Salvaging type 1 error rate control by randomisation tests when your trial is affected by Covid?

Martin Posch (Medical University of Vienna)

September 15, 2020

Trials affected by COVID 19

- Disruptions due to COVID-19 can affect inference in clinical trials in various ways.
- The estimand tested in the trial may change due to
  - change in the patient population (e.g., because certain risk groups do no longer participate)
  - the course of the disease is affected by the impact of COVID-19
  - the efficacy of the treatment is affected (e.g., due to compliance, change in lifestyle)
  - the primary endpoint cannot be assessed as planned.

Interim Data Review EMA

1 26 June 2020
2 EMA/158330/2020 Rev. 1
3 Committee for Human Medicinal Products (CHMP)

4 Points to consider on implications of Coronavirus disease
5 (COVID-19) on methodological aspects of ongoing clinical
6 trials
7

“Sponsors are advised to contemplate an analysis of the accumulating trial data in order to evaluate the implications on recruitment, loss of study participants during the trial, ability to record data and ability to interpret the treatment effect.”

“A more thorough analysis based on blinded review may be warranted, but the use of unblinded data is not recommended. Any analysis that bears the risk, however small, of unblinding should be specified a priori and conducted independently of the Sponsor supervised by an independent Data Monitoring Committee (DMC).”

Interim Data Review FDA

Contains Nonbinding Recommendations

Statistical Considerations for Clinical Trials During the COVID-19 Public Health Emergency

Guidance for Industry

June 2020

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Center for Devices and Radiological Health (CDRH)
Center for Veterinary Medicine (CVM)

"Appropriate participant data to consider when making modifications to the trial to address the impact of COVID-19 include summaries pooled over treatment arms including information on missing data, participant treatment discontinuation or interruptions, participant trial withdrawal, and endpoints."

"Generally, for a blinded trial, modifications based on information that reveals the magnitude of the treatment effect or information presented by treatment arm have the potential to introduce bias."

Trial Adaptations Based on External & Blinded Data

- Disruptions due to COVID-19 can make it unfeasible to continue a trial as planned and adaptations may become necessary.
- What is the potential impact of adaptations on the false positive rates of hypotheses tests?

- It is known that adaptations of clinical trials based on
  - **external data only** has no impact on the type I error rate (given that the adapted trial controls the unconditional type I error rate).
  - **unblinded, comparative data** can lead to a substantial inflation of the type I error rate.
  - **blinded estimates of nuisance parameters** (sample size reassessment) does not inflate the type I error rate.
- What about adaptations that depend on external information and the overall blinded clinical trial data (the full data set without treatment labels)?

Example: START:REACTS Trial

- Randomised, adaptive, multicentre, controlled trial comparing arthroscopic debridement with the InSpace balloon (Stryker, USA) to arthroscopic debridement alone
- Recruitment start: February 2018
- Planned total sample size of 221 with the potential to stop the study for efficacy or futility at a number of interim analyses.
- Participant and assessor blinded
- Primary endpoint: Constant Shoulder Score at 12 month recorded at a hospital out-patient visit.

KUNZ ET AL. 2020, PARSONS ET AL. 2019, METCALFE ET AL. 2020

START:REACTS Trial: Impact of COVID-19

- Delayed Recruitment (cancellation of elective surgery).
- Disruption of follow-up data collection for the primary endpoint (planned appointments not attended)
- Constant Shoulder Score was replaced by the Oxford Shoulder Score determined over the phone
- Assume at the end of the trial a permutation test is performed with the adapted endpoint.

What's the impact on the type I error rate if the decision to change the endpoint was also influenced by inspection of the blinded data set?

Classical Permutation Test

- $\underline{Y}$ . . . outcome data with realisation $\underline{y}$.
- $\underline{Z}$ . . . treatment assignment vector
- Compute the permutation distribution by computing the test statistics $t(\underline{y}, \underline{z})$ for each possible assignment $\underline{z}$.
- p-value: refer the actually observed test statistics to this permutation distribution.
- The permutation test is valid under independence of $\underline{Y}$ and $\underline{Z}$ (i.e., the conditional distribution of $\underline{Y}$ given $\underline{Z} = \underline{z}$ is the same for every consistent $\underline{z}$).

Adaptive Permutation Test

- $\underline{X}$ . . . trial data examined before breaking the blind (may include primary and secondary outcomes)
- Based on $\underline{X}$ it is decided if the primary endpoint is switched from $\underline{Y}$ to an alternative endpoint
- For the selected endpoint a permutation test is performed

Type I error Control

If $\underline{X}$ and $\underline{Z}$ are independent, the permutation test remains valid under adaptations.

- If $\underline{X}$ and $\underline{Z}$ are independent, the permutation test controls the level conditional on the observed data $\underline{X}$.
- Because the adaptations are a function of $\underline{X}$ the level conditional on the adaptation is still α.

The result extends to endpoint adaptations in an interim analysis and to changes in the test statistics.

A Counter Example and the Proper Conclusion

- Three endpoints observed in a blinded way
  1. mortality
  2. serious AIDS events
  3. level of experimental drug in the blood
- Examining the level of study drug unblinds the investigator when looking at the patient level data even if no treatment label is provided.
- He/she can choose from the other two endpoints the one that had a stronger observed treatment effect.
- This increases the type I error rate.

Posch, M., & Proschan, M. A. (2012).

What goes wrong?

- The conclusion we would like to make is that the treatment improved the adapted outcome.
- The null hypothesis tested by the (adaptive) permutation test is that the treatment has no effect on $\underline{X}$, the data used in the adaptation (global null hypothesis), i.e., that $\underline{X}$ and $\underline{Z}$ are independent.
- When the result is statistically significant, the proper conclusion is that the treatment has an effect on at least one of the examined outcomes $\underline{X}$.
- This is a useless conclusion in this counterexample.

Is the problem ameliorated by considering only summary information?

- For instance, knowing only the amount of study drug in the blood combined gives little information.
- However consider the correlation between the level of the study drug in the blood and the two other endpoints: these correlations tell which of the endpoints has the larger t-statistic!
- In settings with additional endpoints even with univariate summary measures the level can be inflated.

The Case of Two Potential Endpoints

The type I error rate is controlled if only two potential endpoints without missing data are examined in a blinded way ($\underline{X}$ consists of 2 variables only).

Assume that based on $\underline{X}$ one endpoint is selected and a permutation test is performed.

- Under the global null hypothesis the type I error is controlled because $\underline{X}$ is independent of Z.
- Assume the treatment has an effect on Endpoint 1 but not Endpoint 2: For a type I error, Endpoint 2 must be selected and its permutation test must be significant. This probability is bounded by α.

What if there is missing data?

- If either endpoint has missing data and missing is informative the number of missing observations can act as a third endpoint
- Inflation is possible

In addition, informative missing data generates bias also without adaptations.

Knowledge of Per-Arm Sample Sizes

If the blinded data of the primary endpoint and the sample sizes in each arm are known at the interim analysis where the endpoint maybe adapted, the permutation test no longer controls the conditional (and overall) type I error rate at level α.

- The per-arm sample sizes can give some information about the treatment effect:
- Example: Binary endpoint, 3 patients per arm, Interim analysis after 3 patients.

P	0	0	0
T	.	.	.

- Solution - Use a stratified permutation test. Consider only those permutations yielding the number of Ts and Cs both *before* and *after* the adaptation.

Adaptive tests controlling the type I error rate after unblinded interim analyses

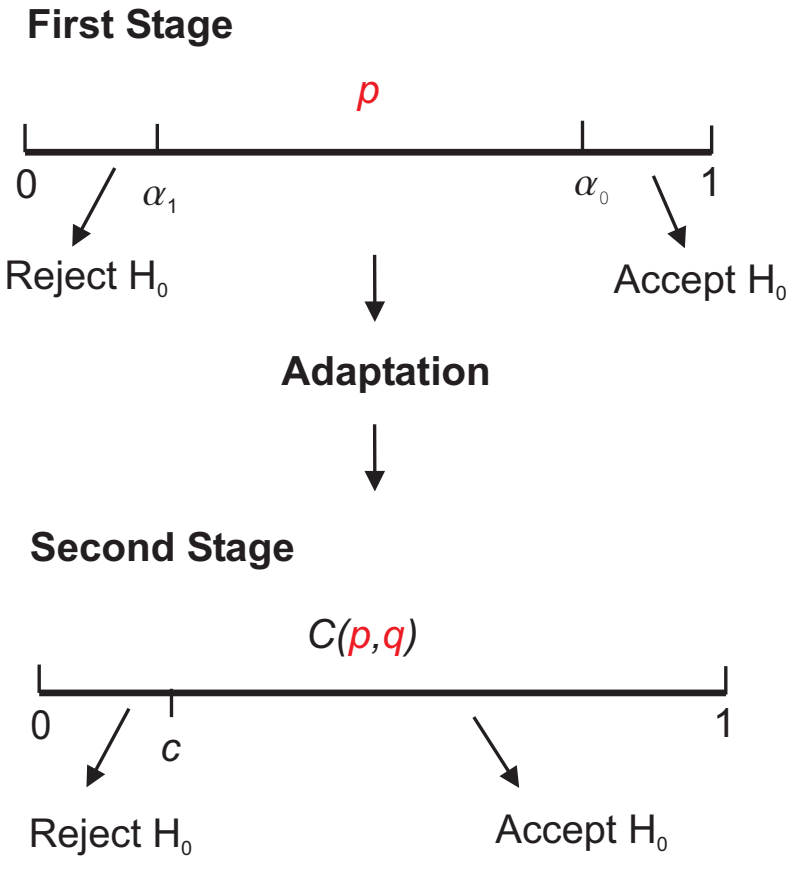
- A trial is performed in two stages
- In an interim analysis the trial may be adapted based on unblinded interim data:
  - adaptations may depend on all (unblinded) interim data including secondary and safety endpoints.
  - the adaptation rule is not preplanned.

How to construct a test that controls the type I error?

- Tests Based on Combination Tests
- Tests Based on the Conditional Error Rate

The Combination Test

(BAUER '89, BAUER & KÖHNE '94)



Planning:

- Fix design for stage 1 (sample sizes, test, ...)
- Fix adaptive combination test (combination function and α_1, α_0 and c)

Stage 1:

- Compute p-value p from Stage-1-data
- Fix design for stage 2 based on data from stage 1

Stage 2:

- Compute p-value q from stage-2-data.
- Reject H_0 iff $C(p, q) \leq c$.

Design Adaptations with Combination Tests

Examples of combination functions:

- Fisher's product test: $C(p, q) = p\,q$
- Weighted inverse normal test:

$$C(p, q) = -\sqrt{w}\,\Phi^{-1}(1-p) - \sqrt{1-w}\,\Phi^{-1}(1-q),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution and $\Phi^{-1}(\cdot)$ its inverse.

CUI, HUNG & WANG '99, LEHMACHER & WASSMER '99

Computation of Stopping Boundaries

The critical values are determined such that the combination test rejects with probability α given the the p-values p and q are independent and uniformly distributed on [0,1], i.e. such that

$$\alpha_1 + P_{H_0}\{\alpha_1 < p \leq \alpha_0, C(p,q) \leq c\} = \alpha.$$

Type I Error Control of the Adaptive Test

- The stage-wise p-values can be, e.g., p-values of randomisation tests (based on the stage-wise data).
- If different hypotheses are tested at the two stages (e.g., because different endpoints are tested) the combination test tests the intersection hypothesis only (“No treatment effect in both endpoints”).
- Using the closed testing principle, one can construct a test controlling the Type 1 error rate for the selected hypothesis.
- For the testing strategy we need to distinguish two scenarios:
  - The alternative endpoint is a pre-specified secondary endpoint and part of a preplanned multiple testing procedure (e.g., a hierarchical test).
  - the alternative endpoint was not part of the original primary analysis

Change of Endpoints

No early stopping ($\alpha_1 = 0, \alpha_0 = 1$)

p, q (p', q')... stage-wise p-values of the original (adapted) endpoint

Alternative EP is pre-planned secondary EP	Alternative EP not part of pre-planned primary analysis
If it is decided to continue with the original endpoint, reject if $$C(p, q) < c$$	
If it is decided to adapt the endpoint, the closed test rejects if	
$$C(p, q') \leq c$$ and $$C(p', q') \leq \alpha.$$	$$C(p, q') \leq c$$ and $$q' \leq \alpha.$$

HOMMEL (2001)

Discussion

- For non-inferiority tests even the blinded data of the primary endpoint reveals information on the treatment allocation.
- Blinded adaptations can also have an impact on bias and MSE of treatment effect estimates

  POSCH ET AL. (2018)
- Adaptations can also impact the type I error rate of tests for secondary endpoints.

Conclusion

- Unrestricted blinded data mining results in a useless conclusion.
- The adapted permutation test tests only the global null hypothesis that there is no effect in none of the endpoints
- Therefore, restrict the set of variables under consideration such that the global null hypothesis is still of interest.
- Or use a multiple test based on the closed testing procedure:
  - If only two endpoints are examined, the standard randomization test controls the type I error rate
  - If more than two variables are inspected either a closed testing procedure can be pre-defined independently of the data
  - or the toolbox of adaptive designs for unblinded IA can be used (even if adaptations are based on blinded data).


Posch, M., & Proschan, M. A. (2012). Unplanned adaptations before breaking the blind. Statistics in Medicine, 31(30), 4146-4153.


Proschan, M., Glimm, E., & Posch, M. (2014). Connections between permutation and t-tests: relevance to adaptive methods. Statistics in Medicine, 33(27), 4734-4742.


EMA (2020), Points to Consider on Implications of Coronavirus Disease (COVID-19) on Methodological Aspects of Ongoing Clinical Trials, available at https://www.ema.europa.eu/.


FDA (2020), Guidance on Conduct of Clinical Trials of Medical Products During COVID-19 Pandemic Guidance for Industry, Investigators, and Institutional Review Boards, available at www.fda.gov.


Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. Biometrical Journal, 43(5), 581-589.


Klinglmüller, F. (2013) Flexible Tests for Clinical Trials Testing Multiple Hypotheses, PhD Thesis, Technical University Vienna


Posch, M., Klinglmüller, F., König, F., & Miller, F. (2018). Estimation after blinded sample size reassessment. Statistical methods in medical research, 27(6), 1830-1846.