Federal Institute
for Drugs
and Medical Devices

SME workshop

Statistical perspectives in regulatory clinical development programmes

Session 3: Statistical considerations in confirmatory clinical trials I

Norbert Benda

Contents

1. **Superiority, non-inferiority and equivalence**
   - Basic principles
   - Important methodological differences
   - Derivation of a non-inferiority margin
   - Issues in non-inferiority trials
2. **Endpoints, effect measures and estimands**
   - Definitions
   - What is an estimand ?
   - Issues and examples
3. **Multiplicity**
   - Introduction
   - Solutions and examples

EMA
Points to consider on switching between superiority and non-inferiority

The European Agency for the Evaluation of Medicinal Products
Evaluation of Medicines for Human Use

London, 27 July 2000
CPMP/EWP/482/99

COMMITTEE FOR PROPRIETARY MEDICINAL PRODUCTS (CPMP)

POINTS TO CONSIDER ON SWITCHING BETWEEN SUPERIORITY AND NON-INFERIORITY

DISCUSSION IN THE EFFICACY WORKING PARTY (EWP)	February 1999/ September 1999
TRANSMISSION TO THE CPMP	September 1999
RELEASE FOR CONSULTATION	September 1999
DEADLINE FOR COMMENTS	December 1999
RE-SUBMISSION TO THE EFFICACY WORKING PARTY	June 2000
APPROVAL BY THE CPMP	July 2000

Points to Consider have been developed to provide advice on selected areas relevant to the development of medicinal products in specific therapeutic fields.

This document will be revised in accordance with the scientific advances made in this area.

EMA Guideline on the choice of the non-inferiority margin

European Medicines Agency
Pre-authorisation Evaluation of Medicines for Human Use

London, 27 July 2005
Doc. Ref. EMEA/CPMP/EWP/2158/99

COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE
(CHMP)

GUIDELINE ON THE CHOICE OF THE NON-INFERIORITY MARGIN

DRAFT AGREED BY THE EFFICACY WORKING PARTY	December 1999 – January 2004
ADOPTION BY COMMITTEE FOR RELEASE FOR CONSULTATION	February 2004
END OF CONSULTATION (DEADLINE FOR COMMENTS)	May 2004
AGREED BY WORKING PARTY	June 2004
ADOPTION BY COMMITTEE	July 2005
DATE FOR COMING INTO EFFECT	January 2006

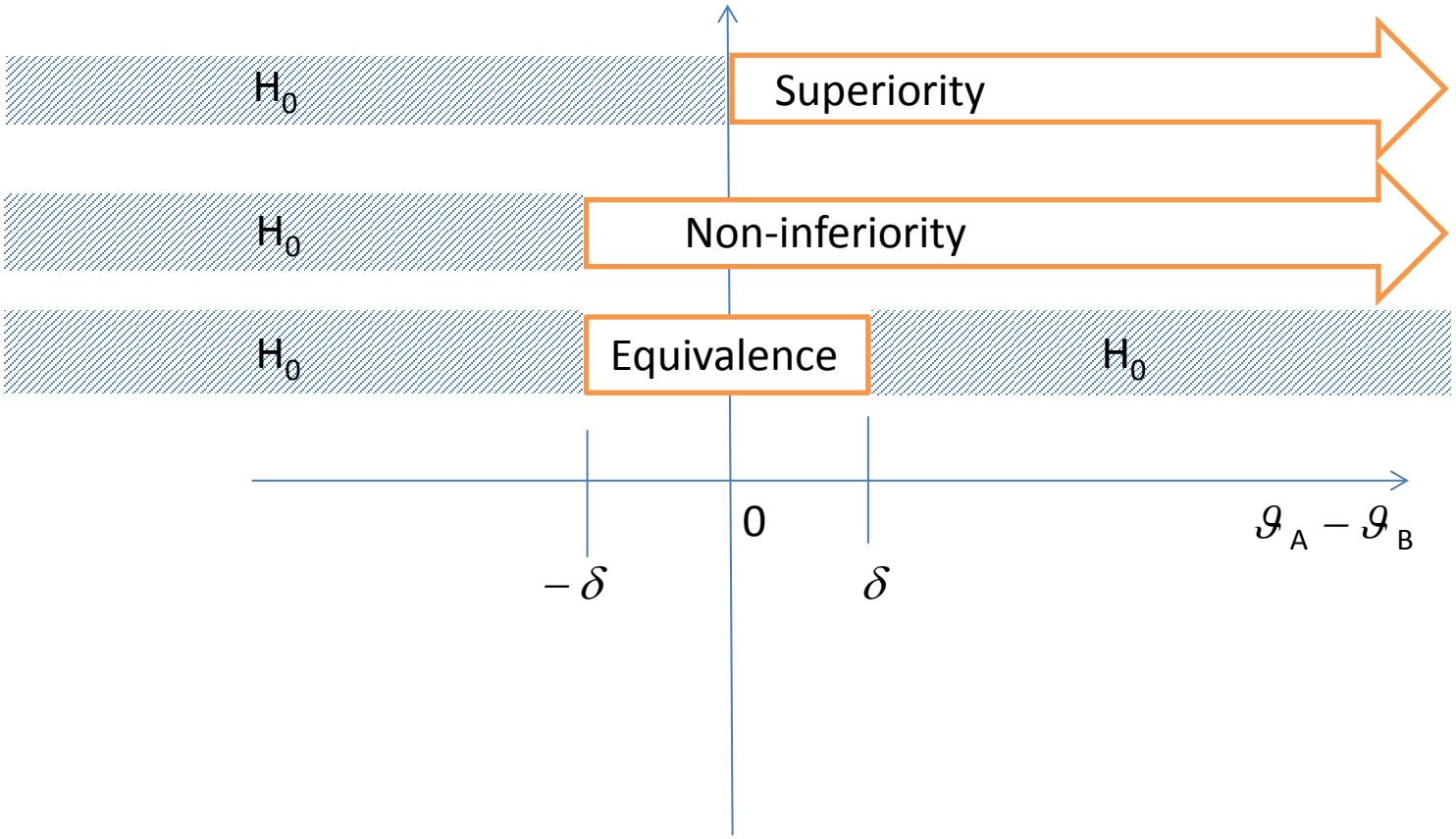
Superiority, non-inferiority, equivalence

- **superiority study**
  - comparison to placebo
    - new drug to be better than placebo
- **non-inferiority study**
  - comparison to an active comparator
    - suggests: new drug as least as good as comparator
    - proofs: new drug not considerably inferior than comparator
- **equivalence study**
  - bioequivalence in generic applications
  - therapeutic equivalence in biosimilars
    - difference between drugs within a given range

Superiority, non-inferiority, equivalence

- **compare parameter ϑ between two treatments**
  - e.g. ϑ = mean change from baseline in Hb1Ac
  - ϑ_A, ϑ_B = mean change for treatment A (new) and treatment B (comparator or placebo)
- **superiority comparison**
  - show: $\vartheta_A > \vartheta_B$
  - reject null hypothesis H_0: $\vartheta_A \leq \vartheta_B$
- **non-inferiority comparison**
  - show: $\vartheta_A > \vartheta_B - \delta$
  - reject null hypothesis H_0: $\vartheta_A \leq \vartheta_B - \delta$
    - δ = non-inferiority margin
- **equivalence comparison**
  - show: $-\delta < \vartheta_A - \vartheta_B < \delta$
  - reject null hypothesis H_0: $\vartheta_A \leq \vartheta_B - \delta$ and $\vartheta_A \geq \vartheta_B + \delta$

Superiority, non-inferiority, equivalence



Confirmatory superiority trial

- **show**
  - new drug better than placebo
  - ➢ use statistical test on null hypothesis H_0: $\vartheta_A \leq \vartheta_B$
  - ➢ significance = reject null hypothesis – conclude superiority
- **validity**
  - type-1 error control:
    - Prob(conclude superiority|no superiority) ≤ 2.5 %
      - false conclusion of superiority ≤ 2.5 %
  - effect estimate relative to placebo $\hat{\vartheta}_A - \hat{\vartheta}_B$
    - unbiased (correct on average)

      or
    - conservative (no overestimation on average)

Confirmatory superiority trial

- **ensure**
  - probability of a false positive decision on superiority should be small ($\leq$ 2.5 %)
    - type-1 error control of the statistical test
    - control for multiple comparisons
  - conservativeness
    - avoid overestimation
      - correct statistical estimation procedure
      - proper missing data imputation
      - proper randomisation
      - etc.

Confirmatory non-inferiority trial

- **show**
  - new drug better than comparator – δ
    - use statistical test on null hypothesis H_0: $\vartheta_A \leq \vartheta_B - \delta$
    - conclusion of non-inferiority (NI) = rejection of null hypothesis
- **validity**
  - type-1 error control:
    - Prob(conclude NI | new drug inferior – δ) $\leq$ 2.5 %
      - false conclusion of NI $\leq$ 2.5 %
  - effect estimate relative to active comparator $\hat{\vartheta}_A - \hat{\vartheta}_B$
    - unbiased (correct on average)
      - no underestimation of an possibly negative effect

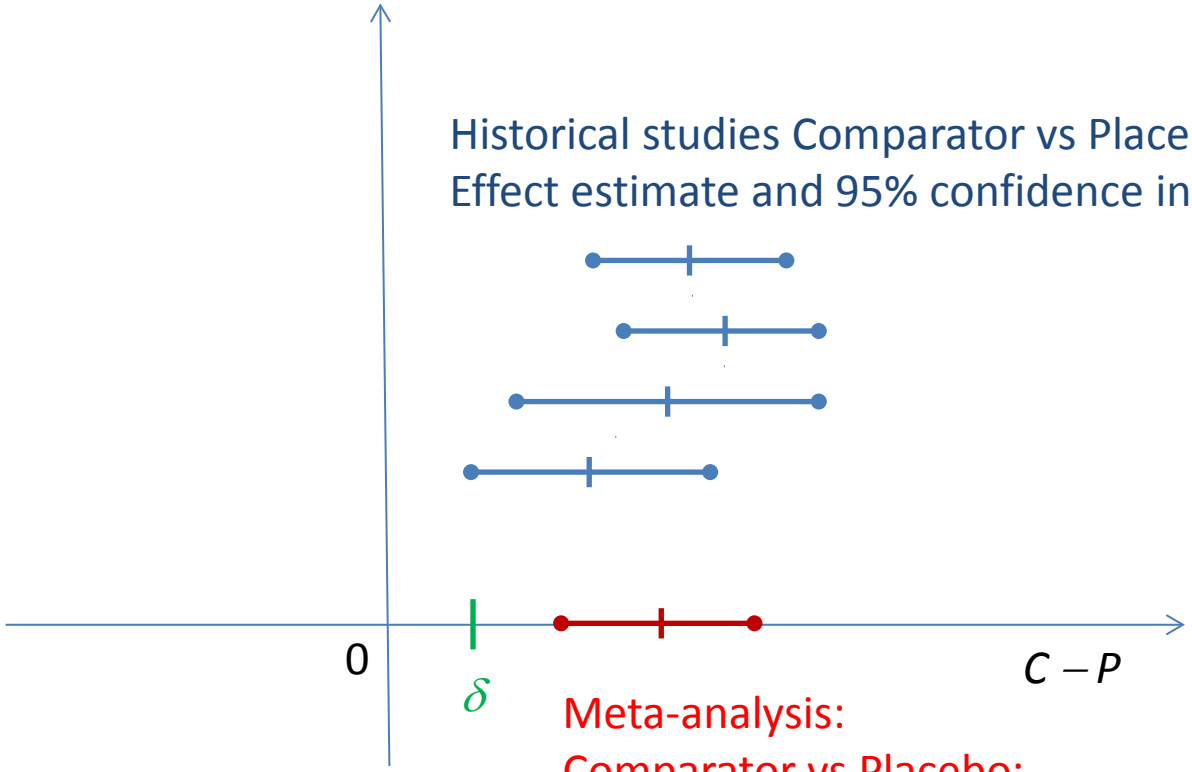
Issues in non-inferiority trials

- **non-inferiority margin δ**
  - clinical justification
    - defined through clinical relevance
  - "statistical" justification
    - defined through comparator benefit compared to placebo
- **sensitivity**
  - NI studies to be designed to detect differences
- **constancy assumption**
  - assumed comparator effect maintained in the actual study
    - relevant population to be tested
    - comparator effect maintained over time
      - control e.g. "*biocreep*" in antimicrobials

Non-inferiority margin

- **clinical justification**
  - clinical relevance
    - involving anticipated risk benefit
- **“statistical” justification**
  - related to putative placebo comparison
    - indirect comparison to placebo using historical data
      - based on estimated difference comparator (C) to placebo (P) C – P
  - use historical placebo controlled studies on the comparator
    - evaluating C – P in a meta-analysis
    - quantifying uncertainty in historical data by using a meta-analysis based 95% confidence interval of C – P
    - define NI margin relative to the lower limit of the confidence interval, e.g. by a given fraction

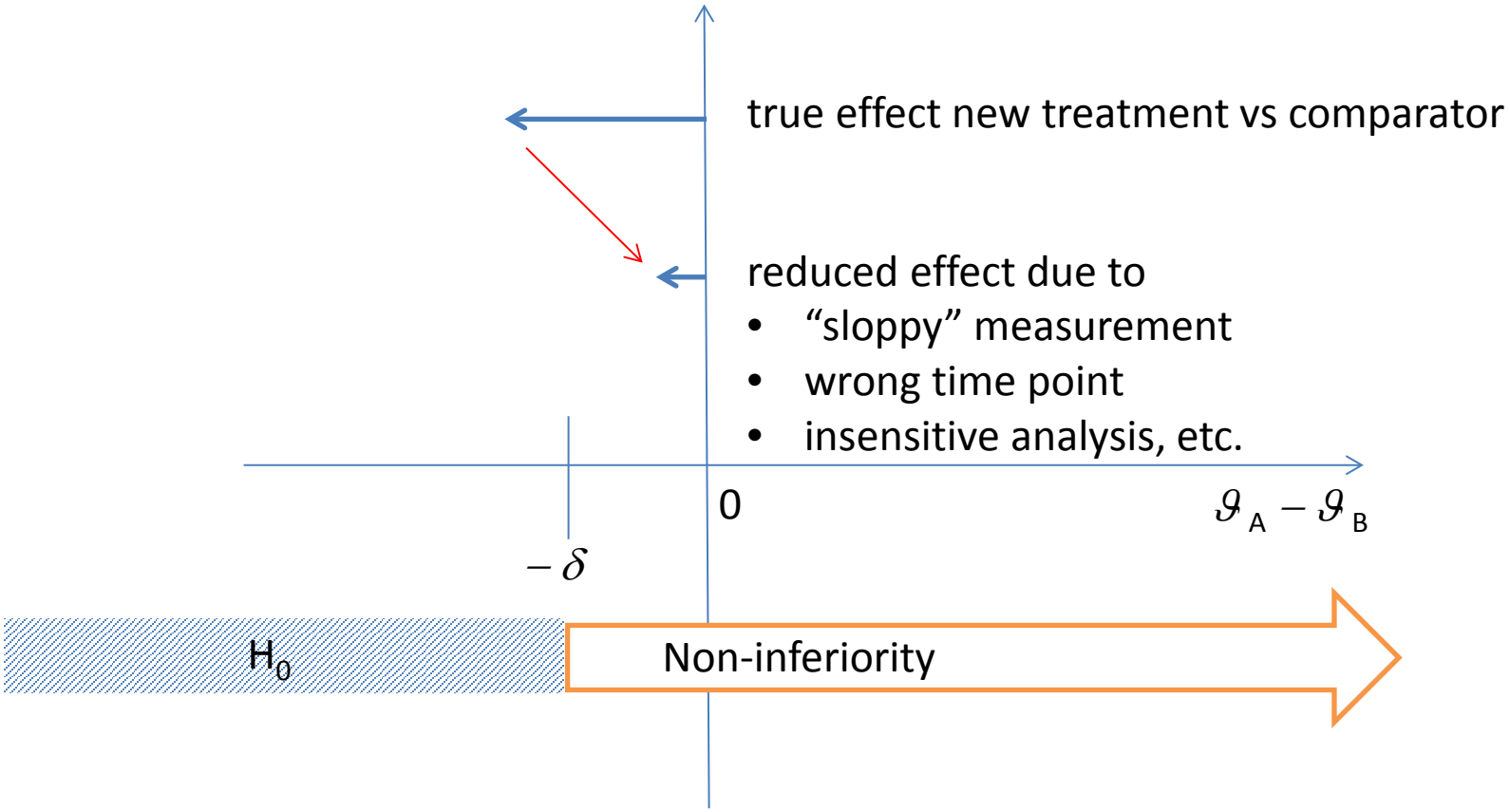
Non-inferiority margin: Statistical justification



Sensitivity of a NI trial

- **lack of sensitivity in a superiority trial**
  - sponsors risk
  - may lead to an unsuccessful trial
- **lack of sensitivity in a NI trial**
  - relevant for approval
  - risk of an overlooked inferiority
    - assume "true" relevant effect $\vartheta_A < \vartheta_B - \delta$
    - insensitive new study
      - e.g. wrong measurement time in treatment of pain
      - estimated effect difference $\hat{\vartheta}_A - \hat{\vartheta}_B \approx 0$
    - study would be a (wrong) success

Insensitive non-inferiority trial



Sensitivity of a NI trial: Toy example

- **historical data from comparator trials**
  - conducted in relevant population of severe cases
  - response rates: comparator 60%, placebo 40%
  - difference to placebo: 20%
    - NI margin chosen for new study = 8%
- **actual NI study (new vs comparator)**
  - conducted in mild and severe population
  - 70% mild – 30% severe
  - assume (e.g.):
    - 100% response expected in mild cases irrespective of treatment
  - expected (putative) response in this population
    - comparator $0.3 \cdot 60\% + 0.7 \cdot 100\% = 88\%$
    - placebo $0.3 \cdot 40\% + 0.7 \cdot 100\% = 82\%$
    - expected (putative) difference to placebo: 6%
  - 8% difference (new vs comparator) would mean
    - new drug inferior to placebo

Sensitivity of a NI trial

- **potential sources of lack of sensitivity in a NI trial**
  - wrong measurement time (too early, too late)
  - wrong or “diluted” population
    - e.g. study conducted in patients with a mild form of the disease, but difference expected in more severe cases
  - lots of missing data + insensitive imputation
    - e.g. missing = failure may be too insensitive
  - insensitive endpoint
    - e.g. dichotimized response less sensitive than continuous outcome (e.g. ACR50 vs ACR score)
  - insensitive measurement (large measurement error)
  - rescue medication
    - e.g. pain
      - primary endpoint VAS pain
      - more rescue medication used for new drug

Equivalence trial

- **Bioequivalence**
  - primary endpoint AUC or Cmax
  - show: $0.8 \leq \text{mean}(AUC_{generic})/\text{mean}(AUC_{originator}) \leq 1.25$
    - symmetric on log-scale:
      - $0.223 \leq \log\left(\text{mean}(AUC_{generic})/\text{mean}(AUC_{originator})\right) \leq 0.223$
    - confirmatory proof given by
      - 90% confidence interval $\subset$ [0.8, 1.25]
      - equivalent to two one-sided 5% tests to proof
        - $\text{mean}(AUC_{generic})/\text{mean}(AUC_{originator}) \leq 1.25$
        - $\text{mean}(AUC_{generic})/\text{mean}(AUC_{originator}) \geq 0.8$
      - increased type-1 error in bioequivalence !
        - 5% one-sided instead of 2.5% one-sided

Equivalence trial

- **therapeutic or PD equivalence**
  - frequently used for biosimilarity
  - demonstration of equivalence by
    - e.g. using 95% confidence interval $\subset$ equivalence range
    - equivalent to two one-sided 2.5% tests
  - usually symmetric equivalence range
    - depending on scale
      - e.g. (0.8, 1.25) is symmetric on log-scale (multiplicative scale)
    - e.g. biosimilarity:
      - If A is biosimilar to B, B should also be biosimilar to A
  - lack of sensitivity issues as in NI trials

ICH Concept Paper on Estimands and Sensitivity Analyses

ICH
harmonisation for better health

Final Concept Paper
E9(R1): Addendum to Statistical Principles for Clinical Trials
on
Choosing Appropriate Estimands and Defining Sensitivity Analyses in Clinical Trials
dated 22 October 2014
Endorsed by the ICH Steering Committee on 23 October 2014

Type of Harmonisation Action Proposed

To develop new regulatory guidance, suggested to be an Addendum to ICH E9, which promotes harmonised standards on the choice of estimand in clinical trials and describes an agreed framework for planning, conducting and interpreting sensitivity analyses of clinical trial data. As with ICH E9, the Addendum will focus on statistical principles related to estimands and sensitivity analysis, not on the use or acceptability of specific statistical procedures or methods. While a variety of mid-stage and late-stage clinical trials may be in scope, the primary focus of the Addendum will be on confirmatory clinical trials.

Statement of the Perceived Problem

- *Incorrect choice of estimand and unclear definitions for estimands lead to problems in relation to trial design, conduct and analysis and introduce potential for inconsistencies in inference and decision making.*

Inferences about the true efficacy and safety profile of a medicinal product are drawn from estimated effects in confirmatory clinical trials. A clinical trial protocol and analysis plan should include a 'golden thread' linking clear trial objectives with selection and prioritisation of endpoints and hypotheses for statistical testing or targets for estimation. These should, in turn, inform details of the trial design, conduct and analysis. In a confirmatory clinical trial data are collected to measure outcomes that quantify the impact of one or more experimental interventions in comparison to a control group, typically over a defined period of time, or until a sufficient number of clinical outcome events have occurred. The trialist is trying to formulate an appropriate and well-defined measure of treatment effect in terms of the data that were intended to be collected. This may then be parameterised, for example to "compare experimental drug X and placebo in terms of improving endpoint Y at time Z for all randomised patients, without regarding adherence to randomised treatment" or to "compare experimental drug X and placebo in terms of improving endpoint Y at time Z for all randomised patients if all patients had remained in the trial and received treatment as planned without rescue medication until time Z". Controversy and confusion exist on the definition and appropriate selection of an appropriate estimand and these two examples should not be taken as preferences or recommendations. These are presented only as illustrations of *estimands*; the property that is to be estimated in the context of a scientific question of interest, to stimulate discussion in generating the addendum.

EMA Guideline on Missing Data

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

2 July 2010
EMA/CPMP/EWP/1776/99 Rev. 1
Committee for Medicinal Products for Human Use (CHMP)

Guideline on Missing Data in Confirmatory Clinical Trials

Discussion in the Efficacy Working Party	June 1999/ November 2000
Transmission to CPMP	January 2001
Released for consultation	January 2001
Deadline for Comments	April 2001
Discussion in the Efficacy Working Party	October 2001
Transmission to CPMP	November 2001
Adoption by CPMP	November 2001
Draft Rev. 1 Agreed by Efficacy Working Party	April 2009
Adoption by CHMP for release for consultation	23 April 2009
End of consultation (deadline for comments)	31 October 2009
Rev. 1 Agreed by Efficacy Working Party	April 2010
Adoption by CHMP	24 June 2010
Date for coming into effect	1 January 2011

This guideline replaces Points to Consider on Missing Data in Clinical Trials (CPMP/EWP/1776/99).

Keywords	*Baseline Observation Carried Forward (BOCF), Generalised Estimating Equations (GEE), Last observation carried forward (LOCF), Missing at random (MAR), Missing completely at random (MCAR), Missing Data, Mixed Models for Repeated Measures (MMRM), Missing not at random (MNAR), pattern mixture models.*

Endpoints and effect measures

- **endpoint**
  - variable to be investigated, e.g.
    - VAS pain measured after x days of treatment
      - possible individual outcomes: 4.2, 6.3, etc.
    - response
      - possible individual outcomes: yes, no
    - time to event (death, stroke, progression or death, etc.)
      - possible individual outcomes: event at 8 months
        censored at 10 months

Endpoints and effect measures

- **effect measure**
  - population parameter that describes a treatment effect, e.g.
    - mean difference in VAS score between treatments A and B
    - difference in response rates
    - hazard ratio in overall survival
  - study result estimates the effect measure
    - observed mean difference
    - difference in observed response rates
    - estimated hazard ratio (e.g. using Cox regression)
  - note:
    - disentangle
      - population effect measure to be estimated
      - observed effect measure as an estimate of this

What is an estimand ?

- **difference between**
  - estimate and estimand ?
    - “d” vs “te”
  - any idea ?

What is an estimand ?

- **estimand = that which is being estimated**
  - latin gerundive *aestimandus* = to be estimated
  - simply speaking: the precise parameter to be estimated

    ceterum censeo parametrum esse aestimandum
  - however:
    - the parameter may not always be given easily
    - may be a (complex) function of other parameters
- **treatment effect estimate may target**
  - effect under perfect adherence: "*de-jure*"
  - effect under real adherence: "*de-facto*"
    - several options possible

Estima*s

- **estimation function (estimator)**
  - statistical procedure that maps the study data to a single value

    (that is intended to estimate the parameter of interest)
- **estimate**
  - value obtained in a given study
- **estimand**
  - parameter to be estimated
    - or a function of estimated parameters to be estimated

Example: Event rate in two treatments

- **event rates**
  - A: 60 %
  - B: 50 %
- **how to measure treatment difference ?**
  - several options
    - Rate difference: 10 %
    - Rate ratio: 60/50 = 1.2
    - Odds ratio: (60/40)/(50/50) = 1.5
    - Hazard ratio resulting from a time-to-event analysis
- **estimand relates to the effect measure**
  - but not only to this !

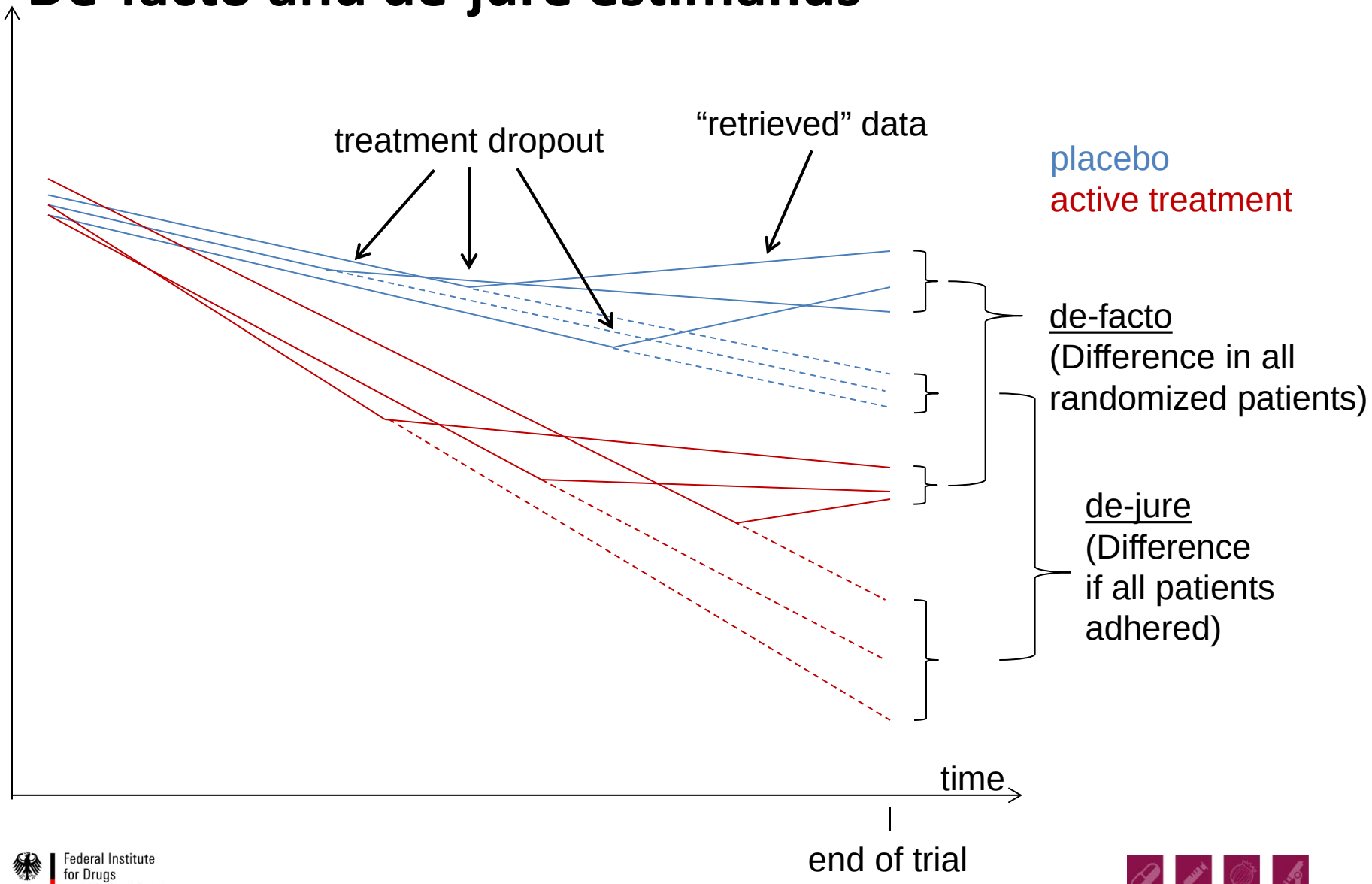
What is an estimand ?

- **estimand**

  = the precise parameter to be estimate

- **related to**
  - endpoint
  - effect measure
    - mean difference, difference between medians, risk ratio, hazard ratio, etc.
  - population
  - time point of measurement, duration of observational period, etc.
  - adherence
    - effect under perfect adherence: "*de-jure*"
    - effect under real adherence: "*de-facto*"

De-facto and de-jure estimands



Estimand issues: Examples

- **rescue medication**
  - e.g. pain, diabetes
    - efficacy if no subject took rescue med (de-jure)
    - efficacy under rescue med (de-facto)
- **quality of life (QoL) in studies with relevant mortality**
  - e.g. oncology
    - QoL in survivors?
- **efficacy and effectiveness under relevant non-adherence**
  - e.g. depression
    - effect if all subjects were adherent

      or
    - effect under actual adherence

Estimands under rescue medication

- **diabetes**
- **primary endpoint: Change in Hb1Ac after 24 weeks**
- **de-facto estimand**
  - Hb1Ac change irrespective of rescue medication use
  - all data used
  - longitudinal model or ANCOVA to estimate
- **de-jure estimand**
  - Hb1Ac change without rescue medication
  - only data until start of rescue medication used
  - longitudinal model on “clean” data

Estimands under rescue medication

- **pain**
- **primary endpoint: Change in VAS pain**
- **de-facto estimand and de-jure estimand**
  - as above
- **severe pain**
  - intake of rescue medication in most patients
  - de-jure estimand not evaluable
  - de-facto estimand insensitive
  - alternative endpoints to be considered
    - amount of rescue medication
    - time to first use of rescue medication

Quality of life in studies with relevant mortality

- **study comparing treatment A and B**
  - primary endpoint
    - survival within one year
  - secondary endpoint
    - quality of life score
  - QoL after death?
    - zero ? - 1 ? – 1000 ? - ∞ ?
  - options discussed
    - death = 0
    - death = lowest rank using a non-parametric analysis
    - rank survival time after QoL in survivors
    - QoL in survivors additional to survival rates
    - joint modelling of survival and QoL

Quality of life and death: Rank analysis

- **rank analysis**
  - death = lowest rank or rank survival time after QoL in survivors
    - assess e.g. medians
    - information loss
    - assessment of clinical relevance may be difficult
    - individual interpretation not given

Quality of life (QoL) in survivors

Simplistic Example

- **sub-populations P1, P2 and P3 with prevalence 1/3 each**
- **compare treatments A and B**

	P1 1/3 of the population	P2 1/3 of the population	P3 1/3 of the population	mean QoL in survivors
A	all die QoL not given	all survive mean QoL = 30	all survive mean QoL = 60	45
B	all die QoL not given	all die QoL not given	all survive mean QoL = 50	50
	A equal to B	A better than B	A better than B	
	A better than or equal to B in all subgroups			**but:** B better than A re. mean QoL

Quality of life in survivors

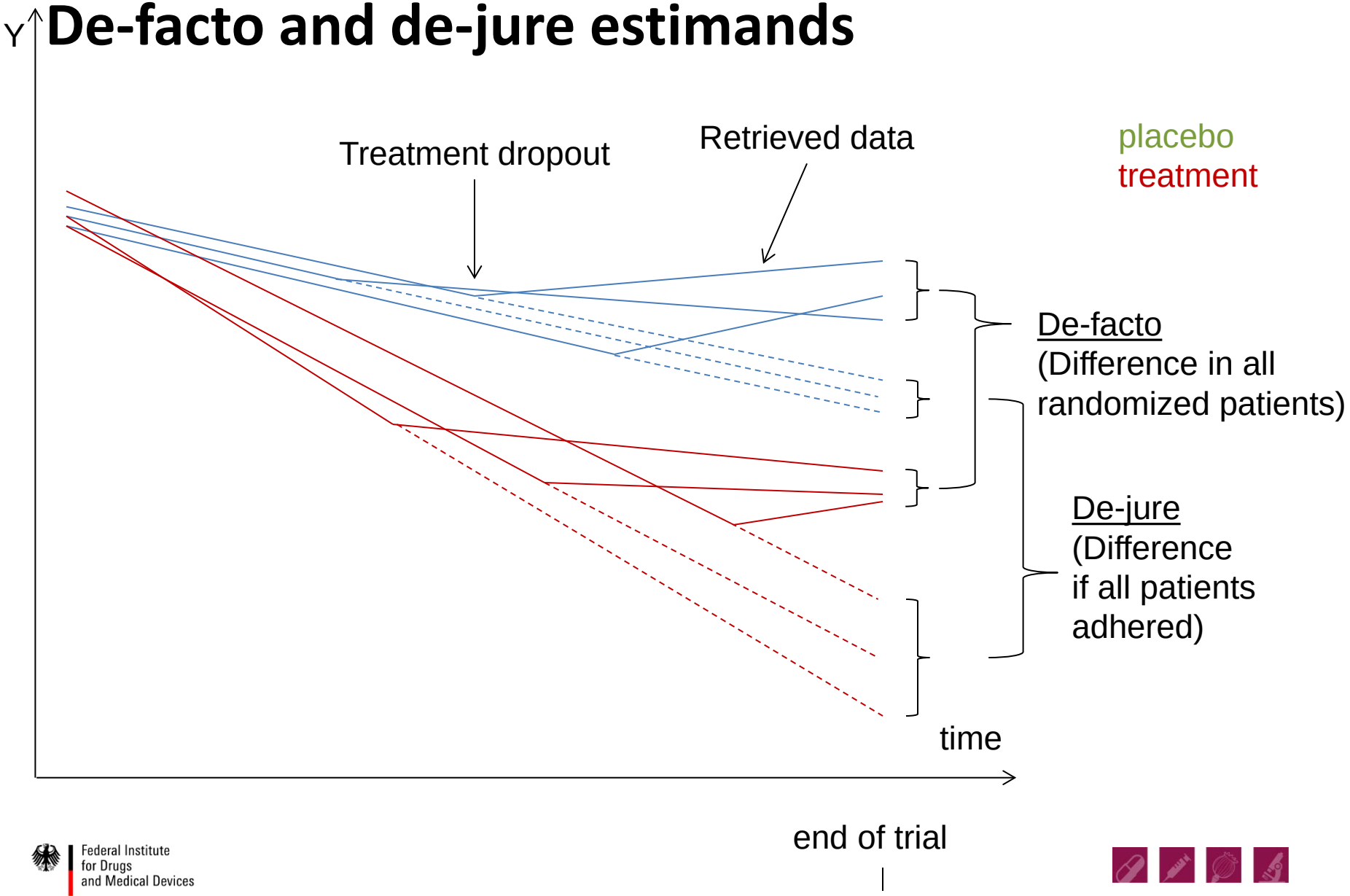
- **treatment difference in survivors**
  - difference in a post-randomization selected population
  - positive overall effect possible despite worse outcome in each patient / subgroup
  - no reasonable estimand
- **survivors cannot be identified upfront**
  - in contrast to *effect in tolerators* in other studies
    - short run-in period to identify tolerators to active treatment
  - mimic effect in those who survive under A and B using
    - causal inference
      - difficult, relying on full identification of instrumental variables
      - not recommended as primary

Different proposals for de-facto and de-jure estimands in the presence of non-adherence

- **de-jure**
  - difference if all patients adhered
  - difference in tolerators
- **de-facto**
  - difference for all randomized patients
  - difference for all randomized patients attributable to the initially randomized treatment
  - difference during adherence
  - difference in AUC during adherence

Mallinckrodt (2013), Carpenter et al (2014)

De-facto and de-jure estimands



De-facto estimands

- **= “treatment policy” estimand**
- **may be difficult to define as a parameter (function)**
  - integration over missingness process
- **in case no “de-facto” data are available**

  (retrieved data, data under rescue med, etc.)
  - difference between de-facto and de-jure can hardly be substantiated
  - analyses targeting de-facto estimands as sensitivity analyses under various assumptions
- **strong de-facto conclusions require de-facto data**
  - patient follow-up after drop-out needed
- **further discussion needed**
  - on applicability of de-facto estimands

Estimands: Summary

- **specification of a relevant estimand first**
  - clarifies study objective
  - needed to define relevant estimation and missing data method
  - impacts study design
- **an estimand includes**
  - assumption on adherence
  - distributional parameter
  - population
- **an estimand**
  - defines the primary analysis
  - different estimands may be used in additional analysis

EMA Points to consider on Multiplicity

The European Agency for the Evaluation of Medicinal Products
Evaluation of Medicines for Human Use

London, 19 September 2002
CPMP/EWP/908/99

COMMITTEE FOR PROPRIETARY MEDICINAL PRODUCTS (CPMP)

POINTS TO CONSIDER ON MULTIPLICITY ISSUES IN CLINICAL TRIALS

DISCUSSION IN THE EFFICACY WORKING PARTY	January 2000
TRANSMISSION TO CPMP	July 2001
RELEASE FOR CONSULTATION	July 2001
DEADLINE FOR COMMENTS	October 2001
DISCUSSION IN THE EFFICACY WORKING PARTY	June 2002
TRANSMISSION TO CPMP	September 2002
ADOPTION BY CPMP	September 2002

Multiplicity and type-1 error control: Example

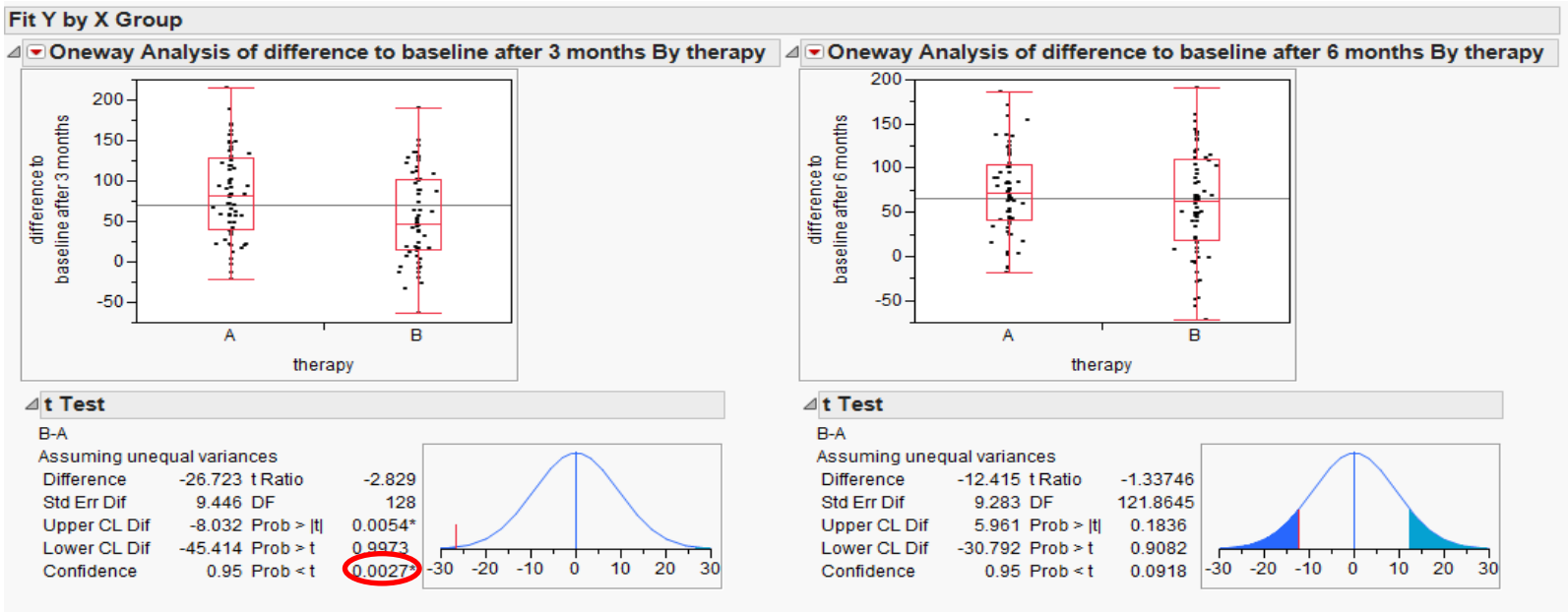
- **clinical trial comparing treatments A and B**
  - primary endpoint: walking distance in 6 minutes (difference to baseline)

    (6-minute-walk-test)
  - statistical test: two sample t-test on
    - null hypothesis H_0: mean (A) = mean (B)
    - obtain p-value
  - *p*-Wert (two-sided)

    = probability to obtain the observed difference (or greater) if in fact the null hypothesis is true (in both directions)

Multiplicity and type-1 error control: Example

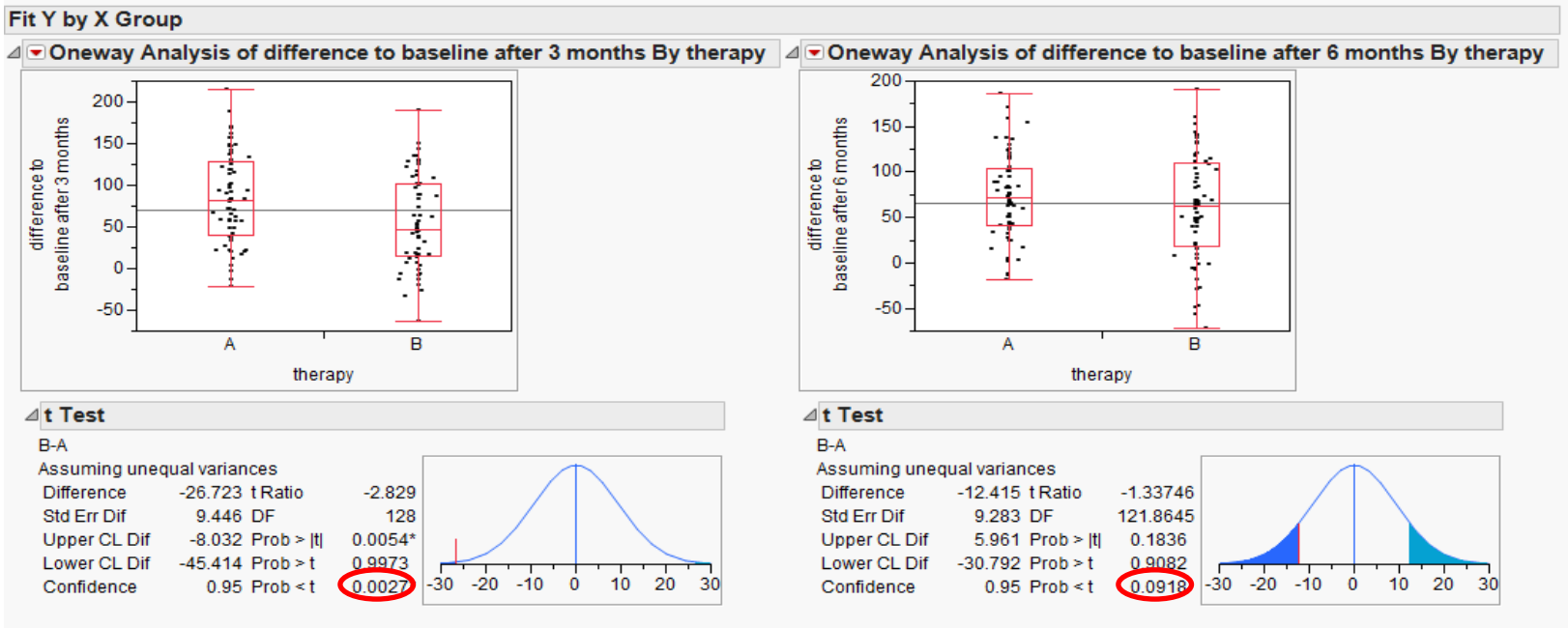
- **clinical trial comparing treatments A and B**
  - statistical test: two sample t-test
    - result: *p* (two-sided) = 0.0027 < 0.05 (5%)
    - small probability to obtain such a result by chance
    - difference declared to be significant



Multiplicity and type-1 error control: Example

- **clinical trial comparing treatments A and B**
  - two comparisons for 6MWT
    - After 3 months: *p*-value = 0.0027
    - After 6 months: *p*-value = 0.0918

    Question: Study successful ?



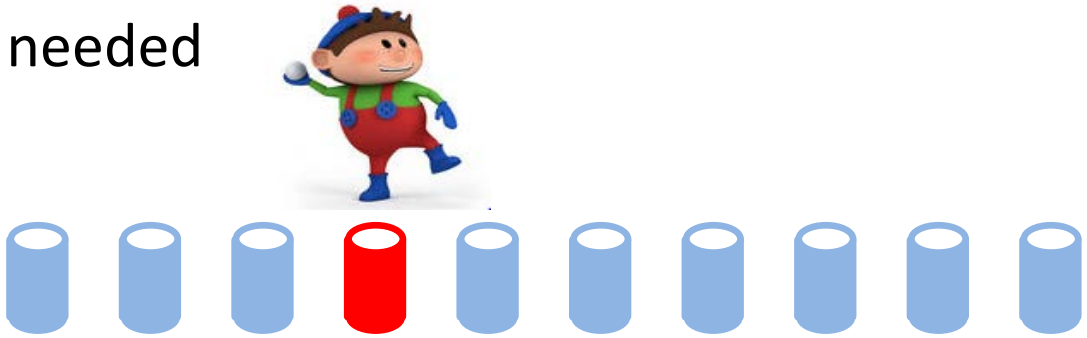
Multiplicity: Example

multiplicity and pre-specification

- post-hoc definition of endpoint of interest (primary endpoint)

  (6MWT after 3 months or 6MWT after 6 months)

  - increases the probability of a false significance
  - is invalid
- pre-specification needed



Multiplicity

Multiplicity

- multiple ways to win
- multiple chances to obtain a significant results due to chance

Example: Study success defined by a significant difference in primary endpoints, e.g.

- progression free survival (PFS)
- overall survival (OS)

- **no adjustment means:**
  - probability of a significant difference in PFS or OS > α if no real difference in PFS or OS
  - increased chances to declare an ineffective treatment to be effective

Multiplicity: Example

Example: Study success defined by a significant difference in primary endpoints,

- progression free survival (PFS)
- overall survival (OS)

Different options to keep type-1 error:

1. **PFS and OS co-primary**
   - both must be significant
2. **Hierarchical testing:**
   - test PFS first, test OS only if PFS is significant (or vice versa)
3. **Adjust α : test PFS and OS with 0.025 each (instead of 0.05) (Bonferroni)**
   - or use a different split (e.g. 0.01 for PFS, 0.04 for OS)
4. **Adjust with more complex methods**
   - "Bonferroni-Holm", "Hochberg", etc.

Multiplicity adjustment: Co-primary endpoints

Example: Study success defined by a significant difference in primary endpoints,

- progression free survival (PFS)
- overall survival (OS)

PFS and OS co-primary

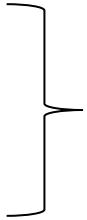
- to be pre-specified in the protocol
- both must be significant
- no valid confirmatory conclusion if only one endpoint is significant
  - e.g. PFS: $p = 0.0000001$, OS: $p = 0.073$
    - "*sorry, you lost*" – no way

Multiplicity adjustment: Hierarchical procedures

Example: Study with three dose groups

Three dose group to be compared with placebo

- pre-specified hierarchical order to test, e.g.
  - dose 3 → dose 2 → dose 1
  - no adjustment of significance level needed
  - if dose 3 significant go forward to dose 2
  - if dose 2 significant go forward to dose 1
  - stop if dose 3 (2) not significant
- no significance can be declared if the procedure has stopped
  - dose 3: $p = 0.07$
  - dose 2: $p = 0.004$
  - dose 1: $p = 0.02$



none of the doses can be declared as successful

Multiplicity: Bonferroni (like) adjustments

Example: Study success defined by a significant difference in

- either progression free survival (PFS)
- or overall survival (OS)

Adjustment needed

- e.g. PFS: $\alpha = 0.025$, OS: $\alpha = 0.025$ (Bonferroni)
- or PFS: $\alpha = 0.01$, OS: $\alpha = 0.04$

- to be pre-specified in the protocol
- α - split influences power depending on the assumptions

Multiplicity: Other adjustments

Example: Study success defined by a significant difference in

- either progression free survival (PFS)
- or overall survival (OS)

E.g. adjustment according to Bonferroni-Holm

- Smaller p-value must be < 0.025
- Larger p-value can be tested at $\alpha = 0.05$
  - PFS: $p = 0.01$, OS: $p = 0.04$ → both significant
  - PFS: $p = 0.04$, OS: $p = 0.04$ → none significant
  - PFS: $p = 0.01$, OS: $p = 0.07$ → PFS significant only
- more powerful than simple Bonferroni
- no corresponding (reasonable) confidence intervals

Sources of multiplicity

- multiple endpoints
- multiple interim looks
- multiple group comparisons
  - dose groups
- multiple (sub-)populations
- multiple analysis methods (tests)
  - all may be valid, but post-hoc selection is not

Multiplicity: Important lessons

- different sources of multiplicity possible
  - complex multiplicity issues when different sources are combined
- different test procedures available for complex multiplicity problems
- pre-specification of multiplicity procedure is paramount
  - post-hoc selection of the multiplicity procedure not valid → no control of type-1 error
- multiplicity adjustment refers to all comparisons that require a confirmatory conclusion
- corresponding confidence intervals may not always be available