I error is $\leq 5\%$ for sure, not only approximately or for "large sample sizes" only. It means we do not rely on approximations. Similarly, for exact 95%-CI, the coverage is $\geq 95\%$ for sure, not only approximately.
- In other words, 'exact' refers to the fact that the statement that we control the type-I error is exact, not approximate. By construction the test may be conservative, hence type-I error $\neq 5\%$, but $< 5\%$.

Unconditional:

- It relates to how probability calculations are performed: without conditioning on the number of events observed in each arms, unlike with Fisher's exact test.

[6] page 23 in *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022)

[7] page 2, same book

How do exact unconditional methods work? (in a nutshell)

- *"P-value: In significance testing, the probability of obtaining a result as extreme or more extreme than that actually observed given that the null hypothesis is true."*[8]
- To define more extreme (in the direction of the alternative hypothesis), there might be several ways (e.g., test statistic, estimate...). Let's choose one (e.g., a clever "score" test statistic).
- The probability of a result as extreme or more can be computed assuming a specific probability for each arm p_1 and p_0.
- The p-value of the unconditional exact tests is the maximum of all possible p-values computed assuming specific values for p_1 and p_0 compatible with the null hypothesis (e.g., max over $p = p_1 = p_0 \in [0, 1]$ if $\mathcal{H}_0 : p_1 - p_0 = 0$), which can be computed "exactly" (using simple binomial distributions).
- Confidence intervals are computed by inverting a series of two-sided hypothesis test. Values not rejected are included in the confidence intervals.

- This is computationally challenging! Because of that these tests have not been used much in the past.[9] But today, computation is no longer an issue.

[8] Senn, Stephen S. Statistical Issues in Drug Development. 3^{rd} Edition, 2021 (page 589)

[9] *And also because "Fisher was bitterly critical of Barnard's proposal for esoteric reasons" as written in Mehta, & Senchaudhuri, Cytel Software Corporation 675 (2003): 1-5.)*

Practice with R! (exercise "Exact.1")

1. Use R and the code provided in Appendix of the SAP [10] to reproduce the confidence intervals and p-values of the main publication presented below [11]. Use the code in appendix section 8.1 for the primary outcome and 8.2 for the secondary outcomes.

	Initial oral antibiotics	Initial intravenous antibiotics	Risk difference (CI*)	$p_{non\text{-}inferiority}$
Primary outcome, clinical sequelae at 6 months				
Main analysis†	0/98	0/84	0 (0·0 to 3·8)	0·012
Per protocol‡	0/81	0/76	0 (0·0 to 4·6)	0·021
Secondary outcomes				
Switch of antibiotics within 28 days due to suspicion of non-acute treatment failure	5/101 (5·0%)	3/91 (3·3%)	1·7% (–5·2 to 8·3)	NA
Recurrent infection within 6 months	0/101	1/91 (1·1%)	–1·1% (–6·2 to 2·7)	NA

2. Compare these results to those obtained using the simple, standard computation (based on standard large sample approximations). E.g., using the code:

```
library(Publish)
tabRes <- rbind(c(5,101-5),c(3,91-3))
rest2x2 <- table2x2(tabRes,stats="rd")
rest2x2
```

[10] SAP available at https://cdn.clinicaltrials.gov/large-docs/25/NCT04563325/SAP_000.pdf

[11] 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.

Reminder: usual large sample approximation

$$\widehat{p}_T = a/(a+b)$$
$$\widehat{p}_C = c/(c+d)$$
Difference: $$\widehat{\delta} = \widehat{p}_T - \widehat{p}_C$$

	Response		
	yes	no	total
Treatment (T)	a	b	a+b $=n_T$
Control (C)	c	d	c+d $=n_C$
total	a+c	b+d	N

Standard error of $\widehat{\delta}$ and 95% confidence interval

$$\widehat{\mathsf{se}}(\hat{\delta}) = \sqrt{ab/(a+b)^3 + cd/(c+d)^3}$$
$$= \sqrt{\widehat{p}_T(1-\widehat{p}_T)/n_T + \widehat{p}_C(1-\widehat{p}_C)/n_C}$$

$$CI_{95\%} = \left[\widehat{\delta} - 1.96\,\widehat{\mathsf{se}}(\hat{\delta}) \;;\; \widehat{\delta} - 1.96\,\widehat{\mathsf{se}}(\hat{\delta})\right]$$

Digression: avoid incompatible results!

- Some researchers choose to present an "exact" p-value computed using a Fisher's exact test together with 95% CI computed using large sample approximation.
- I recommend against this practice, as it can provide inconsistent results.

Example:

	event	no event
Active	5	12
Control	8	3

- $\widehat{p}_1 = 8/11 = 0.73, \quad \widehat{p}_2 = 5/17 = 0.29.$
- $\widehat{\Delta}$= 0.43, with 95%-CI=(0.09 ; 0.77) $\not\ni$ 0.
- p-value from Fisher's exact test is $0.051 > 5\%$.

Here the confidence interval shows a significant result, but not the p-value of Fisher's test. We do not want to take the risk to have to report this kind of inconsistent/incompatible results.

- Exact unconditional methods fortunately provide an alternative framework to avoid incompatible results!

Additional remarks

- Confidence intervals are even more important for non-inferiority studies. Computing them "exactly" or "very precisely" might be important.
- Fisher's exact test cannot be used for non-inferiority, as the null hypothesis is not "no difference", but a difference "small enough to be accepted" (as defined by the non-inferiority margin).
- Exact unconditional testing is therefore very interesting for non-inferiority studies.
- Alternative methods based on "advanced" large sample approximations might also be interesting, especially when the sample size is not small, but risks are close to 0 (or 1). E.g., Miettinen–Nurminen asymptotic score.
- See e.g., Fagerland, et al.[12] or Fay and Hunsberger[13] for additional details.

[12] Fagerland et al. "Recommended confidence intervals for two independent binomial proportions." Statistical Methods in Medical Research 24.2 (2015): 224-254.

[13] Fay and Hunsberger (2021). Practical valid inferences for the two-sample binomial problem. Statistics Surveys, 15:72–110.

Practice with R! (exercise "Exact.2")

1. Use R and to recompute the power 51%, 68% and 82% provided in section 7 of the SAP [14] (see previous slides). You could:
   a. Run the code provided in the webpage, line by line; take the time to understand the rationale for the code and the results.
   b. Change the parameters values as needed (first lines).

2. What is the risk of sequelae is actually 2% both groups? Should you run the code only changing the value of the parameters p1 and p2 or change it a tiny bit more?
3. The main analysis actually included 98 (Oral arm) and 84 (IV arm) patients in each arm. What was the power for these sample sizes, if the risks in each arm were indeed equal to 1%? Can you enumerate all possible results that would have led to significant results, with these sample sizes?

[14] SAP available at https://cdn.clinicaltrials.gov/large-docs/25/NCT04563325/SAP_000.pdf

Digression: ratio or difference in non-inferiority studies?

"If clinical considerations do not dictate the choice of effect measure, then a choice resulting in smaller sample size is reasonable."[15]

Table 1 | Design and data for non-inferiority trial in antibiotic prescribing

	Experimental intervention	Control intervention
Expected treatment failures (%)	5	5
Borderline acceptable treatment failures (%)	10	5
No of participants randomised	400	400
No (%) hypothetical observed treatment failures	24 (6)	20 (5)

KEY MESSAGES

⇒ Investigators in non-inferiority trials should carefully consider the choice of the effect measure used to define the non-inferiority margin (eg, risk difference or risk ratio)

⇒ For unfavourable binary outcomes (eg, treatment failure), defining a non-inferiority margin with the risk difference rather than the risk ratio gives larger power for the same sample size and the same anticipated differences between randomised arms

Table 2 | Analysis of hypothetical data from non-inferiority trial in antibiotic prescribing

Effect measure	Non-inferiority margin	Observed effect (95% CI)	Test of non-inferiority (one sided *v* P=0.025)	Evidence of non-inferiority
Risk difference	10%−5%=5%	1.0% (−2.2% to +4.2%)	P=0.007	Convincing
Risk ratio	10%/5%=2	1.20 (0.67 to 2.14)	P=0.041	Not convincing

CI, confidence interval.

"The discrepancy arises because a risk difference of 5% and a risk ratio of 2 are only simultaneously true if the control risk is exactly 5%. For example, the data in table 1 are consistent with the control risk being only 3% and the experimental risk being 7%: these values fall within the non-inferiority margin of a risk difference of 5%, but outside the non-inferiority margin of a risk ratio of 2." [15]

15 White et al. "Tackling control risk problems in non-inferiority trials." BMJ Medicine 4.1 (2025).

Appendix: non-inferiority frontiers as visual aid

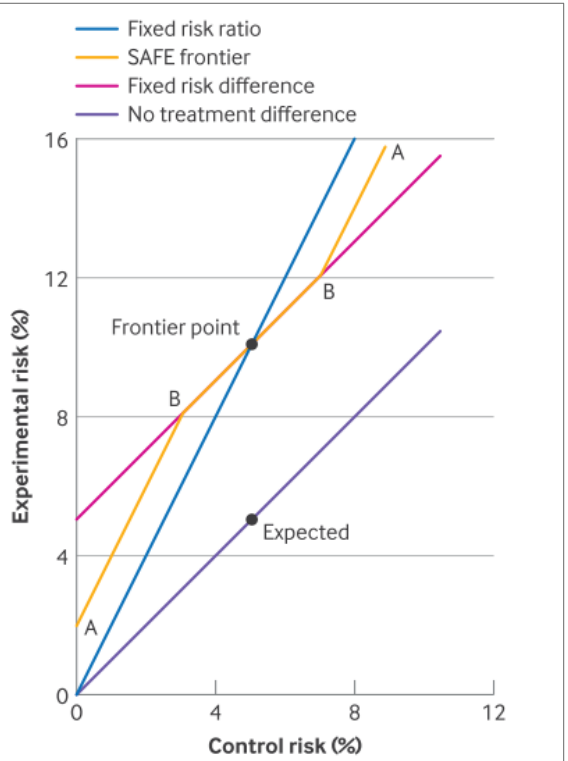


Figure 2 | Three possible non-inferiority frontiers: fixed risk difference frontier, fixed risk ratio frontier, and smooth away from expected (SAFE) frontier (marked A). The SAFE frontier follows the fixed risk difference frontier around the expected control risk, because clinical interpretation of the non-inferiority margin is unchanged in this region, and then bends away smoothly at points B to pass the frontier points A

Non-inferiority frontier graph can help to best understand both:

- what is most clinically relevant to define a non-inferiority margin: the risk difference or the risk ratio (null hypotheses are $\neq$)
- and what leads to sufficient power.

Source: White et al. "Tackling control risk problems in non-inferiority trials." BMJ Medicine 4.1 (2025).

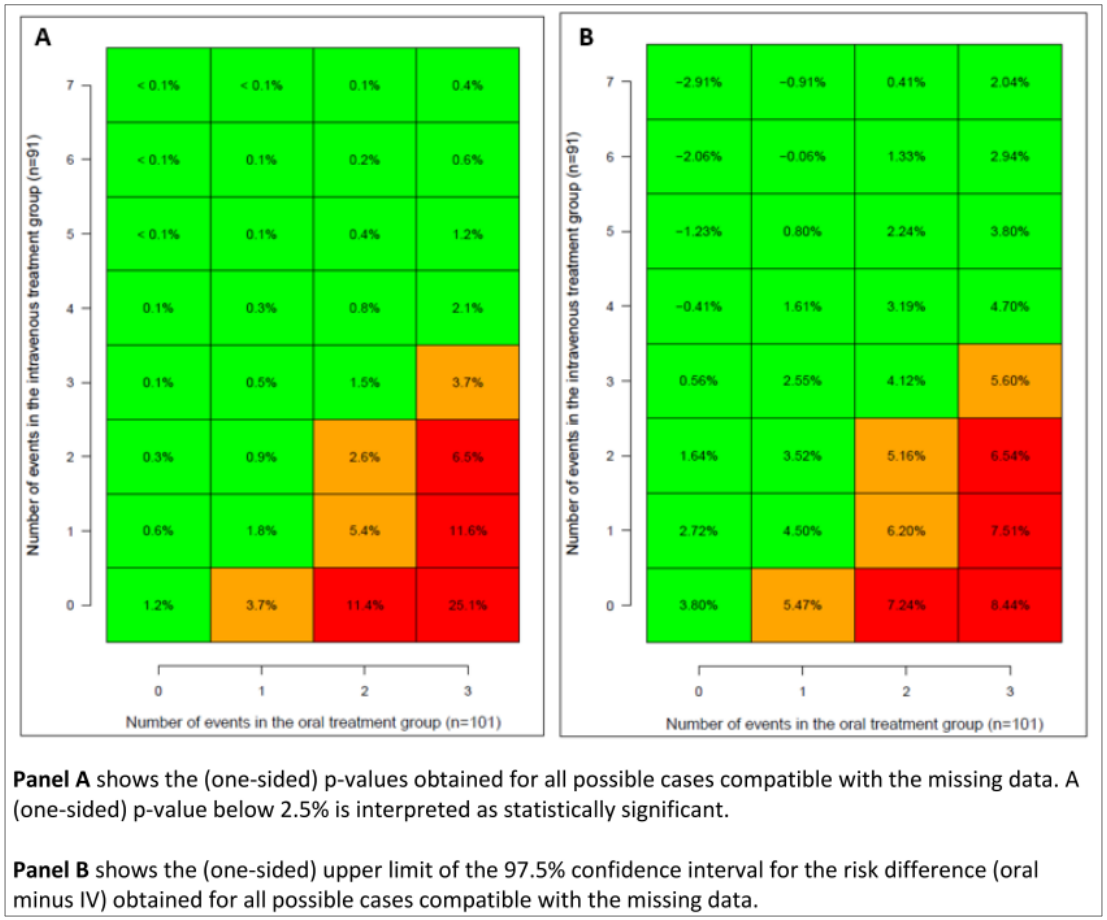
Missing outcome: a simple yet (sometimes) useful sensitivity analysis (1/2)

"The main analysis is based on the 'Main analysis set', which do not include patients for whom the primary endpoint is missing. This choice is thought to be conservative as it is considered very unlikely that a patient with sequelae at 6-months would not attend the follow-up visit at 6-months. This is especially unlikely as telephone interviews focusing on signs of sequelae were planned for all patients not attending the 6 month follow-up. Nevertheless, we will perform the following sensitivity analysis, for completeness. We will report the results (p-value and upper limit of the 97.5% CI) for all possible combinations of the values of missing outcomes in the OT and IVT groups. An graphical representation similar to the 'Enhanced tipping-point display' presented in Liublinska et al. [9, Fig. 2] will be reported"[16] [17]

[16] SAP available at `https://cdn.clinicaltrials.gov/large-docs/25/NCT04563325/SAP_000.pdf`

[17] Interesting reference: Liublinska and Rubin, D. B. (2014). Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. Statistics in Medicine, 33(24):4170–4185.

Missing outcome: a simple yet (sometimes) useful sensitivity analysis (2/2)



Panel A shows the (one-sided) p-values obtained for all possible cases compatible with the missing data. A (one-sided) p-value below 2.5% is interpreted as statistically significant.

Panel B shows the (one-sided) upper limit of the 97.5% confidence interval for the risk difference (oral minus IV) obtained for all possible cases compatible with the missing data.

"An upper limit of less than 5% (the pre-specified noninferiority margin) is interpreted as statistically significant. To facilitate a quick overview of the results, statistically significant results are displayed in green, almost statistically significant (although not statistically significant) are displayed in orange, the others, not statistically significant, in red."[18]

[18] Supplementary material of Nielsen et al. The Lancet Child & Adolescent Health 8.9 (2024): 625-635.

Appendix: useful definitions

Missing Data:

Data that would be meaningful for the analysis of a given estimand but were not collected. They should be distinguished from data that do not exist or data that are not considered meaningful because of an intercurrent event.

Sensitivity Analysis:

A series of analyses conducted with the intent to explore the robustness of inferences from the main estimator to deviations from its underlying modelling assumptions and limitations in the data.

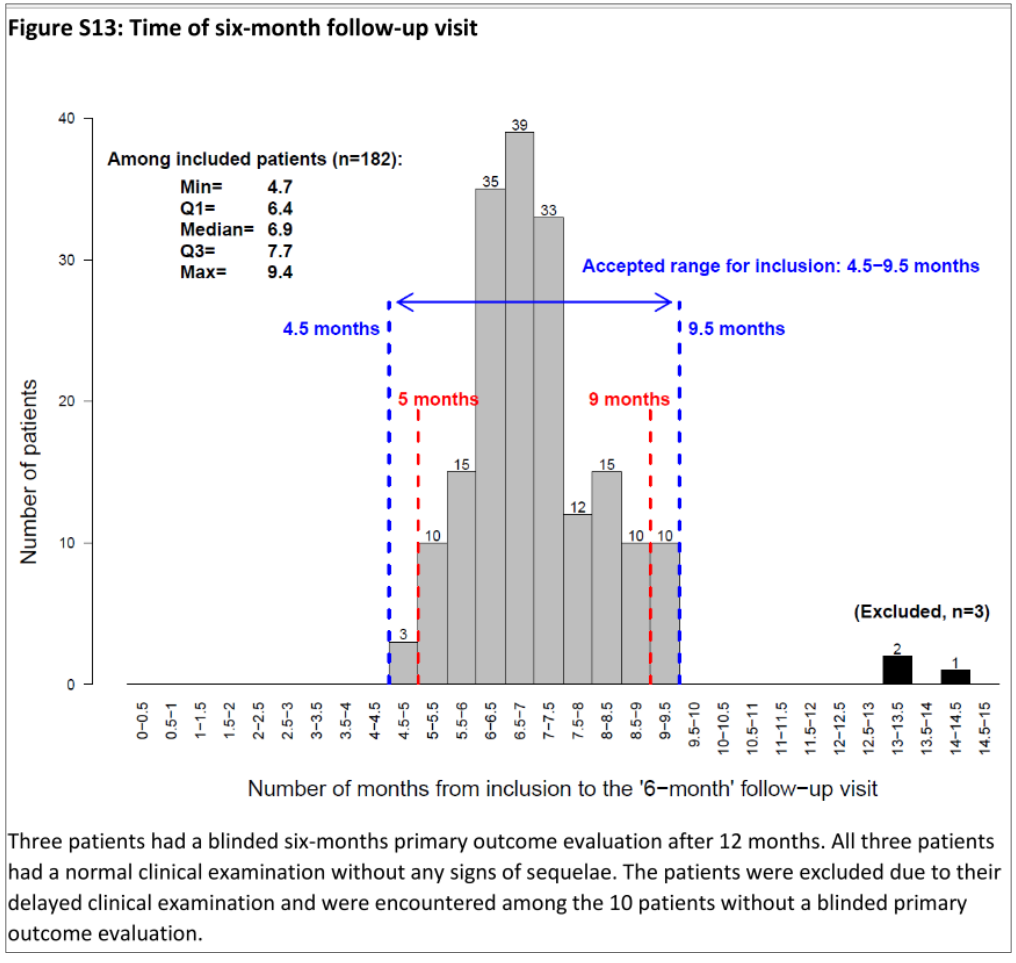
Supplementary Analysis:

A general description for analyses that are conducted in addition to the main and sensitivity analysis with the intent to provide additional insights into the understanding of the treatment effect.

Remark: the terminology "sensitivity analysis" is sometimes misused to refer to a "supplementary analysis".

EMA scientific guidelines: *ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials*, 2020, `https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-s-step-en.pdf`

Digression: accepted range & missing data



Three patients had a blinded six-months primary outcome evaluation after 12 months. All three patients had a normal clinical examination without any signs of sequelae. The patients were excluded due to their delayed clinical examination and were encountered among the 10 patients without a blinded primary outcome evaluation.

- *"The SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) guidelines [3]and ClinicalTrials.gov registry guidelines [14] require outcome definitions to include elements such as the time-point, measurement, analysis metric, method of aggregation and method of assessment."* (Kahan & Jairath. Trials 19, 265 (2018).)
- When visits cannot all occur at the exact same time, it is common to pre-specify an "accepted range".

Anecdote, lessons and tip

Anecdote:

- One of the two statistical reviewers (yes, we had two!) complained that he could not reproduce the main results and that the importance of the exact unconditional was unclear.
- The other wrote *"Appropriate statistical methods have been applied, in line with those pre-specified in the SAP and protocol"* and had very few comments/concerns.

Lessons:

- Not all statistical experts are familiar with all statistical methods.
- One should carefully describe the method that we use; and very carefully when using exact unconditional methods, as many different options exist. *"Readers should not have to infer what was probably done; they should be told explicitly."*[19]

Tip:

- Providing the some R code in appendix of the SAP might help to be transparent and avoid writing lengthy text to explain unnecessary details. The text can focus on the key points.[20]

[19] Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. BMJ 1996; 313: 570–71

[20] In our case, it made it easy to reply to the first statistical reviewer.