Faculty of Health Sciences

Day 4: Multiple testing, Group-sequential trials and miscellaneous topics

Paul Blanche

Section of Biostatistics, University of Copenhagen

November 21, 2025

Outline/Intended Learning Outcomes (ILOs)

Assessing Heterogeneity of Treatment Effect

ILO: outline what it is and recall how it is commonly done.
ILO: recall and exemplify the limitated power.

Multiple testing

ILO: to describe the multiple testing problem and employ basic remedies.
ILO: relate the issue to the categorization of outcomes into primary, secondary and tertiary/exploratory outcomes.
ILO: exemplify when adjustment is needed and when it is not.

Response Adaptive Randomization

ILO: outline what it is and recall arguments to promote its use.
ILO: recall and exemplify its limitations; recall its use is discouraged by many.

Group-sequential trials (GST)

ILO: recall the key elements, strength and limitations of GST
ILO: be prepared to take advice from a statistician to design a GST
ILO: recall examples and use R to perform a few "advanced" calculations

Heterogeneity of Treatment Effect

Heterogeneity of treatment effect (HTE), also called differential treatment effect, is variation in a measure of treatment effect on a scale for which it is mathematically possible that such variation be absent even if the treatment has a nonzero effect. The most commonly used scales for measuring HTE are:

- the original scale for continuous response variables, in which case the treatment effect is the (adjusted) difference in means
- odds ratios, for binary or ordinal outcomes
- hazard ratios, for time-to-event outcomes

How do we actually quantify HTE?

Often, using estimating double differences. That is, comparing the treatment effect in subgroups (e.g., biomarker positive and biomarker negative). That is, essentially using interaction terms.[1]

Source: Frank Harrell's blog `https://www.fharrell.com/post/varyor/`, with a few minor changes in wording (accessed 11/11/2025).

[1] See e.g., Altman & Bland. "Interaction revisited: the difference between two estimates." BMJ 326.7382 (2003): 219.

Why using Odds Ratio?

"Although it is clear that ORs are more difficult to understand than risk ratios (RR) or absolute risk reduction (ARR; risk difference). The reasons for choosing ORs are:"

- *"ORs come directly from logistic models, and logistic models are as likely as any model to fit patterns leading to binary responses. This is primarily because the logistic model places no restrictions on the regression coefficients."*
- *"ORs are capable of being constant over a range of baseline risk all the way from 0 to 1."*
- *"In the multitude of forest plots present in journal articles depicting RCT results, the constancy of ORs over patient types is impressive."*
- *"Unlike RRs, ORs are invariant to the choice of the 'event' vs. the 'no event'. If you interchange event and non-event you would get the reciprocal of the original OR, but the RR would change arbitrarily."*

Source: Text from Frank Harrell, see blog post `https://www.fharrell.com/post/varyor/` (accessed 11/11/2025).

Digression: Predictive vs Prognostic biomarkers

- *"The term biomarker refers to a measurement variable that is associated with disease outcome. It can be a single measurement, such as prostate- specific antigen (PSA) level, or a classifier (signature) computed from measures of numerous other variables, such as OncoType DX recurrence score, which is calculated from the measurements ofthe expression levels of 21 genes."*
- *"There is considerable confusion about the distinction between a predictive biomarker and a prognostic biomarker."*
- *" A prognostic biomarker informs about a likely cancer outcome (eg, disease recurrence, disease progression, death) independent of treatment received."*
- *"A biomarker is predictive if the treatment effect (experimental compared with control) is different for biomarker-positive patients compared with biomarker-negative patients. "*

Source: Ballman *Biomarker: predictive or prognostic?*. Journal of Clinical Oncology 33.33 (2015): 3968-3971.

Power to show HTE is often extremely low!

Example & Exercise 4.1 (think and use R):[2]

Suppose a study is designed to have 80% power to show a treatment effect (e.g., $\delta = 10$ mg/dL mean reduction of blood glucose) while controlling the type-I error at 5% (as commonly done). Further suppose that HTE exists and corresponds to an interaction term between the binary variables 'treatment' and 'biomarker positive', which value is half the size of main effects (e.g., $\delta^+ = 12.5$ and $\delta^- = 7.5$ mg/dL, hence interaction term is $\delta^+ - \delta^- = \delta/2 = 5$ mg/dL). Additionally, we assume that the outcome is quantitative and has the same variability (i.e., standard deviation) in each subgroup.

1. What is its power to show HTE (i.e., interaction term is statistically significant) in an "ideal" study in which half of the patient are biomarker positive and half are biomarker negative? Help: first, the SE of the interaction term is twice that of the main (i.e., average) treatment effect (see appendix in next slide). Second, think about the EZ principle and what you can expect for E(Z) for the interaction term.
2. How many more subjects do we need in the trial to have 80% power to show HTE? Help: remember standard errors are proportional to $1/\sqrt{n}$, hence E(Z) is proportional to $\sqrt{n}$.
3. Among studies that show a significant HTE, what is the expected (average) ratio of estimated HTE size to actual HTE? Help: you can simulate many (e.g., 1 000 000) Z test statistics normally distributed of mean 0.5 and SD= 1 using `z <- rnorm(1e6, 0.5, 1)`. You can then compute the mean among those who are > 1.96 using `mean(z[z>1.96])`.

[2] Slightly adapted from a post from A Gelman, posted on 14/3/2018, available at https://statmodeling.stat.columbia.edu/

Appendix

The SE of the interaction term is twice that of the main (i.e., average) treatment effect. Indeed, $\hat{\delta} = \bar{Y}_n^1 - \bar{Y}_n^0$, hence $\text{SE}(\hat{\delta}) = \sigma\sqrt{2/n}$, with n being the sample size per arm. Let δ^+ and δ^+ denote the treatment effects for 'biomarker positive' and 'biomarker negative' subgroups, respectively. Similarly, we obtain:

$$\text{SE}(\hat{\delta}^+) = \text{SE}(\hat{\delta}^-) = \sigma\sqrt{2/(n/2)} = \sigma\sqrt{4/n}$$

as we assumed that half of the patients are biomarker positive. Then,

$$\begin{aligned}\text{SE}^2(\hat{\delta}^+ - \hat{\delta}^-) &= \text{SE}^2(\hat{\delta}^+) - \text{SE}^2(\hat{\delta}^-) \quad \text{as Var(X+Y)=Var(X)+Var(X), if X independent of Y} \\ &= 2 \cdot \sigma^2 4/n\end{aligned}$$

Hence, we conclude $\text{SE}(\hat{\delta}^+ - \hat{\delta}^-) = 2\sigma\sqrt{2/n} = 2 \cdot \text{SE}(\hat{\delta})$.

Digression: the winner's curse

- Answer to question 3 relates to a well-known phenomenon.

  "Low power has important implications for the interpretation of the results of a given trial. It is well known that conditional on statistical significance, the estimate of the effect size is positively biased. This bias is sometimes called the 'winner's curse', and it is especially large when the power is low.[3]
- Hence it is widely recommended to report the results of all studies, no matter whether the results are statistically significant or not.
- This is increasingly often done, but there is evidence that it has not been done at all in the past. See e.g. figure below.[4]

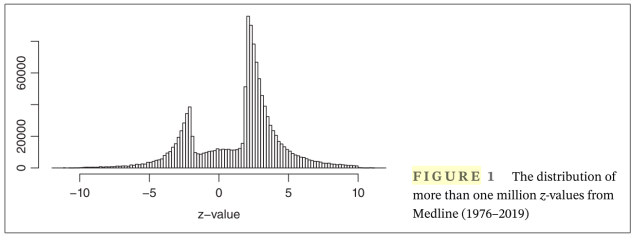


FIGURE 1 The distribution of more than one million z-values from Medline (1976–2019)

[3] van Zwet et al. Statistics in Medicine 40.27 (2021): 6107-6117.

[4] Fig 1 in van Zwet & Cator. Stat Neerl. 2021;1-15.

Digression: on p-values and reproducibility

- If a scientific study reports a discovery with a p-value at or around 0.05, how credible is it? And what are the chances that a replication of this study will produce a similarly 'significant' finding?

[5] Held, Pawel & Schwab (2020), Significance 17(6), 10-11

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Digression: on p-values and reproducibility

- If a scientific study reports a discovery with a p-value at or around 0.05, how credible is it? And what are the chances that a replication of this study will produce a similarly 'significant' finding?

- *''Suppose an original study provides an estimate (assumed to be normally distributed) of the effect size together with a confidence interval from which the p-value can be derived. It is then natural to ask what the power of an identically designed replication study would be if we assume that the effect size found by the original study equals the (unknown) true effect size. This is known as replication power or replication probability. There are in fact two forms of such power: conditional, meaning that the statistical uncertainty of the original effect estimate, as represented by its confidence interval, is ignored; and predictive, meaning that the uncertainty is incorporated. Either way, the replication power depends essentially only on the original p-value and the sample size of the replication study relative to the original study. Let us take the simple case of a replication that uses the same sample size as the original study. If the original p-value was close to the significance level, say 0.05 (or 5%), then both conditional and predictive power turn out to be*

9 / 58

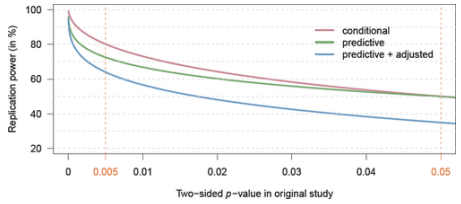
[5] Held, Pawel & Schwab (2020), Significance 17(6), 10-11

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Digression: on p-values and reproducibility

- If a scientific study reports a discovery with a p-value at or around 0.05, how credible is it? And what are the chances that a replication of this study will produce a similarly 'significant' finding?

- *''Suppose an original study provides an estimate (assumed to be normally distributed) of the effect size together with a confidence interval from which the p-value can be derived. It is then natural to ask what the power of an identically designed replication study would be if we assume that the effect size found by the original study equals the (unknown) true effect size. This is known as replication power or replication probability. There are in fact two forms of such power: conditional, meaning that the statistical uncertainty of the original effect estimate, as represented by its confidence interval, is ignored; and predictive, meaning that the uncertainty is incorporated. Either way, the replication power depends essentially only on the original p-value and the sample size of the replication study relative to the original study. Let us take the simple case of a replication that uses the same sample size as the original study. If the original p-value was close to the significance level, say 0.05 (or 5%), then both conditional and predictive power turn out to be only 50%. [...] To gain some intuition why this is so, think of the p-value as a summary measure calculated from the data. If two sets of data have been generated independently but in exactly the same way, the corresponding p-values have the same distribution. The probability that one is larger than the other is then 50%. The original p-value needs to be as small as 0.005 to have a conditional replication power around the usual standard of 80%. This is shown in Figure 1.''*[5]
(Blue curve is computed using a method which aims to adjust for "publication bias".)



9 / 58

[5] Held, Pawel & Schwab (2020), Significance 17(6), 10-11

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Digression: confirmatory versus exploratory research

Exploratory research	Confirmatory research
No hypothesis required/hypothesis can be vague	Clear hypothesis required
Generate new hypothesis from data	Test *a priori* hypothesis with new data
High sensitivity desired, i.e. minimising the risk of false negatives	High specificity desired, i.e. minimising the risk of false positives
Suitable for making new discoveries and finding the unexpected	Suitable for establishing strong evidence and confirming the expected
For example: Testing of new compounds in mice	For example: Assessing the efficacy of a drug in humans

"It is essential to distinguish between exploratory and confirmatory research, and they are equally important to the scientific enterprise. Finding the questions to ask is at least as crucial as answering them, if not more so. But when the two concepts get confused, and the two worlds collide, thea fallout can be disastrous"[6]

To keep in mind!
The main analysis of primary outcomes of clinical trials (and possibly that of key secondary outcomes) is about confirmatory research. Other analyses can often be considered as being about exploratory research.

[6] Schwab & Held. "Different worlds Confirmatory versus exploratory research." (2020) Significance, 8:9.

Outline/Intended Learning Outcomes (ILOs)

Assessing Heterogeneity of Treatment Effect
- ILO: outline what it is and recall how it is commonly done.
- ILO: recall and exemplify the limitated power.

Multiple testing
- ILO: to describe the multiple testing problem and employ basic remedies.
- ILO: relate the issue to the categorization of outcomes into primary, secondary and tertiary/exploratory outcomes.
- ILO: exemplify when adjustment is needed and when it is not.

Response Adaptive Randomization
- ILO: outline what it is and recall arguments to promote its use.
- ILO: recall and exemplify its limitations; recall its use is discouraged by many.

Group-sequential trials (GST)
- ILO: recall the key elements, strength and limitations of GST
- ILO: be prepared to take advice from a statistician to design a GST
- ILO: recall examples and use R to perform a few "advanced" calculations

How to limit the risk false positive findings?

- First option: rigorously adjust for multiple testing, which usually means that we use appropriate methods that control the FWER at 5%.

  Family-wise error rate (FWER): probability of making one or more false discoveries when performing multiple hypotheses tests.

- Second option: do not not compute many p-values and do not make strong claims about 'findings' for which the computation of a p-value was not pre-specified.

Both options make perfect sense and are mentioned in the 'Statistical reporting guidelines' of the New England Journal of Medicine.[7] About the second, a simple but interesting recommendation is:

> d. When no method to adjust for multiplicity of inferences was prespecified in the protocol or SAP, reporting of secondary and exploratory end points should be limited to point estimates of effects with 95% confidence intervals. In such cases, the Methods section should state that the widths of the intervals have not been adjusted for multiplicity and that the intervals may not be used in place of hypothesis testing. The interpretation of these confidence intervals should avoid the language of definitive conclusions used to report statistically significant findings as assessed by formal hypothesis testing. These recommendations supersede previous guidance from the *Journal*, such as that provided in Wang et al. (N Engl J Med 2007;357:2189–94).

[7] Section A.2, accessed 14/11/2025, https://www.nejm.org/author-center/new-manuscripts#statistical-reporting-guidelines

Common sense

Even without rigorously adjusting for multiple testing, we can limit the risk of false positive findings and of making overly strong statement about the level of evidence supported by the data, by pre-specifying what is considered as:

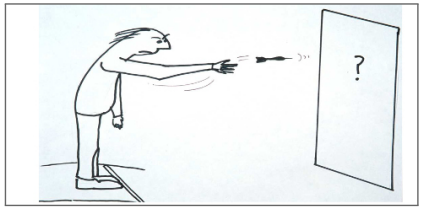
- Primary outcomes (or endpoints, estimands, or main specific research questions)
- (Key) Secondary outcomes
- Tertiary/Exploratory outcomes

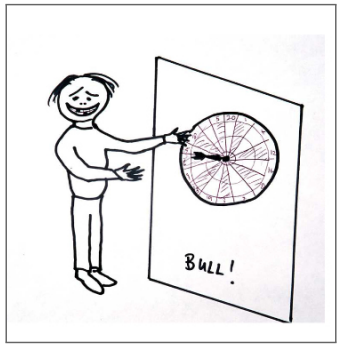
This will, to a large extent, clarify which results should be considered as a results of either a confirmatory or an exploratory analysis. This will already provide a good indication about the level of evidence that can reasonably be claimed for each result.

Recommendation: limit the number of (key) secondary outcomes. Don't select 20!

- When rigorous multiple testing adjustment is performed, the larger the number of hypothesis tests and the lower the power.
- When no adjustment is performed, the less secondary outcomes you list and the more convincing the results will be anyway if some are statistically significant. This is just common sense: multiple testing is less of an issue when few hypothesis tests are performed than when many are performed.
- Even when an outcome is clinically important, if it is associated with a low power, it might be best to consider it as a Tertiary/Exploratory outcome than a secondary outcome.

Again, prespecification matters!





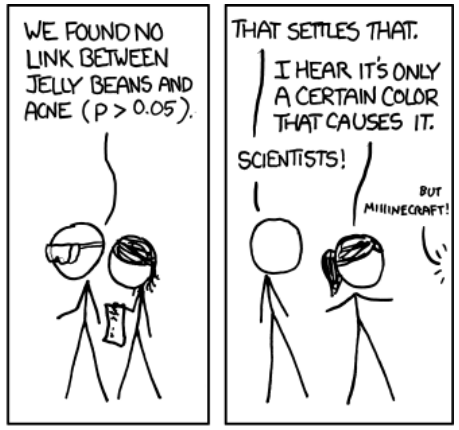
Concluding significance without prespecification is like drawing a dart-board around where the dart lands.

This might be obvious from the cartoon in the following slides. Wouldn't the results be considered substantially more "convincing" if just a few colors, including the "green" color, had been pre-specified as primary and secondary outcomes?

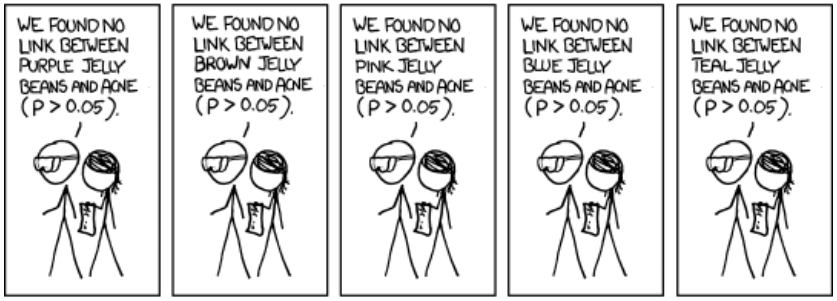
A multiple testing (extreme) example

Are jelly beans associated with acne?

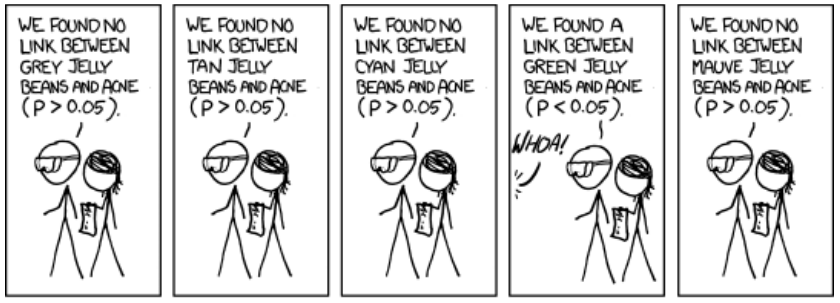
(cartoon from: https://xkcd.com/882/)



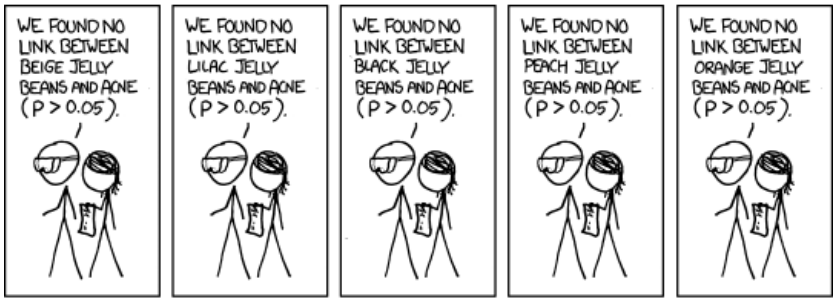
- First test is not significant.
- Move on to other tests.



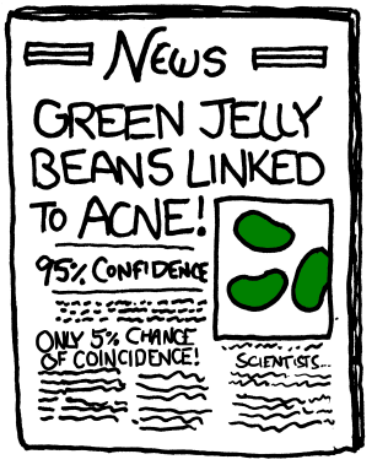
- Five more tests are not significant.
- Move on to other tests.



- Four more tests are not significant, but one is significant (Green!).
- Move on to other tests.



- Five more tests are not significant.
- Stop testing.



- Conclusion, making more than questionable claims about the level of evidence for the discovery.

How to adjust?

Numerous methods exist. See e.g., the recent FDA guideline document[8]. The most commonly used, simplest, and probably the most useful in academia, are:

- The Bonferroni Method. **Principle:** "total" risk of error (FWER) cannot exceed the sum of the errors of each test. **Drawback:** potentially serious loss of power.

  "The most common form of the Bonferroni method divides the available total α (typically 0.05 two-sided) equally among the chosen endpoints. The method then concludes that a treatment effect is significant at the α level for each one of the m endpoints for which the endpoint's p-value is less than α/m. Thus, with two endpoints, the critical α for each endpoint is two-sided 0.025." [8]

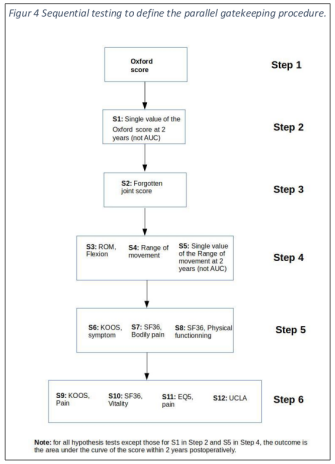
- The Fixed-Sequence Method. **Principle:** "total" risk of error (FWER) cannot exceed that of the error of the first test. **Drawback:** potentially very sensitive to the choice of the sequence!

  "In many studies, testing of the endpoints can be ordered in a specified sequence, often ranking them by clinical relevance or likelihood of success. A fixed-sequence statistical testing procedure tests endpoints in a predefined order, all at the same significance level alpha (e.g., $\alpha = 0.05$), moving to the next endpoint only after a success on the previous endpoint. Such a testing procedure requires (1) prospective specification of the testing sequence and (2) no further testing once the sequence breaks; that is, further testing stops as soon as there is a failure of an endpoint in the sequence to show significance at level alpha (e.g., $\alpha = 0.05$). The appeal of the fixed-sequence testing method is that it does not require any alpha adjustment of the individual tests. Its main drawback is that if a hypothesis in the sequence is not rejected, statistical significance cannot be achieved for the endpoints planned for the subsequent hypotheses, even if they have extremely small p-values." [8]

[8] FDA scientific guidelines: *Multiple Endpoints in Clinical Trials Guidance for Industry*, 2022, https://www.fda.gov/media/162416/download

Parallel gatekeeping procedures

Similarly to the Fixed-Sequence Method are parallel gatekeeping procedures.[9] Below is an example. [10]



- Here, there are 6 steps. Steps 1, 2 and 3 are the same as for the Fixed-Sequence Method.
- At Steps 4, 5 and 6, we adjust using Bonferroni (using $m = 3$, 3 and 4, respectively).
- But, at Step 5, we will use the significance threshold $(k/3) \cdot 5\%$ instead of 5%, where k denotes the number of previously rejected hypotheses at Step 4.
- Similarly, at step 6, we will use the significance $(k/3)(l/3) \cdot 5\%$, where l denotes the number of null hypotheses that have been rejected at step 5.

[9] Dmitrienko, Offen, Westfall. *Gatekeeping strategies for clinical trials that do not require all primary effects to be significant.* Stat Med. 2003 August;22(15):2387-2400.

[10] See SAP available at https://cdn.clinicaltrials.gov/large-docs/40/NCT03396640/SAP_000.pdf

Exercise 4.2

Consider three scenarios (A, B and C), in which we observe the unadjusted p-values in the table below. Which hypotheses do you reject in each scenario?

Outcome	A	B	C
Oxford knee score	0.061	<0.001	<0.001
Oxford knee score at t=24 months	0.012	<0.001	0.016
Forgotten Joint Score	<0.001	<0.001	<0.001
ROM flexion score	0.023	<0.001	<0.001
Range of Movement	<0.001	<0.001	<0.001
Range of Movement at t=24 months	0.055	<0.001	0.032
KOOS symptom score	0.039	<0.001	<0.001
SF36 bodily pain score	<0.001	<0.001	0.010
SF36 physical functionning score	<0.001	<0.001	<0.001
KOOS pain score	0.001	<0.001	0.001
SF36 vitality score	<0.001	0.028	0.015
EQ5D pain score	<0.001	<0.001	0.006
UCLA activity scale	0.048	0.400	0.400

Should we adjust? A debate...

"There is a strong argument for adjusting the level of evidence required for each comparison to guarantee some desired FWER. But there is a counter-argument. If these comparisons had been made in separate trials, no one would be demanding that the investigators adjust for all past and future comparisons (Rothman, 1990)."[11]

[11] Page 190 in *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022).

Should we adjust? A debate...

"There is a strong argument for adjusting the level of evidence required for each comparison to guarantee some desired FWER. But there is a counter-argument. If these comparisons had been made in separate trials, no one would be demanding that the investigators adjust for all past and future comparisons (Rothman, 1990)."[11]

"So which argument is right, the one supporting always adjusting or the one supporting never adjusting for multiple comparisons? Most people choose a middle ground between always adjusting and never adjusting for multiple comparisons. Important considerations guiding the decision are

1. *whether the trial is intended to definitively answer several questions,*
2. *the degree of multiplicity*
3. *whether the questions belong to a related family or are completely separate*
4. *whether there is an appearance of trying to gain from the multiplicity*

(Proschan and Waclawiw, 2000)." [11]

[11] Page 190 in *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022).

Factorial or Multi-arm trials: no adjustment may be accepted

"factorial trials might answer multiple unrelated questions. For example, the Women's Angiographic Vitamin and Estrogen (WAVE) factorial trial evaluated the effects of two very different interventions, hormone replacement therapy and antioxidant vitamins, on progression of heart disease in postmenopausal women with heart disease. In fact, the two components were originally planned as two separate trials and were combined to conserve resources. The separate trials argument of Rothman is compelling for WAVE and many other factorial trials."[12]

Interesting paper[13], abstract is:

"Multi-arm, parallel-group clinical trials are an efficient way of testing several new treatments, treatment regimens or doses. However, guidance on the requirement for statistical adjustment to control for multiple comparisons (type I error) using a shared control group is unclear. We argue, based on current evidence, that adjustment is not always necessary in such situations. We propose that adjustment should not be a requirement in multi-arm, parallel-group trials testing distinct treatments and sharing a control group, and we call for clearer guidance from stakeholders, such as regulators and scientific journals, on the appropriate settings for adjustment of multiplicity."

[12] Page 190 in *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022).

[13] Molloy, White, Nunn, Hayes, Wang & Harrison (2022). *Multiplicity adjustments in parallel-group multi-arm trials sharing a control group: Clear guidance is needed.* Contemporary clinical trials, 113, 106656.

Co-primary outcome: no adjustment needed

"When more than one endpoint is viewed as important in a clinical trial, then a decision must be made as to whether it is desirable to evaluate the simultaneous effects on all endpoints (termed multiple co-primary endpoints), or at least one of the endpoints (termed multiple primary endpoints, or alternative primary endpoints). This decision defines the alternative hypothesis to be tested and provides a framework for approaching trial design. When designing the trial to evaluate the simultaneous effects for all the endpoints, then no adjustment is needed to control the type I error rate. The hypothesis associated with each endpoint can be evaluated at the same significance level that is desired for demonstrating effects on all the endpoints (ICH E-9 Guideline, 1998)."[14]

- Rationale: not needed as "total" risk of error (FWER) cannot exceed that of the error of the test for one of the two co-primary outcomes.

[14] Page 80 in *Fundamental Concepts for New Clinical Trialists*, Scott Evans & Naitee Ting (2015)

Superiority after Non-inferiority: no adjustment needed

IV.1 Interpreting a non-inferiority trial as a superiority trial

If the 95% confidence interval for the treatment effect not only lies entirely above -Δ but also above zero then there is evidence of superiority in terms of statistical significance at the 5% level ($p<0.05$). See Figure 4. In this case it is acceptable to calculate the p-value associated with a test of superiority and to evaluate whether this is sufficiently small to reject convincingly the hypothesis of no difference. There is no multiplicity argument that affects this interpretation because, in statistical terms, it corresponds to a simple closed test procedure. Usually this demonstration of a benefit is sufficient on its own, provided the safety profiles of the new agent and the comparator are similar. When there is an increase in adverse events, however, it is important to estimate the size of the effect to evaluate whether it is sufficient in clinical terms to outweigh the adverse effects.

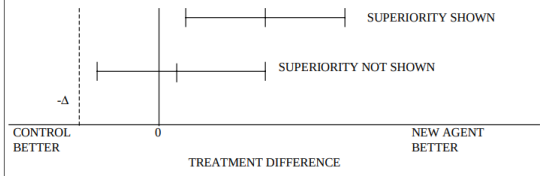


Figure 4: Non-inferiority to superiority

- Rationale: same as the Fixed-Sequence Method (closed test procedure).

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

Outline/Intended Learning Outcomes (ILOs)

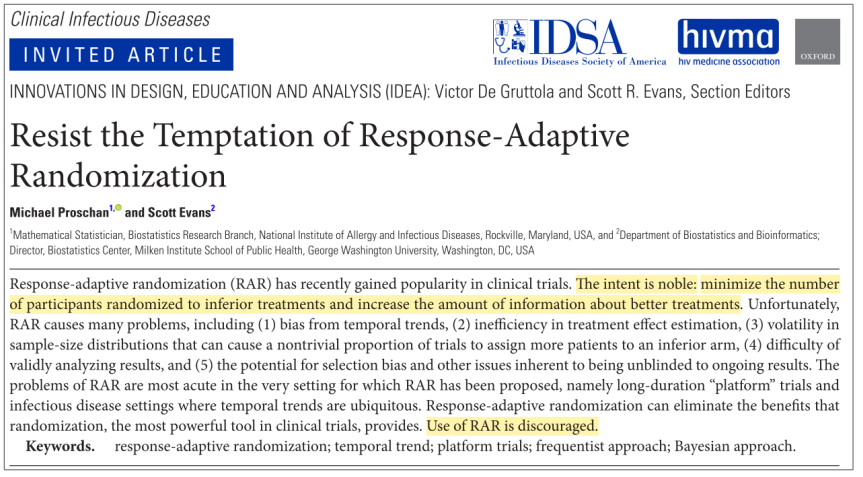
Assessing Heterogeneity of Treatment Effect
- ILO: outline what it is and recall how it is commonly done.
- ILO: recall and exemplify the limitated power.

Multiple testing
- ILO: to describe the multiple testing problem and employ basic remedies.
- ILO: relate the issue to the categorization of outcomes into primary, secondary and tertiary/exploratory outcomes.
- ILO: exemplify when adjustment is needed and when it is not.

Response Adaptive Randomization
- ILO: outline what it is and recall arguments to promote its use.
- ILO: recall and exemplify its limitations; recall its use is discouraged by many.

Group-sequential trials (GST)
- ILO: recall the key elements, strength and limitations of GST
- ILO: be prepared to take advice from a statistician to design a GST
- ILO: recall examples and use R to perform a few "advanced" calculations

A clear take home message



Clinical Infectious Diseases

INVITED ARTICLE

IDSA Infectious Diseases Society of America hivma hiv medicine association OXFORD

INNOVATIONS IN DESIGN, EDUCATION AND ANALYSIS (IDEA): Victor De Gruttola and Scott R. Evans, Section Editors

Resist the Temptation of Response-Adaptive Randomization

Michael Proschan[1] and Scott Evans[2]

[1]Mathematical Statistician, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Rockville, Maryland, USA, and [2]Department of Biostatistics and Bioinformatics; Director, Biostatistics Center, Milken Institute School of Public Health, George Washington University, Washington, DC, USA

Response-adaptive randomization (RAR) has recently gained popularity in clinical trials. The intent is noble: minimize the number of participants randomized to inferior treatments and increase the amount of information about better treatments. Unfortunately, RAR causes many problems, including (1) bias from temporal trends, (2) inefficiency in treatment effect estimation, (3) volatility in sample-size distributions that can cause a nontrivial proportion of trials to assign more patients to an inferior arm, (4) difficulty of validly analyzing results, and (5) the potential for selection bias and other issues inherent to being unblinded to ongoing results. The problems of RAR are most acute in the very setting for which RAR has been proposed, namely long-duration "platform" trials and infectious disease settings where temporal trends are ubiquitous. Response-adaptive randomization can eliminate the benefits that randomization, the most powerful tool in clinical trials, provides. Use of RAR is discouraged.

Keywords. response-adaptive randomization; temporal trend; platform trials; frequentist approach; Bayesian approach.

What is Response Adaptive Randomization (RAR)?

"When the primary outcome of a trial is very short-term such as 24-hour mortality, we could update results after every new patient. Some have argued that we should change assignment probabilities to make the next patient more likely to be given the treatment with better results thus far."[15]

Why is it appealing to many?

"Proponents argue that this so-called response-adaptive randomization is more ethical than conventional randomization techniques. The claim is that fewer patients will be assigned to poorer treatments, especially in multi-armed trials." [15]

[15] Sec. 5.7, page 70, in "Statistical Thinking in Clinical Trials", Proschan, 2022.

Why NOT using RAR?

Five problems RAR may cause: [16]

1. bias from temporal trends
2. inefficiency in treatment effect estimation
3. volatility in sample-size distributions that can cause a nontrivial proportion of trials to assign more patients to an inferior arm
4. difficulty of validly analyzing results
5. the potential for selection bias and other issues inherent to being unblinded to ongoing results

[16] As listed in abstract in Proschan & Evans (2020). *Resist the temptation of response-adaptive randomization*. Clinical Infectious Diseases, 71(11), 3002-3004.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

1. Bias from temporal trends

[...] imagine a clinical trial of 20 patients randomized using treatment:control allocation ratios of 9:1 for the first 10 patients and 1:9 for the second 10 patients. That would be very foolish because nearly all treatment patients are in the first half and all control patients are in the second half of the trial. Any observed difference between treatment and control could easily be explained by a temporal trend. [17]

[17] Proschan & Evans (2020). Clinical Infectious Diseases, 71(11), 3002-3004.

1. Bias from temporal trends

[...] imagine a clinical trial of 20 patients randomized using treatment:control allocation ratios of 9:1 for the first 10 patients and 1:9 for the second 10 patients. That would be very foolish because nearly all treatment patients are in the first half and all control patients are in the second half of the trial. Any observed difference between treatment and control could easily be explained by a temporal trend. [17]

"temporal trends seem especially likely in 2 settings: (1) trials of long duration, such as platform trials in which treatments may continually be added over many years, and (2) trials in infectious diseases such as MERS, Ebola virus, and coronavirus" [17]

[17] Proschan & Evans (2020). Clinical Infectious Diseases, 71(11), 3002-3004.

1. Bias from temporal trends

[...] imagine a clinical trial of 20 patients randomized using treatment:control allocation ratios of 9:1 for the first 10 patients and 1:9 for the second 10 patients. That would be very foolish because nearly all treatment patients are in the first half and all control patients are in the second half of the trial. Any observed difference between treatment and control could easily be explained by a temporal trend. [17]

"temporal trends seem especially likely in 2 settings: (1) trials of long duration, such as platform trials in which treatments may continually be added over many years, and (2) trials in infectious diseases such as MERS, Ebola virus, and coronavirus" [17]

Maybe equally likely in other kinds of trials?

- intervention with learning effect (e.g., surgery or use of new imaging technique)
- "subjective" inclusion criteria (not uncommon in academia?)

[17] Proschan & Evans (2020). Clinical Infectious Diseases, 71(11), 3002-3004.

2. Inefficiency in treatment effect estimation

"The only way to separate the treatment effect from a temporal trend is to compute the treatment effect estimate separately in the 2 halves of the study, average them, and use a stratified variance. Given that 1 group has 9 times the sample size of the other within each half, the variance of that treatment effect estimator is proportional to 1/9 + 1/1 instead of 1/5 + 1/5 with 1:1 randomization; the variance for 9:1 randomization is approximately 2.8 times the variance for 1:1 randomization within each half. In other words, the only way to be confident that the treatment effect estimate is not biased by temporal trends is to use an extremely inefficient estimate that requires 2.8 times as many patients to maintain the same power."[18]

Reminder: Var=SE^2 and it's (almost always) proportional to $1/n$.

[18] Proschan & Evans (2020). Clinical Infectious Diseases, 71(11), 3002-3004.

3. Volatility in sample-size distributions

"Another issue is that RAR results in highly variable per-arm sample sizes, resulting in a nontrivial probability of a larger sample size with the inferior arm than with the superior arm [6]. As noted in reference [6], this negative consequence of RAR has not been appreciated because only the expected sample sizes are typically reported, which obscures the large variability in actual sample sizes per arm." [19] [20]

[19] Proschan & Evans (2020). Clinical Infectious Diseases, 71(11), 3002-3004.

[20] Reference [6] is Thall et al. *Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials.* Ann Oncol 2015; 26:1621–8. 7.

4. Difficulty of validly analyzing results (1/2)

"The Bayesian approach allows seamless analysis of results of a trial that uses RAR. The frequentist approach faces great difficulties in the setting of RAR [..]" [21]

[21] Proschan & Evans (2020). Clinical Infectious Diseases, 71(11), 3002-3004.

4. Difficulty of validly analyzing results (1/2)

"The Bayesian approach allows seamless analysis of results of a trial that uses RAR. The frequentist approach faces great difficulties in the setting of RAR [..]" [21]

"The frequentist approach results in the conclusion that treatment is effective on the basis of how unlikely the observed or better results would be if the treatment truly had no effect. The Bayesian approach, on the other hand, quantifies prior belief about the treatment effect, and then updates that to a posterior belief after observing data." [21]

[21] Proschan & Evans (2020). Clinical Infectious Diseases, 71(11), 3002-3004.

4. Difficulty of validly analyzing results (2/2)

"A major problem is that most analysis methods, such as t tests, tests of proportions, linear models, etc, treat sample sizes in different arms as fixed constants. That is ill advised in a trial using RAR because sample sizes themselves contain important information about whether treatment works." [22]

[22] Proschan & Evans (2020). Clinical Infectious Diseases, 71(11), 3002-3004.

[23] Reference [12] is Wei LJ. Exact two-sample permutation tests based on the randomized play-the-winner rule. Biometrika 1988; 75:603–6)

4. Difficulty of validly analyzing results (2/2)

"A major problem is that most analysis methods, such as t tests, tests of proportions, linear models, etc, treat sample sizes in different arms as fixed constants. That is ill advised in a trial using RAR because sample sizes themselves contain important information about whether treatment works." [22]

"Conditioning on per-arm sample sizes, as we would do for virtually any other form of randomization, conditions away evidence of a treatment benefit. Wei [12] showed that the P value in the original ECMO trial is approximately 0.62 or 0.051 depending on whether we condition on per-arm sample sizes or not." [22] [23]

[22] Proschan & Evans (2020). Clinical Infectious Diseases, 71(11), 3002-3004.

[23] Reference [12] is Wei LJ. Exact two-sample permutation tests based on the randomized play-the-winner rule. Biometrika 1988; 75:603–6)

CONCLUSIONS

Response-adaptive randomization has the potential to nullify many of the advantages of randomization. Disadvantages of RAR include bias in the presence of temporal trends unless one uses an inefficient treatment effect estimate that can result in more, not fewer, patients receiving the inferior treatment (less ethical); analysis issues that could make a substantial proportion of readers question whether trial results are convincing; and problems inherent in loss of blinding. Proponents of RAR point out that bias is less of a problem with less-volatile assignment probabilities and a burn-in period, and that bias can be addressed through modeling. For some people, the perceived ethical advantages of RAR outweigh the above disadvantages, although one can never be confident that all relevant variables have been measured or accounted for correctly in modeling. A properly randomized clinical trial should yield valid inferences without the need to appeal to models to correct for bias. Response-adaptive randomization has noble intent, but introduces serious problems that jeopardize the integrity of a clinical trial.

(Proschan & Evans (2020). Clinical Infectious Diseases, 71(11), 3002-3004)

Outline/Intended Learning Outcomes (ILOs)

Assessing Heterogeneity of Treatment Effect
- ILO: outline what it is and recall how it is commonly done.
- ILO: recall and exemplify the limitated power.

Multiple testing
- ILO: to describe the multiple testing problem and employ basic remedies.
- ILO: relate the issue to the categorization of outcomes into primary, secondary and tertiary/exploratory outcomes.
- ILO: exemplify when adjustment is needed and when it is not.

Response Adaptive Randomization
- ILO: outline what it is and recall arguments to promote its use.
- ILO: recall and exemplify its limitations; recall its use is discouraged by many.

Group-sequential trials (GST)
- ILO: recall the key elements, strength and limitations of GST
- ILO: be prepared to take advice from a statistician to design a GST
- ILO: recall examples and use R to perform a few "advanced" calculations

Phase 3 clinical trials (in the Pharmaceutical industry)

Phase III trials are conducted as the last stage in the drug development process.

Two positive studies are usually required to confirm that a new treatment is superior to the current standard treatment.

Regulators customarily require a hypothesis test to reach significance at the one-sided 2.5% level.

Studies may recruit hundreds, or even thousands, of subjects at a cost of as much as 10k to 50k euros per patient.

The time taken to reach a conclusion eats into the limited patent lifetime remaining to the company developing the drug.

Thus, there are strong incentives to reach an early conclusion for either a positive or negative decision.

Acknowledgments: this slides and several of the following are based of those of Prof. Chris Jennison, from the short course organized by the Danish Society of Biopharmaceutical Statistics, in September 2019.

What are Group sequential trials?

"Group sequential trials (GST) allow researchers to evaluate accumulating data at preplanned time intervals during a trial, without compromising the validity of the final analysis results. These interim looks at the data allow for early stopping of a trial in case of a clear effect of the treatment or a clear lack thereof."[24]

[24] Baayen et al. *Sequential trials in the context of competing risks: Concepts and case study, with R and SAS code.* Statistics in Medicine 38.19 (2019): 3682-3702.

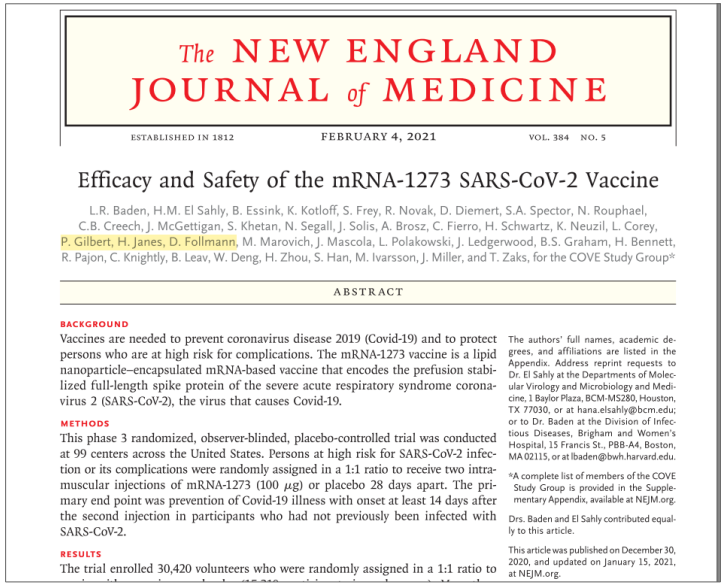
Why GST?

Group sequential trials can be:

- more ethical and efficient (in both academia and industry)
  - stop early if clearly effective / ineffective treatment: patients are better treated faster! (e.g. covid-19 vaccine available earlier)
- economically interesting (mostly in industry)
  - in average, smaller sample size (reduced costs)
  - shorter trial if stop early implies longer period the drug can be marketed with patent protection.

GST are becoming increasingly popular in the pharmaceutical industry and some rumors say that regulatory agencies (EMA, FDA) now tend to consider them as the "default approach". Also increasingly popular in academia, although not very popular in academia yet.

Example: Moderna covid-19 vaccine



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 FEBRUARY 4, 2021 VOL. 384 NO. 5

Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine

L.R. Baden, H.M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S.A. Spector, N. Rouphael, C.B. Creech, J. McGettigan, S. Khetan, N. Segall, J. Solis, A. Brosz, C. Fierro, H. Schwartz, K. Neuzil, L. Corey, P. Gilbert, H. Janes, D. Follmann, M. Marovich, J. Mascola, L. Polakowski, J. Ledgerwood, B.S. Graham, H. Bennett, R. Pajon, C. Knightly, B. Leav, W. Deng, H. Zhou, S. Han, M. Ivarsson, J. Miller, and T. Zaks, for the COVE Study Group*

ABSTRACT

BACKGROUND

Vaccines are needed to prevent coronavirus disease 2019 (Covid-19) and to protect persons who are at high risk for complications. The mRNA-1273 vaccine is a lipid nanoparticle–encapsulated mRNA-based vaccine that encodes the prefusion stabilized full-length spike protein of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes Covid-19.

METHODS

This phase 3 randomized, observer-blinded, placebo-controlled trial was conducted at 99 centers across the United States. Persons at high risk for SARS-CoV-2 infection or its complications were randomly assigned in a 1:1 ratio to receive two intramuscular injections of mRNA-1273 (100 μg) or placebo 28 days apart. The primary end point was prevention of Covid-19 illness with onset at least 14 days after the second injection in participants who had not previously been infected with SARS-CoV-2.

RESULTS

The trial enrolled 30,420 volunteers who were randomly assigned in a 1:1 ratio to

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. El Sahly at the Departments of Molecular Virology and Microbiology and Medicine, 1 Baylor Plaza, BCM-MS280, Houston, TX 77030, or at hana.elsahly@bcm.edu; or to Dr. Baden at the Division of Infectious Diseases, Brigham and Women's Hospital, 15 Francis St., PBB-A4, Boston, MA 02115, or at lbaden@bwh.harvard.edu.

*A complete list of members of the COVE Study Group is provided in the Supplementary Appendix, available at NEJM.org.

Drs. Baden and El Sahly contributed equally to this article.

This article was published on December 30, 2020, and updated on January 15, 2021, at NEJM.org.

(2 interim analyses, Cox model, powered for modest effect VE=1-HR=0.6, found was 0.94)

Statistical framework

Suppose a new treatment (Treatment A) is to be compared to a placebo or positive control (Treatment B).

The treatment effect θ for the primary endpoint represents the advantage of Treatment A over Treatment B.

If $\theta > 0$, Treatment A is more effective.

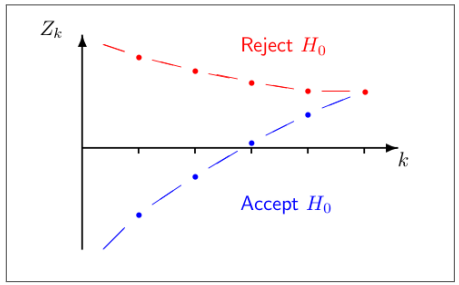
We wish to test the null hypothesis $\mathcal{H}_0 : \theta \leq 0$ against $\theta > 0$ with

- $P_{\theta=0}\{\text{Reject } \mathcal{H}_0\} = \alpha$ (type-I error)
- $P_{\theta=\delta}\{\text{Reject } \mathcal{H}_0\} = 1 - \beta$ (power).

In a group sequential trial, data are examined on a number of occasions, $k = 1, \ldots, K$, to see if an early decision may be possible.
Often, $K = 2$ (one interim analysis).

The main statistical challenge

What we want (most) is to find appropriate boundaries. Typical boundaries for a one-sided test, expressed in terms of standardised test statistics $Z_1, \ldots, Z_K$ (with $K - 1$ interim analyzes), have the form:



Crossing the upper boundary leads to early stopping for a positive outcome, rejecting $\mathcal{H}_0$ in favour of $\theta > 0$. Crossing the lower boundary leads to early stopping for futility ('accept' $\mathcal{H}_0$).

Note: unlike in most observational studies, in properly powered trials it is less controversial to say that we can "accept" $\mathcal{H}_0$.

A bit of theory (sequential distribution theory)

Remarkably, almost always[25], $(\widehat{\theta}_1, \ldots, \widehat{\theta}_K)$ follows, asymptotically (i.e., for large sample sizes), the canonical joint distribution.

For testing $\mathcal{H}_0 : \theta = 0$, the standardised statistic at analysis k is

$$Z_k = \frac{\widehat{\theta}_k}{\mathrm{SE}(\widehat{\theta}_k)} = \widehat{\theta}_k \sqrt{\mathcal{I}_k}$$

For these statistics, asymptotically,

$$(Z_1, \ldots, Z_K) \quad \text{is multivariate normal,}$$
$$Z_k \sim N\left(\theta\sqrt{\mathcal{I}_k}, 1\right), \quad k = 1, \ldots, K$$
$$\mathrm{Cov}\left(Z_{k_1}, Z_{k_2}\right) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}} \quad \text{for } k_1 < k_2$$

We call $\mathcal{I}_k = 1/\mathrm{SE}^2(\widehat{\theta}_k) = 1/\mathrm{Var}(\widehat{\theta}_k)$ the information (at analysis k).

[25] That is, using most of the "usual" statistical methods. See e.g., Theorem 9.3 in *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022) or Jennison & Turnbull: *Group sequential methods with applications to clinical trials*. CRC Press, 1999.

Why is this theory so nice?

In short, whatever

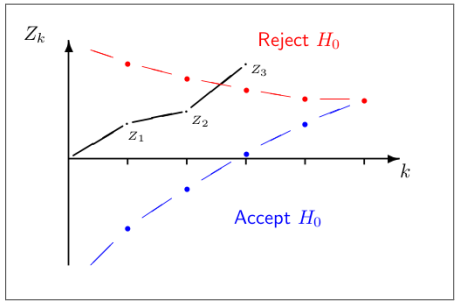
- the kind of data we collect
- the parameter θ of interest
- the standard error $\mathrm{SE}(\widehat{\theta}_k)$ or equivalently the the information $\mathcal{I}_k$,

what is important to know to perform the computation is essentially only how much information is acquired between successive analyses, as this is sufficient to compute the correlation/covariance. This makes the theory widely applicable and make it possible to have "one size fits all" software.

Often (but not always!) $\mathrm{SE}(\widehat{\theta}_k)$ is proportional to $1/\sqrt{n_k}$, hence the information $\mathcal{I}_k$ is proportional to n_k. E.g., $\mathcal{I}_k = n_k/(2\sigma^2)$ or $\mathcal{I}_k = n_k/\{p_C(1-p_C) + p_T(1-p_T)\}$

(First is when comparing the means of quantitative outcomes; outcome measured almost immediately after inclusion. Second when comparing 30-day mortality via a difference in proportion.).

Computation for group sequential tests



Today, probabilities such as $P_\theta\{l_1 < Z_1 < u_1,\, l_2 < Z_2 < u_2,\, Z_3 > u_3\}$ can easily be computed (e.g. `R` package `mvtnorm`).

Combining these probabilities yields type I error rate, power, expected sample size, etc., of a group sequential design.

Boundaries and group sizes can be chosen to define a test with a specific type I error probability and power.

Type-I error, with binding futility boundary:

$$
\begin{aligned}
&P_{\theta=0}\{\text{Reject } \mathcal{H}_0\} = \alpha \\
&= \sum_{k=1}^{K} P_{\theta=0}\{\text{Reject } \mathcal{H}_0 \text{ at analysis } k\} \\
&= \sum_{k=1}^{K} P_{\theta=0}\big\{ \underbrace{l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}}_{\text{Continue until } k\text{-th analysis, as recommended}}, \underbrace{Z_k > u_k}_{\text{Reject}} \big\}
\end{aligned}
$$

with non binding:

$$
= \sum_{k=1}^{K} P_{\theta=0}\big\{ \underbrace{Z_1 < u_1, \ldots, Z_{k-1} < u_{k-1}}_{\text{Continue until } k\text{-th analysis, because no reject}}, \underbrace{Z_k > u_k}_{\text{Reject}} \big\}
$$

Power: same as for type-I error, but $P_{\theta=\delta}$ instead of $P_{\theta=0}$. (Same as type-I error, with binding futility boundary in both cases, as it makes sense to assume that we will stop for futility if the data suggest so.)

Binding or non-binding futility boundaries?

"the investigators sometimes want to have the opportunity to continue the trial while the data suggest stopping for futility. In that case, the futility boundaries are used for guidance only but not for defining a ***binding*** *rule."*[26]

E.g. collecting more data on secondary endpoints can be interesting. It can be allowed if continuing the trial after crossing the futility boundary is not thought unethical.

If a futility boundary is deemed to be non-binding, the type I error rate should be computed ignoring the futility boundary. (worst case scenario)

However, investigators will wish to know power and expected sample size when the futility boundary is obeyed. (expected behavior)

[26] Baayen et al. *Sequential trials in the context of competing risks: Concepts and case study, with R and SAS code.* Statistics in Medicine 38.19 (2019): 3682-3702.

Appendix: quotes recommending non-binding

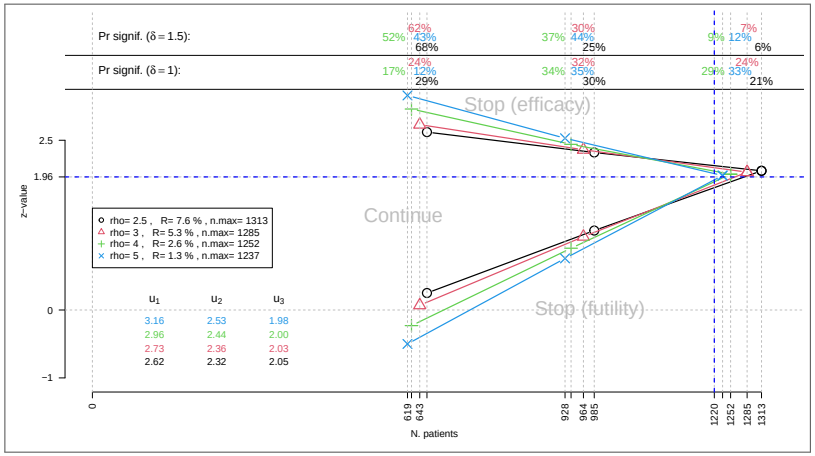
Remark 9.13. *We recommend against adjusting upper (efficacy) boundaries to account for lower futility boundaries because DSMBs often regard futility boundaries as advisory rather than binding.*

"Clinical trials are monitored throughout by a Data and Safety Monitoring Board (DSMB), a group of outside experts who make recommendations to the trial sponsor about whether to continue as is [...] or stop the study. The sponsor makes the final decision, but they almost always accept the recommendations of the DSMB."[27]

Remark: similar recommendations in e.g., Lan, KK Gordon, and David DeMets. *Further comments on the alpha-spending function.* Statistics in Biosciences 1:1 (2009): 95-111.

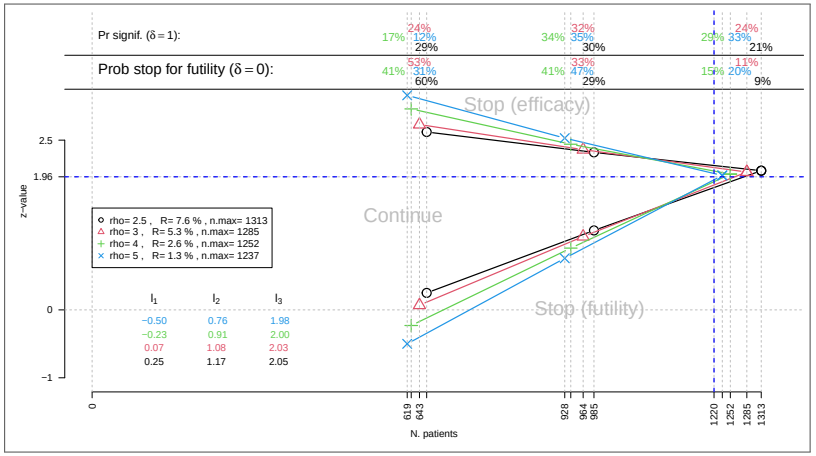
[27] Page 147 in *Statistical Thinking in Clinical Trials*, by Michael Proschan (2022). Remark 9.3 can be found page 173.

Example: some boundary choices (Example & Exercise 3.1, con't)



We plan to have interim analyses after 50% and 75% of the (maximal sample size) data have been collected.

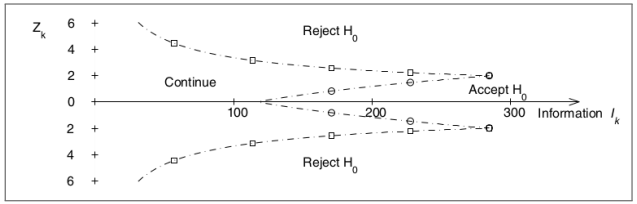
- The larger the chances of stopping early, the larger the maximal sample size to keep the 80% power (see inflation factor R in legend).
- If no interim analysis: one boundary at 1.96, $n = 610 \times 2 = 1220$.
- Chances to stop early not negligible (e.g., 24%+32%=56% with red choice, if treatment effect as large as assumed to power for 80%; and 92% if treatment effect is 50% larger; in that case 62% at half the sample size)



- More ethical than a fixed sample design: we have reasonable chances of stopping early for futility, if the data suggest that the trial will likely **NOT** be positive if we continue after the interim analysis. This would limit the number of (unnecessary) transfers of severely ill patients from the ICU to the treatment facility, which might create more harm than good, if the treatment is not efficacious. (e.g., 53% chance to stop at half the (maximal) sample size with the red choice, if no treatment effect; 33% chances at 75% of (maximal) sample size; hence 86% to stop early, overall.)

Digression: why a one-sided test matters.

- Often, it's 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% (see previous slides).
- In a group sequential trial, it is not. The boundaries would be completely different for a two-sided test, which would correspond to the different objective of showing than one of the two arms is better than the other (no matter which). Using a two-sided test will lead to the decision to continue the trial if the interim data suggest that the experimental treatment might be harmful, as long as there is a decent power to show a statistically significant harmful effect at the end of the trial, if the trial is continued.
- Example of boundary shapes for a two-sided test: [28]

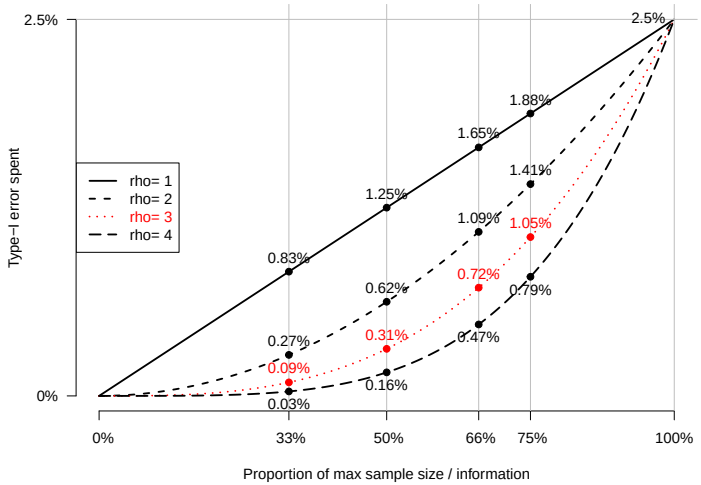


[28] source: Fig 5.1 in Jennison & Turnbull. *Group sequential methods with applications to clinical trials.* CRC Press, 1999.

Further remarks

- Note: $u_K > 1.96$ is to compensate for multiple testing due to multiple looks at the data.
- But we correct in a more clever way than using Bonferonni! (as `qnorm(1-0.025/3)`$= 2.39$)
- As we correct for multiple testing by using "larger" critical values, the power will be too low if we use a maximal sample size (n_{max}) equal to that needed for a fixed design trial (n_{fixed}). Hence the inflation factor R, which tells us that we should use a larger sample size to compensate: $n_{\text{max}} = R \cdot n_{\text{fixed}}$.
- The inflation factor depends on the number of interim analyses, when they happen (i.e., more or less early) and on the choice of shape of the boundary, or equivalently, the so-called α- and β-spending functions. These functions define how much type-I and type-II error we want to spend at each analysis.

Spending functions



Shown are power family spending functions $\alpha(t) = \max\{\alpha t^\rho, \alpha\}$ (Kim and DeMets, 1987). Alternative commonly used spending functions are the so called "Pocock" and "O'Brien & Fleming" spending functions. They lead to designs relatively similar to those with $\rho = 1$ and 3, respectively. See Sect. 7.2.2 in Jennison & Turnbull. *Group sequential methods with applications to clinical trials*, 19

Practice: use R! (Exercise 4.3) (1/2)

1. Run `RDemoGST.R` and recognize the results for "rho=3", especially:
   - the lower and upper boundary values (l_k, u_k), for $k = 1, , 3$.
   - chances of stopping early for both efficacy and futility, at each interim analysis. By the way, do they sum-up to what you expect?
   - inflation factor R.
2. Recompute by yourself these values (at first interim analysis):
   - 53% chance of stopping early for futility, if the treatment does not work.
   - 24% chance of stopping early for efficacy, if the treatment effect indeed corresponds to the value assumed to power the trial.
3. Change the code as relevant and recognize the main results for another "rho" value.

Practice: use R! (Exercise 4.3) (2/2)

4. Let's come back to the design with "rho=3". Remember, we assumed that we expect to enroll approximately 130 patients per year. How long will last the trial, approximately, in each case (stop early or not)?
5. Let's assume that the treatment effect is more modest than expected, say half the efficacy value used to power the trial. What do become the probability of early stopping for efficacy and futility? Does it make sense?
6. Given the long trial duration, we could add an interim analysis after 25% of the patients have been accrued. What would be the pros and cons of this approach? Would it be worth it? Think broadly, e.g., also about logistic and ethical issues...
7. If we decide to have this earlier interim analysis, would it make sense to choose a different "rho" parameter?
8. What about having only 2 interim analyses, at 33% and 66%? Could it be interesting?

Final remarks

- Group-sequential designs offer a lot of interesting options (great flexibility).
- But the design and analysis is not straightforward (especially computation of 95%-CI, unbiased point estimates and p-value after stopping; adjusting for the "random low" or "random high" selection bias is needed; methods exists). Seek advice from a competent statistician!
- Long term outcomes (e.g., weight loss at one-year post-randomization) are often not appropriate for GST. Patients including before the interim analysis who have not yet completed the necessary follow-up to measure the outcome can be problematic (pipeline data).
- Similar methodology may be used to help monitoring safety only.
- "Adaptive designs" are an extension of group-sequential designs that allow to adapt the design of the trial based on the results of the interim analysis (e.g, change inclusion criteria, update maximal sample size, drop or add a arm).
- A new interesting guidelines document has just been released. It covers both group-sequential and adaptive trials.[29] It is worth reading and comments are still welcome!

[29] ICH/EMA harmonised guideline: *Adaptive designs for clinical trials E20*, 2025 (acessed 12/11/2025), https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e20-guideline-adaptive-designs-clinical-trials-step-2b_en.pdf