Die Bedeutung der klinischen Relevanz für die Machbarkeit klinischer Studien – Eine Herausforderung für den Kliniker und Biometriker

Prof. Dr. Dieter Hauschke

Köln, den 18.06.2010

Gliederung

- **VFA ↔ Biometrie ↔ IQWIG**
- **Biometrische ‚Praxis‘**
- **Regulatorische Anforderungen**
- **Klinische Relevanz in COPD-Studien**
  - **Bewertung mittels Konfidenzintervalle**
  - **Überlegungen zur Güte (Power)**
- **Gold Standard Design**
- **Literatur**

Position des IQWIG

IQWiG Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Institute for Quality and Efficiency in Health Care

Allgemeine Methoden

Version 3.0 vom 27.05.2008

Position des IQWIG

Für die Beurteilung der klinischen Relevanz von Studienergebnissen gibt es noch kein breit akzeptiertes methodisches Vorgehen. Nur in wenigen Leitlinien finden sich Hinweise auf die Definition von relevanten beziehungsweise irrelevanten Unterschieden zwischen Gruppen. Ein erster Ansatz zur Beurteilung der klinischen Relevanz von Studienergebnissen ist die Bewertung des Effektschätzers und des dazugehörigen Konfidenzintervalls mithilfe medizinischer Sachkenntnis. Ein formales Relevanzkriterium kann die Beurteilung der (im Falle von günstigen Effekten) unteren Konfidenzgrenze für den Effektschätzer beziehungsweise die Anwendung eines statistischen Tests mit Verschiebung der Nullhypothese zum statistischen Nachweis relevanter klinischer Effekte sein [382]. Eine weitere Möglichkeit liegt darin, ein Relevanzkriterium individuell zu formulieren, zum Beispiel im Sinne einer Responderdefinition [229]. Darüber hinaus spielt die individuelle Einschätzung der Betroffenen eine wesentliche Rolle. Anhaltspunkte hierfür wird im Einzelfall die Darlegung patientenrelevanter Endpunkte liefern. Das Institut nimmt die Bewertung der klinischen Relevanz problemorientiert unter Berücksichtigung dieser Aspekte vor.

Position des IQWIG

IQWiG-Berichte – Jahr: 2009 Nr. 55

Selektive Serotonin- und Noradrenalin-Wiederaufnahmehemmer (SNRI) bei Patienten mit Depressionen

Das Literaturscreening wurde von 2 Reviewern unabhängig voneinander durchgeführt. Nach einer Bewertung der Studienqualität wurden die Ergebnisse der einzelnen Studien nach Prüfsubstanzen und Zielgrößen geordnet gegenübergestellt und beschrieben. Wenn möglich und sinnvoll, wurden Meta-Analysen durchgeführt. Für Ergebnisse kontinuierlicher Skalen war für einen Nutzennachweis zusätzlich zur statistischen Signifikanz des Gruppenunterschieds notwendig, dass der Effekt eine definierte Größe überschritt (Relevanzgrenze Cohen's d / Hedges' g = 0,2).

Position des IQWIG

Tabelle 2 (Fortsetzung): Zusammenfassung zum Vergleich von Duloxetin mit Placebo bzw. SSRI

Zielgröße	Ergebnis der Meta-Analysen bzw. Einzelstudien Gruppenunterschied [95 %-KI]			
	DUL vs. Plc[a]			DUL vs. SSRI[a]
	Kurzzeit-Akuttherapie	Langzeit-Akuttherapie[b]	Rückfallprävention[b]	Kurzzeit-Akuttherapie
Schmerz (VAS)	-4,56 [-6,79; -2,33][e] -0,20 [-0,30; -0,10][g]			

Bei der Interpretation der Effektstärken wurde das Konfidenzintervall zur Relevanzgrenze von 0,2 im Cohen's d / Hedges' g in Relation gesetzt. Lag das Konfidenzintervall vollständig oberhalb der Relevanzgrenze wurde von einer relevanten Effektgröße ausgegangen und ein Nutzenbeleg oder -hinweis attestiert. Lag das Konfidenzintervall teilweise oder komplett unterhalb dieser Grenze, so konnte die Relevanz des vorliegenden Effekts nicht mit ausreichender Sicherheit eingeschätzt werden. Es blieb somit unklar, ob die Effektstärke eine so relevante Größenordnung erreichte, dass von einem Nutzen gesprochen werden konnte. In diesen Fällen blieb der Nutzen unklar und ein Nutzen war nicht belegt.
f: relevante Effektgröße (Cohen's d / Hedges' g)
g: Relevanz des vorliegenden Effekts nicht mit ausreichender Sicherheit einzuschätzen (Cohen's d / Hedges' g)

Position des VFA

Beurteilung klinischer Relevanz von Ergebnissen klinischer Studien

Joachim Röhmel

Bremen, 1. Dezember 2009

Interessenkonflikt: Der Autor wurde vom vfa gegen ein Honorar beauftragt, eine Stellungnahme abzugeben zur Beurteilung klinischer Relevanz von Ergebnissen klinischer Studien unter Berücksichtigung der Praxis des IQWiG, eine Relevanzgrenze für die Effektgröße (ES) bei d=0.2 fest zusetzen und sich dabei auf Cohen (1988) zu berufen.

Position des IQWIG

Kommentar zur Stellungnahme von Prof. Dr. Joachim Röhmel (Bremen)

"Beurteilung klinischer Relevanz von Ergebnissen klinischer Studien"

Prof. Dr. Ralf Bender, Dr. Beate Wieseler, Dr. Guido Skipka,
Dr. Thomas Kaiser, PD Dr. Stefan Lange (IQWiG)

Köln, den 22.03.2010

Position des VFA/Sonofi-Aventis

Gutachten

Methodische Aspekte bei der Festlegung
von Kenngrößen in der
Meta-Epidemiologie klinischer Studien

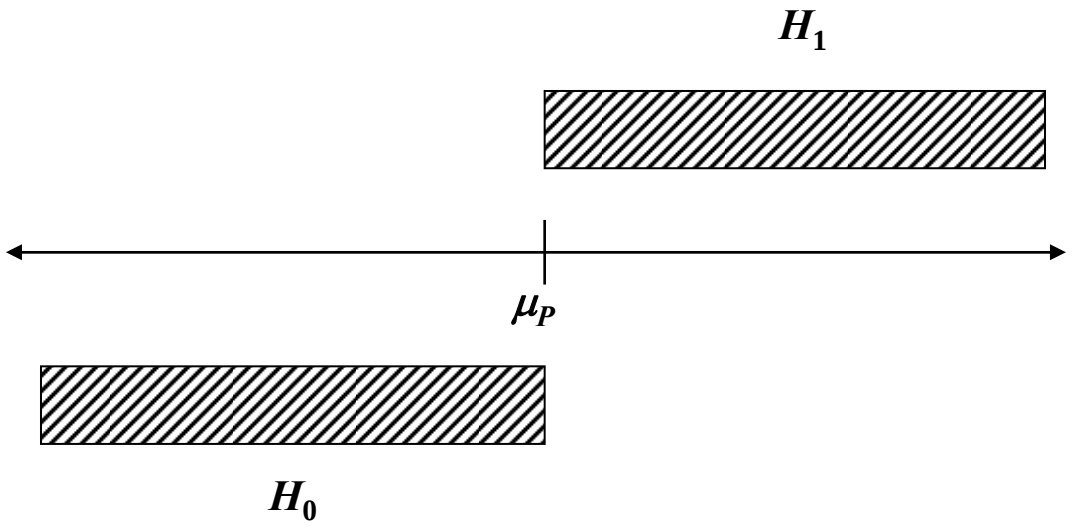
20. April 2009

Prof. Dr. Ulrich Mansmann
IBE, LMU München
mansmann@ibe.med.uni-muenchen.de

Interessenkonflikt: Der Autor wurde von der Firma Sonofi-Aventis gegen die Zahlung eines Honorars beauftragt, die vorhandene methodische Literatur zu meta-epidemiologischen Untersuchungen klinischer Studien zu sichten und im Hinblick auf deren Relevanz für Bewertungen im IQWIG-Abschlußbericht zu *Langwirksame Insulinanaloga zur Behandlung des Diabetes mellitus Typ 2* zu diskutieren.

Statistische Signifikanz

H_0: unterlegen H_1: überlegen



Annahme: höhere Werte der Zielvariablen bedeutet größere Wirksamkeit

Statistische Signifikanz

$$\left[\bar{X}_T - \bar{X}_P - z_{0.975}\,\sigma\sqrt{\frac{2}{n}}, \bar{X}_T - \bar{X}_P + z_{0.975}\,\sigma\sqrt{\frac{2}{n}} \right]$$

$$n = 2(z_{0.975} + z_{0.8})^2 \frac{\sigma^2}{\Delta^2}, \Delta = \mu_T - \mu_P$$

Lehne H_0 ab, falls die untere Grenze des zweiseitigen 95%-KI oberhalb von 0 ist

Biometrische ‘Praxis’

1. **A sample size of 334 in each group will have 80% power to detect a difference in means of 50.000 assuming that the common standard deviation is 230.000 using a two group t-test with a 0.050 two-sided significance level (nQuery)**

2. **A sample size of 334 in each group will have 80% power to detect a minimal important difference in means of 50.000 assuming that the common standard deviation is 230.000 using a two group t-test with a 0.050 two-sided significance level**

Biometrische ‘Praxis’

- **Kontrollierte klinische Studien verfehlen wegen zu geringen Stichprobenumfängen oft ihr Ziel, eine überzeugende Aussage über die Wirksamkeit bzw. Überlegenheit einer Behandlung zu machen.**

- **Statistische Methoden in kontrollierten klinischen Studien führen häufig zu irreführenden und kontroversen Ergebnissen.**

Aktuelle Probleme in kontrollierten klinischen Studien. Die Medizinische Welt (1986), Martin Schumacher

Regulatorische Anforderungen

ICH-E4-Note for Guidance on Dose Response Information to Support Drug Registration (1994)

It should be demonstrated, however, that the lowest dose(s) tested, if these are to be recommended, have a statistically significant and clinically meaningful effect.

Regulatorische Anforderungen

CHMP (2005): Guideline on the Choice of the Non-Inferiority Margin

The next step interpreting a superiority trials is to consider whether the difference from placebo is clinically relevant. This is the clinical judgment stage of the ICH (E10) combination of both statistical reasoning and clinical judgment

Regulatorische Anforderungen

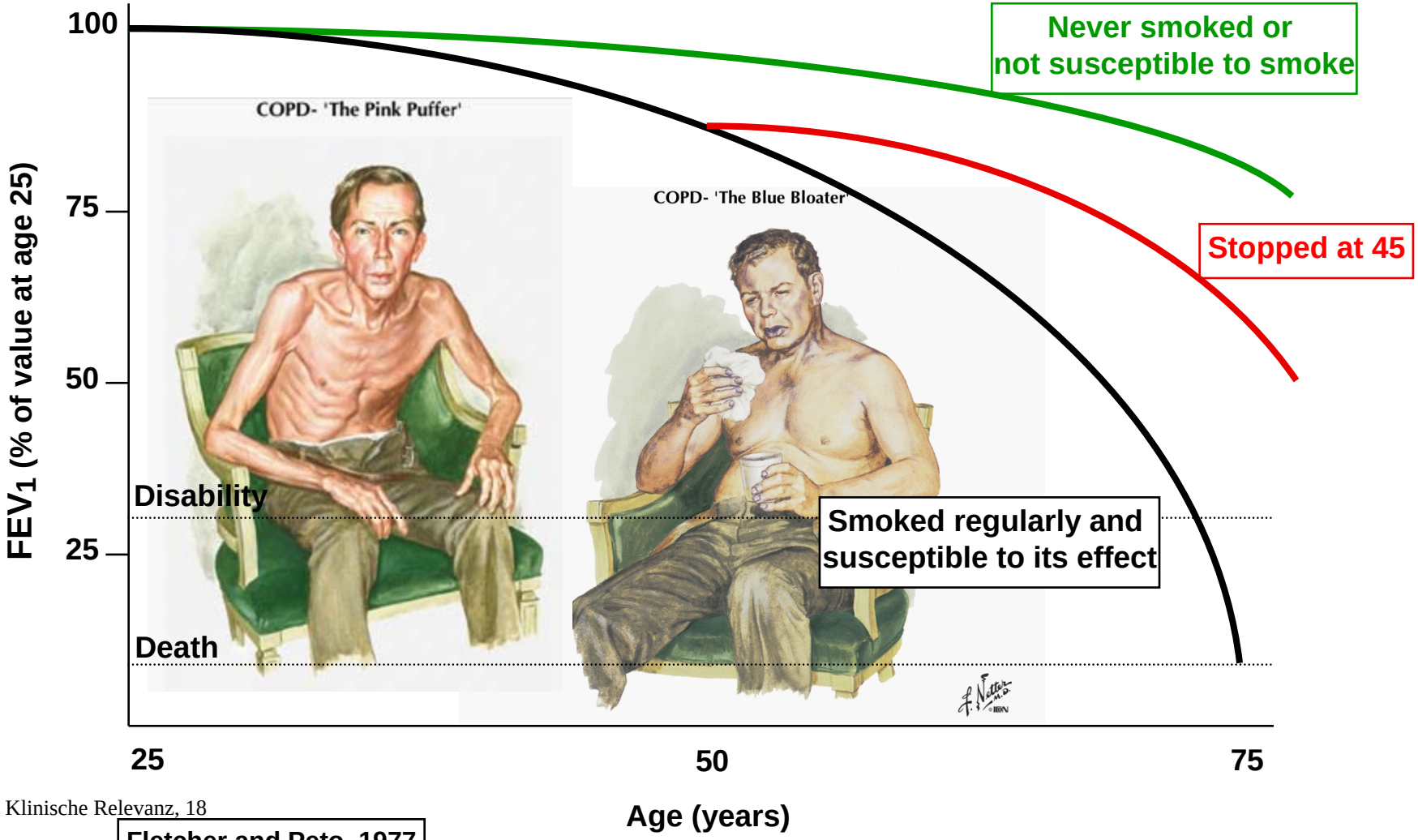
CHMP (2005): Establishing a clinically relevant benefit of the difference between the test product and placebo is accomplished by considering the point estimates of the difference between the test product and placebo and addressing its clinical relevance, either using the original scale or by considering responder rates

Responder Analyse

- **CPMP-Guideline Diabetes Mellitus (2002)**
  - primary variable: glycohaemoglobin (HbA_{1C})
  - primary analysis “should evaluate the difference in evolution from baseline of HbA_{1C}.”
  - “The applicant should also justify the clinical relevance of the effect size observed. One method might be a responder analysis comparing the portion of patients who reached an absolute value of 7%.”

COPD = Chronic Obstructive Pulmonary Disease

Model of Annual Decline in FEV_1



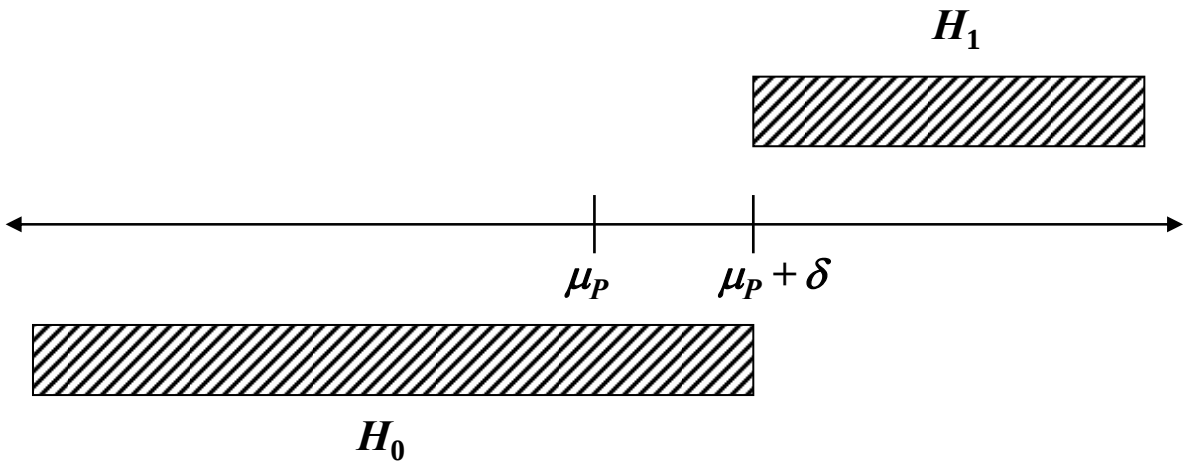
Klinische Relevanz, 18

Fletcher and Peto, 1977

Statistische und Klinische Signifikanz

H_0:nicht relevant überlegen

H_1:relevant überlegen



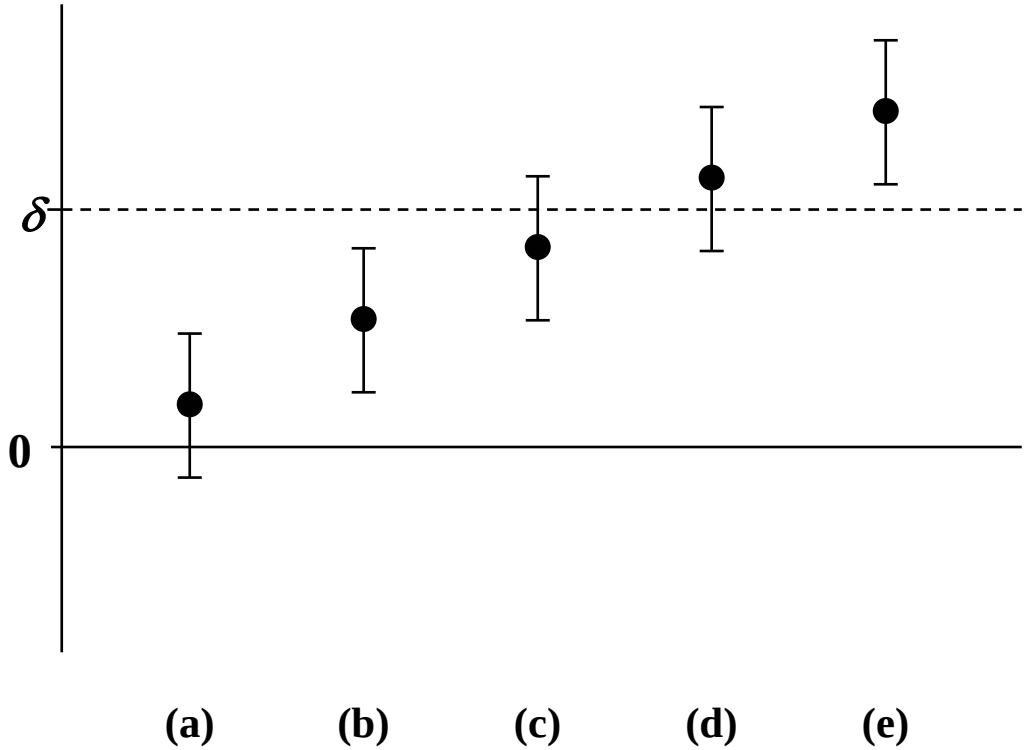
Statistische und Klinische Signifikanz

$$\left[\bar{X}_T - \bar{X}_P - z_{0.975}\sigma\sqrt{\frac{2}{n}}, \bar{X}_T - \bar{X}_P + z_{0.975}\sigma\sqrt{\frac{2}{n}}\right]$$

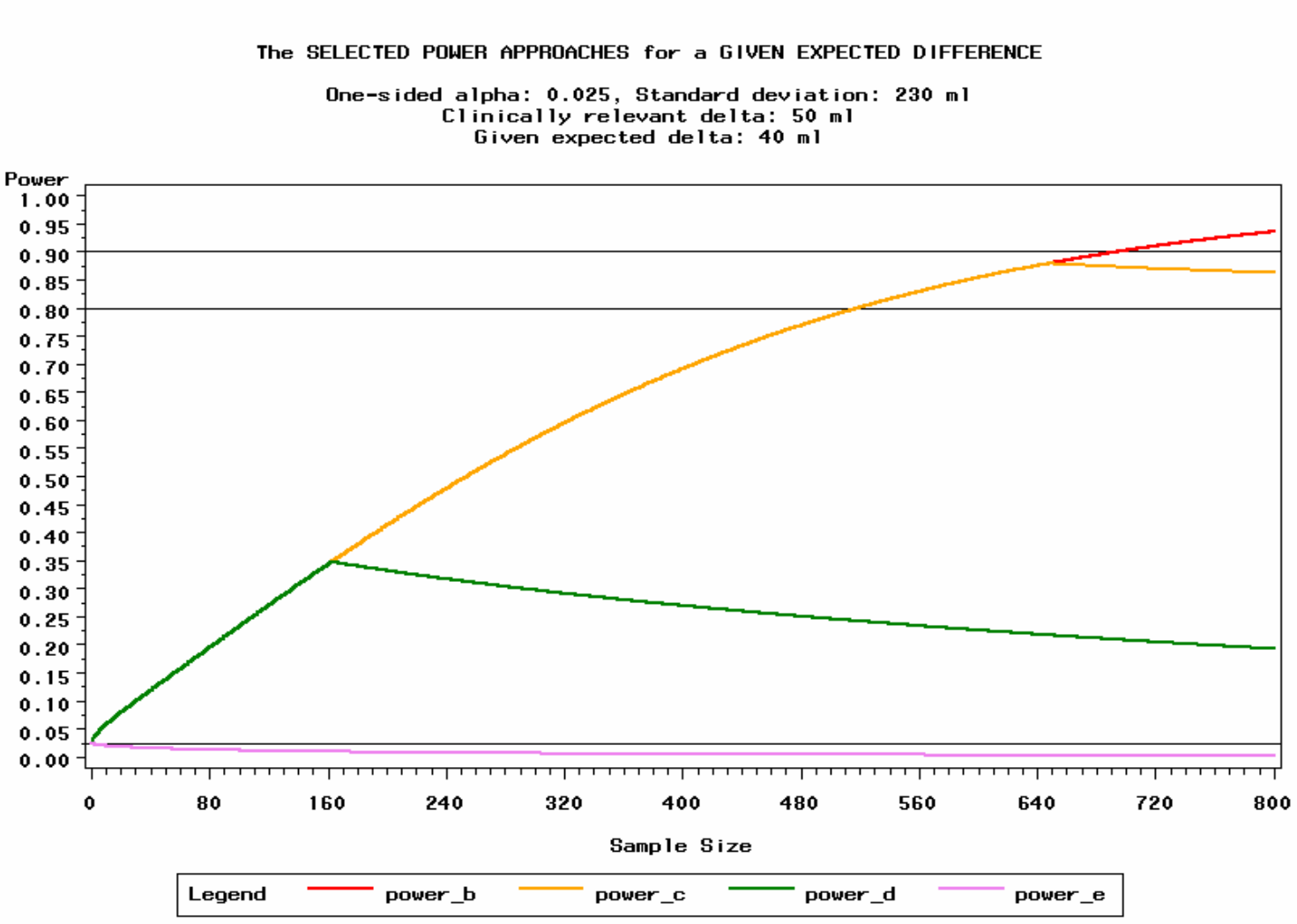
$$n = 2(z_{0.975} + z_{0.8})^2 \frac{\sigma^2}{(\delta - \Delta)^2}, \Delta = \mu_T - \mu_P$$

Lehne H_0 ab, falls die untere Grenze des zweiseitigen 95%-KI oberhalb von δ ist

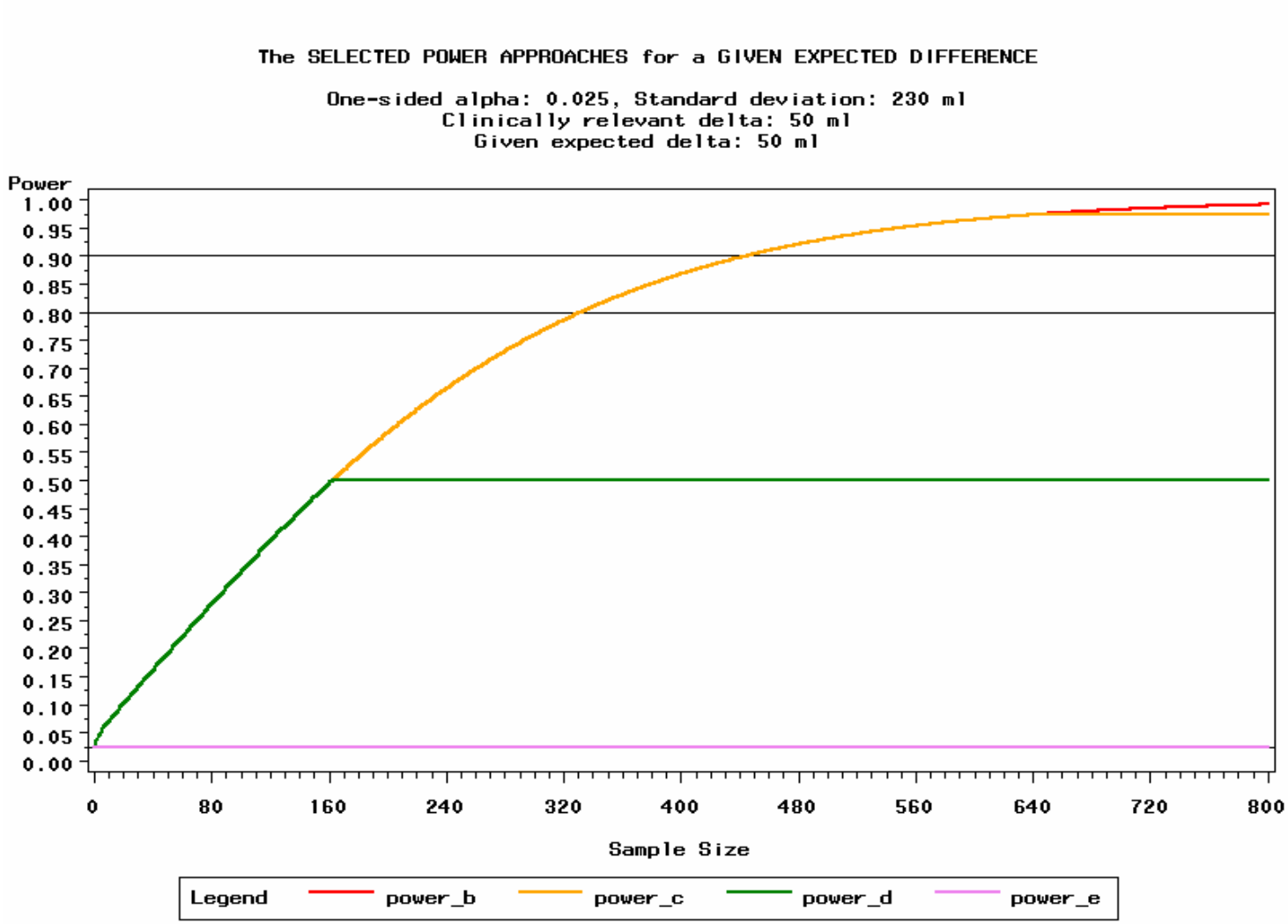
Test – Placebo, zweiseitige 95% Konfidenzintervalle



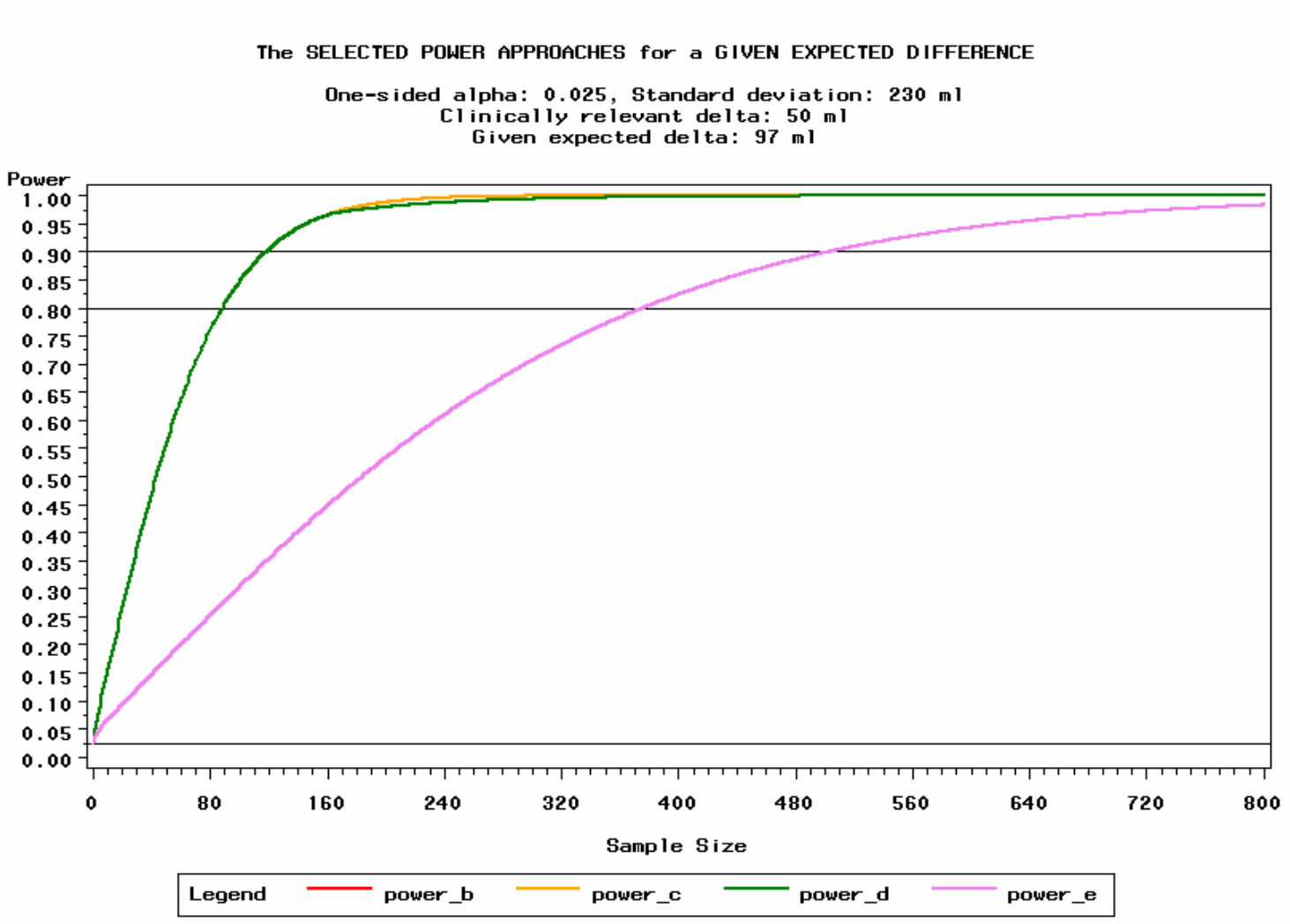
Power für FEV_1



Power für FEV_1



Power für FEV_1



Konklusion

Assessment of clinical relevance by considering point estimates and associated confidence intervals

Meinhard Kieser[1,2,*,†] and Dieter Hauschke[3]

[1] *Department of Biometry, Dr. Willmar Schwabe Pharmaceuticals, Karlsruhe, Germany*
[2] *Medical Biometry Unit, University of Heidelberg, Germany*
[3] *Department of Biometry, Altana Pharma, Konstanz, Germany*

For the proof of efficacy of a new drug in a placebo-controlled clinical trial it is not sufficient merely to demonstrate a statistically significant treatment difference. In recent years, regulatory authorities have strongly recommended assessing additionally whether the observed effect size is also of clinical relevance. This opinion is reflected in various guidelines which are of the utmost importance for the successful approval of a new drug. Clinical relevance can be investigated by responder analyses or by considering the point estimates on the original scale together with the associated confidence intervals. In this paper, we focus on the latter approach and discuss the suitability of different criteria which are commonly applied in medical research.

Keywords: *clinically important difference; clinical relevance; statistical significance; sample size; statistical power*

Konklusion

CHMP (2005) Clinical judgment is applied to assess whether the observed difference from placebo is clinically relevant. The existence of the reference arm (active control) can assist in making this judgment

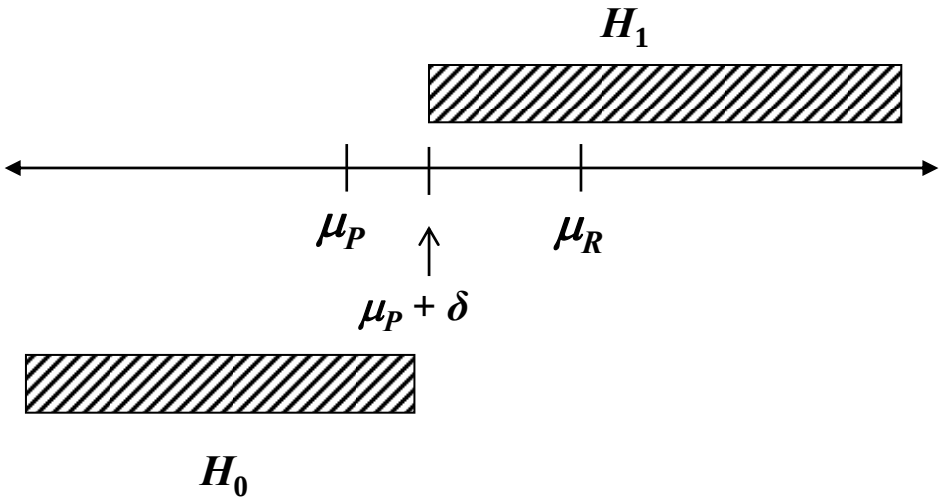
$\Rightarrow$

Hauschke, Pigeot (2006) Proof of Efficacy in the Gold Standard Design (Placebo,Test, Active Control

Gold Standard Design

H_0:nicht relevant überlegen

H_1:relvant überlegen



Gold Standard Design

I.Piegeot & J. Röhmel

- Ulcerative colitis
- Crohn's disease
- Migraine
- Psoriasis
- Depression
- Bipolar Disorders
- Parkinson's Disease
- Allergic Rhino-Conjunctivitis
- Asthma
- Pain
- Schizophrenia
- Multiple Sklerosis

Gold Standard Design

European Medicines Agency
Evaluation of Medicines for Human Use

London, 18 November 2004
CHMP/EWP/2454/02 corr

COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE (CHMP)

GUIDELINE ON CLINICAL INVESTIGATION OF MEDICINAL PRODUCTS INDICATED FOR THE TREATMENT OF PSORIASIS

Gold Standard Design

Parallel group, double-blind, vehicle and active comparator controlled studies are recommended. Only three-arm trials allow comparison of the efficacy and safety of a new agent both with the vehicle and an active comparator, and thereby a proper assessment of benefit/risk ratio. In addition, since response to established topical agents (comparators) is variable, vehicle arm is needed to ascertain assay sensitivity. In designing a non-inferiority study, a non-inferiority margin should be prospectively defined by taking into account both the established efficacy of the vehicle and the efficacy of the active comparator over vehicle (or emollient).

For trials aiming to show superiority of a new drug to the known active treatment, two-arm trials without placebo control are acceptable. If superiority is not shown, non-inferiority can not be claimed due to the lack of a placebo arm as an internal validation.

Gold Standard Design

European Medicines Agency
Evaluation of Medicines for Human Use

London, 26 January 2006
Doc. Ref. CPMP/EWP/252/03 Rev. 1

COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE (CHMP)

DRAFT

GUIDELINE ON CLINICAL MEDICINAL PRODUCTS INTENDED FOR THE TREATMENT OF NEUROPATHIC PAIN

Gold Standard Design

3.2 Therapeutic Confirmatory Studies

Study design

Randomised, double blind, placebo controlled studies are required to establish efficacy in neuropathic pain.

Neuropathic pain is usually present as a chronic situation and the study duration should take this in consideration.

As there is an increasing number of drugs approved for neuropathic pain, in those clinical conditions for which there is an established treatment option a three-arm study (study drug – comparator - placebo) should be provided in order to allow the assessment of comparative efficacy and safety of a new product.

VFA und Biometry und IQWIG

Workshop Meta-Analyses for Drug Assessment

06.09.2010, 55. Annual Meeting of the GMDS

Guido Schwarzer (Working group in systematical reviews), Ralf Bender (IQWIG), Friedhelm Leverkus (Pfizer), Dieter Hauschke (University Freiburg)

Session 1: Incorporating clinical relevance (Dieter Hauschke)

Speaker: Stefanie Thomas, Jürgen Windeler, Joachim Röhmel, Stefan Lange, Armin Koch

Session 2: Indirect Comparisons (Guido Schwarzer)

Speaker: Sybille Sturtz, Ralf Bender, Peter Jüni, Georgia Salanti, Gerd Antes

Literature

Victor (1987) On clinically relevant difference and shifted null hypotheses. *Methods of Information in Medicine* 26, 110-116

Kieser, Röhmel, Friede (2004) Power and sample size determination when assessing the clinical relevance of trial results by responder analysis. *Statistics in Medicine* 23, 3287-3305

Kieser, Hauschke (2005) Assessment of clinical relevance by considering point estimates and associated confidence intervals. *Pharmaceutical Statistics* 4, 101-107

Hauschke, Pigeot (2006) Establishing efficacy of a new experimental treatment in the gold standard design. *Biometrical Journal* 6, 782-786