When convention meets practicality: Combined analysis under the two-trials convention

Dong Xi (Gilead Sciences)
Frank Bretz (Novartis)
Willi Maurer (Novartis)

Wiener Biometrische Sektion Seminar
21 January 2025

Familywise error rate (FWER) control

- FWER: probability to reject at least one true null hypothesis within a trial
  - Strong FWER control at a pre-specified significance level α
- Recently released regulatory guidelines
  - EMA (2017), FDA (2017)
- Common requirements for strong FWER control
  - Prospective analysis plan
  - Careful classification of endpoints for multiplicity adjustment
  - Proper adjustment methods
- Principles apply to multiple endpoints, doses, subgroups, time points, analyses, etc.

Two-trials convention

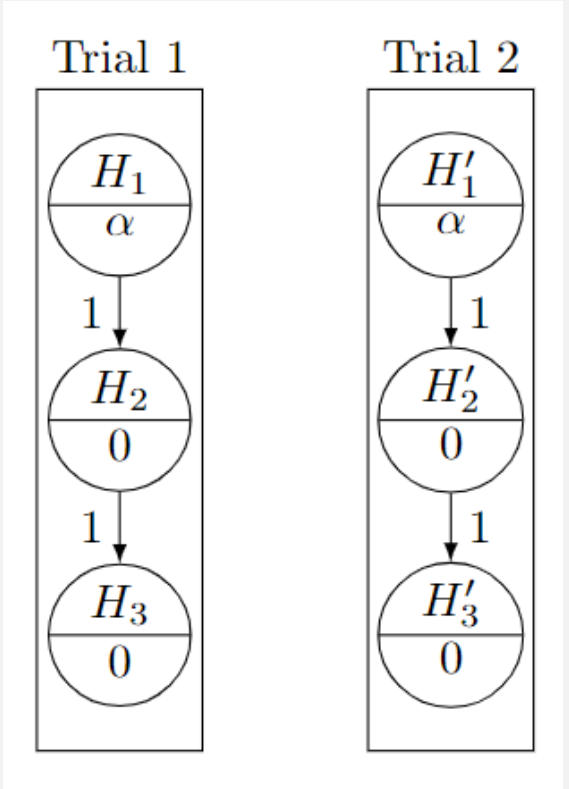
- "Requirement" for two positive confirmatory trials
  - FDA (1998) guideline
  - many examples of diseases under the two-trials convention
  - "replication", "independent substantiation"
- Single trial approvals generally limited to "mortality or irreversible morbidity" settings
  - "statistically very persuasive", "very low p-value"

Hypothetical example

- Two confirmatory trials following the two-trials convention
  - Identically designed and conducted concurrently
  - Primary endpoint E_1
  - Secondary endpoints E_2 and E_3
- Different sample sizes are needed to achieve a certain power (e.g., 80%) for different endpoints
  - E.g., E_2 may require a sample size twice as large as E_1 and E_3 require
- This unbalanced requirements of resources in a single study are amplified under the two-trials convention

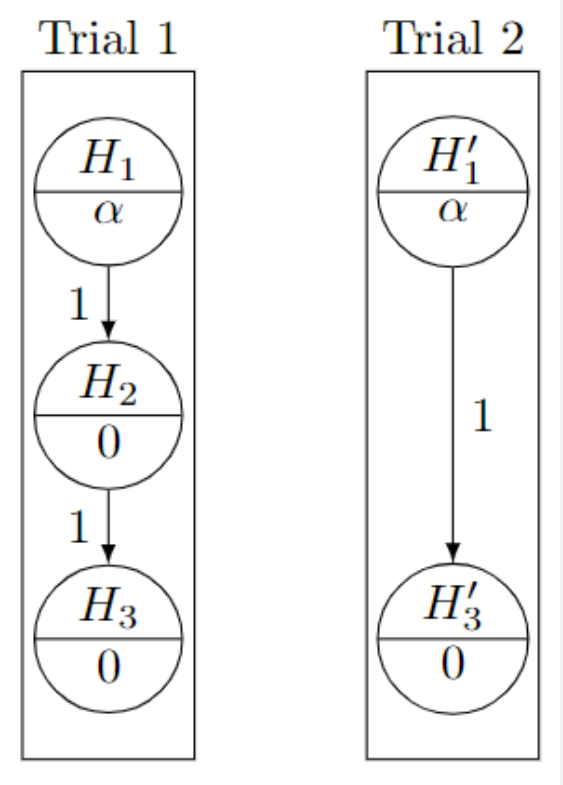
Hypothetical example: Strategy (a)

- Fixed sequence (hierarchical) procedure is applied within each trial
  - FWER controlled at level α within each trial
- E_2 may be underpowered



Hypothetical example: Strategy (b)

- Fixed sequence (hierarchical) procedure is applied within each trial
  - FWER controlled at level α within each trial
- E_2 only tested in Trial 1 with adequate power
  - Trial 1 has a much larger sample size than Trial 2
- No evidence of replicability for E_2
- Statistically inefficient because potentially available data for E_2 from Trial 2 are ignored



Combined analysis to address resource imbalance

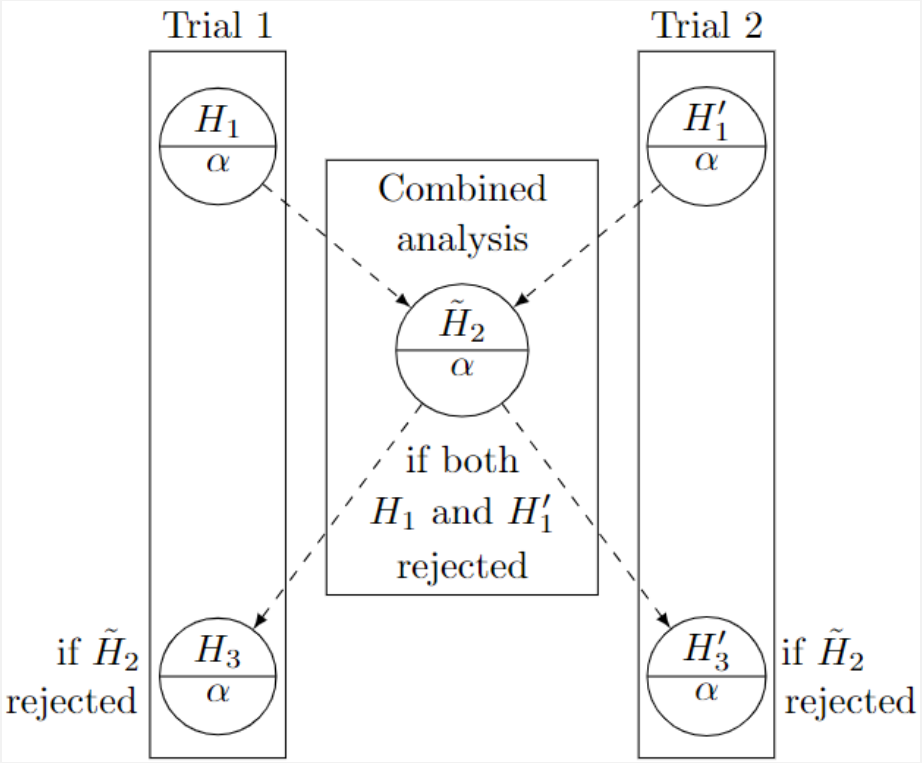
- One solution is to combine data from the two trials for E_2 without doubling the sample size of each trial
- Combining data from two trials increases statistical efficiency
- Different ways to combine
  - Naive pooling
  - Meta-analytic approach using 'trial' as a stratification factor
- "Combinability" needs careful examination to avoid systematic bias/difference

What approaches could be considered for managing multiplicity when data on an endpoint from two or more trials were planned to be pooled, and each trial had multiple endpoints managed?

LaVange (2019)

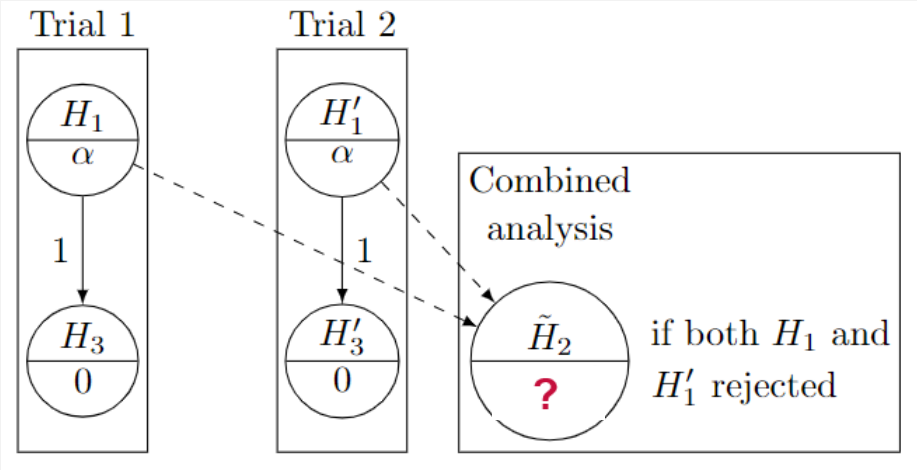
Hypothetical example: Strategy (c)

- Testing E_2 using the combined data
  - Denote the hypothesis as $\tilde{H}_2$
  - Tested only if both H_1 and H_1' are rejected
- Hierarchical structure from E_1 to E_2 then to E_3
  - FWER controlled at level α within each trial
- Testing of H_3 (or H_3') within a particular trial depends on data from the other trial through $\tilde{H}_2$
  - Statistical (homogeneity) and logistical (simultaneity) constraints (Bretz, Maurer, Xi, 2019)



Hypothetical example: Strategy (d)

- Testing $\tilde{H}_2$ using the combined data
  - Tested only if both H_1 and H_1' are rejected
- Hierarchical structure from E_1 to E_3 within each trial
  - FWER controlled at level α within each trial
- No hierarchical structure between E_2 and E_3
  - These endpoints can be tested independently of each other
- At which significance level shall we test $\tilde{H}_2$?



Submissionwise error rate (SWER)
Bretz and Xi (2019)

- SWER: Probability to make a false claim of success for any endpoint, while taking into account that a claim on any endpoint should be based on significant results from associated hypotheses in both trials
  - Depending on the strategy, this is achieved by significance independently in both trials, significance in at least one trial, or in the combined data (Vandemeulebroecke et al., 2023)
- Different focuses compared to FWER
  - FWER concerns false claims about null hypotheses within a trial
  - SWER concerns false claims about endpoints across trials

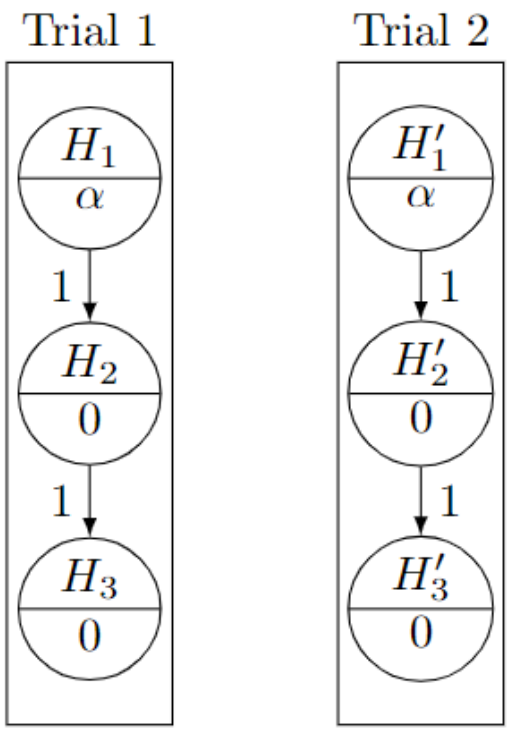
SWER for the hypothetical example

- For E_1, the Type I error to make a success claim is at most α^2 in all four strategies
  - Need to reject both H_1 and H_1' independently at level α
- For E_2 and E_3, regulatory requirements may be less clearly defined
  - Different disease areas may have different requirements
- Two scenarios to make a success claim on a secondary endpoint (e.g., E_2)
  1. Both hypotheses H_2 and H_2' should be rejected
  2. At least one of H_2 and H_2' should be rejected
  - Different implications on SWER

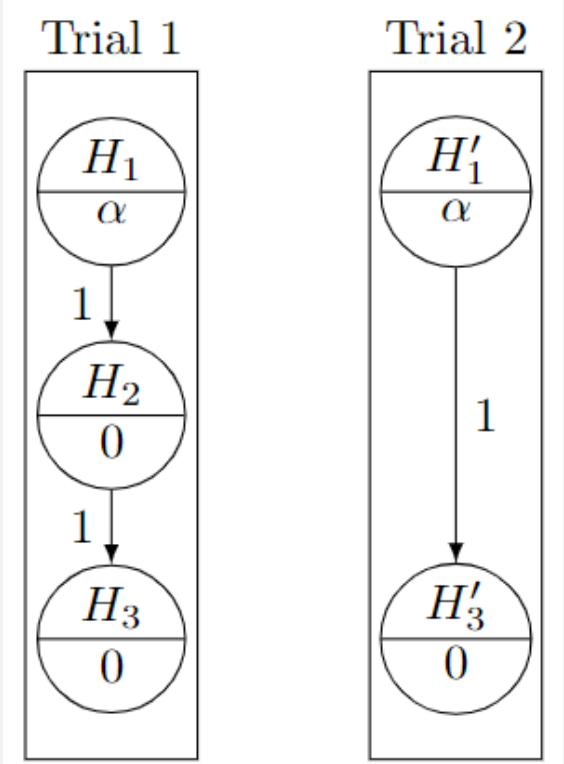
SWER of strategies (a) and (b)

- Type I error to make a false claim on E_2
  - $\leq \alpha^2$ (both H_2 and H_2')
  - $\leq 1-(1-\alpha)^2 = 2\alpha - \alpha^2$ (at least one)
- SWER $\leq \alpha^2$ or $2\alpha - \alpha^2$

Strategy (a)

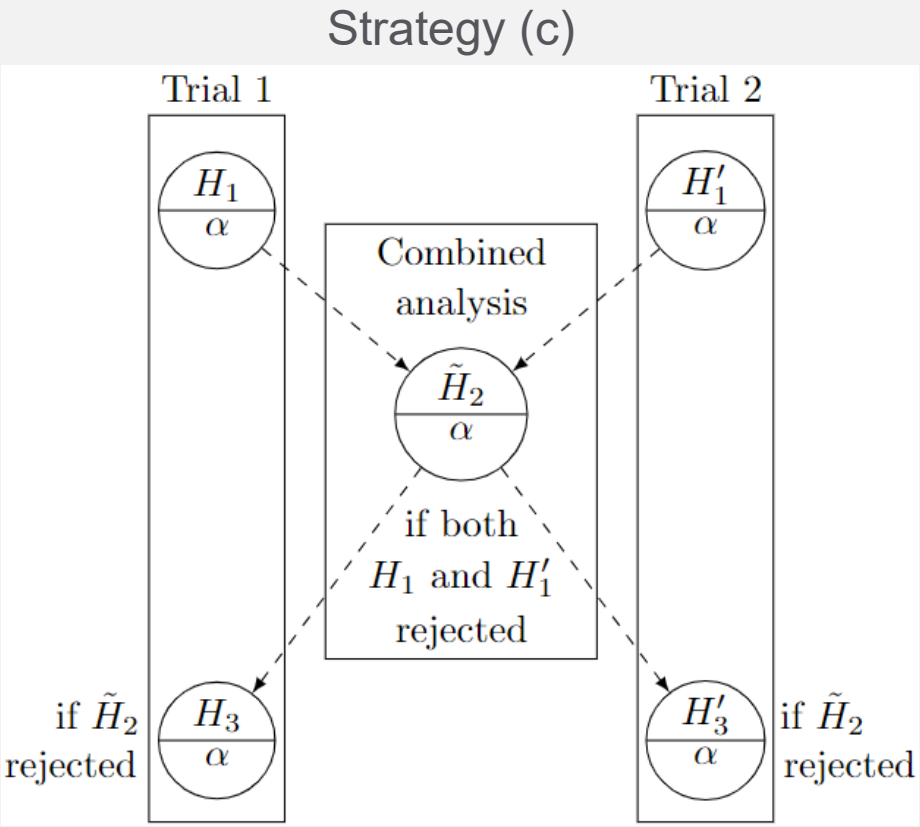


Strategy (b)



- Type I error to make a false claim on E_2
  - $\leq \alpha$
- Type I error to make a false claim on E_3
  - $\leq \alpha^2$ (both H_3 and H_3')
  - $\leq 1-(1-\alpha)^2 = 2\alpha - \alpha^2$ (at least one)
- SWER $\leq \alpha$ or $2\alpha - \alpha^2$

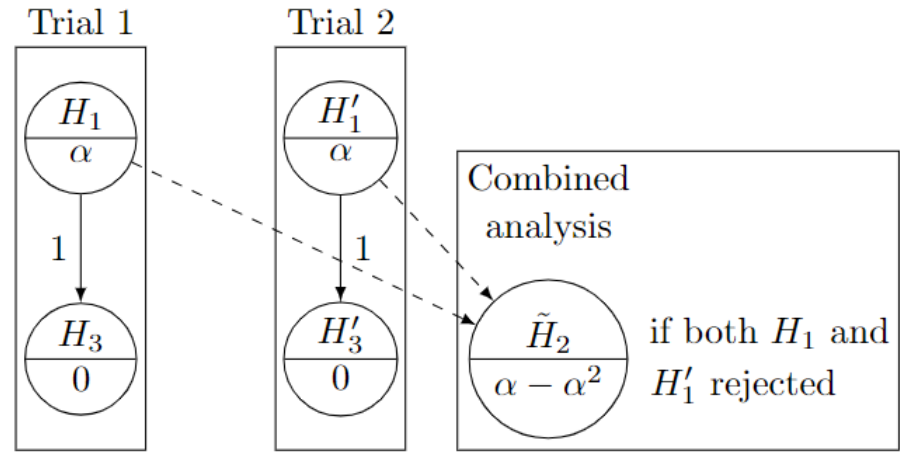
SWER of strategy (c)



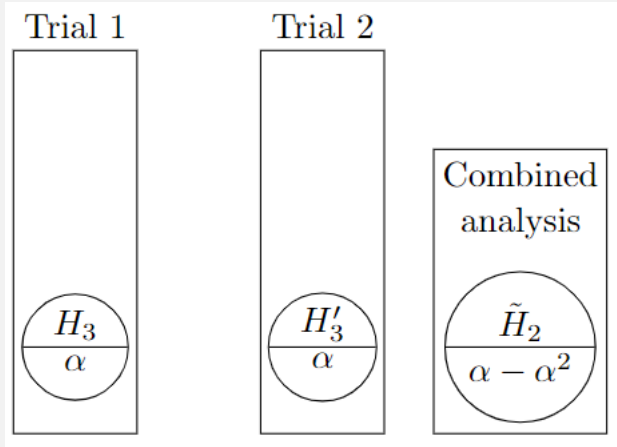
- Type I error to make a false claim on E_2
  - $\leq \alpha$
- Type I error to make a false claim on E_3
  - $\leq \alpha^2$ (both H_3 and H_3')
  - $\leq 1-(1-\alpha)^2 = 2\alpha - \alpha^2$ (at least one)
- SWER $\leq \alpha$ or $2\alpha - \alpha^2$

SWER of strategy (d)

Strategy (d)



After H_1 and H_1' rejected



- Type I error to make a false claim on E_2
  - $\leq \alpha - \alpha^2$
- Type I error to make a false claim on E_3
  - $\leq \alpha^2$ (both H_3 and H_3')
- SWER $\leq \alpha - \alpha^2 + \alpha^2 = \alpha$

- After H_1 and H_1' rejected
  - H_3 and H_3' can be tested each at level α
    - Both have to be rejected for a success claim on E_3
    - Type I error to make a false claim on $E_3 \leq \alpha^2$
  - $\tilde{H}_2$ can be tested at level $\alpha - \alpha^2$
    - Bonferroni split between E_3 (α^2) and E_2 $(\alpha - \alpha^2)$

Summary of error rates for all strategies

- FWER within each trial $\leq \alpha$
- SWER for the primary endpoint $\leq \alpha^2$
- Both hypotheses have to be rejected for a claim on a secondary endpoint

Strategy	Type I error on E_2 ($\leq$)	Type I error on E_3 ($\leq$)	SWER for all endpoints ($\leq$)
(a)	α^2	α^2	α^2
(b)	α	α^2	α
(c)	α	α^2	α
(d)	$\alpha - \alpha^2$	α^2	α

Summary of all strategies

- Strategy (a): conservative SWER control and inadequate power for E_2
- Strategy (b): SWER control at level α but no evidence of replicability for E_2 and inefficient
- Strategy (c): SWER control at level α but trial-level analysis for one trial depends on the other
- Strategy (d): SWER control at level α; sacrifices α^2 for E_2 but allows independent execution of trial-level analysis

Conclusions and other considerations

- When combined data are used for confirmatory testing, Type I error across trials needs to be considered
- Submissionwise error rate (SWER) is an error rate for claims on endpoints across trials
- Calculation of SWER depends on the trial-level FWER-controlling strategies as well as regulatory requirement for making claims on a secondary endpoint
- Application of SWER so far is focusing on identically designed and concurrently conducted trials
  - Null hypotheses for the same endpoint should be true or false simultaneously
- Extensions beyond this setting are of interest for more disease areas

References

- EMA (2017) Guideline on multiplicity issues in clinical trials
- FDA (2017) Multiple Endpoints in Clinical Trials
- FDA (1998) Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products
- LaVange (2019) Statistics at FDA: Reflections on the Past Six Years. Statistics in Biopharmaceutical Research, 11, 1-12
- Bretz, Maurer, Xi (2019) Replicability, Reproducibility, and Multiplicity in Drug Development. Chance, 32(4), 4-11
- Bretz, Xi (2019) Commentary on 'Statistics at FDA'. Statistics in Biopharmaceutical Research, 11, 20-25
- Vandemeulebroecke, Häring, Hua, Wei, Xi (2023) New strategies for confirmatory testing of secondary hypotheses on combined data from multiple trials. In revision.

THANK YOU

Summary of error rates for all strategies

- FWER within each trial $\leq \alpha$
- SWER for the primary endpoint $\leq \alpha^2$
- At least one hypothesis has to be rejected for a claim on a secondary endpoint

Strategy	Type I error on E_2 ($\leq$)	Type I error on E_3 ($\leq$)	SWER for all endpoints ($\leq$)
(a)	$2\alpha - \alpha^2$	$2\alpha - \alpha^2$	$2\alpha - \alpha^2$
(b)	α	$2\alpha - \alpha^2$	$2\alpha - \alpha^2$
(c)	α	$2\alpha - \alpha^2$	$2\alpha - \alpha^2$
(d)	$\alpha - \alpha^2$	$2\alpha - \alpha^2$	$3\alpha - 2\alpha^2$