Confirmatory Adaptive Designs

Franz König and Martin Posch

Medical University of Vienna, Austria
Center for Medical Statistics, Informatics and Intelligent Systems
Franz.Koenig@meduniwien.ac.at
www.meduniwien.ac.at/user/franz.koenig

Statposium17 , Paris, 9 Nov 2017

Content

- Introduction
- European Regulatory Requirements and Experience with Adaptive Designs
- Adaptive Two-Stage Designs
- Adaptive Clinical Trials with Treatment Selection
- Conclusions

Frequentist vs Flexible (Adaptive) Trials

Classical frequentist trials

- details of design and analysis must be prefixed in advance (population, treatments, doses, main and secondary outcome variable(s), analysis strategy, sample sizes,...)
- lack of flexibility to react to information from inside or outside the trial

Medical Statistician:

one who will not accept that Columbus discovered America .... because he said he was looking for India in the trial* plan. (* A cross over trial).

SENN, 1997, P 58

Frequentist vs Flexible (Adaptive) Trials

Classical frequentist trials

- details of design and analysis must be prefixed in advance (population, treatments, doses, main and secondary outcome variable(s), analysis strategy, sample sizes,...)
- lack of flexibility to react to information from inside or outside the trial

Medical Statistician:

one who will not accept that Columbus discovered America ....
because he said he was looking for India in the trial* plan.
(* A cross over trial).

Senn, 1997, p 58

Frequentist vs Flexible (Adaptive) Trials

Classical frequentist trials

- details of design and analysis must be prefixed in advance (population, treatments, doses, main and secondary outcome variable(s), analysis strategy, sample sizes,...)
- lack of flexibility to react to information from inside or outside the trial

Flexible (adaptive) design

- allow for mid-trial design modifications based on all internal and external information gathered at interim analyses without compromising the type I error rate
- To control the type I error rate, the design modifications need **not** be specified in advance.

Pre-Specified Adaptive versus Fully Adaptive Designs

Pre-specified Adaptive Designs

The adaptation rule must be completely pre-specified.

- Group sequential designs
- Blinded sample size reassessment
- Rules for sample size reassessment & treatment selection
- (Bayesian) response adaptive randomization

Schmitz '93, Shun '01, Stallard & Todd '03, Friede & Kieser, Hu & Rosenberger '06, Berry et al. '10, ...

Fully Adaptive Designs

The adaptation rule needs not to be (completely) specified

- Sample size reassessment
- Treatment arm selection
- Population enrichment
- Endpoint selection

Bauer '89,Bauer & Köhne '94, Proschan & Hunsberger'95, Lehmacher & Wassmer '99, Cui et al. '99, Müller & Schäfer '04, Mehta & Pocock '11, ...

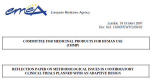
Some History of Adaptive Designs

28 years ago	Bauer: “Multistage Testing with Adaptive Designs”
22 years ago	Proschan & Hunsberger: “Designed Extension of Studies Based on Conditional Power”
10 years ago	EMA Reflection Paper
7 years ago	FDA Draft Guidance (Drugs and Biologics)
2 years ago	FDA Draft Guidance (Devices, CDRH, CBER) - finalized last year

Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls

P. Bauer, F. Bretz, V. Dragalin, F. Koenig, and G. Wassmer.
Featured Article in Statistics in Medicine 35, 325-347, 2016.
http://dx.doi.org/10.1002/sim.6472 (Open Access)
With invited discussion by Hung, Wang and Lawrence; Mehta and Liu; Vollmar; Maurer

What is an Adaptive Design?



European Medicines Agency

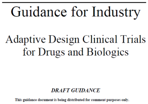
London, 18 October 2007
Doc. Ref. CHMP/EWP/2459/02

COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE
(CHMP)

REFLECTION PAPER ON METHODOLOGICAL ISSUES IN CONFIRMATORY
CLINICAL TRIALS PLANNED WITH AN ADAPTIVE DESIGN

A study design is called "adaptive" if statistical methodology allows the modification of a design element (e.g. sample-size, randomization ratio, number of treatment arms) at an interim analysis with full control of the type I error.

EMA 2007



Guidance for Industry

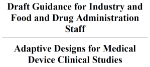
Adaptive Design Clinical Trials
for Drugs and Biologics

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

A study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study.

CBER, CDER FDA 2010



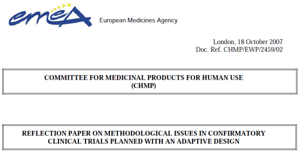
Draft Guidance for Industry and Food and Drug Administration Staff

Adaptive Designs for Medical Device Clinical Studies

A clinical study design that allows for prospectively planned modifications based on accumulating study data without undermining the trial's integrity and validity.

CBER, CDRH, FDA, 2016

Minimal Requirements for Confirmatory Adaptive Trials



European Medicines Agency

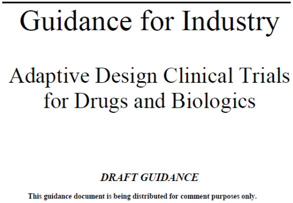
London, 18 October 2007
Doc. Ref. CHMP/EWP/2459/02

COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE
(CHMP)

REFLECTION PAPER ON METHODOLOGICAL ISSUES IN CONFIRMATORY CLINICAL TRIALS PLANNED WITH AN ADAPTIVE DESIGN

"Using an adaptive design implies that the statistical methods control the pre-specified type I error, that correct estimates and confidence intervals for the treatment effect are available, and that methods for the assessment of homogeneity of results from different stages are pre-planned."

EMA reflection paper (2007)



Guidance for Industry

Adaptive Design Clinical Trials for Drugs and Biologics

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

"The chief concerns with these designs are control of the study-wide Type I error rate, minimization of the impact of any adaptation-associated statistical (see section VII.B) or operational bias on the estimates of treatment effects, and the interpretability of trial results."

FDA Draft Guidance (2010)

Where are we now?

- Do sponsors consider adaptive designs in the development plans?
- Which type of adaptive designs are proposed?
- What are frequently identified problems?
- Which issues are still controversial?

European Regulatory Experience with Adaptive Designs

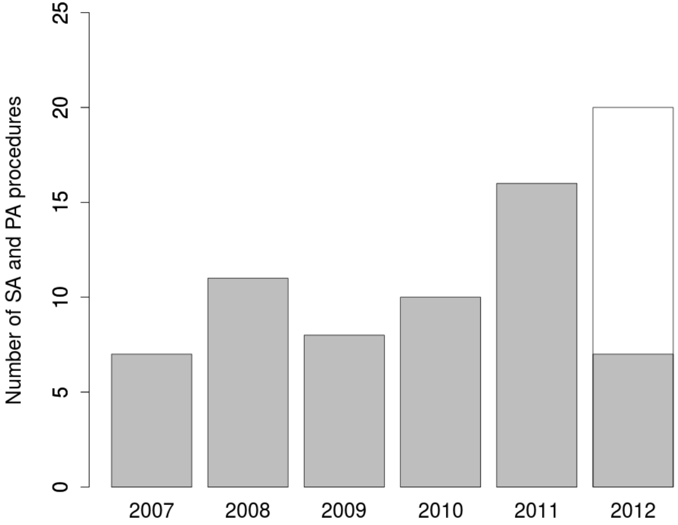
Adaptive Designs in EMA Scientific Advice Procedures

Survey of Scientific Advice/Protocol Assistance Procedures (01/2007- 05/2012)

- Scientific Advice/Protocol Assistance procedures of EMA Scientific Advice Working Party
- Search for Scientific Advice Letters containing terms such as, *adaptive design, flexible design, adaptive interim analysis, ...*
- Exclusion of phase I trials
- 59 procedures identified that contained questions on clinical trials with an adaptive designs
- May not include all procedures addressing adaptive designs (e.g., if sponsors use different terminology).

- A. Elsäßer, J. Regnstrom, T. Vetter, F. Koenig, R. Hemmings, M. Greco, M. Papaluca-Amati, and M. Posch.
  Adaptive clinical trial designs for European marketing authorization: a survey of scientific advice letters from the European Medicines Agency.
  Trials 15, 383, (2014) (Open Access)
  http://dx.doi.org/10.1186/1745-6215-15-383

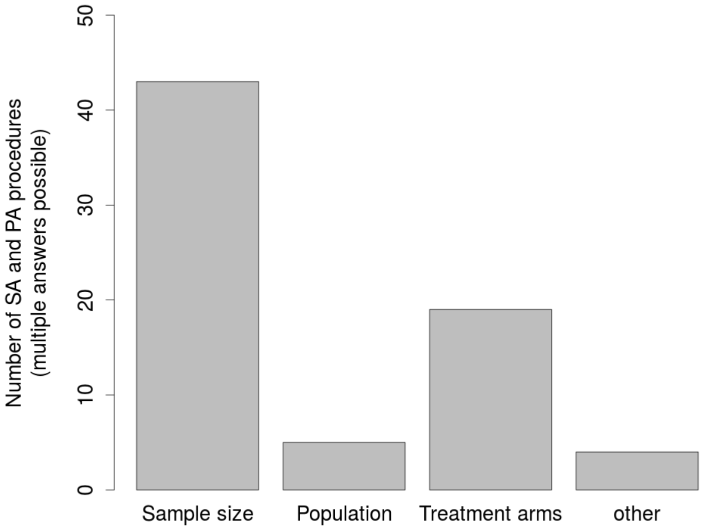
Number of Procedures per Year (n=59)



Types of Clinical Trials

- About 60% rare disease (prevalence of $< 5/10,000$), 1/3 applied for orphan designation
- Indications: About 50% oncology
- About 90% phase III or seamless II/III studies. Additionally, phase II or pediatric studies.
- $\approx 75\%$ proposed as single pivotal trial
- Number of interim analyses:
  $1 \approx 70\%$, $2 \approx 20\%$, $> 2 \approx 5\%$.
- Primary Endpoint: time to event ($\approx 50\%$), binary ($\approx 30\%$), continuous ($\approx 20\%$).

Types of Adaptations (n=59)



What if adaptations and multiplicity issues were ignored ...

- Comparing k treatments against control:
- What is the most extreme inflation of the type I error rate using naively a conventional fixed sample size test at level α at the end of a study if we allow sample size reassessment (and selection) at interim?

Maximum type 1 error inflation:

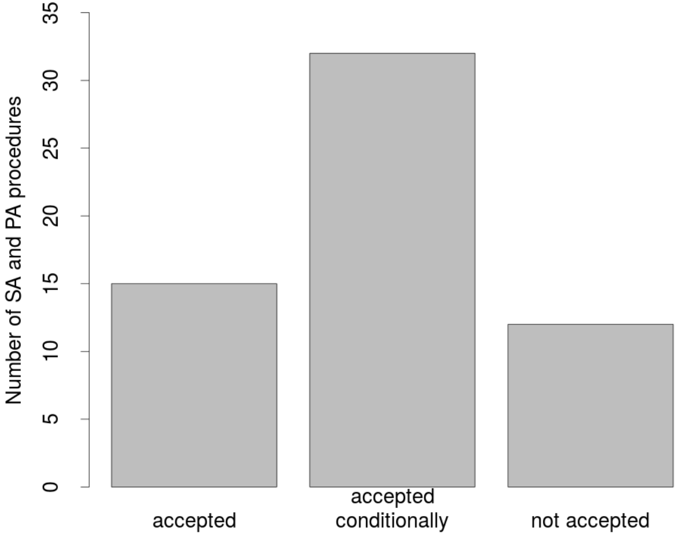
nominal α	$k = 1$ balanced[1]	$k = 1$ unbalanced[2]	$k = 2$ unbalanced[3]
0.05	0.115	0.187	0.289
0.025	0.062	0.106	0.170
0.01	0.027	0.049	0.080

[1] Proschan and Hunsberger 1995

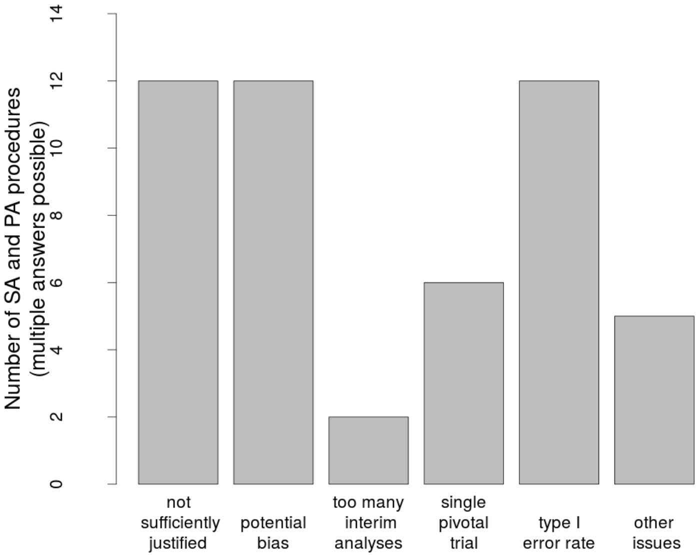
[2] Graf and Bauer 2011

[3] Graf, Bauer and Koenig 2014

Overall Regulatory Response (n=59)



Issues Raised (Years 2009-2012, n=41)



Further Issues Identified in Adaptive Clinical Trial Proposals

- Insufficient sample size for subgroup analyses
- The option for adaptations is not prospectively planned (Post-hoc adaptive trial)
- Issues due to interim analyses (as in group sequential designs)
  - Overrunning
  - Feasibility of interim analyses because of large recruitment rates or delayed endpoints.
  - "Maturity" of survival data in interim analyses
  - Leakage of interim information leading to "silent adaptations", not captured by the statistical methodology. They may result in issues for the interpretability of results.

Estimation

Usually, standard estimators, not accounting for the adaptations, are proposed.

- In general, point estimates of adaptive designs will be biased.
- For specific scenarios, the bias can be quantified by simulations.
- The size of the bias will vary, depending on
  - the type of adaptation and specific adaptation rule,
  - the actual treatment effect(s)
  - nuisance parameters
- Adjusted confidence intervals

Interestingly:

The bias can be smaller for adaptive designs than for fixed sample designs, e.g. multi-arm designs. See "Selection and Bias - Two Hostile Brothers" Bauer et al. 2010

Estimation

Usually, standard estimators, not accounting for the adaptations, are proposed.

- In general, point estimates of adaptive designs will be biased.
- For specific scenarios, the bias can be quantified by simulations.
- The size of the bias will vary, depending on
  - the type of adaptation and specific adaptation rule,
  - the actual treatment effect(s)
  - nuisance parameters
- Adjusted confidence intervals

Interestingly:

The bias can be smaller for adaptive designs than for fixed sample designs, e.g. multi-arm designs. See "Selection and Bias - Two Hostile Brothers" BAUER ET AL. 2010

Type I Error Control

Several approaches seen:

- Adaptive testing procedures (conditional error rate, combination tests)
- “Promising zone” approach.
- Standard analysis not accounting for adaptations.
- Simulation methods to demonstrate type I error control

Type I Error Control

Several approaches seen:

- Adaptive testing procedures (conditional error rate, combination tests)
- "Promising zone" approach.
- Standard analysis not accounting for adaptations.
- Simulation methods to demonstrate type I error control

Adaptive Two-Stage Designs based on Combination Tests

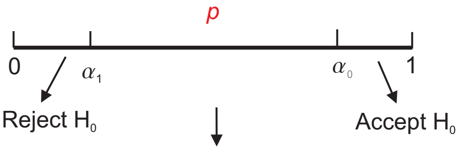
Adaptive Two-Stage Combination Tests

- A trial is performed in two stages
- In an interim analysis the trial may be
  - stopped for futility or efficacy or
  - continued and possibly adapted (sample size, test statistics)
- Adaptation of the design of second stage
  - adaptations depend on all (unblinded) interim data including secondary and safety endpoints.
  - the adaptation rule is not (completely) preplanned.

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

Adaptive Combination Tests (Bauer '89, Bauer & Köhne '94)

First Stage



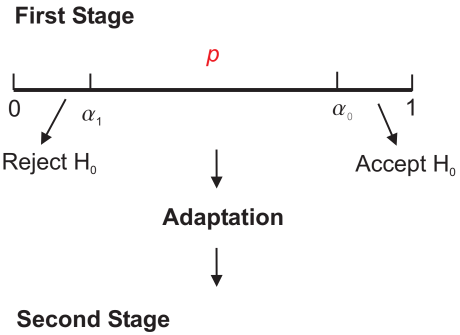
p

0 α_1 α_0 1

Reject H_0

Accept H_0

Adaptive Combination Tests (Bauer '89, Bauer & Köhne '94)



First Stage

p

0 α_1 α_0 1

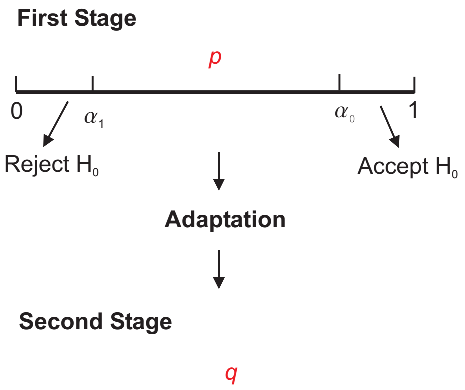
Reject H_0

Accept H_0

Adaptation

Second Stage

Adaptive Combination Tests (Bauer '89, Bauer & Köhne '94)



First Stage

p

0 α_1 α_0 1

Reject H_0

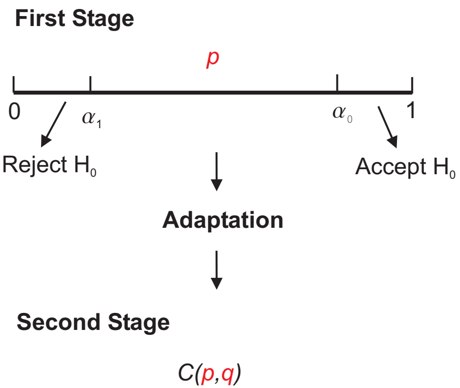
Accept H_0

Adaptation

Second Stage

q

Adaptive Combination Tests (Bauer ’89, Bauer & Köhne ’94)



First Stage

p

0

α_1

α_0

1

Reject H_0

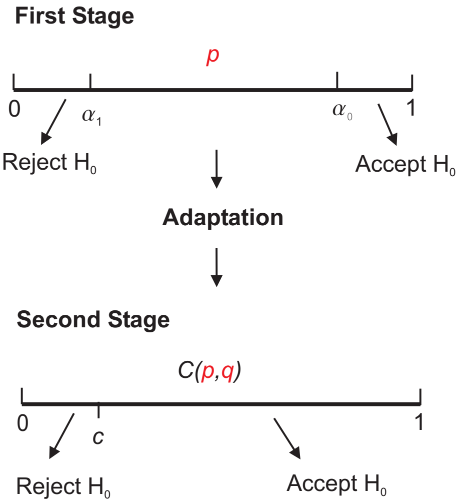
Accept H_0

Adaptation

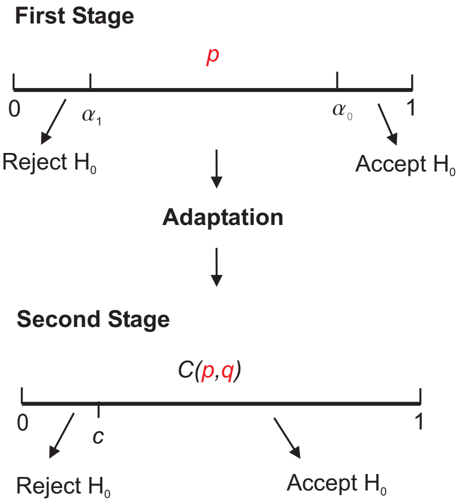
Second Stage

$C(p,q)$

Adaptive Combination Tests (Bauer '89, Bauer & Köhne '94)

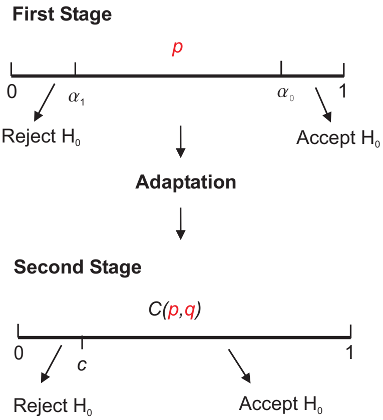


Adaptive Combination Tests (Bauer '89, Bauer & Köhne '94)



Stopping boundaries and combination functions have to be laid down a priori!

Adaptive Combination Tests (Bauer '89, Bauer & Köhne '94)



Planning:

- Fix design (only) for Stage 1
- Fix combination function $C(p,q)$ and critical value c e.g. $C(p,q) = p \cdot q$

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 form Stage 2 data.
- Reject H_0 iff $C(p,q) \leq c$.

Type I error control and combination functions

Type I error control

Type I error rate $\leq \alpha$ if we choose critical value c such that

$$P[p \leq \alpha or C(p,q) \leq c] = \alpha$$

for independent and uniformly distributed p-values p and q.

- *Fisher product test:* $C(p,q) = p \cdot q$

  (Bauer 1989, Bauer & Köhne, 1994)

- *Weighted inverse normal method:*
  $C(p,q) = \Phi(w_1 \Phi^{-1}(p) + w_2 \Phi^{-1}(q))$

  (Lehmacher & Wassmer, 1999)

(Remark: Can use critical values of a group sequential trial with interim information fraction w_1).

Clue of the Adaptive Test

- Do not pool the data of the stages, combine the stage-wise p-values.
- Then the distribution of the combination function under the null does not depend on design modifications
- Hence the adaptive test is still a test at the level α for the modified design!
- Applicable also for multiple looks, multiple hypotheses, ...
- Adaptations can depend on all (unblinded) interim data including secondary and safety endpoints.
- For a control of the type I error rate, one need not pre-specify how the Stage 1 data determine the design of Stage 2.

Numerical Example Product Test

One sample test at level $\alpha = 0.025$ for the mean of (pre-planned) 40 normally distributed observations to test the hypotheses

$$H_0 : \mu = 0 \quad \text{against} \quad H' : \mu > 0$$

- Product test $\alpha_1 = 0.01, \alpha_0 = 1, c = 0.00326$.
- First stage sample size $n_1 = 20$ observations.
- First stage data: mean 3.7, sd 10.9, $p = 0.0727$ (t-test).
- Interim decision: $p > \alpha_1$ continue.
- Second stage: Choose sample size of $n^{(2)} = 30$ observations.
- Second stage data: mean 3.2, sd 9.5, $q = 0.0376$ (t-test).
- Test decision: $p \cdot q = 0.00273 < c$ reject H_0.

Numerical Example Product Test

One sample test at level $\alpha = 0.025$ for the mean of (pre-planned) 40 normally distributed observations to test the hypotheses

$$H_0 : \mu = 0 \quad \text{against} \quad H' : \mu > 0$$

- Product test $\alpha_1 = 0.01, \alpha_0 = 1, c = 0.00326$.
- First stage sample size $n_1 = 20$ observations.
- First stage data: mean 3.7, sd 10.9, $p = 0.0727$ (t-test).
- Interim decision: $p > \alpha_1$ continue.
- Second stage: Choose sample size of $n^{(2)} = 30$ observations.
- Second stage data: mean 3.2, sd 9.5, $q = 0.0376$ (t-test).
- Test decision: $p \cdot q = 0.00273 < c$ reject H_0.

Numerical Example Product Test

One sample test at level $\alpha = 0.025$ for the mean of (pre-planned) 40 normally distributed observations to test the hypotheses

$$H_0 : \mu = 0 \quad \text{against} \quad H' : \mu > 0$$

- Product test $\alpha_1 = 0.01, \alpha_0 = 1, c = 0.00326$.
- First stage sample size $n_1 = 20$ observations.
- First stage data: mean 3.7, sd 10.9, $p = 0.0727$ (t-test).
- Interim decision: $p > \alpha_1$ continue.
- Second stage: Choose sample size of $n^{(2)} = 30$ observations.
- Second stage data: mean 3.2, sd 9.5, $q = 0.0376$ (t-test).
- Test decision: $p \cdot q = 0.00273 < c$ reject H_0.

Numerical Example Product Test

One sample test at level $\alpha = 0.025$ for the mean of (pre-planned) 40 normally distributed observations to test the hypotheses

$$H_0 : \mu = 0 \quad \text{against} \quad H' : \mu > 0$$

- Product test $\alpha_1 = 0.01, \alpha_0 = 1, c = 0.00326$.
- First stage sample size $n_1 = 20$ observations.
- First stage data: mean 3.7, sd 10.9, $p = 0.0727$ (t-test).
- Interim decision: $p > \alpha_1$ continue.
- Second stage: Choose sample size of $n^{(2)} = 30$ observations.
- Second stage data: mean 3.2, sd 9.5, $q = 0.0376$ (t-test).
- Test decision: $p \cdot q = 0.00273 < c$ reject H_0.

Numerical Example Product Test

One sample test at level $\alpha = 0.025$ for the mean of (pre-planned) 40 normally distributed observations to test the hypotheses

$$H_0 : \mu = 0 \quad \text{against} \quad H' : \mu > 0$$

- Product test $\alpha_1 = 0.01, \alpha_0 = 1, c = 0.00326$.
- First stage sample size $n_1 = 20$ observations.
- First stage data: mean 3.7, sd 10.9, $p = 0.0727$ (t-test).
- Interim decision: $p > \alpha_1$ continue.
- Second stage: Choose sample size of $n^{(2)} = 30$ observations.
- Second stage data: mean 3.2, sd 9.5, $q = 0.0376$ (t-test).
- Test decision: $p \cdot q = 0.00273 < c$ reject H_0.

Numerical Example Product Test

One sample test at level $\alpha = 0.025$ for the mean of (pre-planned) 40 normally distributed observations to test the hypotheses

$$H_0 : \mu = 0 \quad \text{against} \quad H' : \mu > 0$$

- Product test $\alpha_1 = 0.01, \alpha_0 = 1, c = 0.00326$.
- First stage sample size $n_1 = 20$ observations.
- First stage data: mean 3.7, sd 10.9, $p = 0.0727$ (t-test).
- Interim decision: $p > \alpha_1$ continue.
- Second stage: Choose sample size of $n^{(2)} = 30$ observations.
- Second stage data: mean 3.2, sd 9.5, $q = 0.0376$ (t-test).
- Test decision: $p \cdot q = 0.00273 < c$ reject H_0.

Numerical Example Product Test

One sample test at level $\alpha = 0.025$ for the mean of (pre-planned) 40 normally distributed observations to test the hypotheses

$$H_0 : \mu = 0 \quad \text{against} \quad H' : \mu > 0$$

- Product test $\alpha_1 = 0.01, \alpha_0 = 1, c = 0.00326$.
- First stage sample size $n_1 = 20$ observations.
- First stage data: mean 3.7, sd 10.9, $p = 0.0727$ (t-test).
- Interim decision: $p > \alpha_1$ continue.
- Second stage: Choose sample size of $n^{(2)} = 30$ observations.
- Second stage data: mean 3.2, sd 9.5, $q = 0.0376$ (t-test).
- Test decision: $p \cdot q = 0.00273 < c$ reject H_0.

Choice of α_1 and n_1

No early rejection $\alpha_1 = 0$

Interim analysis for adaptations and/or futility stopping, only.

- Is there enough information to perform adaptations? What is the precision of effect estimates?

Positive early rejection boundary ($\alpha_1 > 0$)

- What is the probability of an early rejection under different alternative hypotheses?
- Is the interim sample size large enough to obtain sufficient safety data and data on secondary endpoints?
- Is it ethical to continue the trial although a large treatment effect has been observed?
- Is it likely that the interim results after an early rejection are convincing enough to stop the trial?

Choice of α_1 and n_1

No early rejection $\alpha_1 = 0$

Interim analysis for adaptations and/or futility stopping, only.

- Is there enough information to perform adaptations? What is the precision of effect estimates?

Positive early rejection boundary ($\alpha_1 > 0$)

- What is the probability of an early rejection under different alternative hypotheses?
- Is the interim sample size large enough to obtain sufficient safety data and data on secondary endpoints?
- Is it ethical to continue the trial although a large treatment effect has been observed?
- Is it likely that the interim results after an early rejection are convincing enough to stop the trial?

The Issue of Stopping for Futility (Choice of α_0)

Should futility stopping be considered when computing the type I error rate?

- If futility stopping is considered, more liberal rejection boundaries can be applied to control the type I error rate.
- While the type I error rate is controlled, this is due to the assumption that one would have stopped if a futility boundary had been crossed.

The "Inverse-Bonferroni test"

Toss a dice:

1-5: do not perform the clinical trial.

6: Perform the trial at level $6\,\alpha$.

The Issue of Stopping for Futility (Choice of α_0)

Should futility stopping be considered when computing the type I error rate?

- If futility stopping is considered, more liberal rejection boundaries can be applied to control the type I error rate.
- While the type I error rate is controlled, this is due to the assumption that one would have stopped if a futility boundary had been crossed.

The "Inverse-Bonferroni test"

Toss a dice:

1-5: do not perform the clinical trial.

6: Perform the trial at level $6\,\alpha$.

FDA Experience with Stopping for Futility

Lin et al, 2016. CBER Experience with Adaptive Design Clinical Trials

- "One of the **most useful adaptations** in clinical trials is consideration of a study termination for futility ..."
- "There have been **various adaptive designs** proposals ... have attempted **to borrow alpha** from a **binding futility** analysis, most often to increase the frequency or nominal significance level of interim efficacy analyses."
- "Because we have encountered multiple cases in which supposedly **binding futility boundaries** have been crossed and **ignored**, it has been our practise to ask sponsors to **evaluate type I error without accounting for any futility analyses**."

Clinical trials are usually more complex ...

Multiplicity in Adaptive Clinical Trials

Multiplicity arises through

- multiple treatment groups
- multiple endpoints
- multiple subgroups

In Adaptive Clinical Trials, treatment groups, endpoints and subgroups may be dropped or added in interim analyses while controlling the Familywise Type I Error Rate in the strong sense.

Control of the familywise type I error rate in the strong sense:

The probability that any true null hypothesis is rejected is bounded by α, regardles of which and how many null hypotheses are true.

Multiplicity in Adaptive Clinical Trials

Multiplicity arises through

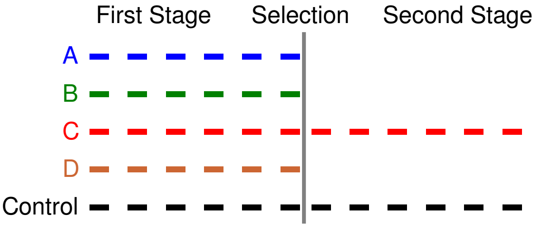
- multiple treatment groups
- multiple endpoints
- multiple subgroups

In Adaptive Clinical Trials, treatment groups, endpoints and subgroups may be dropped or added in interim analyses while controlling the Familywise Type I Error Rate in the strong sense.

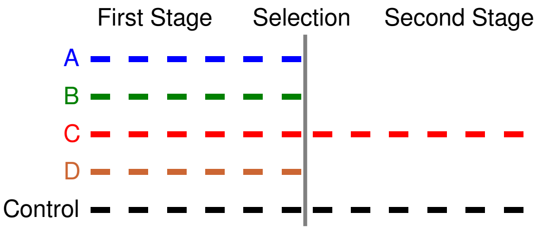
Control of the familywise type I error rate in the strong sense:

The probability that any true null hypothesis is rejected is bounded by α, regardles of which and how many null hypotheses are true.

Adaptive Trials with Treatment Selection

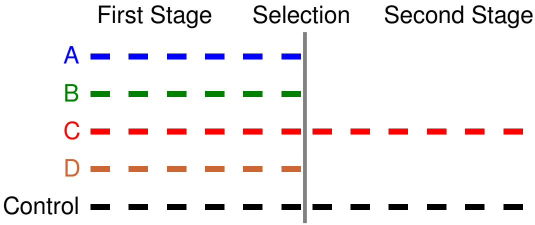


Adaptive Trials with Treatment Selection



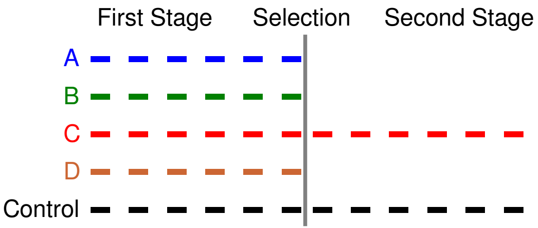
- First stage investigates several doses.
- Second stage treatment(s) are selected based on first stage data.
- Efficacy is demonstrated with data from both stages.

Adaptive Trials with Treatment Selection



- First stage investigates several doses.
- Second stage treatment(s) are selected based on first stage data.
- Efficacy is demonstrated with data from both stages.

Adaptive Trials with Treatment Selection



- First stage investigates several doses.
- Second stage treatment(s) are selected based on first stage data.
- Efficacy is demonstrated with data from both stages.

Advantages of Treatment Selection Designs

- Treatment (dose) selection trial and confirmatory trial incorporated into a single trial
- Saving of sample size

  The first stage data is used in the final test decision.
- Saving of time

  Preparation time for a second trial is spared.

Case Study: Interim Dose Selection

- Seamless phase II/III designs for two pivotal placebo controlled trials of a new chemical entity for the treatment of diabetic nephropathy
- Objectives:
    - Demonstrate superiority in a surrogate marker of kidney disease progression
    - Select two of three initially tested dose strengths based on an interim analysis of the benefit/risk ratio in both trials.
- Pre-planned interim analyses to be performed by an IDMC after 60% of 420 patients had completed 8 weeks of treatment in the first trial.
- Dose selection based on data from both trials using pre-determined criteria for the primary efficacy and safety parameters.
- Proposed type I error rate control: Bonferroni adjustment to control the familywise error rate adjusting the level for two comparisons only.

Case Study: SAWP/CHMP Reply

- The statistical testing procedure was not endorsed, as it was not supposed to control the familywise type I error rate for the three hypotheses initially considered.
- Instead, adaptive combination tests based on the closure principle and adaptive Dunnett test procedures based on the conditional error rate are adequate methods to control the type I error rate.
- The advantage of the proposed design with respect to power should be evaluated as it maybe small.
- Safety evaluation may not be possible to support dose selection at the proposed time of interim analysis.

Issues to address

- Control of the type I error rate *Even if only one experimental arm (and the control) is selected, a multiplicity correction is required.*

  EMA Reflection Paper, 2007, FDA Draft Guidance, 2010

- Interim selection of treatments may introduce bias
  If the treatment with the largest interim effect size is chosen, the effect estimates will be biased.
- Interim data maybe highly variable and lead to selection of the "wrong" treatment arm.
- Unblinding of data at the interim analysis

Control of the Type I error rate

Several procedures have been proposed:

- Methods based on completely predefined adaptation rules
  - Multiplicity adjusted critical values are determined by simulation or numerical integration

  Thall et al. '88, '89, Stallard and Todd '03, Sampson and Sill '05, Magirr et al. '12 ...

- Combine Closure Principle and Adaptive Designs
  - Perform adaptive tests for intersection hypotheses using

  Bauer and Kieser '99, Kieser et al. '99, Hommel 2001, Posch et al. '05, König et al. '08, Bretz et al. '09, Posch et al. '11

Adaptive Designs based on the closure principle

- Selection of treatments may depend on all data collected (also safety data, secondary endpoints)
- Sample sizes may be adapted.
- In principle, pre-specification of the adaptation rules is not required to control the multiple Type I error rate.
    - However, the type of adaptations and the anticipated adaptation rules should be pre-specified
    - Number of adaptations should be limited

Dose selection and efficacy testing

- Parallel group design with $k = 2$ dose groups and a control group (i.e., in total three parallel groups).
- Testing the one sided hypotheses

Dose 1 vs control: $H_{0,1} : \mu_1 \leq \mu_0$ vs. $H_{1,1} : \mu_1 > \mu_0$

Dose 2 vs control: $H_{0,2} : \mu_2 \leq \mu_0$ vs. $H_{1,2} : \mu_2 > \mu_0$

Dose selection and efficacy testing

- After Stage 1 we decide either to
  - go into Stage 2 with BOTH doses or
  - go into Stage 2 with only ONE dose.
- Selection rule unknown before end of Stage 1.
- Choice of sample sizes for Stage 2 depends on selected dose(s) and observed efficiency.
- Regulatory bodies ask for a level $\alpha = 0.025$ test of the intersection hypothesis

$$H_{0,1} \cap H_{0,2} : \mu_1, \mu_2 \leq \mu_0$$

Flexible Closed Test (Bauer & Kieser 1999, Hommel 2001)

- Use flexible two stage test for $H_{0,1} \cap H_{0,2}$,
  e.g. fix a combination test $C(p, q)$ at level α.

- At Stage 1 use a multiplicity adjusted p-value for p
  e.g. p-value of Šidak test

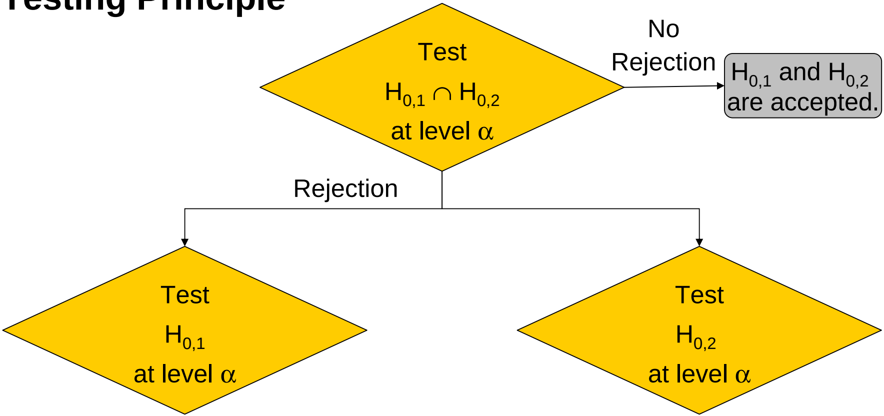
$$p = p_{12} = 1 - [1 - \min(p_1, p_2)]^2$$

- At Stage 2 use the p-value for the selected doses(s):
  - If we select only one, e.g. dose 1, we use $q = q_1$
  - If we select both, we use e.g. Šidak test

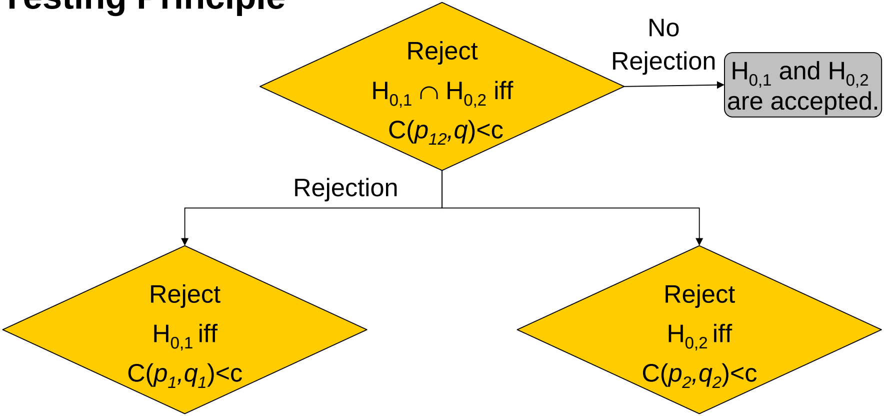
$$q = q_{12} = 1 - [1 - \min(q_1, q_2)]^2$$

- In all cases reject $H_{0,1} \cap H_{0,2}$ iff $C(p, q) \leq c$.

The Closed Testing Principle



Adaptive Closed Testing Principle



Reject $H_{0,1} \cap H_{0,2}$ iff $C(p_{12},q)<c$

No Rejection

$H_{0,1}$ and $H_{0,2}$ are accepted.

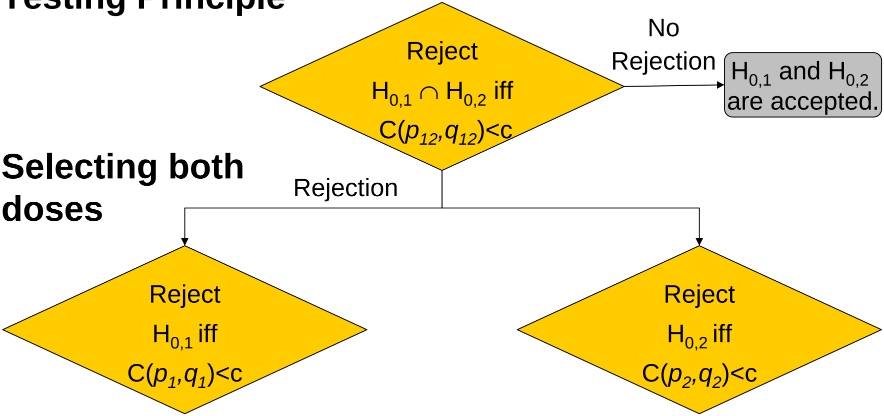
Rejection

Reject $H_{0,1}$ iff $C(p_1,q_1)<c$

Reject $H_{0,2}$ iff $C(p_2,q_2)<c$

Adaptive Closed Testing Principle

Selecting both doses



Reject $H_{0,1} \cap H_{0,2}$ iff $C(p_{12},q_{12})<c$

No Rejection

$H_{0,1}$ and $H_{0,2}$ are accepted.

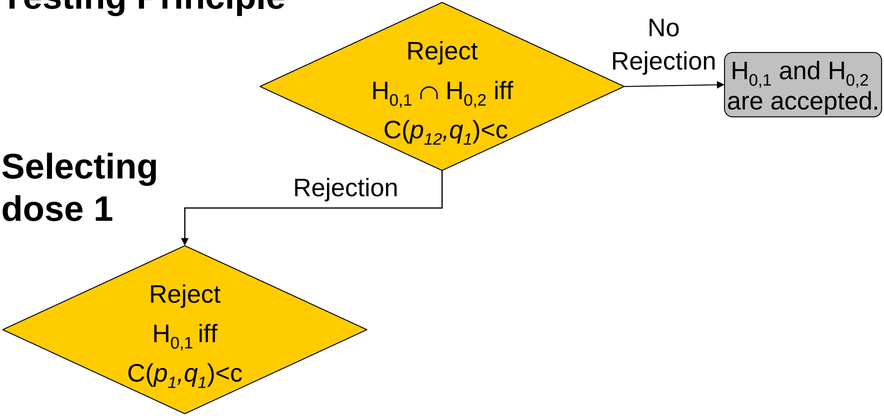
Rejection

Reject $H_{0,1}$ iff $C(p_1,q_1)<c$

Reject $H_{0,2}$ iff $C(p_2,q_2)<c$

Adaptive Closed Testing Principle

Selecting dose 1



Reject

$H_{0,1} \cap H_{0,2}$ iff

$C(p_{12},q_1)<c$

No Rejection

$H_{0,1}$ and $H_{0,2}$ are accepted.

Rejection

Reject

$H_{0,1}$ iff

$C(p_1,q_1)<c$

Power considerations

How to define the “power” if multiple hypotheses are tested in an adaptive trial?

- Probability to reject all hypotheses.
- Probability to reject all selected hypotheses.
- Average power for the selected hypotheses.
- The probability for a particular treatment to be selected and the corresponding hypothesis to be rejected.
- Probability to select and reject any hypotheses.

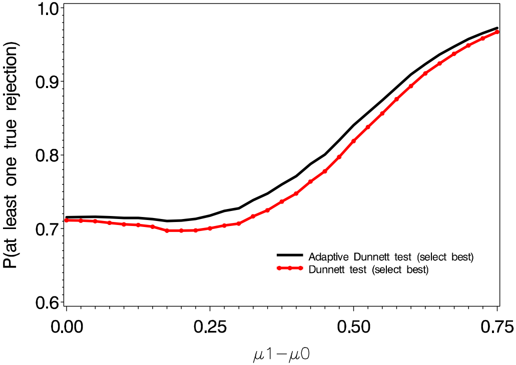
Power considerations

How to define the “power” if multiple hypotheses are tested in an adaptive trial?

- Probability to reject all hypotheses.
- Probability to reject all selected hypotheses.
- Average power for the selected hypotheses.
- The probability for a particular treatment to be selected and the corresponding hypothesis to be rejected.
- Probability to select and reject any hypotheses.

Simulation Study

- Classical approach: Fixed Sample Design with Dunnett Test
- Adaptive design: Adaptive Closed Dunnett Test



- Two treatments versus control

- Normal responses ($\sigma = 1$)

- Total n such that power for single treatment-control comparison is 80% for $\mu_i - \mu_0 = 0.5$

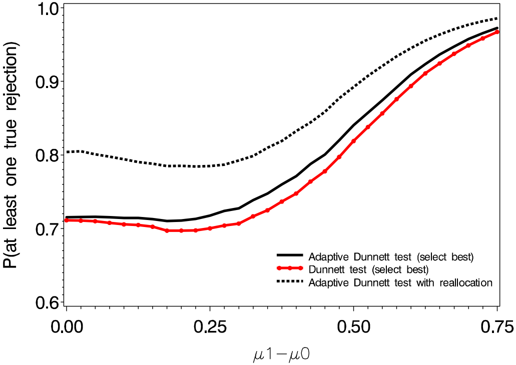
- Interim analysis at $n_1 = n/2$ with selection of "best" treatment

- mean diff. for treatment 2:

$$\mu_2 - \mu_0 = 0.5$$

Simulation Study

- Classical approach: Fixed Sample Design with Dunnett Test
- Adaptive design: Adaptive Closed Dunnett Test



- Two treatments versus control

- Normal responses ($\sigma = 1$)

- Total n such that power for single treatment-control comparison is 80% for $\mu_i - \mu_0 = 0.5$

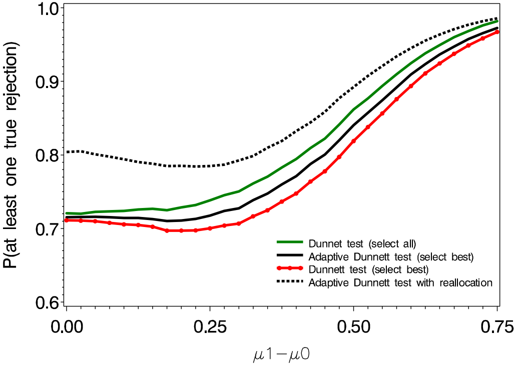
- Interim analysis at $n_1 = n/2$ with selection of "best" treatment

- mean diff. for treatment 2:

$$\mu_2 - \mu_0 = 0.5$$

Simulation Study

- Classical approach: Fixed Sample Design with Dunnett Test
- Adaptive design: Adaptive Closed Dunnett Test



- Two treatments versus control

- Normal responses ($\sigma = 1$)

- Total n such that power for single treatment-control comparison is 80% for $\mu_i - \mu_0 = 0.5$

- Interim analysis at $n_1 = n/2$ with selection of "best" treatment

- mean diff. for treatment 2:

$$\mu_2 - \mu_0 = 0.5$$

Generalizations

- More than two treatments/hypotheses
- More than two stages
- Early Stopping
- Other adaptations (like subgroup selection in adaptive enrichment designs)

Sources for Potential Inflation of Type I Error

Sources of type I error inflation	Means for error control
Early rejection of null hypotheses at interim analyses	Group sequential plans
Adaptation of design features and combination of information across trial stages	Combination of p-values e.g. inverse normal method, Fisher's combination test, conditional error function
Multiple hypotheses testing e.g. with adaptive selection of hypotheses at interim analyses	Multiple testing methodology e.g. closed test procedures

Maurer et al. 2010

All three approaches can be combined

Conclusions

Conclusions (I)

- General inferences about regulatory standards and preferences is difficult
- The assessment depends on the overall quality and the general context:
  - overall drug development program,
  - type of medicinal product
  - indication
  - ....

Conclusions (II)

Questions that should generally be addressed during planning and assessment

1. Is there a good rationale? Have alternative, more standard trial designs been considered?
2. Does the proposal fit well in the context of the development program and the data that will be available for the marketing authorization application?
3. Can the proposal be implemented without important damage to trial integrity?
4. Is the type I error rate controlled?
5. Has the potential bias of treatment effect estimates been evaluated?
6. Is the proposal practical and feasible?

Conclusions (III)

- Adaptive designs seem well accepted if properly planned and implemented
- A range of increasingly complex adaptive designs are proposed, the majority in rare diseases
- Surprisingly, still a lack of methodological knowledge
  - how to achieve type I error control
  - how to assess the efficiency of the design (timing of interim analysis, adaptation rules, power)
- Who should be decide on adaptations at interim, (DMC?, sponsor?, ...)
- Group sequential designs developed in the 70s are now well established - do we still have to wait one decade until the adaptive methodology is common knowledge?

Adaptive Designs

- allow for mid-trial learning and adaptations while strictly controlling the (multiple) type I error rate.
- address different sources of potential multiplicity issues prospectively
- can be more efficient than classical approaches

Writing "online" protocol amendments only if design modification are performed in an ongoing trial will not control the type I error rate at all! Neither post-hoc analysis.

References

See references of

- Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls
  P. Bauer, F. Bretz, V. Dragalin, F. Koenig, and G. Wassmer.
  Statistics in Medicine 35, 325-347, 2016
  http://dx.doi.org/10.1002/sim.6472 (Open Access)
  (179 references)
- Adaptive designs for confirmatory clinical trials
  F. Bretz, F. Koenig, W. Brannath, E. Glimm, and M. Posch
  Statistics in Medicine 28, 1181-1217, 2009
  http://dx.doi.org/10.1002/sim.3538
  (77 references)

Content

- A. Elsäßer, J. Regnstrom, T. Vetter, F. Koenig, R. Hemmings, M. Greco, M. Papaluca-Amati, and M. Posch.
  Adaptive clinical trial designs for European marketing authorization: a survey of scientific advice letters from the European Medicines Agency.
  Trials 15, 383, (2014)
  http://dx.doi.org/10.1186/1745-6215-15-383
  This work has received funding from the Austrian science fund FWF - P23167.
- P. Bauer, F. Bretz, V. Dragalin, F. Koenig, and G. Wassmer.
  Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls.
  Statistics in Medicine - Early View (2015)
  http://dx.doi.org/10.1002/sim.6472
  This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement number FP HEALTH 2013–602552.

Backup Slides for Panel Discussion

Justification of the Adaptive Design

EMA Reflection paper:
"Adaptive designs would be best utilised as a tool for planning clinical trials in areas where it is necessary to cope with difficult experimental situations."

- Is there a need for an adaptive trial?
- Have less complex design options been considered as well and compared to the adaptive design?
- Is the number of interim analysis justified? More than one interim analysis maybe justified in long term clinical trials.
- Is there a need for unblinding?
- Potential advantages of the adaptive design need to be weighed against potential biases and additional complexities.

Simulation Based Procedures for Type I Error Control

Clinical Trial Simulations

"What, if"-scenarios: How do designs and assumptions affect the performance of trials?

- Controllable: doses, regimes, sampling time, study duration, interim analyses, adaptations, ...
- Uncontrollable: drug characteristics (PK/PD), disease progression, drop-outs, unscheduled adaptations: "dealing with the unexpected" as dropping of an unsafe dose, ...

Simulate operating characteristics for specific trial designs :

- Probabilities of "success" (evaluate different power definitions)
- Probabilities for early trial termination (due to safety, efficacy or futility)
- Probabilities to select "best" dose during clinical development
- Impact on effect estimates (bias?) and MSE
- Expected sample sizes
- Demonstration of Type I error rate control

Simulation Based Procedures for Type I Error Control

Type I error estimation by simulation

The adaptive trial is simulated a large number of times under the null hypothesis. The fraction of runs with a rejection of the null hypothesis is calculated.

Straight forward to implement if the trial has

- a single point null hypothesis,
- a fully pre-specified adaptation rule depending on the primary endpoint only,
- no nuisance parameters,
- an adaptation rule that is not too complex such that large number of simulation runs can be performed.

Number of Simulation Runs

Posch et al., 2011; Grieve, 2011

- Precise estimates of the Type I error rate, require large numbers of simulations
- How large? For small sample numbers, a selective choice of seed may lead to biased estimates.

Table: Expected number of seeds to obtain one simulated Type I error rate below 0.025 when the actual error rate is 0.026.

♯ of runs	Expected ♯ of seeds
10^4	4
10^5	43
10^6	8×10^9

Nuisance Parameters and Simulation Studies

It is not sufficient to investigate the global null hypotheses but type I error control has to be shown for

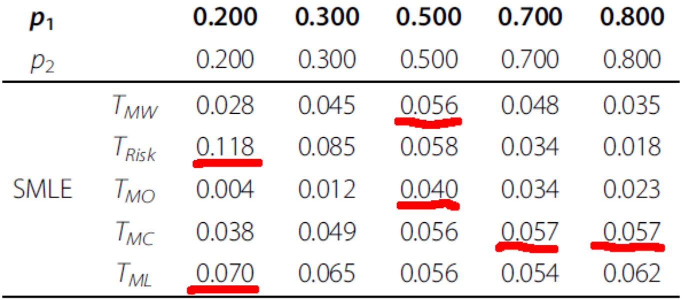
- the global and all intersection null hypotheses
- for all possible (nuisance) parameter values
- all considered adaptation options

For example, one needs to consider

- in multi-armed trials: all combinations of effective and non-effective arms and effect sizes
- in enrichment designs: all combinations of treatment effects in the subgroup and overall population
- with adaptation rules depending on surrogate/safety/secondary endpoints: all effect sizes in these endpoints

Example: Response Adaptive Design

Comparison of rates, n=30, comparison of 6 test statistics for comparison of rates



p_1		**0.200**	**0.300**	**0.500**	**0.700**	**0.800**
p_2		0.200	0.300	0.500	0.700	0.800
	T_{MW}	0.028	0.045	0.056	0.048	0.035
	T_{Risk}	0.118	0.085	0.058	0.034	0.018
SMLE	T_{MO}	0.004	0.012	0.040	0.034	0.023
	T_{MC}	0.038	0.049	0.056	0.057	0.057
	T_{ML}	0.070	0.065	0.056	0.054	0.062

Simulated Type I error (10.000 runs)

Gu & Lee, 2010, Table 11

Challenges of Type I Error Control with Simulations

- Can one sufficiently explore the type I error rate in adaptive clinical trials (relying on an abundance of parameters and assumptions) by simulations?
- Has the worst case scenario with respect to the type I error really been identified?
- Have only scenarios with favourable assumptions been investigated and presented by the sponsor?
- How can one convincingly communicate the results of the very extensive simulation work required?

Summary – Simulations

- In principle, clinical trial simulation is a valid tool to study operating characteristics of clinical trials.
- However, often it may not be feasible to cover the whole relevant parameter space to show FWER in the strong sense by simulations.
- Statistical methods for which type I error control can be demonstrated under less restrictive assumptions (e.g., combination tests, conditional error rate based tests) are preferred.
- Still simulations are valuable to assess the power of adaptive tests.
- To investigate bias and MSE of point estimates, simulation studies are proper tools. Additionally, worst case scenarios for the bias are of interest.

Operational Challenges of Adaptive Designs

Further Operational Challenges of Adaptive Trials (I)

All limitations of group sequential trials apply

- Interim analyses require larger logistic efforts (decision flow; group of people involved in the decision process; high quality data must be provided in relatively short time
- Endpoint (longterm clinical versus surrogacy) - mature enough for interim decisions
- Overrunning / Recruitment stop (ratio recruitment speed versus time until endpoint is observed)
- Need for good data management and monitoring (electronic data capture) to have interim analysis with current data

Further Operational Challenges of Adaptive Trials (II)

Specific challenges in adaptive designs

- Decision Making on adaptations (who, DMC?, experienced members, information flow, sponsor involvement, firewalls ...)
- How much pre-specification is needed?
- Which content should be included in the study protocol, DMC-charta, ...?
- Will adaptations affect other documents (e.g, informed consent if randomisation ratio is changed, doses dropped, ...)
- Immediate implementation of adaptations (e.g., change of randomisiation, population, doses, ..)
- Logistics of drug supply
- Communication of adaptations to health authorities (e.g., ethics committee)
- Un-intended adaptations/modifications to the trial (e.g., change in patient populations or Placebo effect because arm has been dropped)

Challenges in Adaptive Survival Trials

Case Study Survival: Sample Size Reassessment

- Open-label, two-armed, single pivotal phase III study for an anticancer drug in a rare disease
- Objective: To demonstrate superiority of the drug over a standard treatment for the primary endpoint of overall survival.
- Pre-planned adaptive design with two interim analyses (independent data monitoring committee, IDMC) with Haybittle-Peto stopping boundaries
- Interim analyses at 50% and 80% of events, given a fixed overall sample size
- At the second interim analysis, possibility to increase the number of events by 20% if the interim results show a promising but not overwhelming trend (conditional power arguments).
- No increase of the sample size.
- Proposed analysis: inverse normal method

Case Study Survival: SAWP/CHMP Reply

- Design is acceptable from a statistical point of view if the type I error rate is controlled and operational bias is avoided.
- No agreement to the early rejection boundary in the first interim (concerns over the totality of evidence that would be available for a benefit-risk assessment) but agreement to futility stopping.
- Discussion whether primary analysis should be based on the standard fixed sample test statistics. Inverse normal test as sensitivity analysis:
  - If sample size is increased only if a promising interim effect is observed, the fixed sample test controls the type I error rate under certain assumptions ("Promising Zone Approach").
  - The inverse normal method down-weights the second stage treatment effect if the number of events is increased. This is undesirable if the survival curves initially separate but become closer at later time points.
  - A complexity (not explicitly discussed), is the potential inflation of the type I error rate if adaptations are based on information of patients censored at the interim analysis.

Adaptive Tests for Survival Data

- The combination test and the conditional error approach can be extended to survival data and the log-rank test (independent increments property).

  Wassmer 2006, Schaefer & Mueller 2001

- Stagewise p-values are calculated from the events occuring in each stage.
- Caveat: This may lead to biased tests if adaptations are based on covariate information or secondary endpoints of first stage patients censored at the time of the interim analysis. E.g., adaptations based on PFS when the primary endpoint is OS.

  Bauer & Posch, 2001

Type I Error Control in Adaptive Survival Studies

Patients recruited in the first stage maybe still under risk in the second stage.

- Tests based on the independent increments property of the log-rank statistics are in general not valid if adaptations depend on secondary endpoints.

  Posch & Bauer, 2004

- Test procedures where the follow-up time from first stage patients is fixed control the type I error rate, but do not include all events in the test statistics if the trial is extended.

  Jenkins et al. '11, Irle & Schäfer, '12

- Conservative tests based on all observed data are typically strictly conservative. Magirr et al. 2016

Sources for Potential Inflation of Type I Error

Sources of type I error inflation	**Means for error control**
Early rejection of null hypotheses at interim analyses	Group sequential plans
Adaptation of design features and combination of information across trial stages	Combination of p-values e.g. inverse normal method, Fisher's combination test, conditional error function
Multiple hypotheses testing e.g. with adaptive selection of hypotheses at interim analyses	Multiple testing methodology e.g. closed test procedures

Maurer et al. 2010

All three approaches can be combined