

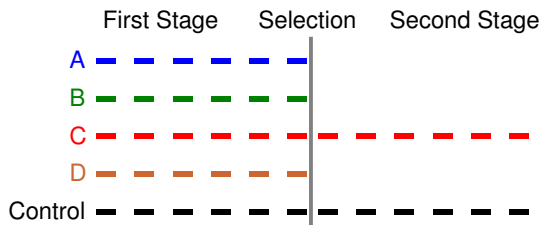
Type I error rate control in adaptive designs for confirmatory clinical trials with treatment selection at interim

Frank Bretz

Statistical Methodology, Novartis

Joint work with Martin Posch (Medical University of Vienna)
and Willi Maurer (Novartis)

Adaptive Designs KOL Lecture Series
February 12, 2010



- Second Stage Trial is planned using first stage data.
- Efficacy is demonstrated with data from **Stages 1 + 2**.
- **Control of the familywise error rate** is required.

EMA REFLECTION PAPER, 2007

The Simulation Approach for Type I error control

For predefined adaptation rules adjusted critical values are determined by simulation or numerical integration under the global null hypothesis.

- Select the “best” THALL ET AL. '88, STALLARD AND TODD, '03, ...
- Bayesian adaptive designs BERRY ET AL. '07, ...

Case Study

- Late development phase of a new drug for a life-threatening disease
- Primary objective: compare two doses of a new compound with standard-of-care for efficacy and safety
- Three parallel groups:
 - **C**: standard-of-care
 - **L**: standard-of-care + **low dose** of new compound
 - **H**: standard-of-care + **high dose** of new compound
- Primary endpoint: 28-day all-cause mortality
- $\theta_C, \theta_L, \theta_H \dots$ mortality rate in the control, the low dose and the high dose group
- One-sided tests of the null hypotheses

$$H_L : \theta_C = \theta_L, \quad H_H : \theta_C = \theta_H \quad \text{against}$$

$$H'_L : \theta_C > \theta_L, \quad H'_H : \theta_C > \theta_H$$

Adaptation Strategy

- Planned total sample size is $N_T = 2100$.
- The first 900 patients are randomized in a 1:1:1 ratio
- Interim analysis after 900 patients are observed
- Discontinue one of the treatments L or H or the entire study based on futility criteria.
- If **both experimental treatments** are continued: randomize the remaining $2100-900=1200$ patients in a **1:1:1 ratio** to the two experimental treatment arms or the control group.
- If **only one experimental treatment** is continued: randomize the remaining 1200 patients in a **1:1 ratio** to the remaining treatment arm or the control group.

Futility Stopping Criteria

- $r_{C,1}, r_{L,1}, r_{H,1}$... interim event rates in the three groups.
- f_1, f_2 pre-specified futility thresholds.
- Drop treatment L for futility if

$$\frac{r_{L,1}}{r_{C,1}} > 1 + f_1.$$

- Drop treatment H for futility if

$$\frac{r_{H,1}}{r_{C,1}} > 1 + f_1$$

or if L is continued and $r_{H,1} - r_{L,1} > f_2$.

The Test Procedure

- Y_i . . . test statistic to test $H_i, i = L, H$
- Reject H_i , if treatment i is continued to the final analysis and $Y_i \geq c$ for some critical value c
- Two choices for Y_i :
 - *max* test
 - *pooled* test
- Choose c to control the Familywise Error Rate (FWER)

The *max* Test

$Y_i = Z_i, i = L, H$, where

$$Z_i = \frac{r_C - r_i}{\sqrt{2\bar{r}_i(1 - \bar{r}_i)}} \sqrt{n_1 + n_2}, \quad i = L, H$$

- $r_L, r_H, r_C \dots$ events rates in the three groups based on all data available in the final analysis.
- $\bar{r}_i = (r_i + r_C)/2, i = L, H$
- $n_1, n_2 \dots$ first and second stage per group sample sizes.

The *pooled* Test

$$Y_i = \begin{cases} \min(Z_i, Z_p) & \text{if both treatments are selected} \\ Z_i & \text{if only one treatment is selected} \end{cases}$$

where Z_p compares the control with the pooled treatment groups:

$$Z_p = \frac{r_C - (r_L + r_H)/2}{\sqrt{2\bar{r}(1 - \bar{r})}} \sqrt{2(n_1 + n_2)}, \quad \bar{r} = (r_C + r_L + r_H)/3$$

Simulation Algorithm

The critical value c is determined by a simulation algorithm:

Step 1 Specify: Test statistics, $f_1, f_2, n_1, \theta = \theta_C = \theta_L = \theta_H$

Step 2 Generate first stage rates drawn from the $B(\theta, n_1)$ distribution. Apply the futility criteria.

Step 3 For the selected treatments

- determine n_2
- generate second stage rates
- compute Y_i .

Step 4 Set $Y_i = -\infty$ for dropped treatments.

Set $Y = \max_{i=L,H} Y_i$.

Step 5 Repeat Steps 2-3-4

Definition of the critical value:

$c = 100(1 - \alpha)$ th percentile of the pool of Y values.

Simulated critical boundaries

For nominal level $\alpha = 0.025$ using 10^6 simulation runs

The simulated critical values depend on the assumed true mortality rate in the control group and the futility thresholds.

Mortality Rate $\theta_C = \theta_L = \theta_H$	Futility Bounds		Max test	Pooled test
	$f_1(\%)$	$f_2(\%)$	c	c
0.1	20	10	2.22	1.96
0.1	-10	-1	2.13	2.08
0.2	20	10	2.21	1.94
0.2	-10	-1	2.11	2.07

- The mortality rate in the control group is unknown
- Possible approach: Use the **observed mortality rate** in the control group.
- Impact on Type I error rate?

Weak Control of the Familywise Error Rate

- Assuming the true mortality rate of the control group is known, the simulation based test under the assumption $\theta_L = \theta_H = \theta_C$ controls the FWER up to Monte Carlo error.
- For **other mortality configurations the FWER could still be larger than α** :

Example:

- Let $\theta_C = 0.2$, $f_1 = 20\%$, $f_2 = 10\%$ and consider the *pooled* test.
- Under the scenario $\theta_L = \theta_C = 0.2$, $\theta_H = 0.1$ the type I error to reject H_L is **2.6%**.

The Impact of Unscheduled Adaptations

Unscheduled dropping of treatments based on safety endpoints may inflate the significance level.

- Assume that in addition to the pre-planned interim analysis after 900 patients a safety analysis after 1500 patients is performed.
- Worst case example: At the safety analysis, we drop a treatment if this increases the conditional probability to reject a null hypothesis given the data observed so far.

Mortality Rate $\theta_C = \theta_L = \theta_H$	Futility Bounds		Type I Error	
	$f_1(\%)$	$f_2(\%)$	<i>Max test</i>	<i>Pooled test</i>
0.1	20	10	2.59	4.31
0.2	20	10	2.59	4.52

Limitations of Simulation Based Procedures

The distribution of the final test statistics under the null hypotheses may depend on **unknown parameters** e.g.,

- the unknown configuration of true and false null hypotheses and the effect sizes of the false null hypothesis.
- the absolute treatment effects, if the adaptation rule depends on interim estimates of absolute treatment effects.
- the joint distribution of data that is correlated with the primary endpoint and used in the adaptations.
- the joint distribution of covariates if the test statistics is adjusted for covariates.

When do Simulation Based Procedures Work?

- If a point null hypothesis is tested and the distribution of the final test statistics is fully specified under the null hypothesis
- If the **least favorable configuration** (LFC) across all nuisance parameters can be identified
 - The Type I error rate at the LFC gives an upper bound for the Type I error rate
 - The resulting test will be strictly conservative.
 - The LFC maybe difficult to identify.

The Adaptive Bonferroni Holm Test

Based on conditional error rates (MÜLLER AND SCHÄFER, 2001) **adaptive level α tests for the intersection hypothesis and the elementary hypotheses** are defined. Reject an elementary hypothesis at multiple level α if both, the intersection and the elementary test reject.

- The conditional error rate approach requires the computation of the conditional error rate for the test of the intersection hypothesis which involves the conditional joint distribution of Z_L, Z_H .
- We propose an alternative test that is based on the **sum of marginal conditional error rates** of the elementary tests at level $\alpha/2$.

POSCH, MAURER, BRETZ, 2009

The Bonferroni Holm Test as Closed Test

Non-Adaptive Fixed Sample Test

$Z_L, Z_H \dots$ Z-statistics for H_L, H_H .

$z_q \dots$ standard normal q-quantiles

- 1 Global Null hypothesis $H_L \cap H_H$

reject if $Z_L \geq z_{1-\alpha/2}$ or $Z_H \geq z_{1-\alpha/2}$

If $H_L \cap H_H$ is rejected:

- 2 Elementary Hypotheses H_L, H_H

Reject H_L if $Z_L \geq z_{1-\alpha}$,

Reject H_H if $Z_H \geq z_{1-\alpha}$

Adaptive tests:

- 1 Global null hypothesis: Adaptive Bonferroni Test
- 2 Elementary null hypotheses: Conditional Rejection Probability Test

Adaptive Test for Elementary Hypotheses

E.g., to test H_H

- if no treatment was dropped, perform the fixed sample test (i.e. reject if $Z_H \geq z_{1-\alpha}$).
- if only treatment H is continued: Let

$$a_H = P(Z_H \geq z_{1-\alpha} | Z_{H,t-})$$

Reject H_H if

$$Z_{H,t+} \geq z_{1-a_H}$$

Adaptive Bonferroni Test for $H_L \cap H_H$

- $(Z_{L,t-}, Z_{H,t-}), \dots$ Z statistics of the observations collected before the adaptation
- $(Z_{A,t+}, Z_{B,t+}), \dots$ Z statistics of the observations collected after the adaptation

The Adaptive Bonferroni Test

- if no treatment was dropped, perform the fixed sample Bonferroni test (i.e. reject if $\max(Z_L, Z_H) \geq z_{1-\alpha/2}$).
- if (say) only treatment H is continued

$$a = P(Z_L \geq z_{1-\alpha/2} | Z_{L,t-}) + P(Z_H \geq z_{1-\alpha/2} | Z_{H,t-})$$

Reject $H_L \cap H_H$ if

$$Z_{H,t+} \geq z_{1-a}$$

Properties of the adaptive Bonferroni Holm Test

The adaptive Bonferroni Holm Test controls the FWER in the strong sense without pre-specification of the selection rule:

- selection may depend on all interim data including safety and secondary endpoints
- selection may depend on external data
- no pre-specification of the analysis time required (continuous monitoring)
- no binding futility stopping rule
- sample sizes and allocation ratios may be adapted

Power Comparison of the Procedures

- After 900 of 2100 patients an interim analysis is performed.
- Control group mortality rate is $\theta_C = 0.2$.
- Futility parameters $f_1 = 10\%$, $f_2 = 1\%$

Rates		<i>max</i> test (%)			<i>pooled</i> test (%)			Adaptive BH (%)		
θ_L	θ_H	P_{any}	P_L	P_H	P_{any}	P_L	P_H	P_{any}	P_L	P_H
0.2	0.14	82	1.5	82	77	1.7	76	80	2.1	80
0.14	0.14	92	82	32	92	83	32	92	84	31

P_{any} ... power to reject any elementary hypothesis

$P_L (P_H)$... power to reject $P_L (P_H)$, respectively.

Conclusion

- Simulation of the FWER under the global null hypothesis in adaptive multi-armed clinical trials is in general insufficient to control the FWER.
- In settings with nuisance parameters the FWER control may depend on specific assumptions on the nuisance parameters.
- Robust alternatives are tests based on conditional error rates and combination tests.
- However, simulations are valuable to assess the power of adaptive tests.

References I



F. Bretz, F. König, W. Brannath, E. Glimm, and M. Posch.
Adaptive designs for confirmatory clinical trials.
Statistics in Medicine, 28:1181 – 1217, 2003.



EMA.

Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design.

EMA Doc. Ref. CHMP/EWP/2459/02, 2007.

available at

<http://www.ema.europa.eu/pdfs/human/ewp/245902enadopted.pdf>.



H. H. Müller and H. Schäfer.

A general statistical principle for changing a design any time during the course of a trial.

Statistics in Medicine, 23:2497–2508, 2004.

References II



M. Posch, W. Maurer, and F. Bretz.

Type I error rate control in adaptive designs for confirmatory clinical trials with treatment selection at interim.

Pharmaceutical Statistics, 2010.

to appear.



N. Stallard and S. Todd.

Sequential designs for phase III clinical trials incorporating treatment selection.

Statistics in Medicine, 22:689–703, 2003.



P. F. Thall, R. Simon, and S. S. Ellenberg.

Two-stage selection and testing designs for comparative clinical trials.

Biometrika, 75:303–310, 1988.