FORSCHUNGSBERICHTE DER ABTEILUNG MEDIZINISCHE BIOMETRIE, UNIVERSITÄT HEIDELBERG

Nr. 49

META-ANALYTISCHE METHODEN FÜR ÄQUIVALENZFRAGESTELLUNGEN

Februar 2005

INSTITUT FÜR MEDIZINISCHE BIOMETRIE UND INFORMATIK

RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG

Forschungsberichte der
Abteilung Medizinische Biometrie, Universität Heidelberg

Nr. 49

Meta-analytische Methoden für Äquivalenzfragestellungen

STEFFEN WITTE

Institut für Medizinische Biometrie und Informatik (IMBI)

der Medizinischen Fakultät der Universität Heidelberg

Heidelberg, Februar 2005

Impressum:

Reihentitel: Forschungsberichte der Abteilung Medizinische Biometrie,
Universität Heidelberg

Herausgeber: Prof. Dr. Norbert Victor

Anschrift: Im Neuenheimer Feld 305, 69120 Heidelberg

Druck: Hausdruckerei der Ruprecht-Karls-Universität Heidelberg

elektronischer Bezug: http://www.biometrie.uni-heidelberg.de

ISSN: 1619-5833

Aus dem Institut für Medizinische Biometrie und Informatik
Abteilung Medizinische Biometrie
(Leiter: Prof. Dr. N. Victor)

Meta-analytische Methoden für Äquivalenzfragestellungen

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Steffen Witte, geb. Ballerstedt
aus
Hameln

April 2003

Dekan: Prof. Dr. Dr. h.c. H.-G. Sonntag

Referent: Prof. Dr. N. Victor

Meinen Eltern,
den Ballerstedts

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

Meta-Analysen sind eine etablierte Methode, um Studienergebnisse zusammenzufassen. Sie werden im Rahmen der medizinischen Forschung zunehmend eingesetzt (Abschnitt 1.3). Das Schlagwort „meta-analysis“ wurde 1989 in MEDLINE aufgenommen. Im Jahre 1990 wurden 273 Meta-Analysen registriert. Zehn Jahre später, im Jahr 2001, waren es bereits 878 (Tabelle 11 auf Seite 99). Meta-Analysen werden vor allem in systematischen Reviews eingesetzt, die im Hinblick auf eine evidenzbasierte Medizin von großem Interesse sind: Systematische Reviews von randomisierten, kontrollierten klinischen Studien werden von der Cochrane Collaboration mit dem größten Evidenzlevel versehen.

Bei fast allen Arbeiten geht es um die Überlegenheit einer Therapie gegenüber einer anderen beziehungsweise gegenüber einer Placebo-Behandlung, so genannte *Überlegenheitsstudien*. Darüberhinaus werden Studiendesigns, bei denen es um den Vergleich zweier aktiver Therapien geht, immer wichtiger. Nicht die Überlegenheit soll dabei geprüft werden, sondern die Gleichartigkeit (Äquivalenz bzw. Nicht-Unterlegenheit) einer Therapie verglichen mit einer anderen (so genannte *Äquivalenzstudien*, Abschnitt 1.5). Diese Studien werden notwendig, wenn etablierte und anerkannte Therapien in einem Indikationsgebiet existieren, deren Wirksamkeit bereits nachgewiesen wurde und davon ausgegangen werden kann, dass diese noch immer gilt („constancy assumption“).

Dass sowohl Meta-Analysen als auch Äquivalenzstudien von steigendem Interesse sind, verdeutlicht Abbildung 1. Dort werden die zum Bezugsjahr 1990 relativen Veränderungen von allen MEDLINE-Einträgen, Meta-Analysen sowie Äquivalenzstudien dargestellt (Tabelle 11 auf Seite 99). Dabei muss darauf hingewiesen werden, dass die tatsächliche Anzahl der Äquivalenzstudien bei weitem höher ist, als hier dargestellt. Denn entsprechende Schlüsselworte gibt es in MEDLINE nicht und auch in Fachkreisen ist keine Suchprozedur bekannt, die alle Äquivalenzstudien finden würde. Dennoch können die hier dargestellten relativen Häufigkeiten einen Eindruck vermitteln, dass die Anzahl an Äquivalenzstudien in den letzten zehn Jahren deutlich zugenommen hat.

Sollen Therapien mittels Meta-Analysen beurteilt werden, ist der Aspekt der Äquivalenz in der Beurteilung der Wirksamkeit und Sicherheit von Bedeutung. Soll die Wirkung einer Therapie beurteilt werden, finden sich häufig keine oder nur qualitativ minderwärti-

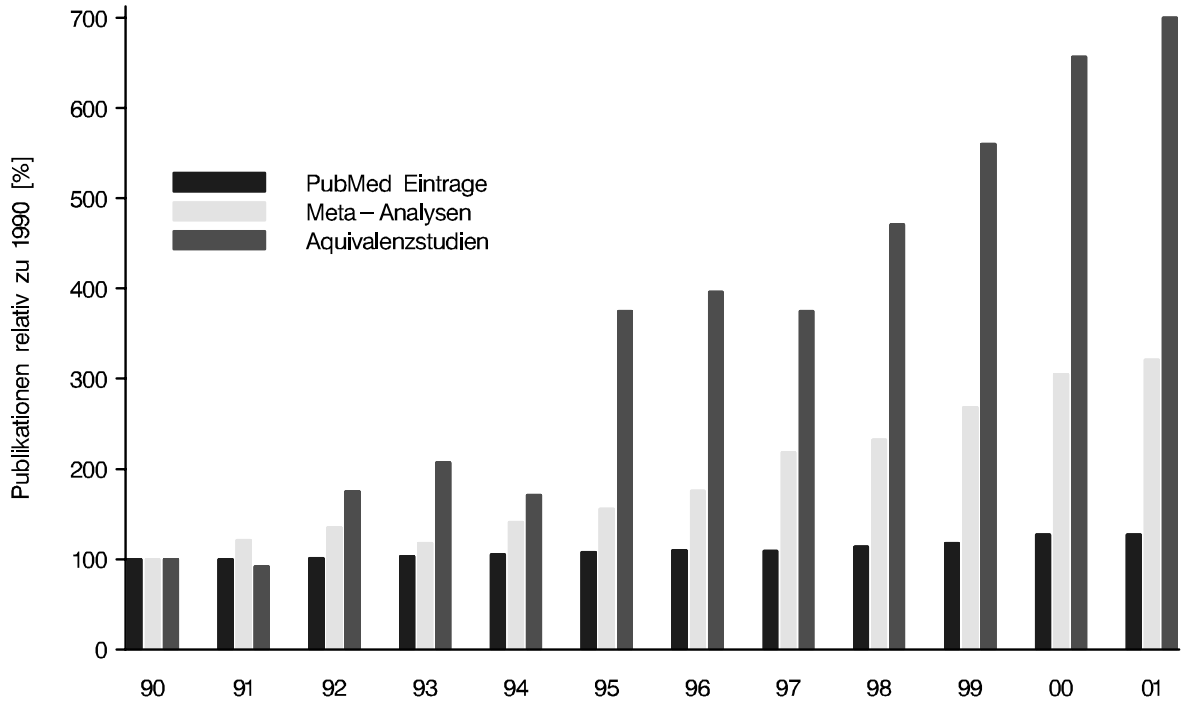


Abbildung 1: Relative Anzahl von MEDLINE-Einträgen, Meta-Analysen, Äquivalenzstudien bezogen auf 1990

ge Placebo-kontrollierte Studien. Um die Wirksamkeit trotzdem beurteilen zu können, können Studien mit aktiven Kontrollen herangezogen werden. Eine Einzelstudie wird jedoch häufig eine zu geringe Power haben (Tabelle 2 auf Seite 28), um *allein* eine angemessene Äquivalenz zu zeigen. Andererseits liefern die Einzelstudien gegebenenfalls widersprüchliche Aussagen. Hier ist die Meta-Analyse das geeignete Mittel, um die Evidenz der zu untersuchenden Therapie adäquat zu beurteilen. Andererseits haben viele Studien das Ziel zu zeigen, dass eine Therapie wirksamer ist als eine andere. Die Sicherheitsparameter werden jedoch nur deskriptiv untersucht, um eine Vergleichbarkeit der Sicherheitsprofile zu postulieren. Soll aber die Idee der Äquivalenztests auch auf die Sicherheitsparameter angewendet werden, so kann eine Meta-Analyse die geeignete Methode sein, um durch die Zusammenfassung mehrerer Studien die Power zu erhöhen. Die Fallzahlplanung der Einzelstudie ist nämlich auf die Hauptzielgröße (Wirksamkeitsparameter) ausgelegt und *nicht* auf eine Äquivalenz der Sicherheitsparameter, so dass diese Fallzahl in der Regel für eine derartige Analyse nicht ausreichen würde.

Es werden also zunehmend Äquivalenzstudien durchgeführt. Dies impliziert, dass auch in systematische Reviews vermehrt Äquivalenzstudien eingehen werden. Diese verwenden

meta-analytische Techniken, aber methodische Arbeiten über die Durchführung solcher Meta-Analysen, die einen Äquivalenznachweis zum Ziel haben, gibt es bisher nicht. In dieser Arbeit soll daher die Situation verdeutlicht, Probleme aufgezeigt und Lösungsansätze vorgeschlagen werden.

1.2 Klinische Beispiele

Im Folgenden sind Beispiele aufgeführt, die Äquivalenzhypothesen untersuchen und meta-analytische Methoden verwenden.

Oxaceprol-Beispiel: Im Rahmen einer Meta-Analyse wurde die Wirksamkeit einer symptomatischen Behandlung von Arthrosen mit Oxaceprol untersucht [115]. Den Anstoß für diese Analyse gab die Arzneimittelkommission der deutschen Ärzteschaft, die eine derartige Analyse als Grundlage für eine Bewertung therapeutischer Leitlinien benötigte [64]. Es wurden jedoch keine randomisierten, *placebo*-kontrollierten klinischen Studien gefunden, die Aufschluss über die Wirksamkeit von Oxaceprol liefern. Stattdessen lagen zwei Äquivalenzstudien Oxaceprol versus Diclofenac und zwei Überlegenheitsstudien Oxaceprol versus Ibuprofen (bei Hüft- und Kniearthrose) vor. Tabelle 1 gibt einen Überblick der Studiencharakteristika. Dabei sprechen positive Werte für Oxaceprol.

Tabelle 1: Studiencharakteristika im Oxaceprol-Beispiel

Studie	A	N	Pop.	ZG	Schätzer und KI	δ_i'	δ_i
Bauer et al. 1999	Diclo	124	PP	LI	0.167 [-0.19, 0.52]	2	0.48
Herrmann et al. 2000	Diclo	219	PP(ITT)	LI	-0.078 [-0.34, 0.19]	2	0.53
Hildebrandt 1995	Ibu	64	ITT	LI	-0.095 [-0.59, 0.40]	-	-
Vagt et al. 1990	Ibu	60	ITT	TPS	-0.444 [-0.95, 0.06]	-	-

A = Medikation der aktiven Kontrollgruppe, N = Gesamtstichprobengröße, Pop. = Auswertungspopulation, ZG = Zielgröße = standardisierte Differenz (positive Werte sprechen für Oxaceprol), KI = 95% Konfidenzintervall, δ_i' = Äquivalenzgrenze auf der Skala des Lequesne-Index, δ_i = Äquivalenzgrenze auf der Skala der standardisierten Differenz, Diclo = Diclofenac, Ibu = Ibuprofen, PP = per Protokoll-Analyse, ITT = intention to treat-Analyse, LI = Lequesne Index, TPS = total pain score

Die Effekte und Konfidenzintervalle sind 0.016 [-0.217, 0.249] für die Wirkung von Oxaceprol versus Diclofenac, -0.264 [-0.616, 0.088] von Oxaceprol versus Ibuprofen und -0.069 [-0.283, 0.144] von Oxaceprol versus Diclofenac *oder* Ibuprofen. Die gesamten Ergebnisse können im Abschlussbericht der Studie nachgelesen werden [114].

In Abschnitt 3.2 auf Seite 40 wird dieses Bespiel für die Auswirkungen der Wahl der Äquivalenzgrenze auf den p-Wert der Äquivalenzuntersuchung verwendet, im Abschnitt 3.3 auf Seite 45 für die Transformation von Äquivalenzgrenzen und im Abschnitt 4.4.3 auf Seite 67 für Sensitivitätsanalysen.

WOMAC-Beispiel: Die symptomatische Arthrosebehandlung ist ein generelles Beispiel für aktiv kontrollierte Studien. Eine aktuelle Meta-Analyse von Deeks et al. untersucht den selektiven COX-2-Hemmer Celecoxib bei Arthrose und rheumatoider Arthritis [27]. Um die Wirksamkeit der Therapien zu beurteilen, wird häufig der WOMAC verwendet, Western Ontario and McMaster Universities Osteoarthritis Index. Diese Zielgröße umfasst eine Schmerzskala, eine Skala für die Steifigkeit und eine für die Funktionalität, sowie einen zusammengefassten Score. Dieses Beispiel wird im Abschnitt 3.5 ausführlich diskutiert.

Asthma-Beispiel: Ebbutt und Frith nennen elf aktiv kontrollierte, doppelt blinde, randomisierte Studien A-K, die ein neues Inhalationsspray mit bisherigen Inhalatoren zur Therapie des Asthma bronchiale vergleichen [31]. Hauptzielgröße ist die 'peak expiratory flow rate' (PEF), hierbei werden Äquivalenzgrenzen von ± 15 l/min angegeben. Da bei allen elf Studien das jeweilige Konfidenzintervall innerhalb der Äquivalenzgrenzen liegt, ist eigentlich keine Meta-Analyse notwendig. Diese Studien werden trotzdem als Beispiel angeführt, da sowohl Ergebnisse der PP-Analyse als auch der ITT-Analyse gegeben sind (Abschnitt 4.4).

Einzelne Reviews: Im Folgenden werden Reviews aufgelistet, die mittels meta-analytischen Techniken Therapien mit aktiven Kontrollen vergleichen. Als Suchbegriffe in MEDLINE beim DIMDI (www.dimdi.de) wurden die Stichworte „effective as“ und „Meta-Analysis“ verwendet. Die Suche lieferte 88 Treffer (Stand: 22. Oktober 2002). Sechs davon wurden als Beispiele ausgewählt. Die ersten drei Beispiele schließen erst aufgrund der Ergebnisse auf Äquivalenz, formulieren dies aber in der Zielsetzung noch nicht. Die letzten drei Beispiele beschreiben schon in der Hypothese die Äquivalenzfrage. Die Beispiele werden kurz in Abschnitt 4.4.1 wieder aufgegriffen:

„Conclusions: We conclude that single-dose amoxicillin is inadequate therapy for uncomplicated cystitis of childhood. Three days of trimethoprim-sulfamethoxazole therapy appears to be *as effective as* conventional length courses of the drug“ (Tran et al. 2001, [96])

„Conclusion: Asymptomatic DVT may be regarded as a reliable surrogate endpoint for clinical outcome in studies investigating thromboprophylaxis in general surgery. LMWH seems to be *as effective and safe as* UFH.“ (Mismetti et al. 2001, [70])

„Results: In five randomized clinical trials comprising 147 patients, enteral nutrition was *as effective as* corticosteroids at inducing a remission (RR = 0.95 (95% confidence interval 0.67, 1.34))“ (Heuschkel et al. 2000, [50])

„Objectives: [...] Is local anaesthesia a safe and effecitve alternative to general anaesthesia? [...] Is day-case *as safe and effective as* inpatient surgery? Is synchronous bilateral hernia repair *as safe and effective as* delayed repair?“ (Cheeks et al. 1998, [21])

„Background: The main hypothesis was that a model with a lower number of antenatal visits, with or without goal-oriented components, would be *as effective as* the standard antenatal-care model in terms of clinical outcomes, perceived satisfaction, and costs.“ (Carroli et al. 2001, [19])

„Introduction: We conducted a systematic review of all published and unpublished trials to determine if celecoxib is *as effective as* other NSAIDs for the treatment of osteoarthritis and rheumatoid arthritis [...].“ (Deeks et al. 2002, [27])

Viele der gefundenen Meta-Analysen kommen aus dem Indikationsbereich der bakteriellen Infektionen. Für diese Erkrankungen hat die EMEA eine spezielle Guideline herausgegeben [34], die speziell auf die Auswertungspopulation eingeht (Abschnitt 4.2). Die Suchstrategie (DT="Meta-Analysis" AND CT DOWN "Bacterial Infections" AND CT="Human" AND effective as) in MEDLINE beim DIMDI dient nicht dazu, *alle* Meta-Analysen über bakterielle Infektionen zu erhalten, die einer Äquivalenzfragestellung nachgehen. Vielmehr sollen einige Beispiele bezüglich eines Indikationsgebietes gefunden werden. Es wurden 13 Treffer erzielt, darunter ist eine Arbeit, die keine Äquivalenzfrage versucht zu beantworten (Stand: 22. Oktober 2002). Unter den 13 Arbeiten sind zwei Literaturübersichtsarbeiten und elf Meta-Analysen. In den meisten Publikationen wird das Ziel der Studie neutral formuliert („... to compare the efficacy and toxicity ...“). Nur zwei Studien formulieren die Zielsetzung wirklich als eine Untersuchung zur Äquivalenz [67, 103]. In einer Publikation werden die Konfidenzintervalle im Sinne einer Äquivalenzprüfung interpretiert, aber es wird trotz breiter Konfidenzintervalle auf eine gleichartige Wirkung geschlossen [103]. In einer weiteren Meta-Analyse wird – obwohl das Ergebnis als „marginal significant“ bezeichnet wird – auf eine etwa gleichartige Wirksamkeit geschlossen („monotherapy was as effective as combination therapy“) [43]. Hier soll eigentlich eine Nicht-Unterlegenheit gezeigt werden, explizit vermerkt wird dies jedoch nicht. Alle übrigen schließen von nicht signifikanten Vergleichen auf die Vergleichbarkeit der Therapien, ohne wenigstens die Konfidenzintervalle entsprechend zu interpretieren. Die Studien werden als Beispiel im Abschnitt 3.2 verwendet.

1.3 Meta-Analysen

Ziele Die Meta-Analyse ist eine Methode, um Ergebnisse aus mindestens zwei bereits durchgeführten Studien zur gleichen medizinischen Frage zusammenzufassen und somit die vorhandene Evidenz zu bündeln.

Im Rahmen der Evidenz basierten Medizin (EbM) bekam die Meta-Analyse eine größere Bedeutung. In der EbM soll individuelle klinische Expertise mit der besten externen Evidenz zusammen die Grundlage medizinischer Entscheidungen sein [81]. Mit systematischen Reviews wird die relevante externe Evidenz zusammengetragen und bewertet. Die Meta-Analyse stellt dabei die Methode für die quantitative Zusammenfassung bereit, obwohl nicht in jedem systematischen Review eine Meta-Analyse möglich ist (beispielsweise wenn nicht genügend Daten vorliegen oder aufgrund erheblicher unerklärter Inkonsistenzen zwischen den Studien). Somit sind Meta-Analysen häufig die Grundlage für Leitlinien und Therapieempfehlungen, die auf der EbM basieren. Meta-Analysen verfolgen insbesondere folgende Ziele:

- Präzisere Schätzung des Behandlungseffekts (Wirkung) und Beurteilung der Wirksamkeit
- Differenzierte Aussage über Subgruppen, die aufgrund niedriger Fallzahlen in Einzelstudien nicht angemessen analysiert werden können
- Beurteilung weiterer Parameter, die wegen der geringen Fallzahl in den Einzelstudien nicht angemessen analysiert werden können (Beispiel: Die Wirksamkeit soll aufgrund der Überlebenszeit beurteilt werden. In den Einzelstudien wurde diese jedoch nur als Nebenzielparameter erfasst.)
- Klärung der Evidenz bei widersprüchlichen Aussagen der Einzelstudien
- Untersuchung und Erklärung der heterogenen Ergebnisse der Einzelstudien
- Verallgemeinerung der bekannten Evidenz auf eine größere Population (Grundgesamtheit)
- Untersuchung von Sicherheitsparametern, insbesondere von seltenen unerwünschten Ereignissen

- Unterstützung des Zulassungsprozesses von Medikamenten (kann eine Meta-Analyse ggf. eine zweite pivotale Phase III Studie ersetzen?)
- Informationsbereitstellung für die Planung einer neuen Studie (zum Beispiel Hypothesen-Generierung oder Effektschätzung für die Definition einer Äquivalenzgrenze für eine Äquivalenzstudie)

Das steigende Interesse an der EbM und die Vielzahl obiger Ziele lässt die Zahl der publizierten Meta-Analysen Jahr für Jahr steigen.

MAP und MAL Um Studienergebnisse quantitativ zusammenzufassen, werden zwei grundlegend unterschiedliche Konzepte verwendet. Entweder die individuellen Studiendaten liegen vor (MAP = meta-analysis of individual patient data) oder die Meta-Analyse stützt sich lediglich auf die publizierten Daten (MAL = meta-analysis of the literature) [89]. In dieser Arbeit wird ausschließlich die zweite weitaus häufigere Form betrachtet. Die Meta-Analyse bleibt jedoch eine Beobachtungsstudie und ist kein Experiment, denn es werden lediglich individuelle oder zusammengefasste Studienergebnisse *beobachtet* [100].

FEM und REM Eine weitere Einteilung betrifft das zugrunde gelegte statistische Modell. Wird der gesuchte Effekt ($\theta \in \mathbb{R}$) als fest angenommen und alle Studien liefern eine Schätzung für diesen Effekt ($\hat{\theta}_i$), so handelt es sich um ein Modell mit festen Effekten (FEM = fixed effects model)

$$\hat{\theta}_i = \theta + \varepsilon_i \sim N(\theta, w_i^{-1}) \tag{1}$$

mit $\varepsilon_i \sim N(0, w_i^{-1})$. Dabei wird die Varianz w_i^{-1} als bekannt vorausgesetzt, obwohl sie in der Praxis in den Einzelstudien geschätzt wird. Wird aber der wahre Effekt einer Studie ($\theta_i = \theta + \nu_i$) selbst als eine Zufallsvariable aufgefasst, so handelt es sich um ein Modell mit zufälligen Effekten (REM = random effects model)

$$\hat{\theta}_i = \theta + \nu_i + \varepsilon_i \sim N(\theta, w_i^{-1} + \tau^2) \tag{2}$$

mit $\nu_i \sim N(0, \tau^2)$ und unbekannter Zwischenstudienvarianz $\tau^2 \in \mathbb{R}$. Man ist üblicherweise an dem Erwartungswert dieser zufälligen Effekte interessiert ($\theta = E(\theta_i)$). Jede Studie liefert eine Schätzung ($\hat{\theta}_i$) für den Effekt (θ_i). Somit gibt es zusätzlich zur Varianz innerhalb einer Studie die Varianz zwischen den Studien (τ^2). Die Konfidenzintervalle sind für θ im REM breiter als im FEM, falls Heterogenität vorliegt. Statistiken, die auf

dem REM beruhen, können auch im FEM verwendet werden, nicht aber umgekehrt. Die Probleme der entsprechenden Statistiken in den beiden genannten Modellen sind genauer bei Ziegler et al. nachzulesen [116]. Ist die Heterogenität in der Meta-Analyse sehr groß und kann durch unterschiedliche Studiendesigns/-bedingungen nicht erklärt werden, sollten auch die Statistiken, die auf dem REM basieren, nicht verwendet werden. In diesem Fall ist generell von einer Meta-Analyse abzuraten.

1.4 Hypothesen und Interpretationen

Zwei Typen Eine häufige Fragestellung in der medizinischen Forschung ist die der therapeutischen Wirksamkeit einer Behandlung bei einer bestimmten Indikation. Klinische Therapiestudien lassen sich in Überlegenheits- bzw. Äquivalenzstudien klassifizieren. Die klassische Überlegenheitsstudie wird üblicherweise in der therapeutischen Forschung angewendet. Ziel dieser Studien ist es zu zeigen, dass die Experimentalgruppe gegenüber der Kontrollgruppe unterschiedlich oder überlegen ist. In der Regel ist die Kontrollgruppe eine Placebokontrolle. In Äquivalenzstudien wird die Frage entgegengesetzt betrachtet – von Interesse ist nicht die Unterschiedlichkeit, sondern die „Gleichheit“ der beiden Therapien. Dies ist zum Beispiel bei der Zulassung eines Generikums von Interesse. „Gleichheit“ kann aber mit empirischer Forschung nicht gezeigt werden; deshalb wird versucht zu zeigen, dass sich die eine Therapie „nicht wesentlich“ von der anderen unterscheidet (Äquivalenz, Abschnitt 1.5). In Abschnitt 3 wird die Definition von „nicht wesentlich“ diskutiert. Je nach Studientyp (Überlegenheitsstudie, Äquivalenzstudie) sind die Hypothesen unterschiedlich zu formulieren.

Hypothesen Abhängig von der Fragestellung ist die Hypothesenbildung für statistische Tests derart zu wählen, dass sich die klinische Arbeitshypothese in der statistischen Alternativhypothese H_1 widerspiegelt. Die statistische Nullhypothese H_0 entspricht dem Komplement. Abbildung 2 zeigt mögliche statistische Nullhypothesen (für einen Parameter), die durch ▬ gekennzeichnet sind. Hierbei geht es immer um den Vergleich zweier Therapien, wobei die Null als Gleichheit der beiden Therapieverfahren zu verstehen ist. Dies trifft auf viele übliche Effekte wie Erwartungswertdifferenz, logarithmiertes Odds Ratio oder relatives Risiko, Risikodifferenz, logarithmierter Erwartungswertquotient zu. Ein Effekt kleiner als Null wird als Unterlegenheit der zu untersuchenden Therapie gegenüber der Kontrolltherapie interpretiert. Die Äquivalenzgrenze wird hierbei mit dem Buchstaben δ bezeichnet. Der Äquivalenzbereich wird durch $-\delta$ und δ beschrieben (Abschnitt 3).

Je nach Wahl des Effektes könnte die Gleichheit der beiden Therapien auch bei Eins liegen wie bei dem Odds Ratio, relativen Risiko oder dem Quotienten aus Erwartungswerten. Bei Bioverfügbarkeitsstudien ist beispielsweise durch die logarithmische Transformation häufig das Verhältnis von geometrischen Mittelwerten der Schätzer für den Therapieeffekt. In diesen Fällen können analoge Überlegungen angestellt und eine Darstellung wie in Abbildung 2 erzeugt werden.

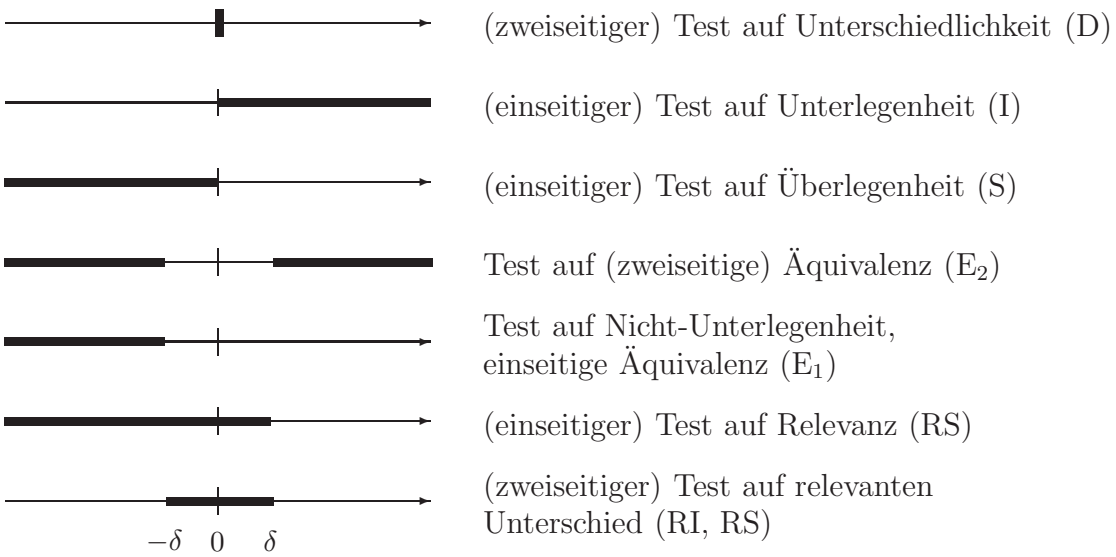


Abbildung 2: Mögliche Nullhypothesen zum statistischen Testen

Terminologie In dieser Arbeit wird einerseits die Hypothese der Nicht-Unterlegenheit und andererseits der zweiseitigen oder echten Äquivalenz betrachtet. Mit Äquivalenz wird hier die Nicht-Unterlegenheit (einseitige Äquivalenz) *oder* die zweiseitige Äquivalenz bezeichnet. Die Arbeitshypothese einer Nicht-Unterlegenheitsstudie lässt sich mit „E nicht relevant schlechter als A“ umschreiben, wobei E die Experimentalgruppe und A die aktive Kontrolle bezeichnet. Ist θ^{EA} der Effekt von E gegenüber A und δ die Nicht-Unterlegenheitsgrenze, so ist die statistische Hypothese

$$H_0^{EA} : \theta^{EA} \leq -\delta \quad \text{vs.} \quad H_1^{EA} : \theta^{EA} > -\delta \tag{3}$$

und speziell mit δ_{clin} (Abschnitt 3.1.1) ist

$$H_0^{clin} : \theta^{EA} \leq -\delta_{clin} \quad \text{vs.} \quad H_1^{clin} : \theta^{EA} > -\delta_{clin}. \tag{4}$$

Mit den Effekten θ^{AP} (Effekt von A gegenüber P, Placebogruppe) und θ^{EP} (Effekt von E gegenüber P) kann man folgende einseitige Hypothese $H_0^{EP}(\lambda)$ mit $\lambda \geq 0$ formulieren. Diese untersucht, ob die Wirkung in der Experimentalgruppe mindestens einen Anteil der Wirkung der aktiven Kontrolle erreicht:

$$H_0^{EP}(\lambda) : \theta^{EP} \leq \lambda\, \theta^{AP} \quad \text{vs.} \quad H_1^{EP}(\lambda) : \theta^{EP} > \lambda\, \theta^{AP} \tag{5}$$

Das ist mit $\theta^{EP} = \theta^{AP} + \theta^{EA}$ äquivalent zu der Schreibweise

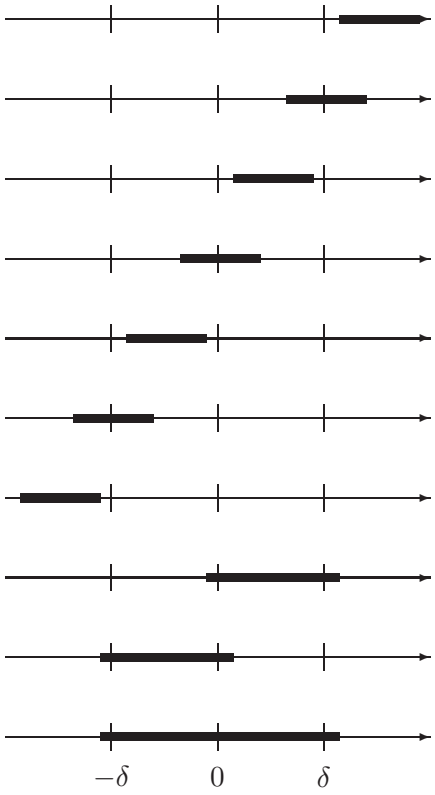
$$H_0^{EP}(\lambda) : \theta^{EA} + (1-\lambda)\, \theta^{AP} \leq 0 \quad \text{vs.} \quad H_1^{EP}(\lambda) : \theta^{EA} + (1-\lambda)\, \theta^{AP} > 0.$$

Mit $\lambda = 1$ ist $H_0^{EP}(\lambda)$ die Überlegenheitshypothese von E gegenüber A und mit $\lambda = 0$ die Überlegenheitshypothese von E gegenüber P und sei mit H_0^{EP} bezeichnet.

$$H_0^{EP} : \theta^{EP} \leq 0 \quad \text{vs.} \quad H_1^{EP} : \theta^{EP} > 0 \tag{6}$$

Mit $\lambda > 1$ könnte man auch eine Hypothese zur relevanten Überlegenheit von E gegenüber A formulieren. Der Vorteil dieser Schreibweise ist zum einen die Allgemeinheit durch die Wahlmöglichkeit von λ und zum anderen, dass die Effekte θ^{EA} und θ^{AP} auch ohne eine Placebo-kontrollierte Studie (E vs. P) geschätzt werden können.

Interpretationen Je nach Fragestellung ist die entsprechende Hypothese zu bilden und der dazugehörige statistische Test auszuwählen. Aufgrund der Äquivalenz von statistischen Testentscheidungen und der Interpretation von Konfidenzintervallen werden die möglichen Intervalle für den entsprechenden Parameter dargestellt und interpretiert. Abbildung 3 zeigt schematisch die möglichen zweiseitigen Konfidenzintervalle, die durch ▬ gekennzeichnet sind [38, 58, 99]. (1-α)-Konfidenzintervalle bedeuten, dass der wahre Effekt mit einer Wahrscheinlichkeit von (1-α) überdeckt wird. (1-α) wird daher als Überdeckungswahrscheinlichkeit (coverage probability) bezeichnet. Ein Überdecken der Null bedeutet ein nicht signifikantes Ergebnis des Tests auf Unterschiedlichkeit ($p > \alpha$). In den Spalten direkt neben den Intervallen sind für die möglichen Tests (RS, S, E_1, E_2, D, I, RI) die entsprechenden Ergebnisse darstellt, wobei jede Spalte einen statistischen Test mit der dazugehörigen Hypothese widerspiegelt. Die Nicht-Unterlegenheit (E_1) kann in sechs von den zehn Fällen aufgrund eines signifikanten Ergebnisses postuliert werden.



RS	S	E_1	E_2	D	I	RI	Interpretation
s	s	s	ns	s	ns	ns	relevante Überlegenheit
ns	s	s	ns	s	ns	ns	sig. Überlegenheit
ns	s	s	s	s	ns	ns	sig. Überlegenheit und Äquivalenz
ns	ns	s	s	ns	ns	ns	Äquivalenz
ns	ns	s	s	s	s	ns	sig. Unterlegenheit und Äquivalenz
ns	ns	ns	ns	s	s	ns	sig. Unterlegenheit
ns	ns	ns	ns	s	s	s	relevante Unterlegenheit
ns	ns	s	ns	ns	ns	ns	Nicht-Unterlegenheit
ns	ns	ns	ns	ns	ns	ns	Unsicherheit
ns	ns	ns	ns	ns	ns	ns	Unsicherheit

Abbildung 3: Interpretation von Konfidenzintervallen, RS = relevante Überlegenheit, S = (statistische) Überlegenheit, E_1 = Nicht-Unterlegenheit, E_2 = zweiseitige Äquivalenz, D = Unterschiedlichkeit, I = (statistische) Unterlegenheit, RI = relevante Unterlegenheit, ns = nicht signifikant und s = signifikant.

Die Interpretation der Studienergebnisse basiert auf den zuvor definierten Hypothesen. Anschließend können jedoch auch post-hoc Interpretationen durchgeführt werden. Dabei ist zu beachten, dass es einige Implikationen gibt; die letzte Spalte stellt nur die „maximale“ Interpretation dar. Zum Beispiel bedeutet der erste Fall, in dem das Konfidenzintervall (KI) sogar jenseits der Relevanzgrenze δ liegt, relevante Überlegenheit der zu untersuchenden Therapie gegenüber der Vergleichsgruppe. Daraus folgen der Reihe nach die signifikante Überlegenheit (gesamtes KI rechts von der Null) und die Nicht-Unterlegenheit (gesamtes KI rechts von $-\delta$). Aus der relevanten Überlegenheit folgt auch die signifikante Unterschiedlichkeit (KI überdeckt nicht die Null). Eine statistisch signifikante Unterschiedlichkeit muss jedoch keinen klinisch relevanten Unterschied bedeuten [99].

Niveau Der maximale Fehler 1. Art wird als Niveau eines Tests (α) bezeichnet. α muss vor der Datenanalyse festgelegt werden, um einen datengesteuerten Einfluss auf die Ergebnisinterpretation zu verhindern. Bei einer zweiseitigen Fragestellung (D) ist ein Niveau von $\alpha = 5\%$ üblich. Gemäß der Guideline ICH-E9 werden $\alpha = 2.5\%$ bei einseitigen Tests vorgeschlagen [56]. Für eine Nicht-Unterlegenheitsstudie bedeutet dies ebenso die Anwendung eines Tests zum Niveau von $\alpha = 2.5\%$ oder bei Symmetrie die Verwendung eines $(1-2\alpha) = 95\%$-Konfidenzintervalls [56]. Um zweiseitige Äquivalenz zu untersuchen, kann ein Niveau von $\alpha = 5\%$ gewählt werden. Dies bedeutet aufgrund des Abschlusstestprinzips, dass ein $(1-2\alpha) = 90\%$-Konfidenzintervall angewendet werden kann [39].

1.5 Äquivalenzstudien

Die randomisierte, Placebo-kontrollierte, doppelt-blinde klinische Studie gilt als Gold-Standard für den Nachweis der Wirksamkeit einer Therapie [77]. Dieser Ansatz ist aber nicht vertretbar, wenn es bereits eine anerkannte Therapie A für die entsprechende Indikation (und Population) gibt, deren Wirksamkeit nachgewiesen ist [80, 92] und außerdem von einer noch immer gültigen Wirksamkeit von A ausgegangen wird („constancy assumption“). Besteht keine Unsicherheit über die Wirksamkeit von A („equipoise“) [29], kann eine Placebo-kontrollierte Studie ethisch problematisch sein. Um dennoch die Wirksamkeit der Therapie zu beurteilen, ist eine Äquivalenzstudie durchzuführen. Temple und Ellenberg unterscheiden weiterhin in symptomatische und kausale Therapien. Sie begründen, dass bei symptomatischen Therapien eher die Möglichkeit besteht, eine randomisierte, Placebo-kontrollierte Studie durchzuführen [93]. Andererseits ist die Entscheidung für eine Placebo-Kontrolle oder eine aktive Kontrolle beim Nachweis der Wirksamkeit abhängig von dem Schweregrad der Erkrankung sowie der Stärke und Variabilität der Wirkung der bekannten Therapie. Bei einer harmlosen Erkrankung könnte auch eine Placebo-Kontrolle gerechtfertigt sein, selbst wenn es bereits eine kausale, wirksame Therapie gäbe. Der Nachweis der Wirksamkeit ist nicht der einzige Grund für die Durchführung einer (Äquivalenz-)Studie. Der direkte Wirksamkeitsvergleich mit einer bekannten Therapie kann Untersuchungsziel sein (dann ist natürlich eine entsprechende Kontrollgruppe in der Studie vorzusehen), aber auch Sicherheitsparameter können im Vordergrund der Studie stehen. Die Argumentationen in diesem Fall sind den oben genannten sehr ähnlich. Die verschiedenen Argumente bezüglich der richtigen Wahl der Kontrollgruppe werden noch immer diskutiert [29, 87, 92, 93]. Simon fasst in seinem Editorial zusammen, dass nur ein

Konsens bei lebensbedrohlichen Indikationen mit bekannter Therapie besteht: In einem solchen Fall wäre eine Placebo-kontrollierte Studie ethisch nicht vertretbar [87]. Die gleichen ethischen Bedenken treten bei dreiarmigen Studien mit Placebo-Kontrolle auf. Eine Studie mit dem Ziel die Wirksamkeit von E sowie die Äquivalenz zu A zu zeigen, sollte dreiarmig angelegt werden (E, A, P) [76]. Trotz der ethischen Bedenken favorisieren einige Autoren aufgrund der statistischen und inhaltlichen Probleme von aktiv kontrollierten Studien die Placebo-kontrollierte Studie [95, 24].

Äquivalenzstudien („equivalence trial“, „active control equivalence trial“ [58, 92], „positive control study“ [58]) sind spezielle Studien mit aktiver Kontrolle, aber nicht alle aktiv kontrollierten Studien müssen Äquivalenzstudien sein: Es ist durchaus denkbar, dass eine neue Therapie einer bestehenden (Standard-)Therapie überlegen zu sein scheint. Dann wäre eine Überlegenheitsstudie das richtige Instrument. Obwohl in der medizinischen Literatur „active controlled trial“ manchmal als Synonym für Äquivalenzstudien steht, werden Äquivalenzstudien mit der statistischen Hypothese definiert (Abbildung 2 auf Seite 11).

Äquivalenzstudien verfolgen insbesondere folgende Ziele [58, 73, 105]:

- (indirekter) Nachweis der Wirksamkeit einer Therapie anhand klinischer Parameter, wenn eine Placebo-kontrollierte Studie nicht durchführbar ist (ein- oder zweiseitige therapeutische Äquivalenz)
- Nachweis der Nicht-Unterlegenheit einer Therapie anhand klinischer Parameter gegenüber einem etablierten Behandlungsverfahren (einseitige therapeutische Äquivalenz)
- (indirekter) Nachweis, dass eine Therapie mindestens einen Effekt der Standardtherapie sichert (einseitige therapeutische Äquivalenz, hier: Relevanz)
- Nachweis der Vergleichbarkeit von Sicherheitsparametern (ein- oder zweiseitige therapeutische Äquivalenz)
- Nachweis der Gleichwertigkeit einer Therapie mit einer anderen Therapie anhand klinischer Parameter (zweiseitige therapeutische Äquivalenz)
- Nachweis der Nicht-Unterlegenheit biologischer Aktivität zweier biologischer Produkte (einseitige Bioäquivalenz)

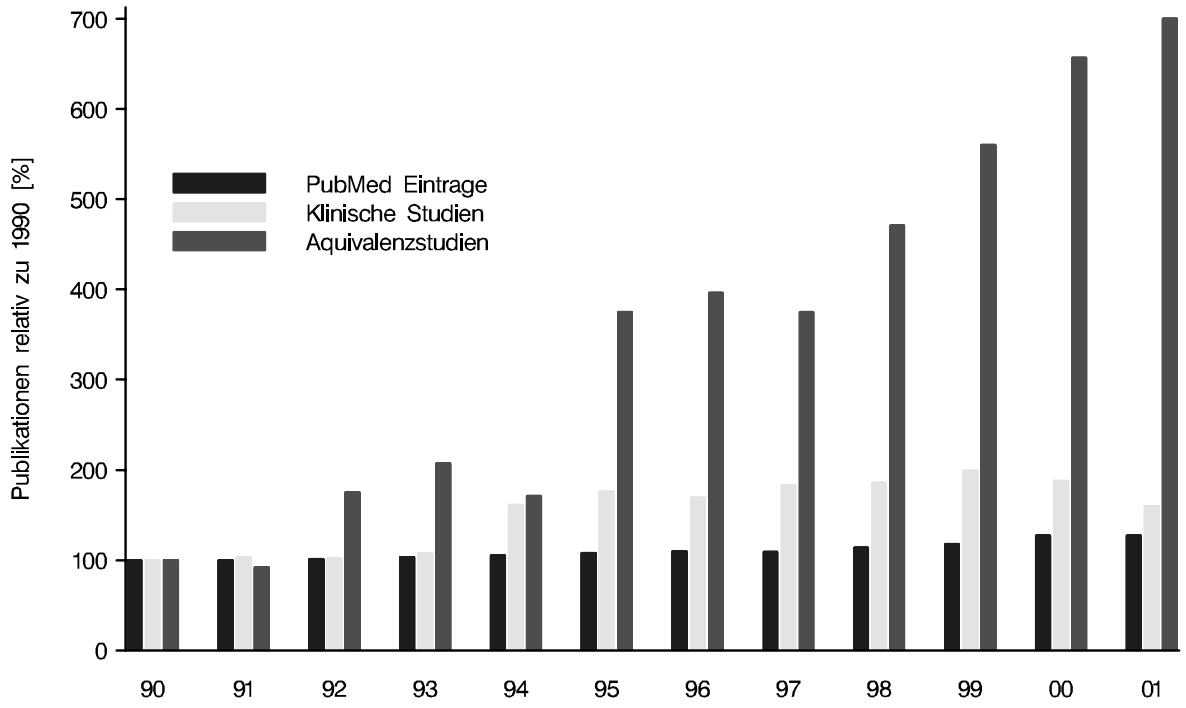


Abbildung 4: Relative Anzahl von MEDLINE-Einträgen, klinischen Studien und Äquivalenzstudien bezogen auf 1990

- Nachweis von ähnlichen pharmakokinitischen Eigenschaften von Generika zur Beurteilung der pharmakologischen Aktivität (*zweiseitige* Bioäquivalenz, da eine erhöhte Bioverfügbarkeit das Sicherheitsprofil ändern kann)

- Verifikation von Modellvoraussetzungen (zum Beispiel Varianzhomogenität, Additivität der Haupteffekte in ANOVA-Modellen, Ausschluss von Carry-over-Effekten, etc. [104])

- Nachweis der Vergleichbarkeit von Subpopulationen (hinsichtlich der Verteilung wichtiger Baseline-Variablen)

Die Anzahl der Veröffentlichungen von Äquivalenzstudien ist in den letzten Jahren deutlich gestiegen. Dies deutet auf ein erhöhtes Interesse in der medizinischen Forschung hin (Abbildung 4, Tabelle 11 auf Seite 99). Dieser Trend kommt zum einen dadurch zustande, dass für viele Indikationen bereits etablierte Therapien vorliegen und somit eine Placebo-kontrollierte Studie ethisch problematisch ist. Zum anderen sind *deutliche* Verbesserungen in der Therapieforschung im Sinne einer Überlegenheit immer schwieriger zu erreichen.

Um die Anzahl der veröffentlichten Äquivalenzstudien in MEDLINE zu erhalten, wurde die Suchstrategie (equivalenc*[All Fields] OR "therapeutic equivalency"[MeSH Terms] OR bioequivalenc*[Text Word] OR non-inferio*[All Fields] OR noninferio*[All Fields]) AND Clinical Trial[ptyp]) in PubMed verwendet. Es ist jedoch davon auszugehen, dass erheblich mehr Äquivalenzstudien publiziert wurden. Einerseits gibt es kein entsprechendes Schlagwort und andererseits ist in Fachkreisen keine akzeptierte Suchstrategie bekannt. Dies wurde durch eine Umfrage per e-mail überprüft. Dennoch kann man die prozentualen Veränderungen verwenden, um einen Eindruck zu bekommen, wie stark die Anzahl der Publikationen auf diesem Gebiet steigt.

1.6 Aufbau der Arbeit

Ausgehend vom Idealfall in Kapitel 2 mit entsprechender Methodik soll im Kapitel 3 auf die Probleme und Lösungen bei der Wahl der Äquivalenzgrenzen eingegangen werden. Kapitel 4 beschäftigt sich mit der Auswahl der Population für die Auswertung der Meta-Analyse. Kapitel 3 und 4 werden jeweils durch ein Beispiel illustriert. In der Diskussion im Kapitel 5 werden unter anderem Bedingungen an Einzelstudien formuliert sowie ein Anforderungskatalog für Meta-Analysen zusammengestellt. Kapitel 7 beinhaltet mehrere Verzeichnisse, einschließlich Symbol- und Abkürzungsverzeichnis.

2 Der Idealfall

Ausgehend von einer Situation, in der alle notwendigen Informationen für die Durchführung einer Meta-Analyse gegeben sind, wird schrittweise von diesem Idealfall abgewichen und die entstehenden Probleme werden diskutiert.

Angenommen es exisitieren $k \geq 2$ Studien zu einer zu untersuchenden Äquivalenzfragestellung mit folgenden Kriterien:

1. gleiche Therapien
2. gleiche Ein- und Ausschlusskriterien
3. gleiche Zielgröße, die nicht Hauptzielkriterium sein muss
4. alle Einzelstudien wurden als Äquivalenzstudien geplant
5. Auswertung mit PP- *und* ITT-Population
6. gleiche Äquivalenzgrenzen
7. Originaldaten von allen Studien liegen vor

Sind alle sieben Kriterien gegeben, kann eine übliche Auswertung für Äquivalenzstudien mit entsprechender Stratifizierung für die Einzelstudien durchgeführt werden.

In der vorliegenden Arbeit soll jedoch keine originaldatenbasierte Meta-Analyse sondern eine publikationsbasierte betrachtet werden, so dass die letzte Bedingung nicht erfüllt ist. Sind die herkömmlichen Verfahren aus der Meta-Analyse geeignet, um die Nicht-Unterlegenheit einer bzw. Äquivalenz zweier Therapien zu zeigen?

Anmerkung: In den folgenden Abschnitten wird mit den Quantilen der Standardnormalverteilung $z_{1-\alpha}$ gearbeitet. Eine Verwendung der t-Verteilung mit $k-1$ Freiheitsgraden ist ebenso möglich und hat sogar bessere Überdeckungswahrscheinlichkeiten [15, 41, 46, 85]. Darauf wurde im Folgenden verzichtet, um die Schreibweise so einfach wie möglich zu gestalten.

2.1 Nicht-Unterlegenheit im FEM

Zunächst wird die Nicht-Unterlegenheit im FEM betrachtet und angenommen, dass die Nicht-Unterlegenheitsgrenze δ bereits definiert ist. Es liegt also ein einseitiges Testproblem mit Stratifizierung nach den einzelnen Studien, die in die Meta-Analyse eingehen, vor. Um dieser Situation gerecht zu werden, kann mit Hilfe des folgenden Satzes eine Teststatistik aus der klassischen Meta-Analyse (Überlegenheitsfragestellung) verwendet werden. Dabei muss lediglich eine Transformation durchgeführt werden.

Satz 2.1 *Sei $T(X_1, \ldots, X_n)$ eine Teststatistik für die Hypothese $\theta \leq 0$ und die Alternative $\theta > 0$, wobei X_i iid. Die Hypothese wird abgelehnt, wenn $T(X_1, \ldots, X_n)$ größer als der kritische Wert $c_{1-\alpha}$ ist. Dann ist $T(X_1+\delta, \ldots, X_n+\delta)$ eine Teststatistik für die Hypothese $\theta \leq -\delta$ und die Alternative $\theta > -\delta$, mit $\delta > 0$, wobei die Hypothese abgelehnt wird, wenn $T(X_1 + \delta, \ldots, X_n + \delta)$ größer als der kritische Wert $c_{1-\alpha}$ ist.*

Beweis Siehe Wellek [104]. □

Die Teststatistik für die Nicht-Unterlegenheit in der t-Test Situation ist zum Beispiel bei [48] gegeben. Für eine Meta-Analyse kann man aus obigem Satz nun folgendes Korollar bilden.

Korollar 2.2 *Sei unter den Bedingungen des FEM $\hat{\theta}_i$ der Effektschätzer aus der i-ten Studie und w_i^{-1} die Varianz von $\hat{\theta}_i$, außerdem $\delta > 0$, dann ist*

$$T_{\theta \leq -\delta} = \frac{\sum w_i \hat{\theta}_i}{\sqrt{\sum w_i}} + \delta \sqrt{\sum w_i}$$

eine Teststatistik für $H_0^{EA} : \theta \leq -\delta$ vs. $H_1^{EA} : \theta > -\delta$. Auf dem Rand der Hypothese gilt $T_{\theta \leq -\delta} \sim N(0,1)$. Die Hypothese kann zum Niveau α verworfen werden, wenn $T_{\theta \leq -\delta} > z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $(1-\alpha)$-Quantil der Standardnormalverteilung ist.

Beweis Die übliche Teststatistik für die Hypotese der klassischen Meta-Analyse $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$ und auch $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$ ist mit dem Schätzer $\hat{\theta}$ für θ und der Standardabweichung $SE = \sqrt{Var(\hat{\theta})}$

$$T_{\theta \leq 0} = T(\hat{\theta}_1, \ldots, \hat{\theta}_k) = \frac{\hat{\theta}}{SE} = \frac{\sum w_i \hat{\theta}_i}{\sum w_i} \sqrt{\sum w_i} = \frac{\sum w_i \hat{\theta}_i}{\sqrt{\sum w_i}} \sim N(0,1).$$

$T_{\theta\leq 0}$ ist auf dem Rand der Hypothese standardnormalverteilt. Da die Varianz w_i^{-1} invariant bezüglich Lokationen ist (für eine Zufallsvariable X und $a \in \mathbb{R}$ gilt $Var(a+X) = Var(X)$ also $Var(\delta + \hat{\theta}_i) = Var(\hat{\theta}_i) = w_i^{-1}$), gilt nun

$$T_{\theta\leq -\delta} = T(\hat{\theta}_1+\delta,\ldots,\hat{\theta}_k+\delta) = \frac{\sum w_i(\hat{\theta}_i+\delta)}{\sqrt{\sum w_i}} = \frac{\sum w_i\hat{\theta}_i + w_i\delta}{\sqrt{\sum w_i}} = T(\hat{\theta}_1,\ldots,\hat{\theta}_k)+\delta\sqrt{\sum w_i}\,.$$

Zusammen mit Satz 2.1 ist somit das Korollar bewiesen. □

Die Hypothese der Nicht-Unterlegenheit $H_0^{EA} : \theta \leq -\delta$ kann eher abgelehnt werden kann, als die einseitige Hypothese $H_0 : \theta \leq 0$, denn die Teststatistik ist um den Term $\delta\sqrt{\sum w_i}$ größer. Es gilt somit $T_{\theta\leq -\delta} > T_{\theta\leq 0}$.

Der p-Wert des Tests berechnet sich wie üblich aus der Verteilungsfunktion der Standardnormalverteilung Φ mit

$$p = 1 - \Phi(T_{\theta\leq -\delta}). \tag{7}$$

Die Hypothese der Nicht-Unterlegenheit beeinflusst die Berechnung eines Konfidenzintervalls nicht; die üblichen Konfidenzintervalle für den Parameter θ können daher verwendet werden. Die Äquivalenz des obigem Tests und dem üblichen Intervall-Inklusions-Methode besteht ebenfalls. Daher ist die Testentscheidung auch anhand der unteren Grenze des Konfidenzintervalls möglich: Die Hypothese $H_0^{EA} : \theta \leq -\delta$ kann verworfen werden, wenn die Grenze des einseitigen nach oben offenen $(1-\alpha)$-Konfidenzintervalls (bzw. die untere Grenze eines zweiseitigen $(1-2\alpha)$-Konfidenzintervalls) L größer als $-\delta$ ist. Die Grenze des einseitigen nach oben offenen Konfidenzintervalls ist

$$\begin{aligned} L &= \hat{\theta} - z_{1-\alpha}\sqrt{\frac{1}{\sum w_i}} = \frac{\sum w_i\hat{\theta}_i}{\sum w_i} - z_{1-\alpha}\sqrt{\frac{1}{\sum w_i}} \\ &= T_{\theta\leq 0}\sqrt{\frac{1}{\sum w_i}} - z_{1-\alpha}\sqrt{\frac{1}{\sum w_i}} = (T_{\theta\leq 0} - z_{1-\alpha})\sqrt{\frac{1}{\sum w_i}}. \end{aligned}$$

Somit ist

$$\begin{aligned} L > -\delta \Leftrightarrow T_{\theta\leq 0} - z_{1-\alpha} > -\delta\sqrt{\sum w_i} \\ \Leftrightarrow T_{\theta\leq 0} + \delta\sqrt{\sum w_i} > z_{1-\alpha} \Leftrightarrow T_{\theta\leq -\delta} > z_{1-\alpha} \end{aligned}$$

und man kann die Intervall-Inklusions-Methode für die Nicht-Unterlegenheit zusammenfassend schreiben als:

$$L > -\delta \Leftrightarrow T_{\theta \leq -\delta} > z_{1-\alpha} \Leftrightarrow H_0^{EA} : \theta \leq -\delta \quad \text{verwerfen}$$

Anmerkung: Die Intervall-Inklusions-Methode wird auch von den ICH Guidelines E3 und E9 empfohlen [55, 56].

2.2 Nicht-Unterlegenheit im REM

Das Modell mit zufälligen Effekten (REM) für Meta-Analysen unterscheidet sich vom FEM insofern, als dass die zugrunde gelegten Effekte θ_i nicht fest sondern zufällig sind, d.h. eine Zufallsvariable mit einer Verteilung darstellen. Wie in Abschnitt 1.3, wird für die Verteilung des Effektes θ_i üblicherweise eine Normalverteilung mit Erwartungswert θ und Varianz τ^2 angenommen. Die Varianz des Effekt-Schätzers $\hat{\theta}_i \sim N(\theta, w_i^{-1} + \tau^2)$ kann vereinfacht als $(w_i^*)^{-1} := w_i^{-1} + \tau^2$ bezeichnet werden. Somit lässt sich das Korollar 2.2 aus dem FEM auf das REM übertragen.

Korollar 2.3 *Sei unter den Bedingungen des REM $\hat{\theta}_i$ der Effektschätzer aus der i-ten Studie und $(w_i^*)^{-1}$ die Varianz von $\hat{\theta}_i$, außerdem $\delta > 0$, dann ist*

$$T^*_{\theta \leq -\delta} = \frac{\sum w_i^* \hat{\theta}_i}{\sqrt{\sum w_i^*}} + \delta \sqrt{\sum w_i^*}$$

eine Teststatistik für $H_0^{EA} : \theta \leq -\delta$ vs. $H_1^{EA} : \theta > -\delta$. Auf dem Rand der Hypothese gilt $T^*_{\theta \leq -\delta} \sim N(0,1)$. Die Hypothese kann zum Niveau α verworfen werden, wenn $T^*_{\theta \leq -\delta} > z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $(1-\alpha)$-Quantil der Standardnormalverteilung ist.

Beweis Im Beweis zum Korollar 2.2 ist lediglich w_i durch w_i^* zu ersetzen. □

Auch bei der notwendigen Ersetzung von τ durch $\hat{\tau}$ bleibt die Verschiebung um δ möglich, sofern $\hat{\tau}$ invariant bezüglich einer Lokation ist. Dies wird von einem Varianzschätzer erwartet. Der üblichen DerSimonian-Laird-Schätzer

$$\hat{\tau}_{DSL} = max\left(\frac{Q - (k-1)}{\sum w_i - \sum w_i^2 / \sum w_i}; 0\right)$$

ist lokationsinvariant. Er besteht im Wesentlichen aus der Heterogenitätsstatistik Q und den Gewichten w_i [28].

$$\begin{aligned} Q(\hat{\theta}_i) &= \sum w_i(\hat{\theta}_i - \hat{\theta})^2 \\ &= \sum w_i \left(\hat{\theta}_i - \frac{\sum w_i\hat{\theta}_i}{\sum w_i}\right)^2 \\ &= \sum w_i \left((\hat{\theta}_i + \delta) - \frac{\sum w_i\hat{\theta}_i}{\sum w_i} - \delta\right)^2 \\ &= \sum w_i \left((\hat{\theta}_i + \delta) - \frac{\sum w_i(\hat{\theta}_i + \delta)}{\sum w_i}\right)^2 = Q(\hat{\theta}_i + \delta) \end{aligned}$$

Somit gilt $Q(\hat{\theta}_i) = Q(\hat{\theta}_i + \delta) \Rightarrow \hat{\tau}_{DSL}(\hat{\theta}_i) = \hat{\tau}_{DSL}(\hat{\theta}_i + \delta)$.

Der p-Wert des Tests berechnet sich wie üblich aus der Verteilungsfunktion der Standardnormalverteilung Φ mit

$$p = 1 - \Phi(T^*_{\theta \leq -\delta}).$$

Die Äquivalenz zwischen der Testentscheidung mit der Teststatistik und der unteren Grenze des Konfidenzintervalls gilt analog zum FEM in Abschnitt 2.1. Die Grenze ist

$$L^* = \hat{\theta} - z_{1-\alpha}\sqrt{\frac{1}{\sum w_i^*}}$$

und es lässt sich zusammenfassend schreiben:

$$L^* > -\delta \Leftrightarrow T^*_{\theta \leq -\delta} > z_{1-\alpha} \Leftrightarrow H_0^{EA} : \theta \leq -\delta \quad \text{verwerfen.}$$

2.3 Äquivalenz im FEM und REM

Analog zum Abschnitt 2.1 kann auch

$$T_{\theta \geq +\delta} = \frac{\sum w_i\hat{\theta}_i}{\sqrt{\sum w_i}} - \delta\sqrt{\sum w_i}$$

als Teststatistik für die Hypothese $H_0 : \theta \geq +\delta$ definiert werden. Mit dem Abschlusstestprinzip können zwei einseitige α-Tests durchgeführt werden: Wenn sowohl $H_0 : \theta \geq +\delta$ als

auch $H_0 : \theta \leq -\delta$ abgelehnt werden, kann auch die Hypothese $H_0 :\mid \theta \mid \geq \delta$ zum Niveau α abgelehnt werden.

Für das REM ist wie in Abschnitt 2.2 lediglich w_i durch w_i^* zu ersetzen.

2.4 Abweichungen vom Idealfall

Die oben genannten Kriterien für den Idealfall können verletzt sein:

1. Verschiedene Therapien: Wie bei jeder Meta-Analyse ist abzuwägen, wie unterschiedlich die Therapieschemata sein dürfen, damit die Einzelstudie eingeschlossen werden kann. In der vorliegenden Arbeit wird darauf nicht weiter eingegangen.

2. Verschiedene Ein- und Ausschlusskriterien: Diese Problematik liegt bei jeder Meta-Analyse vor, die dadurch gegebenenfalls entstehenden Heterogenitäten müssen entsprechend berücksichtigt werden. In der vorliegenden Arbeit wird darauf nicht weiter eingegangen.

3. Die Zielgröße muss erfasst worden sein, da sonst kein Einschluss in die Meta-Analyse erfolgen kann. Wurden nur ähnliche Zielgrößen ermittelt, lässt sich dies gegebenenfalls berücksichtigen (z.B. verschiedene Schmerzscores). Auch dieser Punkt unterscheidet sich nicht von einer gewöhnlichen Meta-Analyse und wird daher nicht weiter betrachtet.

4. Einige der relevanten Studien wurden als Überlegenheitsstudien geplant und durchgeführt (Kapitel 3 und 4).

5. Die Populationsansatz ist unklar oder es liegt nur eine PP- oder ITT-Analyse vor (Kapitel 4).

6. Bei Verwendung von ungleichen oder keinen Äquivalenzgrenzen muss ein Verfahren gefunden werden, mit dem sich Äquivalenzgrenzen für die Meta-Analyse festlegen lassen (Kapitel 3).

3 Die δ-Problematik

Bei Studien mit aktiver Kontrollgruppe A („active controlled trials“) bezeichnet δ die Äquivalenzgrenze bei Äquivalenzstudien bzw. die Nicht-Unterlegenheitsgrenze bei Nicht-Unterlegenheitsstudien. Da man Nicht-Unterlegenheitsstudien als einseitige Äquivalenzstudien auffassen kann (Abschnitt 1.5), wird im Folgenden eine Nicht-Unterlegenheitsgrenze ebenso als Äquivalenzgrenze bezeichnet. Für die Definition von δ gibt es viele Ansätze, 13 Formulierungen davon hat Ng in seiner Arbeit zusammengestellt [73].

Die Wahl der Äquivalenzgrenze δ wird in der Literatur immer wieder diskutiert [45, 48, 73, 79, 111]. Das Problem besteht letztendlich seit der Einführung der Äquivalenztestung in die Biometrie [106].

Bevor verschiedene Definitionen von δ eingeführt werden, muss die Grenze, die für die Fallzahlplanung (Fallzahl-Δ) verwendet wird, von der Äquivalenzgrenze δ abgegrenzt werden. Zur Vereinfachung der Schreib- und Sprechweise kann ohne Einschränkung der Allgemeinheit ein Effekt angenommen werden, der bei Null die Gleichheit der betrachteten Gruppen darstellt (zum Beispiel: Differenz von Erwartungswerten, Risikodifferenz (RD), logarithmiertes Odds Ratio (log(OR)) oder relatives Risiko (log(RR))). Dabei sprechen große Werte für die zu untersuchende Therapieform.

Fallzahl-Δ und Äquivalenz-δ Um in der Planungsphase einer Überlegenheitsstudie die Fallzahl festlegen zu können, muss eine weitere Größe festgelegt werden, die sich auf den Therapieunterschied bezieht. Bei Bock im Kontext des t-Tests heißt diese Größe „zu entdeckende Mindestdifferenz“ und wird mit $\Delta > 0$ bezeichnet [17]. Altman verwendet die Begriffe „clinically relevant difference“ und „difference of interest“ [1]. Eine Fallzahlplanung muss eine Stichprobengröße festgelegen, um mit einer vorgegebenen Wahrscheinlichkeit (1-β) eine Differenz Δ – oder einen größeren Therapieunterschied – zu entdecken. β ist in der Regel 0.1 oder 0.2. Ein Effekt von mehr als Δ bedeutet dabei einen Vorteil für zukünftige Patienten, falls sie die neue Therapie erhalten. Wenn man davon ausgeht, dass die übrigen Planungsannahmen auch in der Studie erfüllt sind und sich ein größerer Therapieunterschied (Punktschätzer) zeigt, ist die tatsächliche Power (= Wahrscheinlichkeit, den Therapieunterschied zu entdecken) größer als die angenomme (1-β).

$$P(\text{verwerfe Hypothese der Überlegenheit}|\theta > \Delta) > (1 - \beta)$$

Wenn Δ als minimale, klinisch relevante Differenz definiert wird, lässt sich ein relevanter

Unterschied mindestens mit der Wahrscheinlichkeit (1-β) aufdecken. Dies bezieht sich auf den statistischen Nachweis eines Unterschieds, nicht auf den Nachweis eines relevanten Unterschieds. Besteht ein kleinerer Unterschied, ist auch die Power kleiner. Der Test auf Unterschiedlichkeit braucht dann jedoch nicht verworfen zu werden. Die Festlegung „$\Delta =$ minimale, klinisch relevante Differenz“ scheint demnach eine sinnvolle Definition zu sein. Ist aber der tatsächliche Effekt *erheblich* größer, so wäre die geplante Fallzahl zu groß, was aus ethischen Gesichtspunkten problematisch sein kann. Daher sollte bei der Fallzahl immer die vermutete Differenz mit in die Überlegungen der Fallzahlplanung einbezogen werden. Sinnvoller wäre die Anwendung von adaptiven Designs. Diese ermöglichen, die Fallzahl während der Studie zu adaptieren [7]. Die minimale, klinisch relevante Differenz stellt jedoch bei Überlegenheitsstudien immer eine Untergrenze dar. Ein kleineres Δ und somit eine größere Fallzahl ist für die Nullhypothese der Gleichheit nicht zu vertreten.

Bei Nicht-Unterlegenheitsstudien ist das Fallzahl-Δ von der Definition der Äquivalenzgrenze δ zu trennen. Es ist

$$P(\text{verwerfe Hypothese der Nicht-Unterlegenheit}|\theta \leq -\delta) \leq \alpha$$

und für das zur Fallzahlplanung einer Nicht-Unterlegenheitsstudie notwendige $\Delta_{NI} > -\delta$ gilt

$$P(\text{verwerfe Hypothese der Nicht-Unterlegenheit}|\theta > \Delta_{NI}) > (1 - \beta).$$

Abbildung 5 zeigt den Fallzahleinfluss auf eine Nicht-Unterlegenheitsstudie. Dabei ist die mittlere Powerfunktion (mittlere Fallzahl) aufgrund von $\Delta_{NI} = 0$ zustande gekommen. Die obere, linke Powerfunktion (hohe Fallzahl) basiert auf $\Delta_{NI} = -\mu 1$ und die untere, rechte (niedrige Fallzahl) auf $\Delta_{NI} = \mu 2$. Bei festem δ sollte jeweils eine Power von 0.8 erreicht werden, wenn der wahre Effekt $\theta = \Delta_{NI}$ ist. Für die praktische Umsetzung kann die Fallzahlformel für den t-Test verwendet werden, indem $\Delta = \delta + \Delta_{NI}$ gesetzt wird.

Um die Unterschiede vom Fallzahl-Δ, Δ_{NI} und der Äquivalenzgrenze δ zu verdeutlichen, sind in Tabelle 2 Fallzahlplanungen für verschiedene Szenarien in der einfachen t-Test Situation aufgeführt. Für den Placebovergleich werden einseitige Hypothesen und ein Niveau von $\alpha = 2.5\%$ vorausgesetzt. Deutlich erkennbar ist, dass für den Nachweis der Nicht-Unterlegenheit eine sehr viel größere Fallzahl (n^{EA}) nötig ist als für den Vergleich mit Placebo (n^{EP} bzw. n^{AP}). Diese Differenz erhöht sich, je größer die beiden Effekte von E und A im Verhältnis zu Placebo sind. So ist zum Beispiel bei tatsächlich gleichen Ef-

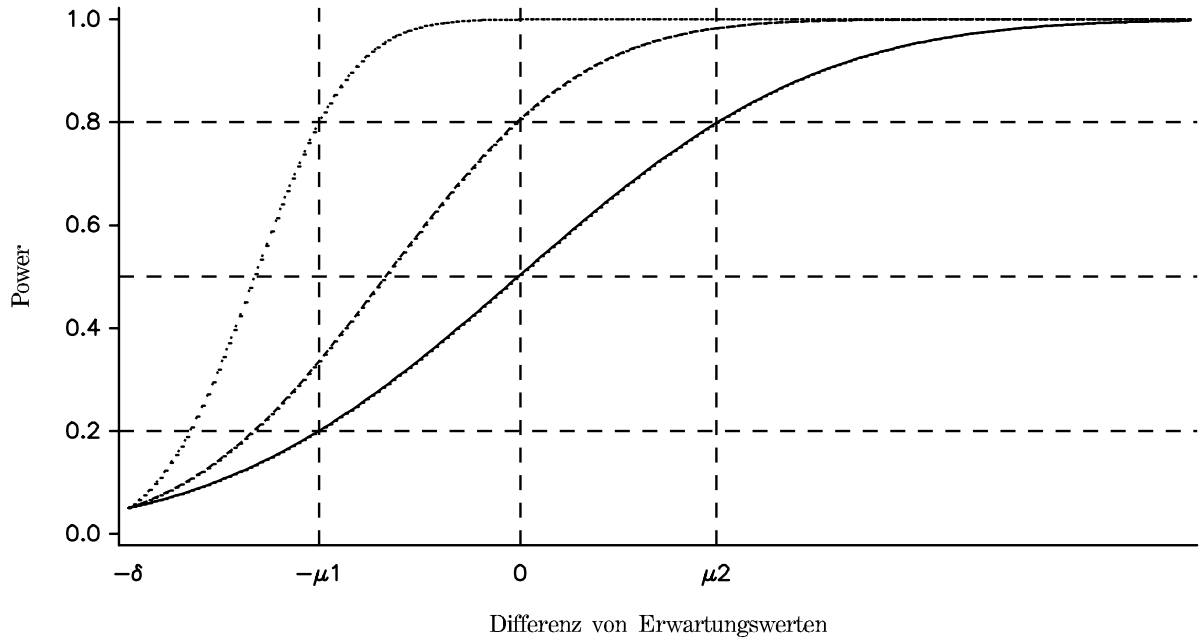


Abbildung 5: Schematische Darstellung von Powerfunktionen für drei verschiedene Stichprobenumfänge für eine Nicht-Unterlegenheitsstudie mit der Äquivalenzgrenze δ. $-\mu 1$, 0 und $\mu 2$ geben die Differenzen von Erwartungswerten wieder.

fekten von der Experimentalgruppe und der aktiven Kontrolle ($\mu^E = \mu^A$, mittlere Spalte) für die Nicht-Unterlegenheitsstudie im Vergleich zur Placebokontolle bei einem Effekt von 3 ($= \mu^E = \mu^A$) die 2.2-fache Fallzahl nötig (128 statt 58). Im Falle eines Effektes von 4 ($= \mu^E = \mu^A$) ist sogar die 3.8-fache Fallzahl nötig (128 statt 34).

Auch bei standardisierten Differenzen, Quotienten usw. sind für Nicht-Unterlegenheitsstudien entprechende Fallzahlen nötig. Nach diesen Erläuterungen zu der oft notwendigerweise hohen Fallzahl soll im Folgenden nun auf die δ-Problematik bei klinischen (Einzel-)Studien eingegangen werden.

3.1 δ-Problematik bei klinischen Studien

Bisher wurde davon ausgegangen, dass die Äquivalenzgrenze δ bekannt ist bzw. bereits festgelegt wurde. Um die Methodik zur Festlegung der Äquivalenzgrenzen für Meta-Analysen zu beurteilen, sind vorerst die entsprechenden Methoden für Einzelstudien zu erörtern. Zunächst soll eine einzelne klinische Studie betrachtet werden, um die wesent-

Tabelle 2: Fallzahlvergleich für die einseitige t-Test Situation bzw. Nicht-Unterlegenheit

Vergleich	$\mu^A = 3,\ \mu^E = 2$	$\mu^A = 3,\ \mu^E = 3$	$\mu^A = 3,\ \mu^E = 4$
E vs. P	$\Delta = 2 \quad n^{EP} = 128$	$\Delta = 3 \quad n^{EP} = 58$	$\Delta = 4 \quad n^{EP} = 34$
A vs. P	$\Delta = 3 \quad n^{AP} = 58$	$\Delta = 3 \quad n^{AP} = 58$	$\Delta = 3 \quad n^{AP} = 58$
E vs. A	$\Delta_{NI} = -1 \quad n^{EA} = 506$	$\Delta_{NI} = 0 \quad n^{EA} = 128$	$\Delta_{NI} = 1 \quad n^{EA} = 58$
	$\mu^A = 4,\ \mu^E = 3$	$\mu^A = 4,\ \mu^E = 4$	$\mu^A = 4,\ \mu^E = 5$
E vs. P	$\Delta = 3 \quad n^{EP} = 58$	$\Delta = 4 \quad n^{EP} = 34$	$\Delta = 5 \quad n^{EP} = 24$
A vs. P	$\Delta = 4 \quad n^{AP} = 34$	$\Delta = 4 \quad n^{AP} = 34$	$\Delta = 4 \quad n^{AP} = 34$
E vs. A	$\Delta_{NI} = -1 \quad n^{EA} = 506$	$\Delta_{NI} = 0 \quad n^{EA} = 128$	$\Delta_{NI} = 1 \quad n^{EA} = 58$

Planungsannahmen: $\delta = 2$, $\alpha = 0.025$, $\beta = 0.2$, $\sigma = 4$ und Erwartungswerte $\mu^P = 0$ (P = Placebogruppe), $\mu^A = 3,\ 4$ (A = aktive Kontrollgruppe), $\mu^E = 2,\ 3,\ 4$ (E = Experimentalgruppe) sowie den Fallzahlen der Paarvergleiche für beide Gruppen n^{AP} , n^{EP} und n^{EA}

lichen Prinzipien für die Wahl der Äquivalenzgrenzen darzustellen. Je nach Methodik erhält δ einen anderen Index, um die verschiedenen Grenzen eindeutig zu benennen. Im Symbolverzeichnis auf Seite 83 sind die Indices zusammengestellt.

Die Festlegung der Äquivalenzgrenzen wird in der Literatur nach verschiedenen Gesichtspunkten beschrieben [16, 102, 111], je nach Intention der Analyse gibt es nicht nur ein δ, sondern mehrere wie bei Relevanzfragestellungen [99]. Es sollten sowohl statistische Gesichtspunkte als auch klinische Kriterien in die Wahl der Äquivalenzgrenze einfließen [57].

3.1.1 Klinische Argumentation

Die Zielsetzungen einer Nicht-Unterlegenheitsstudie (E darf nicht klinisch relevant schlechter sein, als A) ebenso wie die einer Äquivalenzstudie (E und A sollen sich nicht klinisch relevant unterscheiden) verdeutlichen, dass klinische Gesichtspunkte eine wesentliche Rolle bei der Festlegung von δ spielen *müssen*. Diese Grenze soll im Folgenden mit δ_{clin} bezeichnet werden. δ_{clin} beschreibt den kleinsten Wert, der gerade noch einen klinisch wichtigen Unterschied in der Zielgröße darstellt. Oder: δ_{clin}ist der größte Wert, der keinen klinisch wichtigen Unterschied darstellt. Mit δ_{clin} wird die Hypothese (4) getestet.

Jeder Kliniker wird sein eigenes Konzept von einem klinisch wichtigen Unterschied haben [111]. Unter Umständen wird δ_{clin} bei einigen Zielgrößen sehr klein eingeschätzt. So kann

zum Beispiel bei einer Überlebenszeitanalyse jeder gewonnene Lebenstag als klinisch relevant eingestuft werden. In häufig untersuchten Indikationen mit definierten Zielkriterien gibt es bereits etablierte Grenzen. So wird in der „Hypertension-Guideline“ eine klinische Relevanz ab einer Reduktion des systolischen Blutdrucks von 20 mmHg bzw. 10 mmHg des diastolischen Blutdrucks unterstellt [35]. Andererseits könnte man Patienten selbst einschätzen lassen, welcher Unterschied zwischen zwei Therapien für sie noch vertretbar sind. Je nach Ansatz ergeben sich sehr unterschiedliche Grenzen.

Festzuhalten bleibt, dass die klinische Festsetzung von δ_{clin} nicht unproblematisch ist. δ sollte jedoch immer kleiner oder gleich δ_{clin} sein, wobei δ die endgültige Äquivalenzgrenze für die induktive Statistik der (Einzel-)Studie zur Hypothese der Nicht-Unterlegenheit von E gegenüber A darstellt.

3.1.2 Wirksamkeitsargumentation

Ziel jeder Äquivalenzstudie sollte neben der Äquivalenz auch die Wirksamkeit der Therapie E sein, das heißt die Überlegenheit gegenüber Placebo (Hypothese (6)). Siegel bezeichnet das als „at least some efficacy“ [86]. Die allgemeinere Hypothese (5) (Abschnitt 1.4, Seite 12) lautet:

$$H_0^{EP}(\lambda) : \theta^{EP} \leq \lambda\, \theta^{AP} \quad \text{vs.} \quad H_1^{EP}(\lambda) : \theta^{EP} > \lambda\, \theta^{AP}.$$

$\lambda \geq 0$ legt dabei den für E zu sichernden Anteil des Effektes von A gegenüber P fest. Ein geeignetes Studiendesign für diese Hypothese wäre eine dreiarmige Studie mit E, A und einer Placebokontrolle, so dass sich neben dem Vergleich von E und A der Effekt θ^{EP} direkt aus der Studie schätzen ließe. Eine solche dreiarmige Studie ist jedoch in vielen Fällen nicht durchführbar, da entweder eine Placebobehandlung ethisch nicht vertretbar ist oder keine *etablierte* Kontrollgruppe A exisitiert.

Sofern *keine* Placebo-kontrollierte Studie durchgeführt wird, muss δ derart gewählt werden, dass indirekt auf die Überlegenheit von E gegenüber Placebo geschlossen werden kann. Wäre der wahre Effekt der aktiven Kontrolle θ^{AP} bekannt, könnte man δ mit θ^{AP} gleich setzen [73]. Obige Hypothese kann folgendermaßen umgeschrieben werden, wenn $\theta^{EP} = \theta^{AP} + \theta^{EA}$ ist: $H_0^{EP} : \theta^{AP} + \theta^{EA} \leq 0$ und somit auch $H_0^{EP} : \theta^{EA} \leq -\theta^{AP} = -\delta$. Für die Hypothese $H_0^{EP}(\lambda)$ müsste $\delta_\lambda = (1 - \lambda)\, \theta^{AP}$ gesetzt werden.

Gemäß eines Concept Papers der EMEA muss eine neue Therapie (E) nicht unbedingt besser sein als die Standardtherapie (A) [37]. Dort wird δ_λ als ein halb oder ein drittel der Differenz des Effektes der Standardtherapie mit Placebo festgelegt („... to use a delta of one half or one third of the established superiority of the comparator to placebo ...“). Unklar bleibt, was „established superiority“ bedeutet. Hauschke verwendet einen Bruchteil der Differenz der Erwartungswerte $\delta_\lambda = (1 - \lambda)\, \theta^{AP}$ mit $\lambda \in (0, 1)$ und wendet dies auf einen Test der Quotienten von Erwartungswerten an [48].

Üblicherweise ist keine Placebo-Gruppe in Äquivalenzstudien enthalten. Daher muss auf historische Daten zurückgegriffen werden, um den Effekt θ^{AP} zu schätzen. Mindestens eine Studie muss den Effekt θ^{AP} untersucht und die Wirksamkeit belegt haben, da sonst die Wahl des Zielkriteriums und der aktiven Kontrollgruppe nicht gerechtfertigt wäre [113]. Siegel definiert, dass δ dabei nicht größer sein darf als der kleinste wahrscheinliche Effekt der aktiven Kontrolle A. Er spezifiziert dies zum Beispiel mit der unteren Konfidenzintervallgrenze [86]. Wang et al. bezeichnen dies als indirekten Konfidenzintervall-Vergleich (ICIC = „interval confidence interval comparison“) [101]. Auch Wiens beschreibt diese Methode der Wahl von δ ausführlich. Er bezeichnet die Idee als „Putative Placebo“ und verwendet die Bezeichnung $\delta_{pbo}(\lambda)$ [111].

$$\delta_{pbo}(\lambda) := (1 - \lambda)\ \ L^{AP} \tag{8}$$

mit $\lambda \in (0, 1)$ und L^{AP} ist die untere Konfidenzintervallgrenze des Unterschieds θ^{AP} von der aktiven Kontrollgruppe und Placebo. Die Grenze δ_λ bzw. $\delta_{pbo}(\lambda)$ sollten klein genug sein, um aus $\theta^{EA} > \delta$ auch $\theta^{EP} > 0$ schlussfolgern zu können. Wiens zeigt dies für die Definition $\delta_{pbo}(\lambda)$ [111]. Es handelt sich dabei jedoch um einen sehr konservativen Ansatz [93].

Im Folgenden soll eine Methode vorgestellt werden, um ein minimales δ festzulegen. Dies wird mit $\delta_P(0)$ bezeichnet und kann unter bestimmten Voraussetzungen als Grundlage für einen Wirksamkeitsnachweis dienen. Bei der Methode wird additiv vorgegangen: Man muss sich vom Punktschätzer $\hat{\theta}^{AP}$ so weit zur Null hin entfernen, dass es später für die Überlegenheitsaussage von E gegenüber Placebo P ausreicht. Dabei muss nicht wie in (8) ein Bruchteil der unteren Konfidenzintervallgrenze verwendt werden, da dabei auch der Punktschätzer selbst mit verschoben wird. Hierfür gibt es jedoch keinen Grund. Da sich die Unsicherheit eines Punktschätzers in der entsprechenden Standardabweichung ausdrückt, muss ein Anteil der Standardabweichung vom Punktschätzer abgezogen werden.

Satz 3.1 *Es seien folgende Voraussetzungen gegeben: (1) Die Zufallsvariablen innerhalb einer Stichprobe seien unabhängig identisch normalverteilt, (2) mit bekannter Varianz, (3) und über die Zeit sei der Effekt θ^{AP} konstant („constancy assumption"). Sei weiterhin ein (historischer) Schätzer für θ^{AP} durch $\hat{\theta}^{AP}$ gegeben, ebenso der Standardfehler des Schätzers SE^{AP}.*

Wähle nun

$$\delta_P(\lambda) := (1-\lambda)\,\hat{\theta}^{AP} - \left(\sqrt{(1-\lambda)^2 + \frac{(SE^{EA})^2}{(SE^{AP})^2}} - \frac{SE^{EA}}{SE^{AP}}\right)\, z_{1-\alpha}\, SE^{AP},$$

wobei $z_{1-\alpha}$ das $(1-\alpha)$-Quantil der Standardnormalverteilung sowie SE^{EA} die Standardabweichung des Effektschätzers $\hat{\theta}^{EA}$ der aktuellen Äquivalenzstudie bezeichnet. Dann ist der Niveau-α Test auf Nicht-Unterlegenheit von E gegenüber A für die Hypothese $H_0^{EA}(\lambda) : \theta^{EA} \leq -\delta_P(\lambda)$ mit der Statistik

$$T^{EA}(\lambda) = \frac{1}{SE^{EA}}\left(\hat{\theta}^{EA} + \delta_P(\lambda)\right)$$

auch ein einseitiger Test für die Hypothese $H_0^{EP}(\lambda) : \theta^{EP} \leq \lambda\,\theta^{AP}$ zum Niveau α.

Beweis: Die Hypothese $H_0^{EP}(\lambda)$ kann auch als $H_0^{EP}(\lambda) : \theta^{EA} + (1-\lambda)\,\theta^{AP} \leq 0$ geschrieben werden. Eine Teststatistik dafür ist

$$T^{EP}(\lambda) := \frac{1}{SE}\left(\hat{\theta}^{EA} + (1-\lambda)\,\hat{\theta}^{AP}\right),$$

wobei hier $SE := \sqrt{(SE^{EA})^2 + (1-\lambda)^2\,(SE^{AP})^2}$ die Standardabweichung des Kontrastes $\hat{\theta}^{EA} + (1-\lambda)\,\hat{\theta}^{AP}$ ist. Die definierte Äquivalenzgrenze $\delta_P(\lambda)$ lässt sich als

$$\begin{aligned}
\delta_P(\lambda) &= (1-\lambda)\,\hat{\theta}^{AP} - \left(\frac{1}{SE^{AP}}\left(\sqrt{(1-\lambda)^2\,(SE^{AP})^2 + (SE^{EA})^2} - SE^{EA}\right)\right)\, z_{1-\alpha}\, SE^{AP} \\
&= (1-\lambda)\,\hat{\theta}^{AP} - \left(\sqrt{(1-\lambda)^2\,(SE^{AP})^2 + (SE^{EA})^2} - SE^{EA}\right)\, z_{1-\alpha} \\
&= (1-\lambda)\,\hat{\theta}^{AP} - \left(SE - SE^{EA}\right)\, z_{1-\alpha}
\end{aligned}$$

darstellen. Nun kann gezeigt werden, dass $H_0^{EA}(\lambda)$ dann und nur dann verworfen wird, wenn auch $H_0^{EP}(\lambda)$ verworfen wird:

$$
\begin{aligned}
& H_0^{EA}(\lambda) \text{ verwerfen} \\
\Leftrightarrow \quad & T^{EA}(\lambda) > z_{1-\alpha} \\
\Leftrightarrow \quad & \frac{1}{SE^{EA}} \left(\hat{\theta}^{EA} + \delta_P(\lambda) \right) > z_{1-\alpha} \\
\Leftrightarrow \quad & \frac{1}{SE^{EA}} \left(\hat{\theta}^{EA} + (1-\lambda)\, \hat{\theta}^{AP} - (SE - SE^{EA})\, z_{1-\alpha} \right) > z_{1-\alpha} \\
\Leftrightarrow \quad & \frac{1}{SE} \left(\hat{\theta}^{EA} + (1-\lambda)\, \hat{\theta}^{AP} \right) > \frac{SE^{EA}}{SE} z_{1-\alpha} + \frac{1}{SE}(SE - SE^{EA}) z_{1-\alpha} \\
\Leftrightarrow \quad & T^{EP}(\lambda) > z_{1-\alpha} \\
\Leftrightarrow \quad & H_0^{EP}(\lambda) \text{ verwerfen}
\end{aligned}
$$

□

Die Form der Äquivalenzgrenze $\delta_P(\lambda)$ ist die eines Konfidenzintervalls. Es ist lediglich ein Korrekturfaktor nötig (Seite 34). Im Folgenden werden Spezialfälle von $\delta_P(\lambda)$ vorgestellt, indem zusätzliche Annahmen gemacht werden.

Korollar 3.2 *Wähle nun $\lambda = 0$, so dass*

$$\delta_P(0) = \hat{\theta}^{AP} - \left(\sqrt{1 + \frac{(SE^{EA})^2}{(SE^{AP})^2}} - \frac{SE^{EA}}{SE^{AP}} \right) z_{1-\alpha}\, SE^{AP}.$$

Dann ist der Niveau-α Test auf Nicht-Unterlegenheit von E gegenüber A für die Hypothese $H_0^{EA} : \theta^{EA} \leq -\delta_P(0)$ mit der Statistik

$$T^{EA} = \frac{1}{SE^{EA}} \left(\hat{\theta}^{EA} + \delta_P(0) \right)$$

auch ein einseitiger Test für die Hypothese $H_0^{EP} : \theta^{EP} \leq 0$ zum Niveau α.

Beweis: Es ist $\lambda = 0$ in Satz 3.1 einzusetzen. Der Beweis kann jedoch auch mittels Konfidenzintervall durchgeführt werden. Dazu betrachten wir zum einen die untere Konfidenzintervallgrenze L^{EA} für θ^{EA} und L^{EP} als untere Grenze für θ^{EP}. Die Hypothese der Nicht-Unterlegenheit H_0^{EA} zu verwerfen heißt, dass $L^{EA} > -\delta_P(0)$ sein muss. Auf der anderen Seite bedeutet die Hypothese der Überlegenheit von E gegenüber Placebo zu verwerfen, dass $L^{EP} > 0$ ist. Nun gilt:

$$
\begin{aligned}
& L^{EA} = \hat{\theta}^{EA} - z_{1-\alpha} SE^{EA} > -\delta_P(0) \\
\Leftrightarrow \quad & \hat{\theta}^{EA} + \hat{\theta}^{AP} - z_{1-\alpha} \left(SE^{EA} + \left(\sqrt{1 + \frac{(SE^{EA})^2}{(SE^{AP})^2}} - \frac{SE^{EA}}{SE^{AP}} \right) \; SE^{AP} \right) > 0 \\
\Leftrightarrow \quad & \hat{\theta}^{EA} + \hat{\theta}^{AP} - z_{1-\alpha} \left(\sqrt{(SE^{AP})^2 + (SE^{EA})^2} \right) = L^{EP} > 0,
\end{aligned}
$$

mit dem Effektschätzer $\hat{\theta}^{EA} + \hat{\theta}^{AP}$ für $\theta^{EP} = \theta^{EA} + \theta^{AP}$ und der dazugehörigen Standardabweichung $\sqrt{(SE^{AP})^2 + (SE^{EA})^2}$. □

Korollar 3.3 *Zu den Voraussetzungen des Satzes 3.1 soll $\lambda = 0$ und Varianzhomogenität hinzugenommen werden. Außerdem soll nur je eine Studie (E vs. A bzw. A vs. P) gegeben sein und ein Differenzmaß verwendet werden. Dann wird $\delta_P(0)$ zu*

$$
\delta'_P(0) := \hat{\theta}^{AP} - \left(\sqrt{1 + \frac{n^{AP}}{n^{EA}}} - \sqrt{\frac{n^{AP}}{n^{EA}}} \right) \; z_{1-\alpha} \; SE^{AP}
$$

mit

$$
n^{AP} = \frac{2 n^A_{(1)} n^P_{(1)}}{n^A_{(1)} + n^P_{(1)}} \quad \textit{und} \quad n^{EA} = \frac{2 n^E_{(2)} n^A_{(2)}}{n^E_{(2)} + n^A_{(2)}}.
$$

Der Index $_{(1)}$ bezeichne die historische Studie A vs. P und der Index $_{(2)}$ bezeichne die aktuelle Studie E vs. A.

Beweis: Es wurde vorausgesetzt, dass die Zufallsvariablen X_i in jeder Stichprobe unabhängig identisch normalverteilt sind und eine Varianz σ^2 haben: $Var(X_i) = \sigma^2$. Somit ist mit der Definition von n^{AP} und n^{EA}:

$$
\begin{aligned}
Var(\bar{X}^A_{(1)}) = \sigma^2 / n^A_{(1)} \quad & , \quad Var(\bar{X}^P_{(1)}) = \sigma^2 / n^P_{(1)} \\
Var(\bar{X}^E_{(2)}) = \sigma^2 / n^E_{(2)} \quad & , \quad Var(\bar{X}^A_{(2)}) = \sigma^2 / n^A_{(2)} \\
Var(\hat{\theta}^{AP}) &= \frac{\sigma^2}{n^A_{(1)}} + \frac{\sigma^2}{n^P_{(1)}} = \sigma^2 \frac{n^A_{(1)} + n^P_{(1)}}{n^A_{(1)} n^P_{(1)}} = \frac{2\sigma^2}{n^{AP}} \\
Var(\hat{\theta}^{EA}) &= \frac{\sigma^2}{n^E_{(2)}} + \frac{\sigma^2}{n^A_{(2)}} = \sigma^2 \frac{n^E_{(2)} + n^A_{(2)}}{n^E_{(2)} n^A_{(2)}} = \frac{2\sigma^2}{n^{EA}} \\
Var(\hat{\theta}^{AP} + \hat{\theta}^{EA}) &= \frac{2\sigma^2}{n^{AP}} + \frac{2\sigma^2}{n^{EA}} = 2\sigma^2 \frac{n^{AP} + n^{EA}}{n^{AP} n^{EA}}
\end{aligned}
$$

Dann sind die Standardabweichungen gegeben durch

$$SE^{AP} = \sqrt{\frac{2\sigma^2}{n^{AP}}} \text{ und } SE^{EA} = \sqrt{\frac{2\sigma^2}{n^{EA}}} \text{ sowie } SE = \sqrt{2\sigma^2 \frac{n^{AP} + n^{EA}}{n^{AP} n^{EA}}}.$$

Durch Einsetzen der beiden obigen Standardabweichungen in die Formel für $\delta_P(0)$ in Satz 3.1 ist das Korollar bewiesen. □

Korollar 3.4 *Zu den Voraussetzungen des Korollars 3.3 soll gleicher Stichprobenumfang $n^{AP} = n^{EA}$ hinzugenommen werden. Dann wird $\delta_P(0)$ durch einfache Ersetzung zu*

$$\delta''_P(0) := \hat{\theta}^{AP} - (\sqrt{2} - 1)\, z_{1-\alpha}\, SE^{AP}.$$

Die Äquivalenzgrenze $\delta_P(0)$ unterscheidet sich von der unteren Konfidenzintervallgrenze L^{AP} nur durch einen Faktor. Dieser Faktor ist im Satz 3.1 von λ und Standardabweichungen abhängig, in Korollar 3.2 nur von den Standardabweichungen bzw. von Fallzahlen in Korollar 3.3. Dieser Faktor soll als Funktion geschrieben und gedeutet werden.

$$g(\lambda, SE^{AP}, SE^{EA}) := \sqrt{(1-\lambda)^2 + \frac{(SE^{EA})^2}{(SE^{AP})^2}} - \frac{SE^{EA}}{SE^{AP}} \quad \text{mit} \quad \lambda, SE^{AP}, SE^{EA} \in \mathbb{R}^+.$$

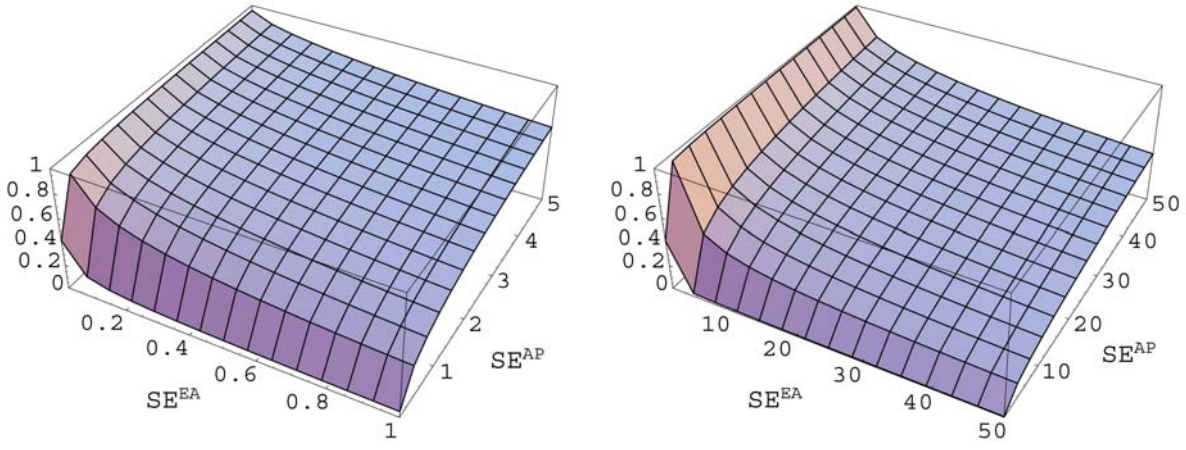


Abbildung 6: Darstellung des Faktors $g(\lambda, SE^{AP}, SE^{EA})$ in $\delta_P(\lambda)$ für $\lambda = 0$

Die Eigenschaften der Funktion $g(\cdot,\cdot,\cdot)$ sind:

1. da $(1-\lambda)^2 \geq 0$ ist auch $g(\cdot,\cdot,\cdot) \geq 0$
2. für $\lambda = 1$ ist $g(\cdot,\cdot,\cdot) = 0$
3. $g(\cdot,\cdot,\cdot) \in [0,(1-\lambda)^2]$
4. $g(\cdot,\cdot,\cdot) \in [0,1] \subset \mathbb{R}$, wenn $\lambda \in [0,1]$
5. für festes λ und SE^{AP} sowie $SE^{EA} \to 0$ gilt $g(\cdot,\cdot,\cdot) \to (1-\lambda)^2$
6. für festes λ und SE^{EA} sowie $SE^{AP} \to 0$ gilt $g(\cdot,\cdot,\cdot) \to 0$
7. für $\lambda = 0$ und $SE^{AP} = SE^{EA}$ ist $g(\cdot,\cdot,\cdot) = \sqrt{2} - 1$
8. Abbildung 6 zeigt Beispiele des Funktionsgraphen

Um die Nicht-Unterlegenheitsgrenze $\delta_P(\lambda)$ für $\lambda \in [0,1]$ zu interpretieren, sollen zum Teil nur die Spezialfälle mit $\lambda = 0$ ($\delta_P(0)$), Varianzhomogenität ($\delta'_P(0)$), bzw. gleichen Stichprobenumfängen ($\delta''_P(0)$) betrachtet werden (Abbildung 8 auf Seite 48):

- Ist die aktive Kontrolle erheblich besser als Placebo, so ist $\hat{\theta}^{AP}$ groß und somit auch $\delta_P(\lambda)$. Damit lässt sich die Überlegenheit von E gegenüber Placebo leicht zeigen.
- Wird die Hypothese (5) und ein $\lambda > 0$ sowie $\lambda \to 1$ verwendet, verkleinert sich Äquivalenzgrenze $\delta_P(\lambda)$. Dadurch ist es schwieriger die Hypothese zu verwerfen. Es gilt $(1-\lambda)\,\delta_P(0) \leq \delta_P(\lambda) \leq \delta_P(0)$ für $\lambda \in [0,1]$.
- Von dem geschätzten Effekt von A gegenüber P wird ein Anteil des Standardfehlers abgezogen. Dieses Vorgehen wird bei der Berechnung von Konfidenzintervallgrenzen angewandt. $\delta_P(0)$ liegt zwischen der Punktschätzung und der unteren Konfidenzintervallgrenze L^{AP}, denn der Faktor $g(\cdot,\cdot,\cdot)$ ist zwischen Null und Eins. Also gilt $\delta_P(\lambda) \in \left[0,\hat{\theta}^{AP}\right)$ und $\delta_P(0) \in \left(L^{AP},\hat{\theta}^{AP}\right)$.
- Der Wertebereich von $\delta_P(0)$ liegt im Intervall $\left(0,\hat{\theta}^{AP}\right)$, denn neben dem obigen Punkt wurde auch die Wirksamkeit von A angenommenen (untere Konfidenzintervallgrenze $L^{AP} > 0$).

- Es gilt $0 \leq \delta_{pbo}(\lambda) := (1-\lambda)\ L^{AP} \leq \delta_P(\lambda) \leq \delta_P(0)$, obwohl mit Satz 3.1 auch die bei WIENS gestellten Bedingungen erfüllt sind. Damit ist $\delta_{pbo}(\lambda)$ zu konservativ für die Hypothesen (5) und (6).

- Bei großen Studien für die Nicht-Unterlegenheit verkleinert sich die Standardabweichung des Schätzers. Aufgrund der Eigenschaft 5 wird daher $\delta_P(0)$ nahe bei der unteren Konfidenzintervallgrenze L^{AP} liegen.

- Das Vertrauen in die Punktschätzung $\hat{\theta}^{AP}$ für sehr große Fallzahlen n^{AP} ist groß: Je größer die Fallzahl für die Ermittlung des Effektes von A ist, desto genauer ist die Schätzung $\hat{\theta}^{AP}$, desto kleiner ist SE^{AP} und desto näher darf $\delta_P(0)$ bei der Schätzung $\hat{\theta}^{AP}$ liegen. Tatsächlich gilt für $n^{AP} \to \infty$ oder $SE^{AP} \to 0\ :\ \delta_P(0) \to \hat{\theta}^{AP}$ (Eigenschaft 6).

- Ausgehend davon, dass $n^{AP} = n^{EA}$ und dass sich die untere Konfidenzintervallgrenze für θ^{AP} der Null beliebig nähert, ist $\delta''_P(0)$ etwa das 0.6-fache der Punktschätzung $\hat{\theta}^{AP}$. Würde die Punktschätzung als „established superiority“ eingestuft, käme das dem Vorschlag des 0.5-fachen des Effektes im Concept Paper [37] sehr nahe.

Fallzahlplanung Für die Fallzahlplanung einer Äquivalenzstudie ist es wesentlich die Äquivalenzgrenze festzulegen. Je kleiner die Grenze, desto größer ist die benötigte Fallzahl n^{EA}. Bei der Bestimmung der Fallzahl für die Äquivalenzstudie besteht ein grundsätzliches Problem, wenn $\delta_P(0)$ verwendet werden soll: Einerseits hängt n^{EA} von $\delta_P(0)$ ab, andererseits wird $\delta_P(0)$ von n^{EA} (bzw. SE^{EA}) beeinflusst. Im Korollar 3.3 wird diese Abhängigkeit besonders deutlich, ist aber auch in der Definition von $\delta_P(\lambda)$ im Satz 3.1 in der Standardabweichung SE^{EA} enthalten. Ursächlich dafür ist, dass mit dem Test sowohl die $\delta_P(0)$-Nicht-Unterlegenheit als auch die Überlegenheit gegenüber Placebo, d.h. die Wirksamkeit gezeigt werden soll. Aufgrund dessen wird sowohl n^{AP} mit dem Konfidenzintervall für A vs. P als auch n^{EA} mit dem Konfidenzintervall für E vs. A in die Formel für $\delta_P(0)$ eingebracht. Es gibt Situationen, in denen dieser Zirkelschluss nicht auftritt oder umgangen werden kann:

- Ist die Äquivalenzfragestellung in einer Studie nicht die Hauptfragestellung, hängt die Fallzahl n^{EA} *nicht* von $\delta_P(0)$ ab. So kann bei dreiarmigen Studien die Fallzahlplanung aufgrund einer Überlegenheitsfragestellung zustande kommen, die Nicht-Unterlegenheit ist dann nur zweitrangig.

- Bei Studien, in denen δ_{clin} kleiner ist als alle $\delta_P(0) \in (L^{AP}, \hat{\theta}^{AP})$, wird δ_{clin} für die Fallzahlplanung verwendet. Es ist davon auszugehen, dass diese Situation in der Praxis allerdings sehr selten auftreten wird. Es würde bedeuten, dass mit der historischen Studie A vs. P nicht nur Signifikanz, sondern sogar Relevanz gezeigt worden wäre.

- Handelt es sich bei der zu planenden Studie um eine Meta-Analyse, so ist die Fallzahl nicht zu wählen. Die bereits durchgeführten Einzelstudien geben die Fallzahl vor. Meta-Analysen sind Beobachtungsstudien [100].

Falls $\delta_P(0)$ für die Fallzahlplanung herangezogen werden muss, kann iterativ bei $\delta_P(0) = \hat{\theta}^{AP}$ begonnen, die dann nötige Fallzahl ermittelt und iterativ für die Berechnung von einem neuen $\delta_P(0)$ zugrunde gelegt werden. Für den Fall der Nicht-Unterlegenheit mit

Tabelle 3: Äquivalenzgrenzen $\delta'_P(0)$ im Fall bekannter und homogener Varianzen

n^{AP}	n^{EA}	niter	L^{AP}	Power	$\delta'_P(0)$
10	.	.	-2.36	.	.
50	121	72	1.71	0.802	3.2
100	72	23	2.67	0.800	4.1
500	53	4	3.96	0.801	4.8
1000	52	3	4.26	0.806	4.9

n^{AP} = Fallzahl aus dem historischen Vergleich von A vs. P, n^{EA} = berechnete Fallzahl für die Nicht-Unterlegenheitsstudie (und damit für den indirekten Wirksamkeitsnachweis E vs. P), niter = Anzahl der Iterationen, L^{AP} = untere Grenze des einseitigen 95% Konfidenzintervalls, $\delta'_P(0)$ aus Korollar 3.3, für alle Berechnungen wurden folgende Parameter zugrunde gelegt: $\alpha = 0.05$, Vorgabe für die Power = $(1 - \beta) = 0.8$, $\sigma = 10$ und ein Punktschätzer $\hat{\theta}^{AP} = 5$.

bekannten sowie homogenen Varianzen (wie im Korollar 3.3) ist das SAS-Programm in Abschnitt 8.3 auf Seite 101 im Anhang gegeben. Im fiktiven Beispiel in Tabelle 3 ist deutlich der Einfluss der Fallzahl der historischen Wirksamkeitsstudie von A zu erkennen: Je größer die Fallzahl n^{AP}, desto genauer ist der Punktschätzer und desto größer darf $\delta_P(0)$ sein, um trotzdem insgesamt mit der Fehlerwahrscheinlichkeit α auf die Wirksamkeit von E zu schließen. Entsprechend verringert sich die Fallzahl für die Nicht-Unterlegenheitsstudie E vs. A. Dabei soll eine Power von 0.8 bei der Identität der beiden zu vergleichenden Therapien erreicht werden. Die fehlenden Werte, die durch einen Punkt gekennzeichnet

sind, entstehen, weil im ersten Fall die Fallzahl $n^{AP} = 10$ nicht ausreicht, um eine Wirksamkeit von A zu zeigen. Dies ist jedoch die Voraussetzung, um A als Kontrollgruppe zu verwenden.

3.2 δ-Problematik bei Meta-Analysen

Für jede Meta-Analyse, die das Ziel verfolgt, Äquivalenz oder Nicht-Unterlegenheit nachzuweisen, ist die Diskussion um die Äquivalenzgrenzen ebenso zu führen wie bei Einzelstudien. Allerdings wird die Problematik dadurch erhöht, dass die der Meta-Analyse zugrunde liegenden Einzelstudien unter Umständen unterschiedliche Äquivalenzgrenzen verwenden. Es gibt keine methodischen Arbeiten, die sich mit diesem speziellen Problem auseinandersetzen. Bei Meta-Analysen treten zusätzlich zu den in Abschnitt 3.1 diskutierten Punkten weitere Probleme auf:

- Definitions-Problem: Muss für eine Meta-Analyse ein δ festgelegt werden? Welche Frage soll damit beantwortet werden (Abschnitt 3.2)?
- δ_i-Problem: Aus verschiedenen Studien können verschiedene Äquivalenzgrenzen δ_i bekannt sein. Wie beeinflussen sie die Wahl von δ (Abschnitt 3.3)?
  - Auswahl-Problem: Welche Studien sollten herangezogen werden?
  - Studientypen-Problem: Wie sollen dabei Überlegenheitsstudien berücksichtigt werden?
  - Fallzahl-Problem: Welchen Einfluss haben dabei Aspekte der Durchführbarkeit bei Einzelstudien, die die Wahl des δ_i beeinflusst haben?
  - Zielgrößen-Problem: Unter Umständen werden geringfügig verschiedene Zielgrößen in den Studien verwendet. Ist eine Anpassung der δ_i möglich, um die Grenzen vergleichen zu können? Sollte die Zielgröße als Einschlusskriterium für die Meta-Analyse gewählt werden?
  - Skalen-Problem: Unter Umständen werden verschiedene Skalen in den Studien verwendet. Ist eine Anpassung der δ_i möglich, um die Grenzen vergleichen zu können? Sollte die Skala als Einschlusskriterium für die Meta-Analyse gewählt werden?
- δ-Entscheidungs-Problem: Welches δ ist letztlich zu verwenden, wenn man zusätzlich die Ansätze aus Abschnitt 3.1 berücksichtigt (Abschnitt 3.4).

Definitions-Problem Eine Meta-Analyse kann als reines Schätzproblem aufgefasst werden. Statistisches Testen im Rahmen der Zusammenfassung von Evidenz zu einer Fragestellung muss nicht durchgeführt werden. Dann stellt sich die Frage, ob es überhaupt notwendig ist, Äquivalenzgrenzen vor Durchführung der Meta-Analyse festzulegen.

Unter der Annahme, dass die Äquivalenzgrenzen feste, nicht zufällige Größen sind, ist die Schätzung der Konfidenzintervalle auch ohne Kenntnis von δ möglich (Abschnitt 2). Eine Entscheidung für oder gegen eine Äquivalenz kann selbstverständlich auch post-hoc getroffen werden. Eine solche Aussage ist jedoch angreifbarer, weil eine datengesteuerte Entscheidung vorgeworfen werden kann. Ist das Konfidenzintervall aber derart klein, dass die Ränder *ohne Zweifel* innerhalb des Äquivalenzbereiches liegen, wäre eine Diskussion um die mehr oder minder exakte Bestimmung von δ hinfällig. Diese Situation wird jedoch selten eintreten.

Eine weitere Möglichkeit der Analyse, ohne die Äquivalenzgrenzen vorab zu spezifizieren, ist die Darstellung von p als Funktion von δ, $p(\delta)$. Damit lässt sich überprüfen, wie sich der p-Wert des Tests auf Äquivalenz abhängig von δ verändert. Hier soll die Nicht-Unterlegenheit als Beispiel dienen, dazu wird lediglich die Formel für den p-Wert verwendet ((7) auf Seite 21).

Für den Test auf Nicht-Unterlegenheit (im REM) ergibt sich

$$p = p(\delta) = 1 - \Phi(T_{\theta \leq -\delta}) = 1 - \Phi\left(T_{\theta \leq 0} + \delta\sqrt{\sum w_i}\right).$$

Die graphische Darstellung dieser Funktion und somit der Einfluss von der Wahl von δ auf den p-Wert für $T_{\theta \leq 0} = -0.675$ und $\sqrt{\sum w_i} = 9.18$ (Oxaceprol-Beispiel für alle vier Studien, Abschnitt 1.2), ist in Abbildung 7 dargestellt.

Falls ein großes δ gewählt wird, ist die Hypothese der Nicht-Unterlegenheits-Fragestellung leicht abzulehnen, denn der dazugehörige p-Wert ist klein. Die Interpretation dieser Funktion ist entsprechend derjenigen von Konfidenzintervallen: Sind auf der x-Achse die in Frage kommenden Werte für δ derart groß, dass der dazugehörige p-Wert immer unter dem vorgegebenen Signifikanzniveau α liegt, so bestätigt sich die Nicht-Unterlegenheit.

Die Schwelle der Testentscheidung ($p = \alpha$) ist dann erreicht, wenn die untere Konfidenzintervallgrenze gerade die Äquivalenzgrenze ist ($-\delta = L$), wie beim Intervall-Inklusions-

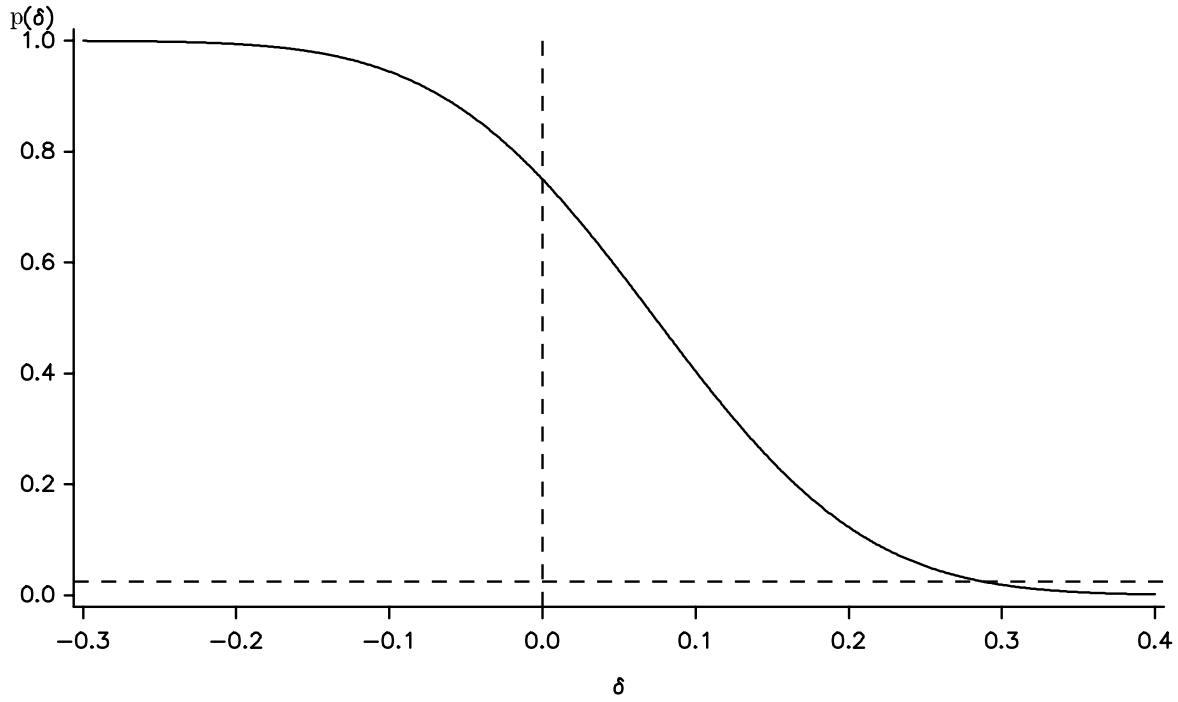


Abbildung 7: Der p-Wert für die Hypothese der Nicht-Unterlegenheit (im REM) in Abhängigkeit von δ im Oxaceprol-Beispiel.

Verfahren im Abschnitt 2.1 beschrieben. Also gilt $p(-L) = \alpha$.

$$\begin{aligned} p(-L) &= 1 - \Phi\left(T_{\theta\leq 0} - L\sqrt{\sum w_i}\right) \\ &= 1 - \Phi\left(T_{\theta\leq 0} - \left(\hat{\theta} - u_{1-\alpha}\sqrt{\frac{1}{\sum w_i}}\right)\sqrt{\sum w_i}\right) \\ &= 1 - \Phi\left(T_{\theta\leq 0} - T_{\theta\leq 0} + z_{1-\alpha}\right) \\ &= 1 - \Phi\left(z_{1-\alpha}\right) = 1 - (1-\alpha) = \alpha \end{aligned}$$

In Abbildung 7 ist daher mit dem Schnittpunkt zwischen der Funktion $p(\delta)$ und der horizontalen Referenzlinie bei 2.5% die untere Grenze des zweiseitigen 95%-Konfidenzintervalls gegeben. Diese graphische Methode kann zur späteren Interpretation der Ergebnisse dienen. Zwar beinhaltet die Darstellung mehr als die Information des Konfidenzintervalls für den Effekt θ, doch zur Interpretation der Ergebnisse wird vor allem der oben beschriebene Schnittpunkt herangezogen. Das Verfahren kann daher ergänzend eingesetzt werden, ist aber nicht sehr mächtig.

Forderung Wie oben dargestellt, kann die Meta-Analyse als reines Schätzproblem aufgefasst und Punktschätzung, Konfidenzintervall sowie Graphiken vom obigen Typ interpretiert werden. Trotzdem sollte δ im Studienprotokoll genau festgelegt und begründet werden. Nur wenn die Meta-Analyse ausschließlich dazu dient, Hypothesen zu generieren oder Effekte zu schätzen, kann darauf verzichtet werden, δ festzulegen. Somit ist spätere Willkür ausgeschlossen und dennoch ist eine weitreichende Interpretation der Ergebnisse möglich. Im Rahmen eines Zulassungsprozesses von Arzneimitteln erübrigt sich die Diskussion um die Notwendigkeit der Festlegung von δ, da durch einen Induktionsschluss von der Verpflichtung in Einzelstudien auch auf die Verpflichtung der Definition von δ bei Meta-Analysen geschlossen werden kann.

Beispiel In den 13 Übersichtsarbeiten im Bereich der bakteriellen Infektionen wurden elf Meta-Analysen identifiziert (Abschnitt 1.2). *Keine* dieser Arbeiten hatte eine Äquivalenzgrenze spezifiziert. Nur in einem Fall wurden die Konfidenzintervalle im Sinne einer Äquivalenzstudie interpretiert. Dort wurde bei einem Konfidenzintervall für das Odds Ratio von [0.7, 1.2] auf eine Äquivalenz geschlossen. Hier ist eine datengesteuerte Interpretation der Ergebnisse nicht ausgeschlossen, zumal beide Autoren dem Pharmakonzern angehören, dessen Produkt untersucht wurde [103].

3.3 δ_i-Problem

Es wird nun vorausgesetzt, dass ein δ für die Meta-Analyse mit einer Äquivalenzfragestellung vor Beginn der Studie festgelegt wird. Zusätzlich zu der klinischen Argumentation aus Abschnitt 3.1.1 und der Wirksamkeitsargumentation aus Abschnitt 3.1.2 ist unter Umständen ein weiterer Punkt zu berücksichtigen: Bereits durchgeführte und publizierte Studien haben Äquivalenzgrenzen δ_i definiert. Dabei ist zu klären, inwiefern diese Informationen berücksichtigt werden müssen. Zunächst werden einige Argumente für unterschiedliche Äquivalenzgrenzen δ_i aufgeführt.

- Die Bestimmung der Grenzen ist ein Prozess, der zum einen objektive Kriterien und zum anderen subjektive Empfindungen sowie Erfahrungen verarbeitet. Jeder Mediziner wird eigene Aspekte und Argumente aufführen; somit werden auch die Ergebnisse leicht differieren. Darüber hinaus ist die Kenntnis über bereits durchgeführte Studien wichtig. Diese können unter Umständen plausible Äquivalenzgrenzen liefern (δ_{clin} im Abschnitt 3.1).

- Die Größenordnung der Grenzen, innerhalb derer man Unterschiede zwischen zwei Therapien als klinisch irrelevant, also äquivalent bezeichnen würde, kann abhängig von der Größenordnung der Zielgröße sein. Ist etwa in einer Population die Zielgröße im Mittel bei 140 Einheiten und in einer anderen bei 100, so könnten zwei Studien (je mit einer der Populationen) ggf. auch zu verschiedenen Festlegungen für die Äquivalenzgrenzen kommen, etwa je $\pm 10\%$, also 14 bzw. 10 Einheiten. Stellt man erhebliche Unterschiede bezüglich der Äquivalenzgrenzen fest, kann unter Umständen die Durchführung der Meta-Analyse zweifelhaft sein, weil die Heterogenität der Äquivalenzgrenzen auch eine Heterogenität der Studien bedeuten kann. Im WOMAC-Beispiel im Abschnitt 3.5 ist keine Abhängigkeit der Äquivalenzgrenze von dem Baselinewert des WOMAC erkennbar.

- Andere Designkriterien könnten die Wahl der Äquivalenzgrenzen ebenfalls beeinflusst haben. So kann etwa eine Nicht-Unterlegenheitsgrenze von δ_i in einer Studie bei einer Beobachtungsdauer von zwei Monaten vertretbar sein, nicht aber bei einer Dauer von einem Monat. Im WOMAC-Beispiel im Abschnitt 3.5 könnte eine derartige Abhängigkeit vorliegen.

- Die verwendeten Äquivalenzgrenzen können aufgrund unterschiedlicher statistischer Hypothesen zustande gekommen sein. In Abschnitt 3.1 wurde auf die verschiedenen Gründe $(\delta_{clin}, \delta_P(\lambda), \delta_P(0))$ für die Wahl der Grenzen bei Äquivalenzstudien hingewiesen. Je nach Fragestellung, die hinter der Äquivalenzgrenze steht, könnten daher verschiedene Grenzen angesetzt werden.

- Äquivalenzstudien benötigen häufig eine recht hohe Fallzahl [38]. Aus diesem Grund wird unter Umständen bei der Planung der Studie schon ein entsprechend großzügiges δ_i gewählt, um den „Erfolg“ der Studie nicht zu gefährden. Auch wenn diese pragmatischen aus der Studienplanung und -durchführung kommenden Einflüsse keine Rolle spielen sollten, geschieht dies in der Praxis doch.

Minimum-Lösung Eine Lösung für die Wahl eines δ, dass die bekannten Äquivalenzgrenzen berücksichtigt, ist die Verwendung des Minimums aller verfügbaren m Äquivalenzgrenzen, also

$$\delta_{min} := min(\delta_1, \ldots, \delta_m). \tag{9}$$

Die Minimum-Lösung ist eine eher konservative Methode, um zu einem δ zu gelangen. Das heißt die Hypothese der Äquivalenz wird eher nicht abgelehnt: Je kleiner die Äquivalenzgrenzen gewählt werden, desto schwieriger ist es, dieses strenge Kriterium zu erreichen. Es ist jedoch nicht sicher, dass die Minimum-Lösung konservativ ist, da vielleicht nur wenige oder nur recht große δ_i bekannt sind. Sicher ist, dass δ_{min} nicht allein stehen darf, andere Aspekte müssen bei der Festlegung von δ beachtet werden (Abschnitt 3.4).

Sind in Studien Fallzahl-Δ_i gegeben, können diese als Obergrenze dienen. Damit könnten die Δ_i mit in die Definition von δ_{min} aufgenommen werden, da die Fallzahl-Δ_i eher zu groß sein werden.

Ist das δ_i einer speziellen Studie extrem klein, stellt sich die Frage, ob dieses δ_i aus der Ermittlung von δ_{min} auszuschließen ist. Da in einem solchen Fall die Definition von δ_{min} post-hoc beeinflusst werden kann, ist die Situation wenn möglich zu vermeiden. Handelt es sich jedoch um einen nicht nachvollziehbar kleinen Wert, kann mit einer guten Begründung der entsprechende Wert ignoriert werden. Diese Argumentation spricht für die Verwendung von δ_{Q1}.

Q1-Lösung Eine weitere Lösung für die Wahl eines δ ist die Verwendung des 1. Quartils aller verfügbaren m Äquivalenzgrenzen, also

$$\delta_{Q1} := Q1(\delta_1, \ldots, \delta_m). \tag{10}$$

Diese Alternative ist robust gegen Ausreißer, also gegen Studien, die extrem kleine Äquivalenzgrenzen verwenden. Erst bei $m \geq 5$ kann ein Unterschied zwischen δ_{min} und δ_{Q1} entstehen. Im Folgenden sollen δ_{min} und δ_{Q1} gleichwertig betrachtet werden.

Die Definitionen von δ_{min} und δ_{Q1} werfen einige Probleme auf, die schon in Abschnitt 3.2 genannt wurden und im Folgenden behandelt werden.

Auswahl-Problem Welche m Äquivalenzgrenzen sollen für δ_{min} bzw. δ_{Q1} verwendet werden? Der engste in Frage kommende Kreis von Äquivalenzgrenzen wird durch die Einzelstudien beschrieben, die in die Meta-Analyse eingebracht werden. All diese Äquivalenzstudien, die ein δ_i liefern, sollten in (9) bzw. (10) eingehen. Im zweiten Schritt sollten diejenigen Studien mit berücksichtigt werden, die aufgrund von mangelnden Informationen ausgeschlossen werden mussten. Auch diese δ_i sollten berücksichtigt werden.

Ein weiterer Aspekt ist, wie mit Studien umgegangen werden soll, die andere Therapieformen evaluieren oder ein anderes Design verwenden. Sicher ist, dass die gleiche Zielgröße

verwendet werden muss, über alle anderen Kriterien (Indikation, Therapieform, Studiendesign) bleibt zu diskutieren.

Bei der Auswahl der Äquivalenzgrenzen ist die praktische Durchführbarkeit zu beachten: Der Arbeitsaufwand würde unter Umständen sehr groß, wenn alle publizierten Äquivalenzgrenzen mit der entsprechenden Zielgröße für δ_{min} bzw. δ_{Q1} relevant wären. Eine andere Auswahlmethode, bei der etwa alle weiteren Äquivalenzgrenzen δ_i in (9) bzw. (10) berücksichtigt werden, die den Studienplanern der Meta-Analyse zur Verfügung stehen, würde die Möglichkeit der Selektion und Manipulierbarkeit geben und somit die Meta-Analyse angreibar machen. Aus Gründen der Praktikabilität ist daher die Menge derjenigen Studien sinnvoll, die aufgrund der Ein- und Ausschlusskriterien der Meta-Analyse ohnehin für die Meta-Analyse in Frage kämen.

Eine weitere Klasse von Studien, die bei der Auswahl mit berücksichtigt werden sollten, sind Studien die die Ermittlung von minimal klinisch relevanten Differenzen zum Ziel haben. Um diese Art der Studien zu finden, ist es notwendig, die Literatursuche entsprechend auszuweiten. Das WOMAC Beispiel in Abschnitt 3.5 liefert drei solcher Studien.

Studientypen-Problem Sind Überlegenheitsstudien unter den ausgewählten Einzelstudien, gibt es zwei Möglichkeiten für die Minimum-Lösung. Zum einen könnten diese Studien bezüglich der Definition von δ_{min} ignoriert werden, da kein δ_i explizit definiert wurde. Zum anderen könnten sie als einseitige Überlegenheitsstudie und somit als Spezialfall einer Nicht-Unterlegenheitsstudie mit $\delta_i = 0$ aufgenommen werden. Dies würde zu einem $\delta_{min} = 0$ führen. Die zweite Möglichkeit hätte folgenden Grund: Wurden bereits erste Überlegenheitsstudien von E gegenüber A durchgeführt, wäre es nicht mehr vertretbar, von der Meta-Analyse weniger zu fordern. In der Praxis jedoch ist es häufig gerade umgekehrt. Es wurden in der Vergangenheit viele Überlegenheitsstudien durchgeführt und man interpretierte die Ergebnisse post-hoc im Sinne einer Äquivalenzfragestellung. Durch diese Überlegenheitsstudien würde δ_{min} auf Null gesenkt, obwohl das in diesem Fall nicht gerechtfertigt ist. Bei einer Nicht-Unterlegenheitsfragestellung für die Meta-Analyse sollten daher Überlegenheitsstudien bei der Festlegung von δ_{min} ignoriert werden.

Für echte bzw. zweiseitige Äquivalenzfragestellungen *müssen* Überlegenheitsstudien aus der Definition (9) ausgeschlossen werden, da eine zweiseitige Äquivalenz mit $\delta = 0$ statistisch nie gezeigt werden könnte.

Fallzahl-Problem Ein Mangel an Power aufgrund der hohen benötigten Fallzahl, wie er in Einzelstudien vorkommt, sollte bei der Meta-Analyse keine Rolle spielen. Da sich aber die Fallzahl ggf. in den δ_i niederschlägt, sind die entsprechenden Äquivalenzgrenzen unter Umständen für eine realistische und klinisch noch zu vertretende Äquivalenz zu hoch. Da aber δ_{min} bzw. δ_{Q1} nur *eine* Stufe im Entscheidungsprozess für die Wahl eines δ darstellen und sowohl das Minimum als auch das 1. Quartil von zu hohen δ_i unbeeinflusst bleiben, können diese potentiellen Einflüsse unberücksichtigt bleiben.

Zielgrößen- und Skalen-Problem Leicht unterschiedliche Zielgrößen können unter Umständen in einer Meta-Analyse zusammengeführt werden, beispielsweise bei Schmerzscores, die aus verschiedenen einzelnen Komponenten bestehen. Sofern diese Entscheidung für die Meta-Analyse gefällt und im Protokoll fixiert wurde, ist die Verwendung entsprechender Äquivalenzgrenzen δ_i aus diesen Studien unproblematisch. Jedoch wird dieses Problem oft mit dem Skalen-Problem zusammen auftreten: Die Einzelstudien verwenden unterschiedliche Skalen für die Zielgröße. Dann sind alle diejenigen Studien für δ_{min} bzw. δ_{Q1} zu verwenden, deren Äquivalenzgrenzen sich transformieren lassen. Für eine Transformation von einem δ_i', das in der Studie festgelegt wurde, auf ein δ_i für die Formel (9) bzw. (10) sind unter Umständen Parameterschätzungen nötig. Diese Parameterschätzungen müssen in der Publikaton genannt sein. Mitunter werden diese in den Einzelstudien jedoch nicht genannt (das Fehlen könnte sogar der Ausschlussgrund sein). Andererseits wurde bereits darauf eingegangen, dass diese Studien mit in die Definition von δ_{min} bzw. δ_{Q1} aufgenommen werden sollten. In diesem speziellen Fall müssen die Studien aus der Berechnung von δ_{min} bzw. δ_{Q1} aufgrund der nicht möglichen Transformation ausgeschlossen werden. Mitunter können aber entsprechende Parameterschätzungen aus vergleichbaren Studien verwendet werden. Auch hier sind konservative Ansätze zu verfolgen.

Um die Transformation an einem Beispiel zu erläutern, soll das Oxaceprol-Beispiel herangezogen werden (Tabelle 1 auf Seite 5): Bei einer stetigen Zielgröße, bei der die standardisierte Differenz als Effekt θ_i verwendet werden soll, müssen auch Äquivalenzgrenzen (δ_i' auf der Skala der Zielgröße) entsprechend mit dem Varianzschätzer aus der Studie transformiert werden. Somit können ursprünglich gleiche Äquivalenzgrenzen (δ_i') nach der Transformation (δ_i) unterschiedlich sein: Im Oxaceprol-Beispiel ist $\delta_1' = 2$ [6] und $\delta_2' = 2$ [49] aber $\hat{\sigma}_1 = 4.2$ und $\hat{\sigma}_2 = 3.8$. Somit ist $\delta_1 = \delta_1'/\hat{\sigma}_1 = 2/4.2 = 0.48$ und $\delta_2 = \delta_2'/\hat{\sigma}_2 = 2/3.8 = 0.53$.

3.4 Entscheidungprozess für die Äquivalenzgrenzen

Zur Definition der Äquivalenzgrenzen wurden bisher einzelne Konzepte vorgestellt und deren Vor- und Nachteile diskutiert. Die verschiedenen Ansätze wurden mit δ_{clin}, $\delta_{pbo}(\lambda)$, $\delta_P(\lambda)$, $\delta_P(0)$ und δ_{min} bzw. δ_{Q1} bezeichnet. Zu diskutieren bleibt, ob *ein* δ für die induktive Statistik der Meta-Analyse festgelegt werden kann.

Auf der einen Seite ist eine neue inhaltliche Auseinandersetzung mit den Äquivalenzgrenzen auch vor der Durchführung einer *Meta-Analyse* mit einer Äquivalenzfragestellung notwendig. Klinische Experten sollten die notwendige Expertise einbringen (δ_{clin}). Zum anderen muss immer berücksichtigt werden, dass zumindest auf eine bessere Effektivität gegenüber Placebo geschlossen werden kann ($\delta_P(0)$). Außerdem ist das know-how und die Bedingungen der Einzelstudien, welches sich in δ_{min} bzw. δ_{Q1} ausdrückt, zu berücksichtigen. Darüberhinaus hat eine Meta-Analyse aufgrund der Erhöhung der Power die Möglichkeit auch kleine Effektunterschiede aufzudecken; somit erhöht sich ebenso die Power, eine Äquivalenz oder Nicht-Unterlegenheit zu zeigen. Daher sollte die Meta-Analyse die Möglichkeit besitzen, die strengste Form zu zeigen, so dass

$$\delta = min(\delta_{clin}, \delta_P(0), \delta_{Q1}) \tag{11}$$

gesetzt werden kann (oder δ_{min} statt δ_{Q1}). $\delta_{pbo}(\lambda)$ taucht deshalb in der Definition (11) nicht auf, weil die Definition von $\delta_{pbo}(\lambda)$ zu restriktiv ist. Erfüllt wird die Zielsetzung (Überlegenheit gegenüber Placebo) auch mit $\delta_P(0)$ (Abschnitt 3.1). Wichtig ist bei (11) die Unabhängigkeit von Restriktionen, die bei Einzelstudien durch die Fallzahl bedingt sein können. Für die Meta-Analyse sollte ein faires und davon unbeeinflusstes δ festgesetzt werden.

Im Gegensatz zu (11) kann ein hierarchisches Vorgehen empfohlen werden, bei dem die verschiedenen δ und damit die Hypothesen a-priori in eine Rangordnung gebracht werden. Eine Hypothese kann getestet werden, wenn die vorherige bereits abgelehnt wurde. Wenn auf jeder Stufe ein Niveau-α-Test verwendet wird, hält die Prozedur das globale Niveau α ein. Folgende Reihenfolge wird vorgeschlagen: $\delta_P(0) \rightarrow \delta_{clin} \rightarrow \delta_{Q1} \rightarrow \delta_P(\lambda)$, das heißt:

1. $H_0^{EP} : \theta^{EP} \leq 0$
2. $H_0^{EA} : \theta^{EA} \leq -\delta_{clin}$
3. $H_0^{EA} : \theta \leq -\delta_{Q1}$
4. $H_0^{EP}(\lambda) : \theta^{EP} \leq \lambda\, \theta^{AP}$.

3.5 Beispiel: WOMAC in Arthrosestudien

Einführung Arthrose ist eine degenerative Gelenkerkrankung mit hoher Prävalenz, die mit Funktionsbeeinträchtigungen und chronischen Schmerzen einhergeht. In diesem Krankheitsgebiet werden vor allem symptomatische Therapien geprüft. Es gibt eine Reihe von etablierten Medikamenten, die in vergleichenden Studien als Standardtherapie verwendet werden (zum Beispiel Naproxen, Diclofenac oder Ibuprofen). Deshalb sind häufig Äquivalenzstudien nötig (Abschnitt 1.2).

Im folgenden Beispiel soll der validierte WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index) als Zielgröße herangezogen werden [10, 11]. Der WOMAC wird für klinische Studien im Bereich der Knie- und Hüftgelenksarthrose empfohlen [36, 20]. Der Evaluierungsbogen wird von den Patienten selbst ausgefüllt. Er besteht aus den Dimensionen Schmerz (5 Fragen), Steifigkeit (2 Fragen) sowie Funktionalität (17 Fragen). Dabei ist zu beachten, dass es verschiedene Versionen gibt, die unterschiedliche Skalen verwenden: VAS (=visual analog scale, 0-100mm), Likert Skala (0-4) oder eine 11er Skala (0-10). Alle folgenden Angaben wurden aus Gründen der Vergleichbarkeit auf die VAS transformiert (Skalenproblem).

Für einen Gesamtscore werden von Bellamy fünf Alternativen der Aggregation genannt [13], von denen nur einer („Total WOMAC Score“) regelmäßig Verwendung findet und ein anderer („Normalized WOMAC Score“) nur selten erwähnt wird [90]. Beim „Total WOMAC Score“ werden alle 24 Fragen gleich gewichtet, die drei Dimensionen Schmerz, Steifigkeit und Funktionalität also je nach Anzahl der Fragen pro Dimension gewichtet. Beim „Normalized WOMAC Score“ werden zunächst die Scores der drei Komponenten ermittelt. Diese werden dann in einem zweiten Schritt gleich gewichtet zusammengefasst. Im Folgenden wird vor allem die Schmerz-Subskala verwendet.

Als Beispiel soll eine Meta-Analyse mit drei Einzelstudien von Deeks et al. herangezogen werden, die die Wirksamkeit und Sicherheit von Celecoxib für die Behandlung von Arthrose und rheumatoider Arthritis untersucht [27].

Methoden (i) Um δ_{clin} festlegen zu können, wurde Herr Professor Scharf im Rahmen einer Gonarthrose-Studie befragt. (ii) Um δ_{min} bzw. δ_{Q1} zu definieren, wurde versucht die Äquivalenzgrenzen aus den Studien zu ermitteln. Dazu wurden wenn möglich die einzelnen Publikationen und der Bericht im Rahmen des amerikanischen Zulassungsantrages verwendet. (iii) Zur Ermittlung von $\delta_P(\lambda)$ sowie $\delta_{pbo}(\lambda)$ wurde eine randomisier-

te, Placebo-kontrollierte Studie von MAKAROWSKI et al. verwendet, die *auch* Naproxen nach zwölf Wochen untersuchte und den WOMAC als Zielgröße verwendete [68]. (iv) Als weitere Informationsquelle wurden Guidelines und Richtlinien herangezogen. (v) Eine Literatursuche in MEDLINE diente als Grundlage über bereits verwendete oder empfohlene Äquivalenzgrenzen: WOMAC[All Fields] AND (Review[ptyp] OR Meta-Analysis[ptyp] OR Clinical Trial[ptyp] OR clinically[All Fields] OR relevant[All Fields] OR minimal[All Fields]). Verwendete Fallzahl-Δ wurden zusätzlich als Obergrenze für eine Äquivalenzgrenze berücksichtigt.

Ergebnisse (i) Es konnte $\delta_{clin} = 8$ definiert werden, das heißt eine Veränderung von 20% bei einer Ausgangsgröße von 40 Punkten (SCHARF 2002, persönliche Mitteilung). (ii) Keine der drei Studien nennt eine Äquivalenzgrenze, so dass kein δ_{min} bzw. δ_{Q1} gemäß (9) bzw. (10) definiert werden konnte. (iii) Die Studie von MAKAROWSKI et al. liefert $\hat{\theta}^{AP} = 9.95$ (Differenz der Mittelwerte) mit $SE^{AP} = 2.05$ und einem 95% Konfidenzintervall [5.9, 14.0] (rekalkuliert). Die Meta-Analyse von DEEKS et al. schätzt $\hat{\theta}^{EA} = 0.1$ (Differenz der Mittelwerte), $SE^{EA} = 3.7$, und ein 95% Konfidenzintervall [-7.2, 7.3], wobei hohe Werte für Celecoxib sprechen. Daher ist bei einem α-Level von 2.5%: $\delta_{pbo}(0) = L^{AP} = 5.9$ und $\delta_P(0) = 9.08$. Des Weiteren ist mit $\lambda = 0.5$: $\delta_{pbo}(0.5) = 0.5\ L^{AP} = 2.95$ und $\delta_P(0.5) = 4.52$. Die Ergebnisse sind in Tabelle 4 und in Abbildung 8 dargestellt. (iv) Die beiden indikationsspezifischen (Draft-)Guidelines geben keine Empfehlungen zu möglichen Äquivalenzgrenzen [20, 36], ebensowenig sind Hinweise im „User's Guide“ für den WOMAC zu finden [13].

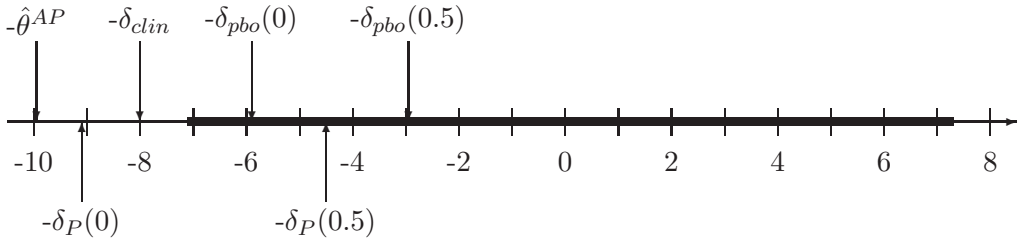


Abbildung 8: Verschiedene δ-Definitionen der WOMAC Schmerz-Subskala (0-100) mit einem 95%-Konfidenzintervall

(v) Unter den 71 gefundenen Artikeln (Stand: 15. Mai 2002) waren zwei Doppelpublikationen, 33 Studien mit dem Ziel der Überlegenheit, acht Validierungsstudien, vier Korrelationsstudien sowie 13 Studien mit anderen Zielsetzungen (zum Beispiel: einarmige follow-up Studien). Bei drei der verbleibenden elf Studien war der WOMAC nur sekundäre Zielgröße. Bei zwei weiteren Studien wurde kein δ angegeben, obwohl es sich um Studien handelte,

Tabelle 4: Ermittlung von verschiedenen δ für den WOMAC

Kennzahl		WOMAC Schmerz-Subskala
$\bar{x}^A_{(1)}$, $\bar{x}^P_{(1)}$		37.80, 47.75
$\sigma^A_{(1)}$, $\sigma^P_{(1)}$		16.0, 15.5
$n^A_{(1)}$, $n^P_{(1)}$		118, 117
$\hat{\theta}^{AP}$		9.95
SE^{AP}		2.05
$\hat{\theta}^{EA}$		0.1
SE^{EA}		3.7
$\delta_P(0.5)$	(gemäß Satz 3.1)	4.52
$\delta_P(0)$	(gemäß Korollar 3.2)	9.08
$\delta'_P(0)$	(gemäß Korollar 3.3)	- (*)
$\delta''_P(0)$	(gemäß Korollar 3.4)	8.55
$\delta_{pbo}(0)$	(gemäß (8))	5.90
$\delta_{pbo}(0.5)$	(gemäß (8))	2.95

Die Messgrößen sind auf eine (0-100)-Skala transformiert, A = Naproxen, P = Placebo, E = neue Therapieform, die Daten der placebokontrollierten Studie: Makarowski et al. [68], (*) die Vereinfachung in Korollar 3.3 kann für Einzelstudien, nicht aber für Meta-Analysen verwendet werden.

die das Hauptziel verfolgten, eine neue Therapie mit einer aktiven Standardtherapie zu vergleichen. Es blieben sechs Publikationen, die Informationen über Äquivalenzgrenzen lieferten. Die Spezifizierung einer Äquivalenzgrenze wäre aufgrund der Interpretationen in weiteren neun Studien angebracht gewesen, da es sich um mehrarmige Studien handelt die zum Teil in der Diskussion vergleichende Aussagen trafen („effective as“, „similar to“, „comparable to“).

In der Publikation von Beaupré et al. wurde mit einem Verweis auf die Arbeit von Takeda und Wessel [91] $\delta_1 = 10$ als minimal klinisch relevant eingestuft [8]. In einer weiteren Arbeit wurde eine Äquivalenzgrenze der WOMAC Schmerz-Subskala angegeben: $\delta_2 = 12$ (im Original: 60 mm auf der Summenskala 0-500 mm, [61]). Singer et al. definierten die Äquivalenzgrenze auf der Skala der Teststatistik des Mann-Whitney-Tests [88]. Diese besteht im Wesentlichen aus der Wahrscheinlichkeit, ob eine Messgröße in der einen Gruppe größer ist als in der anderen. Das Konzept ist für die Verwendung von Scores durchaus sinnvoll, jedoch schwierig zu interpretieren und nicht zu transformieren. Die Nicht-Unterlegenheitsgrenze wird mit 0.36 beziffert. Den Autoren nach entspricht dies

einer standardisierten Differenz von -0.5, tatsächlich jedoch einer *nicht* standardisierten Differenz von -0.5. Für die Berechnung wird auf Colditz et al. verwiesen [23]. Dort ist eine Umrechnung von Differenzen von Mittelwerten auf die Skala der Teststatistik des Mann-Whitney-Tests vorgestellt. Demnach entspricht unter Annahmen der Normalverteilung eine standardisierte Differenz von -0.5 einem Wert von 0.31. Aufgrund der Unsicherheit der Daten wird diese Arbeit im Weiteren nicht berücksichtigt.

Des Weiteren gibt es drei Studien, die sich *systematisch* mit der Ermittlung von klinisch relevanten Differenzen auseinandersetzen. Alle drei Arbeiten basieren auf dem gleichen Prinzip: Der WOMAC wird zum Ausgangszeitpunkt sowie nach einigen Wochen erfasst. Zusätzlich wird die klinische Veränderung auf einer Likert-Skala abgefragt (Patienteneinschätzung [3, 4, 33] oder Arzteinschätzung [33]). Der Unterschied zwischen den Gruppen „no change“ und „slightly worse“ wird als minimal klinisch relevante Verschlechterung angesehen. Der Unterschied zwischen den Gruppen „no change“ und „slightly better“ wird als minimal klinisch relevante Verbesserung betrachtet. Als minimal klinisch relevante Verschlechterung (mögliche Äquivalenzgrenze einer Nicht-Unterlegenheitsstudie) für die WOMAC Schmerz-Subskala ergibt sich $\delta_3 = 9.7$ [33], $\delta_4 = 11.0$ [3] und $\delta_5 = 6.4$ [4]. Das entspricht prozentualen Veränderungen in Bezug auf die Ausgangswerte von 14.9%, 22.8% und 13.9%. Die Daten lassen keinen Schluss auf einen Zusammenhang mit dem Publikationsjahr, der Beobachtungsdauer oder dem Baselineniveau zu. Dennoch ist ein Zusammenhang zwischen der Dauer und der minimal klinisch relevanten prozentualen Veränderung möglich (Abbildung 9).

Die Daten mit den wichtigsten Studiencharakteristika sind in Tabelle 5 dargestellt. In Tabelle 12 auf Seite 100 sind die Daten für den Gesamtscore sowie die übrigen Subskalen ersichtlich.

Einige Studien nennen eine Differenz oder Rate für die Ermittlung der Fallzahl, wie beispielsweise $\Delta = 25\%$ für die WOMAC Schmerz-Subskala bei Bellamy et al. [12]. Die Annahme basiert auf Baselinedaten einer anderen Studie [11]: 48 Punkte bedeutet $\Delta = 12$, was mit δ_2 übereinstimmt. Eine derartige Veränderung wird in diesem Kontext als klinisch wichtig angesehen („... the detection of the clinically important between drug difference ...“). Das heißt aber nicht, dass ein kleinerer Wert nicht auch schon von klinischer Bedeutung sein könnte. Im Rahmen einer post-hoc Power Analyse bezeichnen Bellamy et al. tatsächlich in der gleichen Publikation eine Veränderung von 20% als klinisch relevant. Das wiederum würde zu einem $\Delta = 9.6$ oder mit den Baselinedaten der eigenen

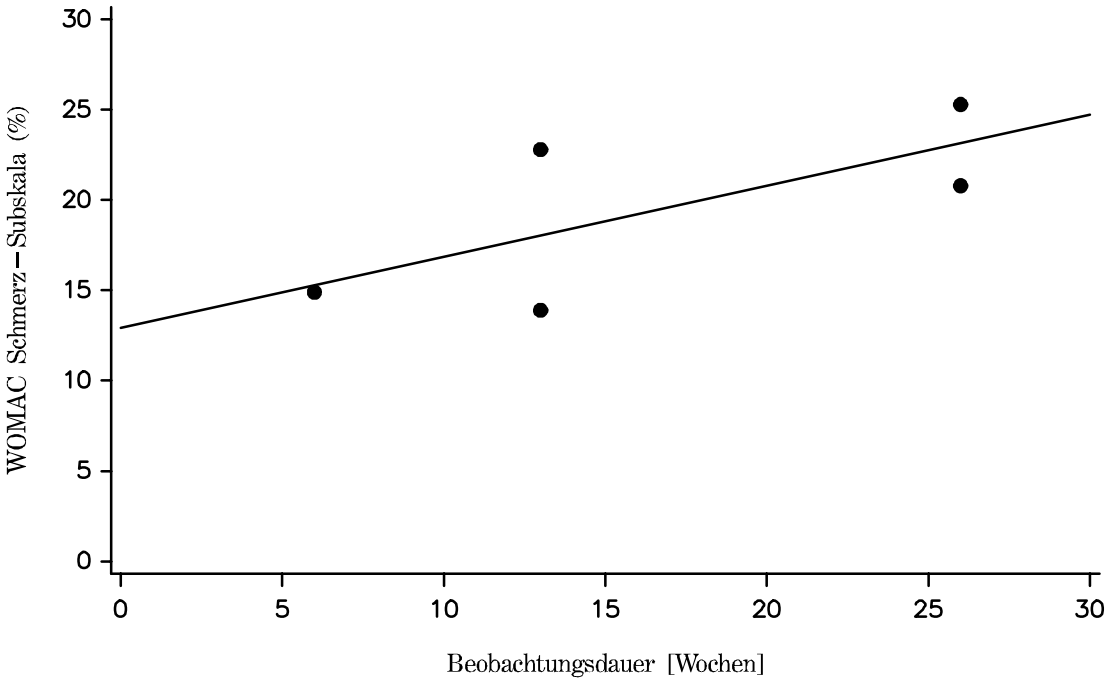


Abbildung 9: Aus 5 Studien ermittelte (untere) Äquivalenzgrenzen für die WOMAC Schmerz-Subskala: Beobachtungsdauer und Veränderung in % gegenüber Ausgangswert

Studie (46 Punkte) sogar zu $\Delta = 9.2$ führen. Auch in einer anderen Studie wurde für die Fallzahlberechnung die Veränderung um $\Delta = 20\%$ für die WOMAC Schmerz-Subskala verwendet [75]. MAKAROWSKI et al. basierten ihre Fallzahlplanung der Placebovergleiche auf der WOMAC Schmerz-Subskala und setzten dafür eine Differenz von $\Delta = 14.1$ an.

Entscheidungsprozess für WOMAC Schmerz-Subskala Äquivalenzgrenzen aus den Einzelstudien sind nicht gegeben, jedoch existieren *systematische* Arbeiten, die sich mit dem Problem der Identifikation von klinisch relevanten Differenzen auseinandersetzen. Aus den Ergebnissen lässt sich zwar die Minimum-Lösung gemäß Formel (9) nicht direkt ableiten, doch können die anderen oben genannten Studien verwendet werden: Das heißt $\delta_{min} = min(\delta_1, ..., \delta_5) = min(10, 12, 9.7, 11, 6.4) = 6.4$ bzw. $\delta_{min} = 13.9\%$. Die Angaben zum Fallzahl-Δ liegen darüber, beeinflussen δ_{min} demnach nicht. Somit kann gemäß Formel (11)

$$\delta = min(8, 9.08, 6.4) = 6.4$$

festgelegt werden.

Tabelle 5: Nicht-Unterlegenheitsgrenzen für WOMAC Schmerz-Subskala

	Indikationen	Behandlungen	Dauer [Wochen]	Design	i	δ_i (%)
[8]	nach Knieersatz	SE + CPM SE + SB SE + -	26	RCT N=120 einfach-blind	1	10 (20.8)
[61]	OA(Knie) OA(Hüfte)	Nimesulide Naproxen	26	RCT N=370 doppelt-blind	2	12 (25.3)
[33]	OA(Knie) OA(Hüfte)	Rofecoxib 12.5mg Rofecoxib 25mg Ibuprofen 2400mg Placebo	6	2 RCTs N=1545 doppelt-blind	3	9.7 (14.9)
[3]	OA(Knie) OA(Hüfte)	aktive PT passive PT	13	prospektiv N=122	4	11.0 (22.8)
[4]	OA(Knie) OA(Hüfte)	aktive PT passive PT	13	prospektiv N=192	5	6.4 (13.9)

% = prozentuale Veränderung bezüglich Ausgangswert, PT = Physiotherapie, OA = Arthrose, Ibu = Ibuprofen, SE = standard exercise, SB = Sliderboard, CPM = continuous passive motion

Der Wert 6.4 ist deutlich geringer als die anderen, daher könnte die Verwendung von δ_{Q1} sinnvoll sein. $\delta_{Q1} = Q1(10, 12, 9.7, 11, 6.4) = 9.7$. Dann wäre

$$\delta = min(8, 9.08, 9.7) = 8.$$

Diskussion Soll in einer klinischen Studie oder einer Meta-Analyse die WOMAC Schmerz-Subskala als Hauptzielgröße verwendet werden, anhand dessen die Äquivalenz verschiedener Therapien untersucht werden soll, sollte die Nicht-Unterlegenheitsgrenze nicht größer als 9.08 sein. Ansonsten ist selbst bei Nicht-Unterlegenheit die Schlussfolgerung der Wirksamkeit *nicht* gesichert. Aufgrund der verschiedenen Hypothesen aus Abschnitt 1.4 müssen auch die einzelnen δ betrachtet werden. Im Studienprotokoll der

zu planenden Einzelstudie oder Meta-Analyse sollten daher die Komponenten $\delta_{clin} = 8$, $\delta_{min} = 6.4, \delta_{Q1} = 9.7, \delta_P(0) = 9.08, \delta_P(0.5) = 4.52$ festgelegt und später im Bericht diskutiert werden.

In der aktuellen Meta-Analyse von DEEKS et al. wird der WOMAC als Zielgröße verwendet, eine Definition der Äquivalenzgrenze fehlt jedoch [27]. Es wird aufgrund der Konfidenzintervalle auf eine Äquivalenz geschlossen, *ohne* vorher festgelegt zu haben, was als äquivalent zu bezeichnen ist („The confidence intervals around the point estimates of efficacy were reasonably narrow, which means that it is unlikely that there were clinically important differences."). Letztlich würde auch bei einer a-priori-Definition der Äquivalenzgrenze die Interpretation in diesem Fall genauso ausfallen, weil die Konfidenzintervalle sehr klein sind. Dennoch ist in jedem Fall eine Festlegung der Äquivalenzgrenze sinnvoll, denn die Interpretation wird dadurch glaubwürdiger und auch in weniger deutlichen Fällen nicht angreifbar.

$\delta_{pbo}(0)$ ist zu konservativ und braucht nicht weiter betrachtet zu werden. Wenn nun ein hierarchisches Vorgehen gewählt wird, so können die Hypothesen $H_0^{EA} : \theta^{EA} \leq -\delta_P(0)$ (also $H_0^{EP} : \theta^{EP} \leq 0$), $H_0^{EA} : \theta^{EA} \leq -\delta_{clin}$ und $H_0^{EA} : \theta \leq -\delta_{Q1}$ verworfen werden. Demnach ist Celecoxib signifikant besser als Placebo. Da in diesem Fall alle von DEEKS et al. verwendeten Studien dreiarmige Studien mit Placebokontrolle waren, konnte ein direkter Vergleich durchgeführt werden, der natürlich – wenn möglich – dem indirekten Vergleich vorzuziehen ist (siehe Seite 15). Die Meta-Analyse kam zum gleichen Ergebnis: Die Überlegenheit von Celecoxib gegenüber Placebo konnte gezeigt werden. Außerdem ist die untere Konfidenzintervallgrenze etwas oberhalb von $-\delta_{clin}$, welches die Wirksamkeit mit einer klinischen Nicht-Unterlegenheit unterstützt. Da aber die untere Konfidenzintervallgrenze der Meta-Analyse deutlich *kleiner* ist als $-\delta_P(0.5)$, kann die Hypothese $H_0^{EP}(0.5) : \theta^{EP} \leq 0.5\ \theta^{AP}$ nicht verworfen werden: Es kann demnach *nicht* gezeigt werden, dass Celecoxib mindestens die Hälfte des etablierten Effektes der Vergleichsgruppe (hier: Naproxen) sichert.

4 Die Wahl der Auswertungspopulation

In vergleichenden klinischen Studien stellt sich oft die Frage, welche Daten in die Analyse aufgenommen werden sollen. Häufig werden Patienten nicht protokollgerecht in die Studie eingeschlossen, behandelt oder beobachtet. In diesen Fällen ist zu entscheiden, ob und wie die Daten der Patienten in die Analyse eingehen. Friedman et al. unterscheiden „exclusions“ (Patienten scheiden vor der Randomisierung aus der Studie aus) und „withdrawals“ (Patienten können wegen Protokollverletzungen nicht ohne weiteres in die Studie aufgenommen werden). „Exclusions“ können zwar den induktiven Schluss auf die Grundgesamtheit verändern, beeinflussen aber nicht die vergleichende Analyse, daher werden diese Fälle im Folgenden nicht weiter betrachtet. „Withdrawals“ teilen Friedman et al. in vier Gruppen ein [42]. (i) „Ineligibility“: Patienten, die die Ein- oder Ausschlusskriterien nicht erfüllen, also eigentlich gar nicht hätten eingebracht werden dürfen. (ii) „Nonadherence“: Patienten, die nicht in der ihnen zugewiesenen Gruppe bleiben: Gruppenwechsler oder Patienten die abbrechen („dropouts“). (iii) „Poor quality or missing data“: Patienten, die Daten mit schlechter Qualität beitragen oder fehlende Werte liefern. (iv) „Competing events“: Patienten mit konkurrierenden Ereignissen, die eine weitere Beobachtung unmöglich machen. Nicht alle Autoren befürworten diese Einteilung, beispielsweise werden die Begriffe „dropout“ und „withdrawal“ unterschiedlich belegt. An Beispielen sollen nun die Situationen verdeutlicht werden, die zu Unterschieden zwischen verschiedenen Populationen führen. Die ersten drei Situationen werden bei Windeler als „dropouts“ bezeichnet und die folgenden als „withdrawal“ [112]. Gemäß der Bezeichnung von Friedman et al. handelt es sich in allen Fällen um „withdrawals“; die genauere Einteilung ist jeweils in Klammern hinter der Beschreibung aufgeführt [42].

- Patient verstirbt vor Erhebung des Zielkriteriums („competing event“)
- Patient verstirbt nach Einschluss, aber vor Beginn der Therapie („competing event“)
- Patient erscheint nicht zur Kontrolluntersuchung, Zielkriterium konnte nicht erfasst werden („lost to follow-up“, „poor quality data“)
- Bewertung des Zielkriteriums liefert einen Ausreißer („poor quality data“)
- Patient oder Arzt bricht die Therapie aufgrund von unerwünschten Ereignissen ab („nonadherence“)

- Patient wechselt von einem Therapiearm zu einem anderen („nonadherence“)
- Ein Ausschlusskriterium wird erst nach Einschluss des Patienten festgestellt („ineligibility“)
- Eingangsdiagnose wurde nicht korrekt festgestellt („ineligibility“)
- Patient oder Arzt wendet die Therpie nicht protokollgerecht an („nonadherence“)
- Patient wendet unerlaubte Begleittherapie an („nonadherence“)

Um die aufgeführten Situationen adäquat zu berücksichtigen, gibt es vor allem zwei Methoden: Die PP-Analyse und die ITT-Analyse. Wenn die entsprechenden Fälle aus der Analyse ausgeklammert werden, spricht man von einer per Protokoll Analyse (PP-Analyse, PP-Population). Bei Verwendung der PP-Analyse im Vergleich zur Situation ohne Protokollverletzungen ergeben sich folgende Aspekte: Die *Fallzahl* sinkt, der *Effekt* wird eher größer geschätzt (eventuell Überschätzung, aufgrund von Patienten, die die Studie mangels Wirksamkeit verlassen), die *Interpretation des Effektes* folgt einem explanatorischen Ansatz [31, 84], die *Studienvarianz* wird in der Regel geringer, *Verzerrungen* können aufgrund von Interaktionen der Therapiegruppe mit der entsprechenden Situation auftreten (Höhe und Richtung sind unklar [42]), die *Strukturgleichheit* kann verletzt werden (d.h. Gruppen werden unterschiedlich), die *Power* wird in der Regel sinken (vor allem aufgrund der Fallzahlreduktion [56]), *Konfidenzintervalle* können breiter werden (vor allem aufgrund der Fallzahlreduktion [31]) und die *Gesamtinterpretation* wird eher konservativ für Äquivalenzstudien [31, 65].

Eine ITT-Analyse (ITT = intention to treat) liegt vor, wenn alle Fälle entsprechend der ihnen zugewiesenen Therapiegruppe analysiert werden. Die genaue Definition des ITT-Prinzips ist aber je nach Autor leicht unterschiedlich [31]. So gibt es die strikte Regel: „analysiere so, wie randomisiert wurde“ („as randomized“) (zum Beispiel [1, 60]). Oft wird jedoch an die Population aus praktischen Erwägungen eine Bedingung gestellt: „mindestens eine Therpieanwendung und/oder mindestens ein Zielkriterium muss erfasst worden sein“ [59]. In der Guidance ICH-E9 wird diese Auswertungspopulation „full analysis set“ genannt und soll unter Berücksichtigung der praktischen Möglichkeiten der ITT-Population möglichst gut entsprechen [56].

ITT wird im Kontext mit dem pragmatischen (im Gegensatz zum explanatorischen) Studienansatz diskutiert, da die Analyse von Studiendaten nach dem ITT-Prinzip eher

der späteren klinischen Praxis entspricht, in der das Therapieverfahren eingesetzt werden soll. Somit liefert dieser Ansatz konkrete Hinweise auf das spätere Handeln, erklärt aber nicht unbedingt die Zusammenhänge zwischen Therapie und Wirkung korrekt [40, 112]. Bei Verwendung der ITT-Analyse im Vergleich zur Situation ohne Protokollverletzungen ergeben sich folgende Aspekte: Die *Fallzahl* ist (fast) wie geplant, der *Effekt* wird eher kleiner geschätzt, da die Gruppen einander ähnlicher werden [31, 58], oder bleibt konstant [31], die *Interpretation des Effektes* folgt einem pragmatischen Ansatz [84], die *Studienvarianz* wird in der Regel höher, *Verzerrungen* werden größtenteils vermieden, die *Strukturgleichheit* bleibt bestehen [31], die *Power* wird in der Regel kleiner (vor allem aufgrund Varianzerhöhung [56]), *Konfidenzintervalle* werden im Allgemeinen eher breiter und die *Gesamtinterpretation* wird für Überlegenheitsstudien eher konservativ [31, 56] und kann für Äquivalenzstudien liberal sein [65, 78].

Tabelle 6: Unterschiede zwischen einer PP- und einer ITT-Analyse

Kriterium	Unterschied
Fallzahl:	$N(\text{PP}) \leq N(\text{ITT})$
Effektschätzung:	$\lvert\hat{\theta}(\text{PP})\rvert \geq \lvert\hat{\theta}(\text{ITT})\rvert$
empirische Varianz:	$\hat{\sigma}^2(\text{PP}) \leq \hat{\sigma}^2(\text{ITT})$
Varianz des Effektschätzers:	$SE(\text{PP}) \geq SE(\text{ITT})$
Konfidenzintervallbreite:	$\lvert\text{KI}\rvert(\text{PP}) \geq \lvert\text{KI}\rvert(\text{ITT})$

Tabelle 6 zeigt die Unterschiede zwischen einer PP- und einer ITT-Analyse, wobei die angegebenen Relationen nicht unbedingt feststehen, sondern lediglich Tendenzen aufzeigen sollen. So berücksichtigt zum Beispiel die Argumentation ($N(\text{PP}) \leq N(\text{ITT}) \Rightarrow SE(\text{PP}) \geq SE(\text{ITT})$) von Garrett nicht den Einfluss der Varianzreduktion [45]. Der Fallzahleinfluss wird jedoch in der Regel größer sein, daher die obige Zusammenfassung. Im Folgenden wird zunächst eine einzelne klinische Studie dargestellt, um die wesentlichen Prinzipien für die Wahl der Auswertungspopulation zu beschreiben.

4.1 Populationen bei Überlegenheitsstudien

Für Überlegenheitsstudien hat sich der Ansatz der ITT-Analyse durchgesetzt und wird von vielen Autoren propagiert [1, 40, 42, 60, 72, 84, 112]. Sir Austin Bradford Hill hat den Begriff im Jahre 1961 in der 7. Auflage von „Principles of medical Statistics“

eingeführt [52]. Feinstein schlägt vor, die ITT-Analyse nur als *eine* Form zu verwenden, nicht aber allein. So können therapeutische und/oder prognostische Unterschiede entdeckt werden, die sonst verdeckt würden. Das führt unter anderem aus regulatorischer Sicht zu der Forderung, dass neben der ITT-Analyse auch eine PP-Analyse als Sensitivitätsanalyse vorzulegen ist und abweichende Ergebnisse untersucht werden müssen [56, 72].

Im Allgemeinen wird die Verwendung einer ITT-Analyse konservativ sein, da sich durch Einbeziehung der Protokoll-Verletzer und „withdrawals“ die zu vergleichenden Gruppen ähnlicher werden [58].

4.2 Populationen bei Äquivalenzstudien

Über die Verwendung der Auswertungspopulationen bei Äquivalenzstudien wurde bisher wenig geschrieben. Die Idee des pragmatischen Studienansatzes bleibt auch für die Äquivalenzstudie erhalten. Somit ist grundsätzlich auch für Äquivalenzstudien eine ITT-Analyse sinnvoll [65, 113], das Prinzip aber besonders sorgfältig umzusetzen [63]. Aber bei genügend „dropouts“ wäre es immer möglich, eine Äquivalenz auf der Grundlage einer ITT-Population zu zeigen, nicht jedoch mit der PP-Population. Da die Fallzahl würde dabei entsprechend sinken und damit zu sehr breiten Konfidenzintervallen führen.

Einige der Konsequenzen einer ITT-Analyse bei Äquivalenzstudien auf das Ergebnis sind komplementär zu denen bei den Überlegenheitsstudien [60, 112], so dass auf die Beschreibung der Auswertungspopulation besonders geachtet werden muss [113]. Denn die Verwendung der ITT-Analyse ist aufgrund der Angleichung der Gruppen bei Äquivalenzstudien *nicht* konservativ [18, 56, 65]. Im Extremfall könnte eine Äquivalenz mit dem ITT-Prinzip gezeigt werden, wenn alle Patienten trotz ihrer Randomisierung immer nur eine (oder keine) Therapie anwenden, völlig unabhängig von dem wirklichen Unterschied der beiden Therapien. Die PP-Analyse weist ebenfalls Kritikpunkte auf: Der Ausschluss eines Patienten aus der PP-Population kann die Ergebnisse der zu vergleichenden Gruppen angleichen (Varianzreduktion). Dies kann zu liberalen Entscheidungen führen. Beispiel: In der Experimentalgruppe wirkt die Behandlung auf alle Patienten gleichartig. In der Vergleichsgruppe ist dies für die Hälfte der Patienten der Fall. Bei der anderen Hälfte liegt kein positiver Effekt vor; dort kommt es zu vielen „dropouts“. Die PP-Analyse vergleicht somit zwei homogene Gruppen mit gleichem Erfolg und schließt auf Äquivalenz, die aber tatsächlich nicht vorliegt. Dass die PP-Analyse nicht die alleinige Methode der Wahl sein kann, haben auch Röhmel und Garrett in ihren aktuellen Arbeiten festgestellt [45, 78].

Wenn auch die meisten Autoren, die sich mit ITT und Äquivalenzstudien beschäftigen, keine klare Regel angeben, wie bezüglich der Auswertungspopulation bei Äquivalenzstudien zu verfahren ist, so wird doch immer wieder darauf hingewiesen, dass sowohl eine ITT-Analyse als auch eine PP-Analyse durchzuführen ist. Abweichungen der beiden Ergebnisse müssen diskutiert werden [58, 65]. Eine eindeutige Aussage kann nur getroffen werden, wenn sich die Analyseergebnisse der PP- und der ITT-Analyse im Wesentlichen *nicht* unterscheiden. In der „Note for Guidance“ für die Evaluation von neuen antibakteriellen Produkten werden drei Auswertungspopulationen empfohlen: PP-, ITT- sowie eine modifizierte ITT-Population [34]. Die Ergebnisse der drei Analysen sollen konsistent sein. Dies wird auch in anderen Arbeiten der EMEA dargestellt: „In a non-inferiority trial, the full analysis set and the PP analysis set have equal importance and their use should lead to similar conclusions for a robust interpretation.“ [38] und „However, in an equivalence or non-inferiority trial use of the full analysis set is generally not conservative and its role should be considered very carefully.“ [56].

4.3 Populationen bei Meta-Analysen - Überlegenheit

Bei der Analyse einer Originaldaten basierten Meta-Analyse weist das „Cochrane Reviewers' Handbook“ explizit auf die Verwendung des ITT-Prinzips hin [94]. Als Qualitätskriterium für eine publikationsbasierte Meta-Analyse wird im Handbuch nicht ausdrücklich verlangt, eine ITT-Analyse der Einzelstudien durchzuführen. Die unterschiedliche Berücksichtigung von fehlenden Werten und Protokollverletzern in den Einzelstudien wird erwähnt, der Einfluss ist aber noch nicht klar belegt [82]. Grundsätzlich wird der ITT-Ansatz aus den Einzelstudien übernommen, so dass es als Qualitätskriterium gilt, wenn die eingeschlossenen Studien nach dem ITT-Prinzip analysiert wurden.

4.4 Populationen bei Meta-Analysen - Äquivalenz

Über die Problematik der Auswahl und Behandlung der Populationsansätze bei Meta-Analysen mit Äquivalenzhypothesen gibt es bisher keine methodischen Arbeiten. Die Situation ist im Prinzip vergleichbar mit der einzelnen Äquivalenzstudie: Eine ITT-Analyse ist bei Meta-Analysen zwar grundsätzlich sinnvoll, ist unter Umständen aber liberal und hält somit das α-Niveau nicht ein. Daher ist die ITT-Analyse allein nicht ausreichend und es bietet sich die PP-Analyse als Sensitivitätsanalyse an (vergleiche Abschnitt 4.2).

Für die publikationsbasierte Meta-Analyse kommt erschwerend hinzu, dass in der Regel in den Publikationen nur *eine* Analyse präsentiert wird, entweder aus Platzgründen oder weil nur eine Analyse durchgeführt wurde. In den Meta-Analysen von Carroli et al. und von Deeks et al. wird beispielsweise die ITT-Analyse jeweils als Qualitätskriterium zu Beurteilung der Einzelstudien verwendet [19, 27]. Dies ist kein adäquates Vorgehen. Ziel dieses Abschnittes ist es, unter anderem dies Vorgehen kritisch zu betrachten, verschiedene Szenarien zu beschreiben, Methoden bereit zu stellen und Empfehlungen zur Durchführung von Meta-Analysen zu formulieren.

Auswahl Grundsätzlich sollte – analog zum Auswahl-Problem in Abschnitt 3.3 – die Auswertungspopulation der Einzelstudien nicht als Ausschlusskriterium für die Meta-Analyse verwendet werden. Da zu einem klinischen Problem häufig sowohl Überlegenheits- als auch Äquivalenzstudien zu erwarten sind, ist mit unterschiedlichen Auswertungspopulationen zu rechnen.

4.4.1 Szenarien

Im Allgemeinen ist pro Studie nur die Analyse einer Auswertungspopulation bekannt: Entweder (i) eine ITT-Analyse oder (ii) eine PP-Analyse. (iii) Nur in Einzelfällen werden beide Ergebnisse bekannt sein. Auf die unterschiedlichen Möglichkeiten eine ITT-Analyse durchzuführen soll hier nicht eingegangen werden. Folgende Szenarien sind denkbar:

1. Idealfall: für *alle* Studien stehen die Ergebnisse aus der PP- *und* der ITT-Analyse zur Verfügung (nur (iii)).

2. Für *einzelne* Studien sind die Ergebnisse der PP- *und* ITT-Analyse bekannt ((i) und/oder (ii) und (iii)).

3. Für *alle* Studien sind entweder die Ergebnisse einer PP- *oder* einer ITT-Analyse bekannt ((i) und (ii)).

4. Das Prinzip der Analyse ist teilweise *nicht* bekannt.

5. Für alle Studien sind *nur* die Ergebnisse der ITT-Analyse bekannt (nur (i)).

6. Für alle Studien sind *nur* die Ergebnisse der PP-Analyse bekannt (nur (ii)).

Beispiel (1. Szenario) In den 13 Übersichtsarbeiten im Bereich bakterielle Infektionen (Abschnitt 1.2) wurde nur in zwei Fällen kritisch auf die Problematik der Auswertungspopulation eingegangen [5, 103]. Barza et al. führen zunächsteine PP-Analyse mit einem binären Endpunkt durch. Dazu wurden die Daten aus den Originalpublikationen verwendet. In einem zweiten Schritt haben sie außerdem die Anzahl der randomisierten Patienten in die Analyse mit aufgenommen. Somit wurden alle „withdrawals“ als Therapieversager gewertet, was dem ITT-Prinzip entspricht [5]. Auch bei Mismetti et al. wurde eine ITT-Rekonstruktion vorgenommen [70].

Beispiel (1. Szenario) Wasilewski et al. untersuchten einen Wirkstoff ihres eigenen Arbeitgebers und haben mit zwei Studien eine Meta-Analyse durchgeführt und präsentieren sowohl eine ITT- als auch eine PP-Analyse. Bei einem der beiden Endpunkte („clinical response“) war der Punktschätzer bei der PP-Analyse etwas größer als bei der ITT-Analyse (4.4 statt 2.5, „termination visit“) und das Konfidenzintervall für die Differenz der Erfolgsraten deutlich breiter (27 statt 10.6). Diese Beobachtungen stimmen mit den Aussagen in Tabelle 6 auf Seite 57 überein. Bei dem anderen Endpunkt („bacteriological response“) blieben sowohl Punktschätzer als auch Konfidenzintervall in der PP- und der ITT-Analyse etwa gleich [103].

Beispiel (1. Szenario) Ebbutt und Frith nennen in einer methodischen Arbeit über Äquivalenzstudien ein Beispiel über Therapien bei Asthma mit elf Studien, wobei in jedem Fall eine PP-Analyse als auch eine ITT-Analyse dargestellt sind (Asthma-Beispiel, Abschnitt 1.2, Seite 6) [31].

Beispiel (2. Szenario) Im Oxaceprol-Beispiel (Tabelle 1, Abschnitt 1.2) sind zweimal die Ergebnisse der ITT-Analyse und zweimal die Ergebnisse der PP-Analyse gegeben [114]. Von einer Studie ist bekannt, dass beide Analysen durchgeführt wurden, so dass bei Erhalt der Ergebnisse das 2. Szenario vorliegen würde.

Beispiel (3. Szenario) Bei Tran et al. werden zehn PP-Analysen und zwölf ITT-Analysen verwendet [96].

Beispiel (4. Szenario) Bei Michael et al. wird bei allen zehn Studien die Auswertungspopulation als unbekannt bezeichnet [69].

4.4.2 Methoden

Innerhalb einer Meta-Analyse kann es zu Verzerrungen kommen, wenn unterschiedliche Populationsansätze zusammengefasst werden. Diese potentielle Heterogenität muss im Rahmen der Meta-Analyse untersucht und berücksichtigt werden. Im Studienprotokoll ist die Art der Analyse dieser möglichen Heterogenität festzulegen. Es sollten dabei sowohl analytische als auch graphische Methoden verwendet werden. Eine Beurteilung der Heterogenität ist aber bei den Szenarien 5. und 6. in der Aufstellung auf Seite 60 nicht möglich. In den beiden Fällen sind nur schwächere Aussagen zu treffen; ein endgültiger induktiver Schluss ist nicht möglich.

Um die Vorgehensweise bei den verschiedenen Szenarien festzulegen, werden verschiedene Techniken im Folgenden kurz genannt und erläutert.

ITT-Rekonstruktion In einem Spezialfall kann eine ITT-Analyse rekonstruiert werden: Bei einer binären Zielvariable kann eine ITT-Analyse durchgeführt werden, indem die Anzahl der randomisierten Patienten als Basis für die Schätzung der Raten verwendet wird. Somit werden die aus der PP-Analyse herausgenommenen Patienten automatisch als Therapieversager gewertet. Das Vorgehen der „ITT-Rekonstruktion“ kann und sollte dazu verwendet werden, das ITT-Prinzip umzusetzen. Somit ist bei gegebener PP-Analyse und bekannter Anzahl der randomisierten Patienten sowohl eine PP- als auch eine ITT-Analyse möglich. Bei komplexeren Analysen ist dagegen eine Rekalkulation nicht möglich.

Graphische Methoden Zur Darstellung der Unterschiede zwischen den Analysen (PP oder ITT) bieten sich fast alle Graphiktypen der Meta-Analyse an, indem die Ergebnisse aus den verschiedenen Auswertungspopulationen unterschiedlich gekennzeichnet werden. Aus der Abbildung muss klar erkennbar sein, ob eine Studie ein oder zwei Analysen für die Graphik liefert. So kann man Forest- bzw. Konfidenzintervall-Plots [110], Funnel-Plots [66] (Abbildungen 10 und 11), Radial-Plots bzw. Galbraith-Plots [44] verwenden. Zur Darstellung von beiden Ergebnissen kann der l'Abbé-Plot verwendet werden [62, 98].

Analytische Methoden Neben den graphischen Methoden zur Beurteilung, ob die Populationsansätze Heterogenität verursachen, sollten analytische Methoden verwendet werden. Wenn alle Studien genau einen Schätzer liefern (3. Szenario, oder 1. und 2. Szenario mit Auswahl je einer Analyse), können die bekannten Heterogenitätstests genutzt werden. Cochran's Q ist der am häufigsten angewendete Heterogenitätstest [22]. Andererseits kann der Egger-Test verwendet werden [32, 109, 83]. Dieser wurde ursprüglich als ein

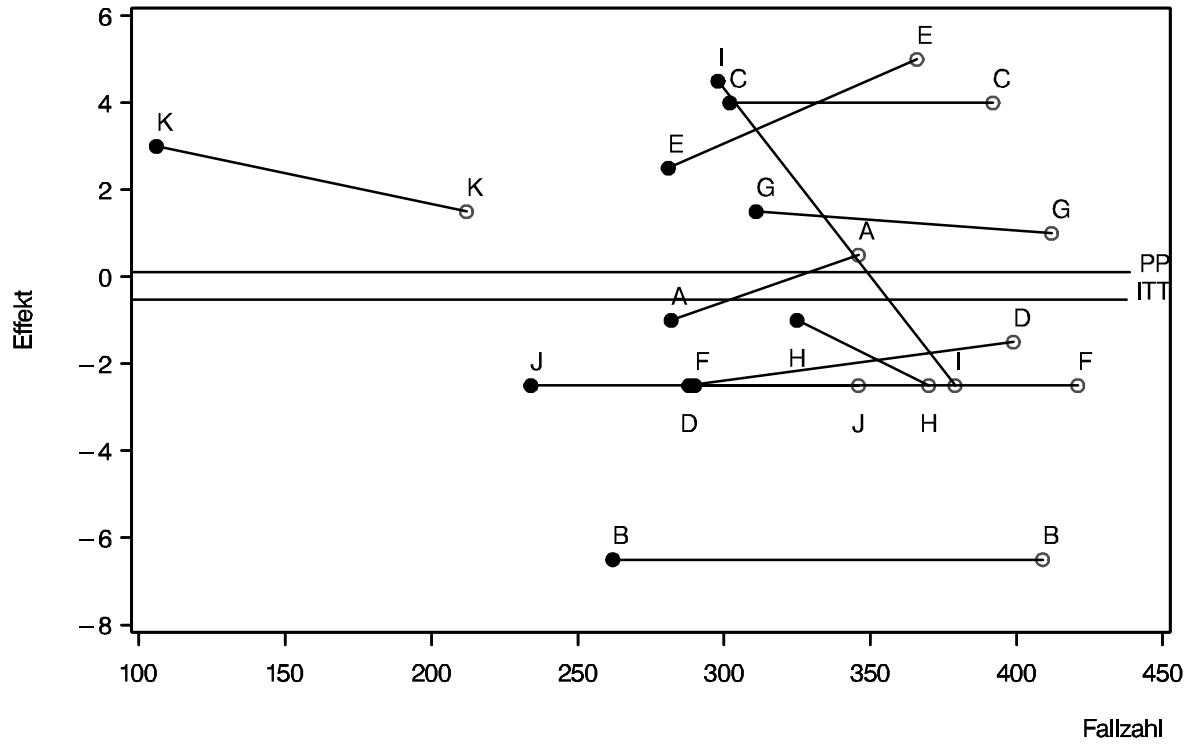


Abbildung 10: Funnel-Plot mit bivariater Darstellung für das Asthma-Beispiel: • = PP-Analyse und ∘ = ITT-Analyse der Studien A-K. Die Referenzlinien zeigen die Meta-Analyse-Schätzer für die PP- bzw. die ITT-Analyse.

Test auf „publication bias“ eingeführt und nutzt eine lineare Regression im Radial-Plot. Ein dritter Test, der auf Rangkorrelationen beruht, wurde von Begg und Mazumdar vorgeschlagen [9, 83].

Deskriptive Methoden Wird eine Meta-Analyse betrachtet in der der überwiegende Teil der Einzelstudien mit dem gleichen Populationsansatz ausgewertet wurde, so können die Beiträge q_i zur Heterogenitätstatistik $Q = \sum q_i$ deskriptiv interpretiert werden. Bei deutlichen Beiträgen durch die Studien mit dem selteneren Populationsansatz kann von einer Heterogenität ausgegangen werden. Diese Heterogenität kann aufgrund des anderen Populationsansatzes verursacht worden sein. Eine Zusammenführung der Studien ist dann nicht ohne weiteres sinnvoll.

Sensitivitätsanalysen Als Sensitivitätsanalyse sollten sowohl die Ergebnisse der PP-Analysen als auch die Ergebnisse der ITT-Analysen zusammengefasst werden. Das entspricht dem Vorgehen auch bei anderen Subgruppenanalysen. Abweichungen der beiden Ergebnisse sind zu interpretieren. Eine andere Möglichkeit der Sensitivitätsanalyse ist

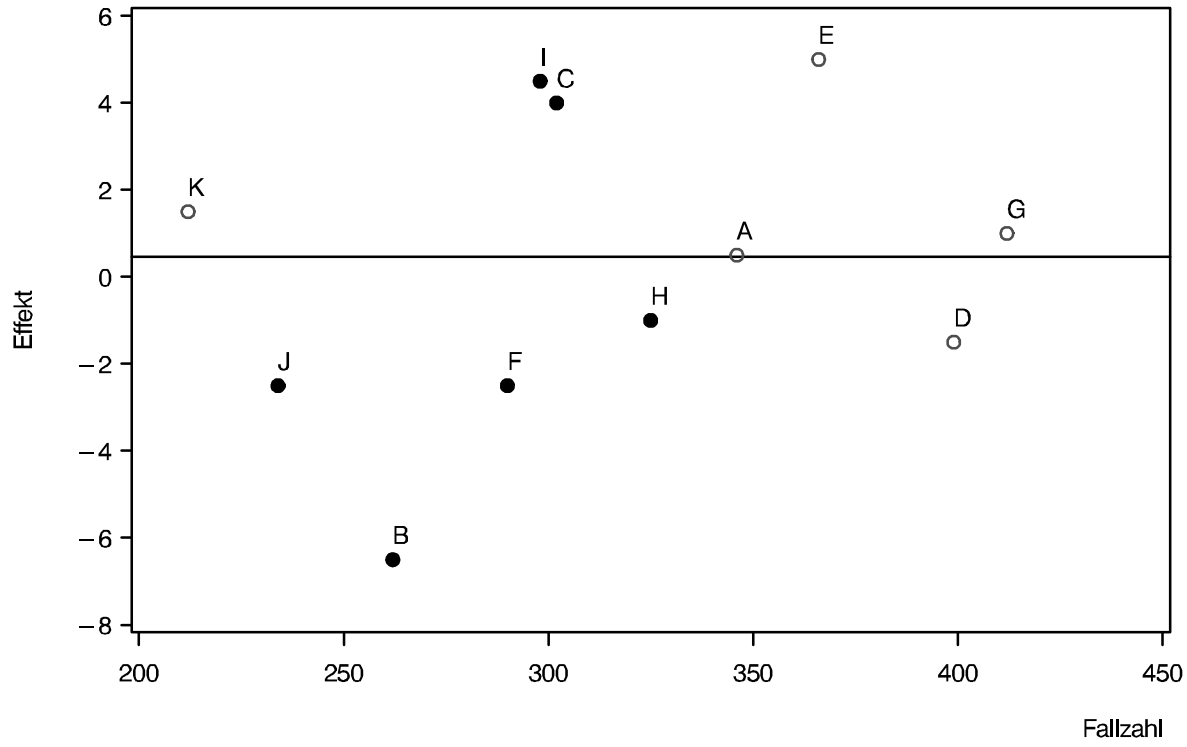


Abbildung 11: Funnel-Plot für das 3. Szenario, oder eine zufällige Auswahl der Analyse (Asthma-Beispiel), • = PP-Analyse, ∘ = ITT-Analyse

möglich, wenn zum Teil Ergebnisse von PP- und ITT-Analysen vorliegen (2. und ggf. 4. Szenario in der Aufstellung auf Seite 60). Es können zunächst die PP-Analysen mit den restlichen Studien zusammen ausgewertet und in einem zweiten Schritt alle vorliegenden ITT-Analysen mit den restlichen Studien analysiert werden. Auch hier sind Abweichungen entsprechend zu diskutieren. Im Idealfall (1. Szenario in der Aufstellung auf Seite 60) sind die beiden vorgestellten Techniken identisch. Sensitivitätsanalysen liefern für eine Meta-Analyse mehrere einzelne Analysen. Indem die Ergebnisse miteinander verglichen werden, kann auf Homogenität oder Heterogenität durch die Verwendung von verschiedenen Populationsansätzen geschlossen werden .

Meta-Regression Derjenige Effekt, der durch verschiedene Populationen verursacht wird, lässt sich mit einer Meta-Regression schätzen. Das FEM kann als

$$\hat{\theta}_i = \theta + \varepsilon_i \sim N(\theta, w_i^{-1})$$

mit unbekanntem $\theta \in \mathbb{R}$ und $\varepsilon_i \sim N(0, w_i^{-1})$ geschrieben werden. Die Varianz w_i^{-1} wird

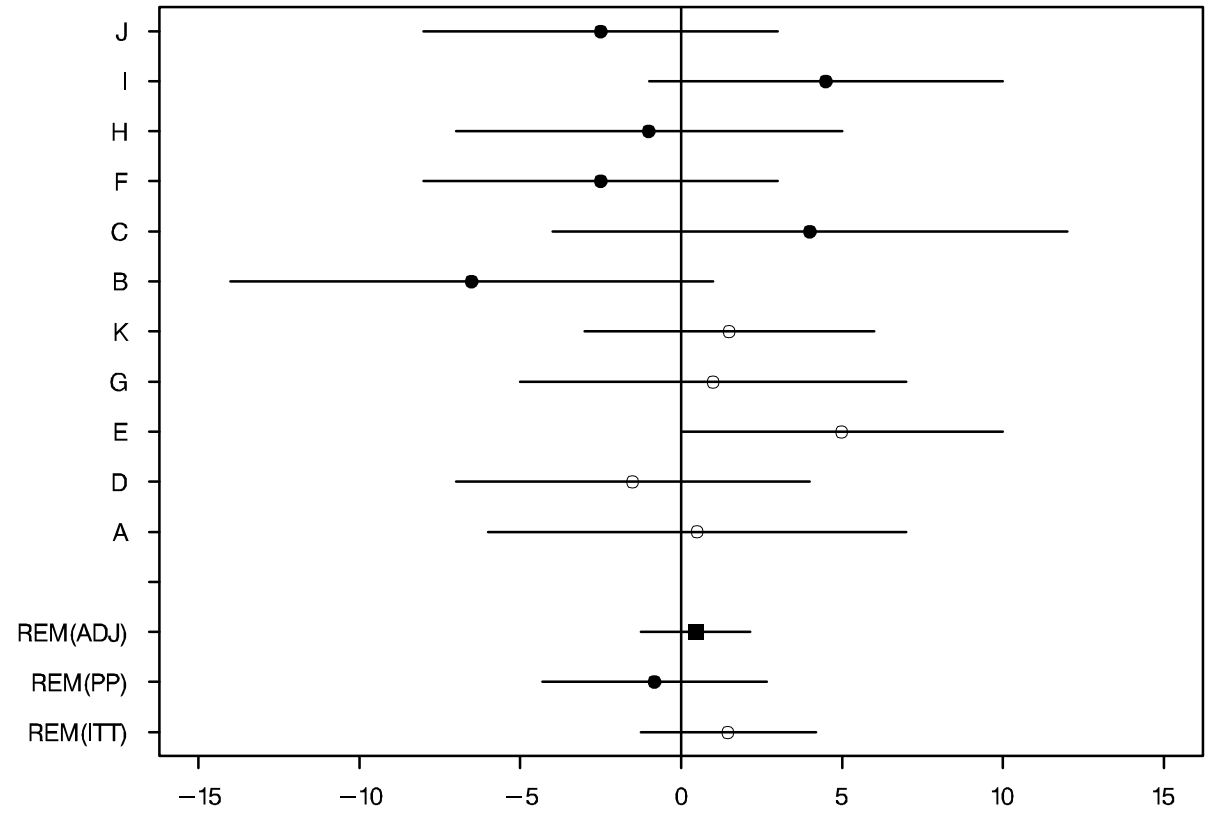


Abbildung 12: Forest-Plot für das 3. Szenario, oder eine zufällige Auswahl der Analyse (Asthma-Beispiel), getrennt nach PP-Analysen (PP) und ITT-Analysen (ITT), sowie einem mittleren Effekt (ADJ), • = PP-Analyse, ∘ = ITT-Analyse

dabei als bekannt vorausgesetzt, obwohl sie in der Praxis in den Einzelstudien geschätzt wird. Das REM kann als

$$\hat{\theta}_i = \theta + \nu_i + \varepsilon_i \sim N(\theta, w_i^{-1} + \tau^2)$$

mit $\nu_i \sim N(0, \tau^2)$ und unbekanntem $\tau^2 \in \mathbb{R}$ geschrieben werden. $\hat{\tau}^2$ bezeichne einen Schätzer der Zwischenstudienvarianz. Nun kann eine zusätzliche Kovariable auf Studienebene x_{1i} hinzugenommen werden. x_{1i} steht dabei für die entsprechende Population mit $x_{1i} = 0$ für die PP- und $x_{1i} = 1$ für die ITT-Analyse. Es ist

$$\hat{\theta}_i = \theta + \beta_1 x_{1i} + \nu_i + \varepsilon_i \sim N(\theta + \beta_1 x_{1i}, w_i^{-1} + \tau^2)$$

mit unbekanntem $\beta_1 \in \mathbb{R}$. Der Index 1 soll andeuten, dass auch noch weitere Kovariablen in das Modell aufgenommen werden können. Dieses Modell kann mit statistischer

Standardsoftware gelöst werden. Den Programmcode in SAS 6.11 bzw. 6.12 für das REM hat NORMAND veröffentlicht [74]. WHITEHEAD hat in ihrem Buch die Erweiterung für die Meta-Regression erläutert [109]. Aufgrund der Verwendung der neuen SAS Version 8 ergeben sich leichte Modifikationen (Abschnitt 8.4).

Bivariate Analyse Liegen zwei Zielgrößen pro Studie vor, so ist eine bivariate Modellierung möglich. In einer Meta-Analyse mit Studien die PP- und ITT-Analysen liefern kann eine solche Modellierung angewendet werden (1. Szenario). VAN HOUWELINGEN hat in seiner Übersichtsarbeit die bivariate Modellierung beschrieben [98]. Das Modell mit zufälligen Effekten für alle $i = 1, \ldots, k$ als

$$\begin{pmatrix} \hat{\theta}_{0i} \\ \hat{\theta}_{1i} \end{pmatrix} = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} + \begin{pmatrix} \nu_{0i} \\ \nu_{1i} \end{pmatrix} + \begin{pmatrix} \varepsilon_{0i} \\ \varepsilon_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}, \begin{pmatrix} w_{0i}^{-1} + \tau_0^2 & \sigma_{01} \\ \sigma_{01} & w_{1i}^{-1} + \tau_1^2 \end{pmatrix} \right), \quad (12)$$

geschrieben werden, wobei das Subscript 0 für die PP-Analyse und 1 für die ITT-Analyse steht. Die Parameter θ_0 und θ_1 bezeichnen die festen Effekte, ν_{0i} und ν_{1i} bilden den Vektor der zufälligen Effekte, ε_{0i} und ε_{1i} stehen für die Fehlerterme. τ_0^2 und τ_1^2 bezeichnen die Zwischenstudienvarianzen, w_{0i}^{-1} und w_{1i}^{-1} stellen die Varianzen innerhalb der Studien dar. Da die Ergebnisse aus der PP-Analyse ($\hat{\theta}_{0i}$) und der ITT-Analyse ($\hat{\theta}_{1i}$) abhängig sind, muss auch die Kovarianz σ_{01} modelliert werden. Die Parameter $\theta_0, \theta_1, \tau_0^2, \tau_1^2$ sowie σ_{01} sind unbekannt und müssen geschätzt werden. w_{0i}^{-1} und w_{1i}^{-1} werden als bekannt vorausgesetzt und können aus den Studien verwendet werden.

Für die technische Umsetzung des bivariaten Modells mit SAS wird

$$\begin{pmatrix} \nu_{0i} \\ \nu_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \sigma_{01} \\ \sigma_{01} & \tau_1^2 \end{pmatrix} \right) \quad \text{und} \quad \begin{pmatrix} \varepsilon_{0i} \\ \varepsilon_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} w_{0i}^{-1} & 0 \\ 0 & w_{1i}^{-1} \end{pmatrix} \right)$$

gesetzt, so dass für dieses gemischte lineare Modell `PROC MIXED` angewendet werden kann.

Mit dem bivariaten Modell ist es möglich alle Studiendaten zu verwenden. Sowohl Meta-Analyse-Schätzer für die PP-Analyse und ITT-Analyse als auch der Unterschied kannn simultan unter Berücksichtigung der Kovarianz ermittelt werden. Schätzer und Konfidenzintervalle können zum Beispiel mit SAS berechnet und interpretiert werden (Abschnitt 8.5).

Boostrap-Verfahren Eine weitere Methode der Auswertung und somit Schätzung von Parametern ist das Bootstrap-Verfahren, eine sogenannte Resampling-Technik. Dabei wer-

den für jede Studie mit einer Wahrscheinlichkeit von 0.5 entweder die Ergebnisse der PP- oder der ITT-Analyse ausgewählt. Liegt in einer Studie nur die PP- oder ITT-Analyse vor, wird diese immer ausgewählt. Die Einzelstudie kann aber im Gegensatz zu einer bivariaten Auswertung in der Analyse verbleiben. Anschließend wird eine Meta-Regression zur Schätzung des „Intercept“ (Meta-Analyse-Schätzer für die PP-Population bei 0/1-Kodierung: 0 = PP, 1 = ITT) und des Populationseffektes (Meta-Analyse-Schätzer für den Unterschied zwischen ITT- und PP-Schätzer) gerechnet. Ein gemeinsamer Meta-Analyse-Schätzer sowie ein Bootstrap-Konfidenzintervall können aus den einzelnen Schätzern ermittelt werden. Dazu wird die empirische Standardabweichung der Mittelwerte aus der Bootstrapstichprobe verwendet [26].

Permutationsverfahren Neben dem Bootstrap kann eine weitere Art einer Resampling-Technik verwendet werden, ein sogenanntes Permutationsverfahren. Es wird jede Möglichkeit von Kombinationen aus PP- und ITT-Analysen ausgewählt, so dass gleich viele PP- als auch ITT-Analysen in die Meta-Analyse einfließen. Bei ungerader Anzahl soll eine PP-Analyse mehr in die Meta-Analye eingehen. In jedem der Fälle wird eine Meta-Regression gerechnet, um den „Intercept“ und den Populationseffekt zu schätzen. Mit k Studien, in denen jeweils die Ergebnisse aus der PP-Analyse *und* der ITT-Analyse bekannt sind, gibt es folgende Anzahl von Möglichkeiten:

$$\binom{k}{k/2} \quad \text{für } k \text{ gerade bzw.} \quad \binom{k}{(k+1)/2} \quad \text{für } k \text{ ungerade}$$

Darüberhinaus können Studien vorliegen, bei denen nur eine der beiden Analysen angegeben ist. Meta-Analyse-Schätzer und ein Permutations-Konfidenzintervall können aus den einzelnen Schätzern wie beim Bootstrap-Verfahren ermittelt werden.

4.4.3 Vorgehensweisen und Beispiele

Im Abschnitt 4.4.2 wurden mehrere Methoden für Meta-Analysen unter Berücksichtigung von verschiedenen Populationsansätzen vorgestellt. Tabelle 7 stellt Methoden für die auf Seite 60 angeführten Szenarien zusammen.

Um die Vorgehensweisen zu verdeutlichen, sollen zwei Beispiele herangezogen werden.

Oxaceprol-Beispiel Hätte Hildebrandt neben der PP- auch die ITT-Analyse publiziert, könnten folgende (Sensitivitäts-)Analysen durchgeführt werden (Tabelle 8): Nur die

Tabelle 7: Vorgehensweisen der Berücksichtigung von verschiedenen Populationsansätzen bei den verschiedenen Szenarien

Szenario	mögliche Analysetechniken
1. Szenario	Sensitivitätsanalyse, Bootstrap- und Permutationsverfahren, bivariate Analyse
2. Szenario	Sensitivitätsanalyse, Bootstrap- und Permutationsverfahren, Meta-Regressionen
3. Szenario	Sensitivitätsanalyse, Meta-Regression
4. Szenario	Sensitivitätsanalyse, Meta-Regression
5. Szenario	-
6. Szenario	-

PP-Analysen (Analyse 1.1), nur die ITT-Analysen (Analyse 1.2). Weiterhin könnten die Studien, für die beide Analysen bekannt sind, einmal mit den PP-Analysen (Analyse 2.1) und einmal mit den ITT-Analysen als Meta-Regression ausgewertet werden (Analyse 2.2).

Tabelle 8: Sensitivitätsanalysen bzw. Meta-Regressionen zur Untersuchung der Heterogenität aufgrund von verschiedenen Auswertungspopulationen im Oxaceprol-Beispiel

Studie	bekannt	Analyse 1.1	Analyse 1.2	Analyse 2.1	Analyse 2.2
[6]	PP	PP	-	PP	PP
[49]	PP+ITT	PP	ITT	PP	ITT
[51]	ITT	-	ITT	ITT	ITT
[97]	ITT	-	ITT	ITT	ITT

Asthma-Beispiel Das Asthma-Beispiel ist aussagekräftiger als das Oxaceprol-Beispiel: Zum einen weil es sich um elf statt vier Studien handelt und zum anderen weil stets sowohl die Ergebnisse einer ITT-Analyse als auch einer PP-Analyse bekannt sind (1. Szenario). Die Situation ist mit einer einzelnen Äquivalenzstudie vergleichbar. Die *Sensitivitätsanalysen* liefern $\hat{\theta} = 0.0844$ für den Therapieeffekt unter Verwendung der PP-Analysen und $\hat{\theta} = -0.5727$ unter der Verwendung der ITT-Analysen. Die Konfidenzintervalle überlappen sich deutlich, was darauf hinweist, dass der Unterschied zwischen den beiden Analysen nicht besonders deutlich ist (Tabelle 9). Außerdem überdecken beide Intervalle die

Null. Aussagekräftiger ist eine *bivariate Analyse*: $\hat{\theta} = 0.1111$ (für PP-Population) und $\hat{\theta} = -0.5151$ (für ITT-Population) sowie eine Differenz zwischen ITT- und PP-Analyse von -0.6263. Der Unterschied zwischen den Ergebnissen der beiden Analysen ist nicht signifikant. Die Zwischenstudienvarianzen sind mit $\hat{\tau}_0^2 = 1.1899$ und $\hat{\tau}_1^2 = 1.5212$ sehr ähnlich. Die Kovarianz der beiden Gruppen ist mit $\sigma_{01} = 5.815$ sehr hoch, was ebenfalls für eine gute Übereinstimmung der Ergebnisse spricht. Deutlich wird dies auch bei der *graphischen Analyse*: In Abbildung 10 auf Seite 63 sind die Schätzer beider Analysen dargestellt und miteinander verbunden. Auffallend unterschiedlich sind nur die Ergebnisse von Studie I.

Tabelle 9: Analyseergebnisse des Asthma-Beispiels

Auswertung	Population	$\hat{\theta}$	Konfidenzintervall	$\hat{\tau}^2$
alle Studien:				
PP-Analyse	PP	0.0844	[-1.743, 1.9116]	$\hat{\tau}^2 = 0.0385$
ITT-Analyse	ITT	-0.5727	[-2.341, 1.1957]	$\hat{\tau}^2 = 1.6502$
Bivariate Analyse	PP	0.1111	[-1.807, 2.0294]	$\hat{\tau}_0^2 = 1.1899$
Bivariate Analyse	ITT	-0.5151	[-2.240, 1.2101]	$\hat{\tau}_1^2 = 1.5282$
Bivariate Analyse	ITT-PP	-0.6263	[-2.135, 0.8826]	$\hat{\sigma}_{01} = 5.8915$
zufällige Auswahl:				
nur PP-Analysen	PP	-0.6289	[-3.458, 2.2000]	$\hat{\tau}^2 = 2.9619$
nur ITT-Analysen	ITT	1.5188	[-.8316, 3.8692]	$\hat{\tau}^2 = 0.0000$
Meta-Regression	PP	-0.6072	[-3.055, 1.8404]	
Meta-Regression	ITT	1.5188	[-.8316, 3.8692]	$\hat{\tau}^2 = 0.0000$
Meta-Regression	ITT-PP	2.1260	[-1.267, 5.5193]	
Resampling Techniken:				
Bootstrap-Verfahren	PP	0.0552	[-2.116, 2.2263]	
Bootstrap-Verfahren	ITT	-0.5661	[-2.685, 1.5530]	$\hat{\tau}^2 = 0.9905$(*)
Bootstrap-Verfahren	ITT-PP	-0.6213	[-4.357, 3.1140]	
Permutationsverfahren	PP	0.0662	[-1.101, 2.4260]	
Permutationsverfahren	ITT	-0.5795	[-2.634, 1.4745]	$\hat{\tau}^2 = 0.9685$(*)
Permutationsverfahren	ITT-PP	-0.6458	[-4.116, 2.8248]	

„zufällige Auswahl“ bedeutet, dass für dieses Beispiel für jede der elf Studien mit der Wahrscheinlichkeit 0.5 entweder die PP- *oder* die ITT-Analyse ausgewählt wurde. $\hat{\tau}^2$: Zwischenstudienvarianz, $\hat{\sigma}_{01}$: Kovarianz im bivariaten Modell, (*) Median = 0

Zur Erläuterung soll eine zufällige Auswahl der Populationsansätze getroffen werden, um an diesem Beispiel eine Meta-Regression zu rechnen und zu interpretieren: Für jede der elf Studien wird mit der Wahrscheinlichkeit 0.5 entweder die PP- *oder* die ITT-Analyse ausgewählt. Die *Sensitivitätsanalysen* für die beiden Populationsansätze liefern $\hat{\theta} = -0.6289$ (PP-Analyse mit den Studien B, C, F, H, I, J) bzw. $\hat{\theta} = 1.1588$ (ITT-Analyse mit den Studien A, D, E, G, K). Durch die zufällige Auswahl ist jetzt der Effekt in der ITT-Analyse größer. Das macht deutlich, dass sich sehr unterschiedliche Ergebnisse ergeben können, falls nur eine der Analysen publiziert wird. Mit einer Meta-Regression lässt sich der Unterschied zwischen den beiden Populationsansätzen schätzen sowie ein Konfidenzintervall berechnen (Tabelle 9). Auch hier liegt kein signifikanter Unterschied vor. Unabhängig vom gewählten Verfahren liegen die Konfidenzintervalle innerhalb der von Ebbutt angegebenen Äquivalenzgrenzen von ± 15. Die Ergebnisse sind graphisch in Abbildung 12 auf Seite 65 dargestellt.

Die beiden Resampling-Techniken (Bootstrap mit 10000 Wiederholungen) liefern Punktschätzer, die der bivariaten Analyse sehr ähnlich sind. Die Konfidenzintervalle sind etwas breiter für die PP- und ITT-Analyse, aber deutlich breiter bei dem Unterschied (ITT-PP). Da in jedem einzelnen Resampling-Schritt eine Meta-Regression gerechnet wird, kann jeweils nur eine Zwischenstudienvarianz $\hat{\tau}^2$ geschätzt werden. Hier weichen Mittelwert (0.9905 bzw. 0.9685) und Median (0) aus den Resampling-Stichproben sehr voneinander ab. Der Mittelwert ist dem Ergebnis aus der bivariaten Analyse bei weitem näher und scheint daher der bessere Schätzer zu sein. Eine endgültige Aussage kann jedoch anhand dieses Beispiels nicht gegeben werden.

Das kommentierte SAS-Programm zum Asthma-Beispiel für die bivariate Analyse sowie die Meta-Regression sind im Anhang in den Abschnitten 8.4 und 8.5 ab Seite 102 ausführlich dargestellt.

5 Diskussion

Meta-Analysen und Äquivalenzstudien werden in der Medizinischen Forschung immer häufiger verwendet. Meta-analytische Methoden, um Äquivalenzhypothesen zu untersuchen, wurden jedoch bisher nicht behandelt.

Generelles Unabhängig vom Studientyp können bereits durchgeführte Überlegenheits- und Äquivalenzstudien Evidenz zu einer klinischen Fragestellung liefern. Eine Selektion bezüglich des Studientyps ist daher nicht empfehlenswert, vielmehr bestimmen im Wesentlichen die Therapiegruppen, Zielgrößen, Studiendesign sowie Datenqualität den Einschluss in eine Meta-Analyse. Wenn aber Überlegenheitsstudien mit einem signifikanten Ergebnis vorliegen (E besser als A) und ebenso Äquivalenzstudien (E äquivalent zu A), sollte diese Heterogenität genau untersucht werden. Ergibt sich aufgrund der Meta-Analyse eine Äquivalenz der beiden Therapieverfahren, während eine Einzelstudie eine signifikante Überlegenheit einer Behandlungsform gezeigt hat, bedarf dieses ggf. widerprüchliche Ergebnis ebenso einer genauen Interpretation.

In der vorliegenden Arbeit konnte gezeigt werden, dass die in der Meta-Analyse üblichen Verfahren im FEM und REM [74, 98, 107] auch für Äquivalenzhypothesen angewendet werden können (Kapitel 2). Das Intervall-Inklusions-Verfahren hat auch für Meta-Analysen seine Gültigkeit. Die Interpretationen in den genannten Arbeiten sind daher methodisch gerechtfertigt. In den Reviews werden Äquivalenzgrenzen und Auswertungspopulationen jedoch nicht adäquat berücksichtigt. Das Intervall-Inklusions-Verfahren ist auf einen statistischen Test übertragbar. Daher ist es sinnvoll die in dieser Arbeit hergeleitete Teststatistik und den dazugehörigen p-Wert in einer Meta-Analyse anzugeben. Es gilt, wie auch für einzelne Äquivalenzstudien, dass Teststatistik und p-Wert nicht ohne Konfidenzintervall präsentiert werden sollten.

Äquivalenzgrenzen Um p-Werte zu berechnen und Konfidenzintervalle zu interpretieren ist die Definition einer Äquivalenzgrenze notwendig. Allerdings kann es in Einzelfällen ausreichend sind, die Meta-Analyse als Schätzproblem ohne Berücksichtigung von Äquivalenzgrenzen zu betrachten. Sollen jedoch Therapieempfehlungen abgegeben werden, ist ein entscheidungstheoretischer Ansatz nötig. Es wird gefordert, dass Äquivalenzgrenzen für Einzelstudien a-priori festgelegt [57] und im Protokoll festgeschrieben werden [37], auch wenn die Nicht-Unterlegenheit in einer Überlegenheitsstudie nur eine mögliche Interpretation darstellt [38]. Die Bestimmung der Äquivalenzgrenze δ stellt bei der Studien-

planung von Äquivalenz- und Nicht-Unterlegenheitsstudien ein wesentliches Problem dar: Zum einen ist die Definition von δ medizinisch schwierig und sie hat hat erheblichen Einfluss auf die spätere Interpretation der Ergebnisse. Zum anderen resultieren praktische Probleme, wie beispielsweise ein Einfluss auf die Fallzahl.

In der vorliegenden Arbeit wurden verschiedene Konzepte vorgestellt, die unterschiedliche Werte für δ liefern werden. Diese Äquivalenzgrenzen können je nach Intention und Fragestellung eingesetzt werden. Wie an einigen Beispielen in dieser Arbeit aufgezeigt, ist in den Publikationen jedoch häufig keine Angabe zu Äquivalenzgrenzen zu finden. Im CONSORT-Statement, den Richtlinien zur Publikation von randomisierten Studien [2], wird nur auf die Beschreibung der Hypothesen eingegangen. Das schließt δ *implizit* mit ein. Um die Publikationsqualität zu erhöhen und anschließend Meta-Analysen zu vereinfachen, ist es erforderlich, dass in Einzelstudien (i) Äquivalenzgrenzen definiert und (ii) zusammen mit den Hypothesen in den Publikationen genannt werden. (iii) Die Grenzen und Hypothesen müssen begründet werden. Im CONSORT-Statement sollte dies *explizit* genannt werden. Des Weiteren sind analog die Hypothesen und Äquivalenzgrenzen für eine Meta-Analyse festzulegen und ebenso zu publizieren. Ein entsprechender Hinweis sollte daher im QUOROM-Statement, den Richtlinien zur Publikation von Meta-Analysen [71], aufgenommen werden.

Soll eine neue Therapie eingeführt werden, sollte die Wirksamkeit im Placebovergleich gezeigt werden. Eine neue Therapie, die zwar der Standardtherapie nicht unterlegen, aber schlechter als Placebo ist, wäre von geringem klinischen Nutzen. Daher sollte die Äquivalenzgrenze δ nicht größer sein als $\delta_P(0)$. Diese Problematik hat auch Wiens beschrieben [111], es handelt sich dabei jedoch um einen sehr konservativen Ansatz [93]. Hauck und Anderson haben 1999 ein δ eingeführt, um die Wirksamkeit indirekt zeigen zu können [47]. Dabei handelt es sich um das kleinste δ mit dieser Eigenschaft (Abschnitt 3.1.2). Dieser Ansatz wurde auch von Whitehead in einem Vortrag diskutiert [108]. Durrleman und Chaikin verwenden den Effekt θ^{EP}, der aus historischen Daten und der aktuellen Studie geschätzt wird [30]. Diese Informationen finden sich in dem beschriebenen Ansatz in $\delta_P(0)$ wieder. Der Vorteil liegt darin, dass θ^{EA} wie üblich geschätzt und interpretiert werden kann. Durch den Vergleich mit $\delta_P(0)$ wird indirekt gegen Placebo getestet. Die im Satz 3.1 verwendeten Annahmen sind so allgemein, dass die Äquivalenzgrenze $\delta_P(\lambda)$ auch in praktischen Situationen angewendet werden kann. Im Gegensatz zu Wiens wurde in der vorliegenden Definition von $\delta_P(\lambda)$ zugelassen, dass der Stichprobenumfang und

die Varianzen in den beiden Studien unterschiedlich sein dürfen. Ebenso können mehrere Studien zusammen einen Schätzer für θ^{AP} liefern. So kann ein Meta-Analyse-Schätzer für den Vergleich zwischen A und P verwendet werden, für den Effekt ebenso wie für die Varianz des Schätzers. Auch andere Autoren haben darauf hingewiesen, dass in diesem Kontext eine Meta-Analyse verwendet werden sollte [47, 54, 101, 108]. Damit kann nicht nur eine bessere Punktschätzung angegeben sondern auch die Variabilität zwischen den einzelnen Studien (A vs. P) berücksichtigt werden. Schwierigkeiten der „constancy assumption“ können damit abgeschwächt werden. Wenn der Effekt θ^{AP} über die Zeit nicht konstant ist („constancy assumption“ verletzt [73]), so ist zu vermuten, dass sich mit der Zeit die Wirksamkeit der aktiven Kontrolle A verschlechtert [101], beispielsweise aufgrund von Resistenzen [14, 25]. Auch eine derzeit übliche Begleit- oder Basistherapie kann den Effekt der aktiven Kontrolle beeinflussen [93]. Die Konsequenz wäre ein zu großes und damit antikonservatives $\delta_P(0)$. Dies bedeutet, dass mit einer Wahrscheinlichkeit von mehr als α auf eine Wirksamkeit der neuen Therapie geschlossen würde, obwohl dies nicht gerechtfertigt wäre. Ein Ausweg, der auch regulatorischen Belangen gerecht würde, könnte die Einführung eines Faktors λ wie für $\delta_P(\lambda)$ sein, der $\delta_P(0)$ verkleinert [73, 101]. Andererseits könnte trotz der Hypothese $H_0^{EP} : \theta^{EP} \leq 0$ die Äquivalenzgrenze $(1 - \lambda)\ \delta_P(0)$ verwendet werden. Der Sicherheitsfaktor $(1 - \lambda)$ ist eine ad-hoc Lösung, falls die „constancy assumption“ *nicht* erfüllt ist. Ein Zeiteffekt der Wirkung von A wird jedoch nicht berücksichtigt.

Neben der Wirksamkeitsargumentation ist andererseits der Einfluss der Einzelstudien zu berücksichtigen. Die dort definierten δ_i liefern wichtige Informationen über die Äquivalenzgrenze. Der Mittelwert von δ_i als zusätzliche Äquivalenzgrenze wäre zu groß, weil unter Umständen viele der verwendeten δ_i zu groß angesetzt wurden. Der Median kann ebenso zu groß sein. Deshalb wird das Minimum oder das 1. Quartil als heuristische Lösung vorgeschlagen. Ein Äquivalenztest mit δ_{Q1} kann die übrigen Äquivalenztests mit $\delta_P(0)$ und δ_{clin} unterstützen.

Das empfohlene hierarchische Vorgehen (Abschnitt 3.4, Seite 46), um die Hypothesen zu testen, verpflichtet zur klaren Definition aller verwendeten Äquivalenzgrenzen. Die hier vorgeschlagene a-priori-Reihenfolge wird in der Regel auch die Reihenfolge der Grenzen sein, so dass kaum Powerverlust zu erwarten ist (Abbildung 8, Seite 48). Mit diesem Vorgehen können die vier Hypothesen schrittweise getestet und interpretiert werden. Zu beachten ist neben der Forderung ($\delta \leq \delta_P(0)$), dass auch $\delta_P(0)$ nicht allein getestet werden

sollte: Eine zwar wirksame, aber deutlich schlechtere Therapie als die Standardtherapie ist nicht erstrebenswert [86].

In dieser Arbeit wurden Äquivalenzhypothesen untersucht. Es gibt jedoch auch den Fall, dass hauptsächlich Nicht-Unterlegenheitsstudien durchgeführt wurden, um die Wirkungen einer Therapie zu untersuchen. Wenn die Einzelstudien stets einen positiven Effekt für die Therapie im Gegensatz zu einer aktiven Kontrolle zeigen, kann aus der Nicht-Unterlegenheits- eine Überlegenheitshypothese werden. Das Ziel einer Meta-Analyse kann demnach sein, die Nicht-Unterlegenheitsgrenze bis auf Null zu verkleinern. Da Meta-Analysen durch die Zusammenfassung von Studienergebnissen die statistische Power erhöhen, kann ein solches Ziel realistisch sein. Jedoch ist zu beachten, dass mit einer entsprechend großen Fallzahl beziehungsweise Anzahl an Studien jeder noch so kleine Unterschied signifikant werden kann, die Relevanz des Effektes jedoch in Frage zu stellen ist [99]. Darüber hinaus wird eine Nicht-Unterlegenheitsstudie mit einem $\delta \to 0$ zu einer einseitigen Überlegenheitsstudie, während bei zweiseitiger Äquivalenz aus $\delta \to 0$ eine Einpunkt-Alternative wird. Diese ist statistisch nicht zu zeigen.

Ein prinzipielles Problem von Äquivalenzstudien ist, dass die Äquivalenz eines wahren Effektes von $-\delta$ (bei Nicht-Unterlegenheitsstudien) bzw. $\pm\delta$ (bei zweiseitigen Äquivalenzstudien) nicht gezeigt werden kann, obwohl ein Effekt von $\pm\delta$ noch als klinisch akzeptabel gilt. Dieses Problem ist auch mit Meta-Analysen nicht zu lösen.

Im Zusammenhang mit der Äquivalenzgrenze δ werden an eine Meta-Analyse folgende Forderungen gestellt:

1. Alle zu verwendenden Äquivalenzgrenzen $(\delta, \delta_P(0), \delta_{clin}, \delta_{Q1}, \delta_{min}, \delta_P(\lambda))$ bzw. deren Berechnungsmethode und die dazugehörigen statistischen Hypothesen sind im Protokoll der Meta-Analyse festzulegen und zu begründen.

2. Keine Äquivalenzgrenze δ darf $\delta_P(0)$ übersteigen.

3. Es empfiehlt sich ein hierarchisches Vorgehen mit $\delta_P(0) \to \delta_{clin} \to \delta_{Q1} \to \delta_P(\lambda)$. Weder $\delta_P(0)$ zum Nachweis der Wirksamkeit noch δ_{clin} zum Nachweis der klinischen Äquivalenz sollten *allein* verwendet werden. Wenn nur *eine* Äquivalenzgrenze δ definiert werden soll, ist $\delta \leq \delta_P(0)$ und $\delta \leq \delta_{clin}$ zu wählen, beispielsweise $\delta = min(\delta_P(0), \delta_{clin}, \delta_{Q1})$.

4. Bei Meta-Analysen, die im Zulassungsprozess von Medikamenten verwendet werden und somit von regulatorischem Interesse sind, ist zu empfehlen, die Äquivalenzgrenzen mit den entsprechenden Behörden abzustimmen [31].

5. In δ_{min} bzw. δ_{Q1} sollten alle δ_i der Studien eingehen, die in die Meta-Analyse einfließen *könnten*. Treten deutliche Unterschiede in den δ_i auf, sind diese zu diskutieren und mögliche Ursachen zu eruieren. Die verschiedenen δ_i können den einzelnen Studiencharakteristika gegenübergestellt werden. Diese Unterschiede können zu Heterogenität in der Meta-Analyse führen, da in den Einzelstudien sehr unterschiedliche Bedingungen vorliegen. Überlegenheitsstudien sollten *nicht* mit $\delta_i = 0$ in die Definition von δ_{min} bzw. δ_{Q1} einbezogen werden.

Ergänzend kann der p-Wert eines Tests auf Nicht-Unterlegenheit wie in Abbildung 7 auf Seite 40 in Abhängigkeit von der Äquivalenzgrenze δ graphisch dargestellt werden. Damit kann der Einfluss der Äquivalenzgrenze auf den p-Wert direkt beurteilt werden. Eine ähnliche Darstellung in Abhängigkeit des Faktors λ findet sich bei Whitehead [108].

Populationen Die Populationsproblematik für Äquivalenzstudien verschärft sich für Meta-Analysen dadurch, dass in den Publikationen in der Regel nur eine der Analysen dargestellt und die genaue Definition häufig nicht angegeben wird. Daher sollten (i) Einzelstudien mit einer Äquivalenzhypothese die Hauptzielgrößen mit einer PP- *und* einer ITT-Population analysieren, (ii) die Definitionen von PP und ITT genau beschreiben und (iii) die Ergebnisse zur Verfügung stellen. Wenn sich PP- und ITT-Population nicht unterscheiden, ist dies explizit zu erwähnen. Idealerweise sollten beiden Analysen in einer Publikation dargestellt werden. Aus Platzgründen kann ein Verweis auf Internetseiten mit den weiteren Ergebnissen publiziert werden. In der vorliegenden Arbeit konnte gezeigt werden, dass die Auswirkung des Populationsansatzes auf den Effekt mit Verwendung beider Analysen beurteilt werden kann. Somit ist als Qualitätskriterium für eine Einzelstudie die Durchführung beider Analysen zu nennen. Es reicht daher nicht, nur das ITT-Prinzip in den Einzelstudien anzuwenden, was einige Autoren als alleiniges Qualitätskriterium ansehen [19, 27]. Eine Übersichtsarbeit in hochrangigen Zeitschriften (BMJ, JAMA, Lancet, NEJM) hat gezeigt, das zum Teil nicht nach dem ITT-Prinzip ausgewertet wird, obwohl die Auswertung als ITT-Analyse bezeichnet wurde (15 der untersuchten 119 Arbeiten) [53]. Der derzeitige Qualitätsstandard von Studienergebnissen ist daher noch nicht ausreichend.

In einer Meta-Analyse sind wie auch in Einzelstudien beide Populationsansätze (PP und ITT) zu berücksichtigen. Dazu werden entweder die Originaldaten verwendet, in der Publikation sind beide Analysen gegeben, eine Rekonstruktion der fehlenden Analysen ist möglich oder die Daten können von den Autoren beschafft werden. Der Einfluss der Populationsansätze lässt sich durch Sensitivitätsanalysen einschätzen. Der Populationseffekt sollte jedoch besser direkt geschätzt werden – entweder mit einer bivariaten Analyse, mit einer Meta-Regression oder mit Resampling-Techniken. Ist der Einfluss der Populationsansätze groß oder sogar signifikant, ist von einer einheitlichen Interpretation abzusehen. Die Methode der Wahl, um den Populationseffekt zu schätzen, ist die bivariate Analyse, in die alle Informationen eingehen. Liegt bei einigen wenigen Studien nur eine Analyse vor, könnten diese Studien bei einer bivariaten Analyse nicht verwendet werden. Es bieten sich hierbei Resampling-Techniken an, bei denen die entsprechende Analyse immer mit dem gleichen Analyseergebnis in die Auswertung eingeht. Analoge, zu vergleichende Vorgehensweisen aus klinischen Studien oder Meta-Analysen gibt es nicht.

Im Zusammenhang mit den Auswertungsansätzen werden an eine Meta-Analyse folgende Forderungen gestellt:

1. Im Protokoll der Meta-Analyse sollten die analytischen und graphischen Methoden beschrieben werden, um eine potentielle Heterogenität aufgrund verschiedener Auswertungsansätze zu untersuchen.

2. Es sollte versucht werden, für jede Studie sowohl die Ergebnisse der PP- als auch der ITT-Analyse zu erhalten. Es ist ggf. eine ITT-Rekonstruktion durchzuführen.

3. Die durch verschiedene Populationen verursachte Heterogenität sollte mindestens mittels einer Sensitivitätsanalyse untersucht und kritisch diskutiert werden.

4. Sofern die Voraussetzungen erfüllt sind, sollte eine bivariate Analyse durchgeführt werden.

5. Unterscheiden sich die Analyseergebnisse der PP- und ITT-Analyse erheblich oder ist der Unterschied im Rahmen der bivariaten Analyse (oder Meta-Regression) sogar signifikant, sollten die Ursachen untersucht werden. Eine globale Aussage über beide Populationen lässt sich dann nicht treffen.

6. Die vorgestellten Resampling-Techniken sind insbesondere dann zu empfehlen, wenn in einigen wenigen Studien nur eine PP- *oder* ITT-Analyse vorliegt.

Ausblick Die Populationsproblematik ist auch bei Einzelstudien mit Äquivalenzhypothesen noch nicht vollständig gelöst, weitere Forschungen auf diesem Gebiet sind notwendig. Darauf aufbauend kann die Auswirkung auf Meta-Analysen beurteilt werden. Ideales Instrument wäre der Vergleich von publikationsbasierten und originaldatenbasierten Meta-Analysen, um den Informationsverlust bezüglich der Populationsansätze und der Aggregation von Daten beurteilen zu können. Darüber hinaus könnte die Voraussetzung der bekannten Varianzen im Satz 3.1 aufgegeben und eine entsprechende Äquivalenzgrenze ermittelt werden. Wenn zur Planung einer Einzelstudie eine Äquivalenzgrenze $\delta_P(\lambda)$ verwendet wird, ist diese Grenze auch für die Fallzahlplanung relevant. Andererseits wird die Grenze von der Fallzahl beeinflusst. Für den einfachen Fall bekannter Varianzen wurde in dieser Arbeit eine iterative Lösung vorgeschlagen. Erweiterungen auf allgemeinere Fälle sind noch umzusetzen. Des Weiteren ist die problematische Annahme der konstanten Effekte („constancy assumption“) noch näher zu untersuchen und gegebenenfalls neue Methoden zu erarbeiten.

6 Zusammenfassung

Eine Meta-Analyse ist eine statistische Methode, Ergebnisse verschiedener Einzelstudien zusammenzufassen. Von zunehmender Wichtigkeit in der klinischen Forschung sind Äquivalenzfragestellungen: Wirken zwei Behandlungen ähnlich, oder ist eine neue Therapie nicht wesentlich schlechter als die etablierte Therapie?

Die üblichen statistischen Methoden einer Meta-Analyse lassen sich auf Äquivalenzfragestellungen übertragen. In der vorliegenden Arbeit wurde eine modifizierte Teststatistik entwickelt und ein entsprechender p-Wert angegeben. Dabei wurde die Äquivalenz zum Intervall-Inklusions-Verfahren gezeigt.

Zur Bestimmung der Äquivalenzgrenzen wurden bekannte Techniken aus der Theorie der Äquivalenzstudien für Meta-Analysen angepasst und in einem Beispiel dargestellt. Eine Äquivalenzgrenze $\delta_P(0)$ wurde so hergeleitet, dass die Wirksamkeit einer Therapie indirekt nachgewiesen werden kann. Darüber hinausgehende Fragestellungen sollten hierarchisch untersucht werden. Neben der klinischen Äquivalenz (δ_{clin}) wurden zwei Möglichkeiten hergeleitet, die Äquivalenzgrenzen der Einzelstudien in die Meta-Analyse einfließen zu lassen: Die Minimum-Lösung δ_{min} sowie die Quartil-Lösung δ_{Q1}. Desweiteren lässt sich mit $\delta_P(\lambda)$ testen, ob die neue Therapie einen Anteil der Wirkung der etablierten Therapie erreicht.

Um unterschiedliche Auswertungspopulationen (per-protocol, PP oder intention-to-treat, ITT) in einer Meta-Analyse zu berücksichtigen, wurden verschiedene Methoden vorgestellt, diskutiert und an einem Beispiel illustriert. Empfehlenswert ist insbesondere eine bivariate Methode, in die alle Studienergebnisse eingehen. Resampling-Techniken empfehlen sich insbesondere dann, wenn in einigen Einzelstudien nur eine der Analysen (PP oder ITT) vorliegt.

Fazit: Äquivalenzfragen lassen sich mit meta-analytischen Methoden adäquat beantworten. Für Meta-Analysen und Einzelstudien wurde ein Anforderungskatalog für die Publikationsqualität und die biometrischen Methoden zusammengestellt.

7 Verzeichnisse

7.1 Abbildungsverzeichnis

7.2 Tabellenverzeichnis

7.3 Symbolverzeichnis

k	Anzahl der Einzelstudien einer Meta-Analyse	19
L	untere Konfidenzintervallgrenze	21
L^*	untere Konfidenzintervallgrenze im REM	23
L^{AP}	untere Konfidenzintervallgrenze von θ^{AP}	30
L^{EP}	untere Konfidenzintervallgrenze von θ^{EP}	32
λ	reelle Zahl, Anteil	12
$N(\cdot, \cdot)$	Normalverteilung	9
n^{AP}	Fallzahl der Studie A vs. P	26, 33
n^{EA}	Fallzahl der Studie E vs. A	26, 33
n^{EP}	Fallzahl der Studie E vs. P	26
P	Placebogruppe	12
p	p-Wert	21
$p(\delta)$	p-Wert in Abhängigkeit von δ	39
Φ	p-Wert	21
Q	Heterogenitätsstatistik nach COCHRAN	23, 23
SE	Standardabweichung des Effektschätzers $\hat{\theta}$	20, 31
SE^{EA}	Standardabweichung des Effektschätzers $\hat{\theta}^{EA}$	31
SE^{AP}	Standardabweichung des Effektschätzers $\hat{\theta}^{AP}$	31
σ_{01}	Kovarianz zwischen PP- und ITT-Analyse	66
τ	Zwischenstudienvarianz	9
τ_0^2, τ_1^2	Varianzkomponenten der bivariaten Analyse	66
$\hat{\tau}_{DSL}$	ein Schätzer der Zwischenstudienvarianz	22
θ	globaler Effekt	9
θ_i	wahrer Effekt der i-ten Studie	9
$\hat{\theta}_i$	Effektschätzer der i-ten Studie	9
θ^{AP}	Effekt der aktiven Kontrolle gegenüber Placebo	12
θ^{EA}	Effekt der neuen Therapie gegenüber der aktiven Kontrolle	11
θ^{EP}	Effekt der neuen Therapie gegenüber Placebo	12

7.4 Abkürzungsverzeichnis

A	aktive Kontrollgruppe	11
DIMDI	Deutsches Institut für Medizinische Dokumentation und Information	6
E	Experimentalgruppe	11
EbM	Evidenz basierte Medizin	8
EMEA	The European Agency for the Evaluation of Medicinal Products	7
FEM	Modell mit festen Effekten (fixed effects model)	9
ICH	International Conference on Harmonisation	22
ITT	intention to treat (population)	56
MAL	publikationsbasierte Meta-Analyse (MA on literature)	9
MAP	originaldatenbasierte Meta-Analyse (MA on patient data)	9
P	Placebogruppe	12
PP	per protocol (population)	56
REM	Modell mit zufälligen Effekten (random effects model)	9
SAS	Statistical Analysis Software	37
WOMAC	Western Ontario and McMaster Universities Osteoarthritis Index	6

7.5 Literaturverzeichnis

[1] Altman D.G. (1991) Practical Statistics for Medical Research. **1**. Chapman and Hall, London, New York, Tokyo, Melbourne, Madras.

[2] Altman D.G., Schulz K.F., Moher D., Egger M., Davidoff F., Elbourne D., Gotzsche P.C. und Lang T. (2001) The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern.Med* **134**, 663-694.

[3] Angst F., Aeschlimann A. und Stucki G. (2001) Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis Rheum* **45**, 384-391.

[4] Angst F., Aeschlimann A., Michel B.A. und Stucki G. (2002) Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities. *J Rheumatol* **29**, 131-138.

[5] Barza M., Ioannidis J.P., Cappelleri J.C. und Lau J. (1996) Single or multiple daily doses of aminoglycosides: a meta-analysis. *Br Med J* **312**, 338-345.

[6] Bauer H.W., Klasser M., von Hanstein K.L., Rolinger H., Schladitz G., Henke H.D., Gimbel W. und Steinbach K. (1999) Oxaceprol is as effective as diclofenac in the therapy of osteoarthritis of the knee and hip. *Clinical Rheumatology* **18**, 4-9.

[7] Bauer P., Brannath W. und Posch M. (2001) Flexible two stage designs: An Overview. *Methods of Information in Medicine* **40**, 117-121.

[8] Beaupré L.A., Davies D.M., Jones C.A. und Cinats J.G. (2001) Exercise combined with continuous passive motion or slider board therapy compared with exercise only: a randomized controlled trial of patients following total knee arthroplasty. *Phys Ther* **81**, 1029-1037.

[9] Begg C.B. und Mazumdar M. (1994) Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50**, 1088-1101.

[10] Bellamy N., Buchanan W.W., Goldsmith C.H., Campbell J. und Stitt L.W. (1988) Validation study of WOMAC: a health status instrument for measuring clinically

important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol* **15**, 1833-1840.

[11] Bellamy N., Kean W.F., Buchanan W.W., Gerecz-Simon E. und Campbell J. (1992) Double blind randomized controlled trial of sodium meclofenamate (Meclomen) and diclofenac sodium (Voltaren): post validation reapplication of the WOMAC Osteoarthritis Index. *J Rheumatol* **19**, 153-159.

[12] Bellamy N., Buchanan W.W., Chalmers A., Ford P.M., Kean W.F., Kraag G.R., Gerecz-Simon E. und Campbell J. (1993) A multicenter study of tenoxicam and diclofenac in patients with osteoarthritis of the knee. *J Rheumatol* **20**, 999-1004.

[13] Bellamy N. (1995) WOMAC Osteoarthritis Index: A User's Guide. London, Ontario.

[14] Benninger M.S., Sedory Holzer S.E. und Lau J. (2000) Diagnosis and treatment of uncomplicated acute bacterial rhinosinusitis: summary of the Agency for Health Care Policy and Research evidence-based report. *Otolaryngol Head Neck Surg* **122**, 1-7.

[15] Berkey C.S., Hoaglin D.C., Mosteller F. und Colditz G.A. (1995) A random-effects regression model for meta-analysis. *Stat Med* **14**, 395-411.

[16] Blackwelder W.C. (2002) Showing a treatment is good because it is not bad. when does "noninferiority" imply effectiveness? *Control Clin Trials* **23**, 52-54.

[17] Bock J. (1998) Bestimmung des Stichprobenumfangs für biologische Experimente und kontrollierte klinische Studien. Oldenbourg, München.

[18] Califf R.M. (1998) A perspective on the regulation of the evaluation of new antithrombotic drugs. *Am J Cardiol* **82**, 25P-35P.

[19] Carroli G., Villar J., Piaggio G., Khan-Neelofur D., Gulmezoglu M., Mugford M., Lumbiganon P., Farnot U. und Bersgjo P. (2001) WHO systematic review of randomised controlled trials of routine antenatal care. *Lancet* **357**, 1565-1570.

[20] CDER (1999) Guidance for Industry, Clinical Development Programs for Drugs, Devices, and Biological Products Intended for the Treatment of Osteoarthritis (OA) (Draft Guidance). *CDER* .

[21] Cheek C.M., Black N.A., Devlin H.B., Kingsnorth A.N., Taylor R.S. und Watkin D.F. (1998) Groin hernia surgery: a systematic review. *Ann R Coll Surg Engl* **80 Suppl 1**, S1-80.

[22] Cochran W.G. (1954) The combination of estimates from different experiments. *Biometrics* **10**, 101-129.

[23] Colditz G.A., Miller J.N. und Mosteller F. (1988) Measuring gain in the evaluation of medical technology. The probability of a better outcome. *Int J Technol Assess Health Care* **4**, 637-642.

[24] Collier J. (1995) Confusion over use of placebos in clinical trials. *Br Med J* **311**, 821-822.

[25] Crome P. und Bruce-Jones P. (1992) Infection in the elderly: studies with lomefloxacin. *Am J Med* **92**, 126S-129S.

[26] Davison A.C. und Hinkley D.V. (1997) Bootstrap methods and their applications. Cambridge University Press, Cambridge.

[27] Deeks J.J., Smith L.A. und Bradley M.D. (2002) Efficacy, tolerability, and upper gastrointestinal safety of celecoxib for treatment of osteoarthritis and rheumatoid arthritis: systematic review of randomised controlled trials. *Br Med J* **325**, 619-623.

[28] DerSimonian R. und Laird N.M. (1986) Meta-Analysis in Clinical Trials. *Controlled Clinical Trials* **7**, 177-188.

[29] Djulbegovic B. (2001) Letter to the Editor. *Ann Intern Med* **135**, 62-63.

[30] Durrleman S. und Chaikin P. (2003) The use of putative placebo in active control trials: two applications in a regulatory setting. *Stat Med* **22**, 941-952.

[31] Ebbutt A.F. und Frith L. (1998) Practical issues in equivalence trials. *Stat Med* **17**, 1691-1701.

[32] Egger M., Davey S.G., Schneider M. und Minder C. (1997) Bias in meta-analysis detected by a simple, graphical test. *Br Med J* **315**, 629-634.

[33] Ehrich E.W., Davies G.M., Watson D.J., Bolognese J.A., Seidenberg B.C. und Bellamy N. (2000) Minimal perceptible clinical improvement with the Western Ontario and McMaster Universities osteoarthritis index questionnaire and global assessments in patients with osteoarthritis. *J Rheumatol* **27**, 2635-2641.

[34] EMEA (1997) Note for guidance on evaluation of new anti-bacterial medicinal products. *EMEA* **CPMP/EWP/558/95**.

[35] EMEA (1998) Note for guidance on clinical investigation of medicinal products in the treatment of hypertension. *EMEA* **CPMP/EWP/238/95 Rev. 1**.

[36] EMEA (1998) Points to Consider on Clinical Investigation of Medicinal Products used in the Treatment of Osteoarthritis. *EMEA* **CPMP/EWP/784/97**.

[37] EMEA (1999) Concept Paper on the Choice of Delta. *EMEA* **CPMP/EWP/2158/99**.

[38] EMEA (2000) Points to Consider on Switching between Superiority and Non-Inferiority. *EMEA* **CPMP/EWP/482/99**.

[39] EMEA (2001) Note for guidance on the investigation of bioavailability and bioequivalence. *EMEA* **CPMP/EWP/QWP/1401/98**.

[40] Fisher L.D., Dixon D.O., Herson J., Frankowski R.K., Hearron M.S. und Peace K.E. (1990) Intention to Treat in Clinical Trials. 331-350. Marcel Dekker, New York.

[41] Follmann D.A. und Proschan M.A. (1999) Valid inference in random effects meta-analysis. *Biometrics* **55**, 732-737.

[42] Friedman L.M., Furberg C.D. und DeMets D.L. (1998) Issues in data analysis. **3**, 284-322. Springer, New York.

[43] Furno P., Bucaneve G. und Del Favero A. (2002) Monotherapy or aminoglycoside-containing combinations for empirical antibiotic treatment of febrile neutropenic patients: a meta-analysis. *Lancet Infect Dis* **2**, 231-242.

[44] Galbraith R.F. (1988) A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* **7**, 889-894.

[45] Garrett A.D. (2003) Therapeutic equivalence: fallacies and falsification. *Stat Med* **22**, 741-762.

[46] Hartung J. und Knapp G. (2001) Strategien zur Beurteilung eines Behandlungseffektes mit Hilfe einer Meta-Analyse. *Inform Biom Epidemiol Med Biol* **32**, 44-59.

[47] Hauck W.W. und Anderson S. (1999) Some issues in the design and analysis of equivalence trials. *Drug Inform J* **33**, 109-118.

[48] Hauschke D. (2001) Choice of delta: A special case. *Drug Inform J* **35**, 875-879.

[49] Herrmann G., Steeger D., Klasser M., Wirbitzky J., Fürst M., Venbrocks R., Rohde H., Jungmichel D., Hildebrandt H.D., Parnham M.J., Gimbel W. und Dirschedl H. (2000) Oxaceprol is a well-tolerated therapy for osteoarthritis with efficacy equivalent to diclofenac. *Clinical Rheumatology* **19**, 99-104.

[50] Heuschkel R.B., Menache C.C., Megerian J.T. und Baird A.E. (2000) Enteral nutrition and corticosteroids in the treatment of acute Crohn's disease in children. *J Pediatr Gastroentero Nutr* **31**, 8-15.

[51] Hildebrandt H.D. (1995) Therapie von Gon- und Coxarthrosen. Klinischer Vergleich von Oxaceprol und Ibuprofen. *Zeitschrift für Allgemeinmedizin* **71**, 1742-1748.

[52] Hill A.B. (1961) Clinical trials. **7**, 259. The Lancet Ltd., London.

[53] Hollis S. und Campbell F. (1999) What is meant by intention to treat analysis? Survey of published randomised controlled trials. *Br Med J* **319**, 670-674.

[54] Holmgren E.B. (1999) Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. *J Biopharm Stat* **9**, 651-659.

[55] ICH (1995) Note for guidance on structure and content of clinical study reports (ICH-E3). *EMEA* **CPMP/ICH/137/95**.

[56] ICH (1998) Note for guidance on statistical principles for clinical trials (ICH-E9). *EMEA* **CPMP/ICH/363/96**.

[57] ICH (2000) Note for guidance on choice of control group in clinical trials (ICH-E10). *EMEA* **CPMP/ICH/364/96**.

[58] Jones B., Jarvis P., Lewis J.A. und Ebbutt A.F. (1996) Trials to assess equivalence: the importance of rigorous methods. *Br Med J* **313**, 36-39.

[59] Katz R. (1991) Regulatory View. Use of Subgroup Data for Determination of Efficacy. 251-263. Raven Press, New York.

[60] Kay R. (1995) The principle of intent-to-treat in comparative trials. *Herpes* **2**, 90-92.

[61] Kriegel W., Korff K.J., Ehrlich J.C., Lehnhardt K., Macciocchi A., Moresino C. und Pawlowski C. (2001) Double-blind study comparing the long-term efficacy of the COX-2 inhibitor nimesulide and naproxen in patients with osteoarthritis. *Int J Clin Pract* **55**, 510-514.

[62] L'Abbé K.A., Detsky A.S. und O'Rourke K. (1987) Meta-analysis in clinical research. *Ann Intern Med* **107**, 224-233.

[63] Lange S. und Windeler J. (1997) Das Konzept der therapeutischen Äquivalenz. *Medizinische Klinik* **92**, 215-220.

[64] Lasek R. und Müller-Oerlinghausen B. (2001) Degenerative Gelenkerkrankungen. *Arzneiverordnung in der Praxis, Therapieempfehlungen der Arzneimittelkommission der deutschen Ärzteschaft* , 1-27.

[65] Lewis J.A. und Machin D. (1993) Intention to treat–who should use ITT? *Br J Cancer* **68**, 647-650.

[66] Light R.J. und Pillemer D.B. (1984) Summing up: The science of reviewing research. Harvard University Press, Cambridge, Mass.

[67] Loenhout-Rooyackers J.H., Keyser A., Laheij R.J., Verbeek A.L. und van der Meer J.W. (2001) Tuberculous meningitis: is a 6-month treatment regimen sufficient? *Int J Tuberc Lung Dis* **5**, 1028-1035.

[68] Makarowski W., Zhao W.W., Bevirt T. und Recker D.P. (2002) Efficacy and safety of the COX-2 specific inhibitor valdecoxib in the management of osteoarthritis of the hip: a randomized, double-blind, placebo-controlled comparison with naproxen. *Osteoarthritis Cartilage* **10**, 290-296.

[69] Michael M., Hodson E.M., Craig J.C., Martin S. und Moyer V.A. (2002) Short compared with standard duration of antibiotic treatment for urinary tract infection: a systematic review of randomised controlled trials. *Arch Dis Child* **87**, 118-123.

[70] Mismetti P., Laporte S., Darmon J.Y., Buchmüller A. und Decousus H. (2001) Meta-analysis of low molecular weight heparin in the prevention of venous thromboembolism in general surgery. *Br J Surg* **88**, 913-930.

[71] Moher D., Cook D.J., Eastwood S., Olkin I., Rennie D. und Stroup D.F. (1999) Improving the quality of reports of meta-analysis of randomised controlled trials: the QUOROM statement. *Lancet* **354**, 1896-1900.

[72] Newell D.J. (1992) Intention-to-treat analysis: implications for quantitative and qualitative research. *Int J Epidemiol* **21**, 837-841.

[73] Ng T.-H. (2001) Choice of delta in equivalence testing. *Drug Inform J* **35**, 1517-1527.

[74] Normand S.-L.T. (1999) Tutorial in Biostatistics Meta-Analysis: Formulation, Evaluating, Combining, and Reporting. *Stat Med* **18**, 321-359.

[75] Petrella R.J., DiSilvestro M.D. und Hildebrand C. (2002) Effects of hyaluronate sodium on pain and physical functioning in osteoarthritis of the knee: a randomized, double-blind, placebo- controlled clinical trial. *Arch Intern Med* **162**, 292-298.

[76] Pigeot I., Schäfer J., Röhmel J. und Hauschke D. (2003) Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Stat Med* **22**, 883-899.

[77] Pocock S.J. (1983) The justification for randomized controlled trials. **1**, 50-65. Wiley, Chichester.

[78] Röhmel J. (1998) Therapeutic equivalence investigations: statistical considerations. *Stat Med* **17**, 1703-1714.

[79] Röhmel J. (2001) Statistical considerations of FDA and CPMP rules for the investigation of new anti-bacterial products. *Stat Med* **20**, 2561-2571.

[80] Rothman K.J. und Michels K.B. (1994) The continuing unethical use of placebo controls. *The New England Journal of Medicine* **331**, 394-398.

[81] Sackett D.L., Rosenberg W.M., Gray J.A., Haynes R.B. und Richardson W.S. (1996) Evidence based medicine: what it is and what it isn't. *Br Med J* **312**, 71-72.

[82] Schulz K.F., Chalmers I., Hayes R.J. und Altman D.G. (1995) Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *J Am Med Ass* **273**, 408-412.

[83] Schwarzer G., Antes G. und Schumacher M. (2002) Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Stat Med* **21**, 2465-2477.

[84] Senn S. (1997) Intention to treat. In: Statistical issues in drug development. 153-160. John Wiley & Sons, Chicester, New York, Winheim, Brisbane, Singapore, Toronto.

[85] Sidik K. und Jonkman J.N. (2002) A simple confidence interval for meta-analysis. *Stat Med* **21**, 3153-3159.

[86] Siegel J.P. (2000) Equivalence and noninferiority trials. *American heart journal* **139**, S166-S170.

[87] Simon R. (2000) Editorial: Are placebo-controlled clinical trials ethical or needed when alternative treatment exist? *Ann Intern Med* **133**, 474-475.

[88] Singer F., Mayrhofer F., Klein G., Hawel R. und Kollenz C.J. (2000) Evaluation of the efficacy and dose-response relationship of dexibuprofen (S(+)-ibuprofen) in patients with osteoarthritis of the hip and comparison with racemic ibuprofen using the WOMAC osteoarthritis index. *Int J Clin Pharmacol Ther* **38**, 15-24.

[89] Stewart L.A. und Parmar M.K.B. (1993) Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet* **341**, 418-422.

[90] Stucki G., Meier D., Stucki S., Michel B.A., Tyndall A.G., Dick W. und Theiler R. (1996) Evaluation einer deutschen Verison des WOMAC (Western Ontario and McMaster Universities) Arthrose Index. *Z Rheumatol* **55**, 40-49.

[91] Takeda W. und Wessel J. (1994) Acupuncture for the treatment of pain of osteoarthritic knees. *Arthritis Care Res* **7**, 118-122.

[92] Temple R. und Ellenberg S.S. (2000) Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. *Ann Intern Med* **133**, 455-463.

[93] Temple R. (2002) Policy developments in regulatory approval. *Stat Med* **21**, 2939-2948.

[94] The Cochrane Collaboration (2002) Cochrane Reviewers' Handbook. **4**. Update Software, Oxford.

[95] Tramer M.R., Reynolds D.J., Moore R.A. und McQuay H.J. (1998) When placebo controlled trials are essential and equivalence trials are inadequate. *Br Med J* **317**, 875-880.

[96] Tran D., Muchant D.G. und Aronoff S.C. (2001) Short-course versus conventional length antimicrobial therapy for uncomplicated lower urinary tract infections in children: a meta- analysis of 1279 patients. *J Pediatr* **139**, 93-99.

[97] Vagt C.W., Kaiser T. und Leineweber G. (1990) Wirksamkeitsvergleich der oralen Therapie mit Oxaceprol versus Ibuprofen bei Gonarthrose und Coxarthrose. *Rheuma* **10**, 263-267.

[98] Van Houwelingen H.C., Arends L.R. und Stijnen T. (2002) Advanced methods in meta-analysis: multivariate approach and meta- regression. *Stat Med* **21**, 589-624.

[99] Victor N. (1987) On clinically relevant differences and shifted null hypotheses. *Methods Inf Med* **26**, 109-116.

[100] Victor N. (1995) The Challenge of Meta-Analysis: Discussion indications and contraindications for Meta-Analysis. *J Clin Epidemiol* **48**, 5-8.

[101] Wang S.J., Hung H.M. und Tsong Y. (2002) Utility and pitfalls of some statistical methods in active controlled clinical trials. *Controlled Clinical Trials* **23**, 15-28.

[102] Wang S.J. und James Hung H.M. (2003) Assessing treatment efficacy in noninferiority trials. *Control Clin Trials* **24**, 147-155.

[103] Wasilewski M.M., Johns D. und Sides G.D. (1999) Five-day dirithromycin therapy is as effective as seven-day erythromycin therapy for acute exacerbations of chronic bronchitis. *J Antimicrob Chemother* **43**, 541-548.

[104] Wellek S. (1994) Statistische Methoden zum Nachweis von Äquivalenz. **1**. Gustav Fischer Verlag, Stuttgart, Jena, New York.

[105] Wellek S. (2000) Proof of equivalence as a new issue in confirmatory statistics. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie* **28**, 25-34.

[106] Westlake W.J. (1972) Use of confidence intervals in analysis of comparative bioavailability trials. *J Pharm Sci* **61**, 1340-1341.

[107] Whitehead A. und Whitehead J.A. (1991) A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in medicine* **10**, 1665-1677.

[108] Whitehead A. (2002) Statistical issues in determining the effect of a new treatment from active-control trials. *23rd Annual Conference, The International Society for Clinical Biostatistics, Dijon* .

[109] Whitehead A. (2002) Meta-Analysis of controlled clinical trials. **1**. Wiley, Chichester.

[110] Whitehead A. (2002) A confidence interval plot. In:Meta-Analysis of controlled clinical trials. **1**, 183-186. Wiley, Chichester.

[111] Wiens B.L. (2002) Choosing an equivalence limit for noninferiority or equivalence studies. *Control Clin Trials* **23**, 2-14.

[112] Windeler J. (1993) Das Intention-to-treat-Prinzip in klinischen Arzneimittelprüfungen. *Arzneimitteltherapie* **11**, 103-111.

[113] Windeler J. und Trampisch H.J. (1995) Empfehlungen zur Durchführung von Studien zur therapeutischen Äquivalenz. *Inform Biom Epidemiol Med Biol* **4**, 350-355.

[114] Witte S. und Victor N. (2000) Meta-Analyse zur Effektivität von S-Adenosylmethionin und Oxaceprol für die Behandlung von Arthrosen. *Forschungsberichte der Abteilung Medizinische Biometrie* **36**, 1-108. Abteilung Medizinische Biometrie, Heidelberg.

[115] Witte S., Lasek R. und Victor N. (2002) Wirksamkeit von Ademethionin und Oxaceprol für die Behandlung von Arthrosen: Eine Meta-Analyse. *Der Orthopäde* **31**, 1058-1065.

[116] Ziegler S., Koch A. und Victor N. (2001) Deficits and remedy of the standard random effects methods in meta- analysis. *Methods Inf Med* **40**, 148-155.

7.6 Eigene Arbeiten

Folgende Publikationen, Vorträge und Poster stehen im Zusammenhang mit dieser Arbeit:

Witte S. und Victor N. (2000) Meta-Analyse zur Effektivität von S-Adenosylmethionin und Oxaceprol für die Behandlung von Arthrosen. *Forschungsberichte der Abteilung Medizinische Biometrie* **36**, 1-108. Heidelberg, Abteilung Medizinische Biometrie.

Witte S., Lasek R. und Victor N. (2002) Wirksamkeit von Ademethionin und Oxaceprol für die Behandlung von Arthrosen: Eine Meta-Analyse. *Der Orthopäde* **31**(11), 1058-1065.

Witte S. (2002) Äquivalenz und Meta-Analyse - Offene Probleme und erste Gedanken (Vortrag). 2. Workshop der Projektgruppe "Methodik systematischer Reviews" der GMDS, Freiburg, 22. Februar 2002

Witte S. (2002) Meta-analysis using noninferiority trials (Poster). 23rd annual conference of International Society of Clinical Biostatistics (ISCB), Dijon, 9.-13. September 2002

Victor N. und Witte S. (2002) Meta-Analysen von Äquivalenzstudien (Vortrag). Kolloquium: Statistische Methoden in der empirischen Forschung, Berlin, 17. Dezember 2002

Witte S. und Victor N. Meta-analysis to investigate equivalence (Vortrag). 50. Biometrisches Kolloquium, Heidelberg, 16.-19. März 2004

Witte S. und Victor N. (2004) Some problems with the investigation of noninferiority in meta-analysis. *Methods of Information in Medicine* (im Druck)

8 Anhang

8.1 Anzahl von Studientypen in MEDLINE

Tabelle 11: Anzahl der verschiedenen Studientypen in MEDLINE: 1990 – 2001

Jahr	MEDLINE	Clinical Trial	RCT	TE	TE+BE	MA
1990	395705	11605	6701	11	28	273
1991	397132	11981	7134	14	26	333
1992	400290	11933	7539	23	49	371
1993	408064	12461	8209	32	58	323
1994	417575	18789	9462	31	48	386
1995	428568	20483	10351	53	105	427
1996	438190	19684	10133	53	111	481
1997	434706	21256	10283	46	105	596
1998	452817	21697	10644	49	132	636
1999	467951	23158	11239	71	157	734
2000	506074	21853	10579	96	184	835
2001	506186	18675	9817	91	192	878

RCT = randomisierte klinische Studie, TE = therapeutische Äquivalenz, Suchstrategie in PubMed: "Clinical Trial"[PTYP] AND "therapeutic equivalence"[MeSH], TE+BE = TE und Bioäquivalenz, Suchstrategie in MEDLINE: (equivalenc*[All Fields] OR "therapeutic equivalency"[MeSH Terms] OR bioequivalenc*[Text Word] OR non-inferio*[All Fields]) AND Clinical Trial[ptyp]), MA = Meta-Analyse

8.2 Äquivalenzgrenzen für WOMAC

Tabelle 12: Äquivalenzgrenzen für WOMAC

i	1	2	3	4	5	min
Referenz	[8]	[61]	[33]	[3]	[4]	
untere Schranken:						
WOMAC A	10	12	9.7	11.0	6.4	6.4
	(20.8%)	(25.3%)	(14.9%)	(22.8%)	(13.9%)	(13.9%)
WOMAC B			10.0	5.1	2.9	2.9
			(15.2%)	(11.1%)	(6.2%)	(6.2%)
WOMAC C			9.3	13.3	10.3	9.3
			(14.6%)	(27.7%)	(22.2%)	(14.6%)
Normalized WOMAC			9.7	9.8	8.8	8.8
			(14.9%)	(27.2%)	(17.3%)	(14.9%)
Total WOMAC			9.9	12.9	9.6	9.6
			(15.4%)	(26.9%)	(20.7%)	(15.4%)
obere Schranken:						
WOMAC A	10			7.5	8.3	7.5
	(20.8%)			(15.5%)	(18.0%)	(15.5%)
WOMAC B				7.2	10.1	7.2
				(15.6%)	(21.7%)	(15.6%)
WOMAC C				6.7	8.0	6.7
				(13.9%)	(17.3%)	(13.9%)
Normalized WOMAC				7.1	6.5	6.5
				(14.9%)	(14.0%)	(14.0%)
Total WOMAC				6.7	8.2	6.7
				(14.0%)	(17.7%)	(14.0%)

% = Veränderung vom Ausgangswert, A = Schmerz Subskala, B = Steifigkeit, C = Funktionalität

8.3 Fallzahlplanung für eine Nichtunterlegenheitsstudie

Das folgende SAS-Programm dient der Fallzahlplanung für den einfachen Fall der Nicht-Unterlegenheit mit bekannten sowie homogenen Varianzen und der Verwendung einer Nicht-Unterlegenheitsschranke $\delta'_P(0)$ aus Korollar 3.3 (Tabelle 3 auf Seite 37).

```
DATA samplesize;
  INPUT alpha power0 n1 sigma1 xaxp1 @@;
  DATALINES;
  0.050 0.8   10 10 5  0.050 0.8   50 10 5  0.050 0.8  100 10 5
  0.050 0.8  500 10 5  0.050 0.8 1000 10 5;
RUN;

DATA test(KEEP = n1 n2 power alpha power0 delta lcl1 xaxp1 niter);
  SET samplesize;
  *** Parameter ***;
  n2 = .; niter = .; lev = 1-alpha;
  ua = PROBIT(lev); ub = PROBIT(power0);
  se1 = SQRT(2) * sigma1 / SQRT(n1); lcl1 = xaxp1 - ua * se1;
  IF lcl1<0 THEN PUT "WARNING: Aktive Kontrolle nicht wirksam! "/_ALL_;
  ELSE DO;
    *** Startwerte ***;
    power = 0;
    n2 = CEIL((ua + ub) ** 2 * sigma1 ** 2 * 2 / xaxp1 ** 2);
    delta = xaxp1 - (SQRT(1+n1/n2) - SQRT(n1/n2)) * ua * se1;
    shift = delta * SQRT(n2) / SQRT(2) / sigma1;
    power = 1 - PROBNORM(ua - shift);
    niter = 1;
    *** Iterationen ***;
    DO WHILE(power < power0);
      niter + 1; n2 + 1;
      delta = xaxp1 - (SQRT(1+n1/n2) - SQRT(n1/n2)) * ua * se1;
      shift = delta * SQRT(n2) / SQRT(2) / sigma1;
      power = 1 - PROBNORM(ua - shift);
    END;
  END;
RUN;
```

8.4 Beispiel zur Meta-Regression mit SAS

Das Asthma-Beispiel aus Abschnitt 1.2 auf Seite 6 soll für eine Meta-Regression (REM) verwendet werden. Die Ergebnisse der Meta-Regression sind in Tabelle 9 auf Seite 69 dargestellt. Der SAS Code zur Auswertung lehnt sich an die Arbeiten von WHITEHEAD [109] und VAN HOUWELINGEN [98] an. Im ersten Schritt werden die Originaldaten aus der Publikation in den Datensatz `meta` eingelesen.

```
DATA meta;
  INPUT trial$ drug n_itt n_pp lcl_itt ucl_itt lcl_pp ucl_pp;
  DATALINES;
  A 1 346 282  -6  7  -8  6
  B 1 409 262 -13  0 -14  1
  C 1 392 302  -3 11  -4 12
  D 2 399 288  -7  4  -9  4
  E 2 366 281  -0 10  -3  8
  F 2 421 290  -7  2  -8  3
  G 2 412 311  -5  7  -5  8
  H 2 370 325  -8  3  -7  5
  I 2 379 298  -8  3  -1 10
  J 2 346 234  -7  2  -8  3
  K 2 212 106  -3  6  -3  9
  ;
RUN;
```

Im zweiten Schritt sind einige Variablen zu generieren: `study` beinhaltet die Studienidentifikation, `mean` den Effektschätzer $\hat{\theta}_i$, `value` die dazugehörige Varianz $Var(\hat{\theta}_i) = w_i^{-1}$ und `x_1i` die Kovariable x_{1i}. Die Variablen `n`, `theta_i`, `var_i`, `lower` und `upper` beziehen sich auf eine zufällige Auswahl von PP- oder ITT-Analyse je Studie. Die Variablen `row`, `col` und `value` sind notwendig, um später im `RANDOM`-Statement die als fest angenommene Varianz definieren zu können.

```
DATA meta;
  SET meta;
  mean_itt =  (ucl_itt + lcl_itt) / 2;
  mean_pp  =  (ucl_pp  + lcl_pp ) / 2;
  var_itt  = ((ucl_itt - lcl_itt) / 4) ** 2;
  var_pp   = ((ucl_pp  - lcl_pp ) / 4) ** 2;
```

```
  x_1i = (RANUNI(12345) < 0.5);
  pop  = x_1i - 0.5;
  IF x_1i THEN DO;
    n       = n_itt;
    theta_i = mean_itt;
    var_i   = var_itt;
    lower   = lcl_itt;
    upper   = ucl_itt;
  END;
  ELSE DO;
    n       = n_pp;
    theta_i = mean_pp;
    var_i   = var_pp;
    lower   = lcl_pp;
    upper   = ucl_pp;
  END;
  study    = _n_;
  row      = _n_;
  col      = _n_;
  value    = var_i;
  estimate = var_i;
  int      = 1;
RUN;
```

Die Meta-Regression (REM) kann mit `PROC MIXED` durchgeführt werden, somit erhält man die Schätzer $\hat{\theta}, \hat{\beta}_1$ und $\hat{\tau}^2$ (Tabelle 9 auf Seite 69). Vorher muss ein Datensatz erstellt werden, der die Studienvarianzen in der Variablen `estimate` und außerdem für die Zwischenstudienvarianz τ^2 einen Startwert enthält.

```
DATA covvars;
  estimate = 1;
RUN;

DATA covvars (KEEP = estimate);
  SET covvars meta;
RUN;
```

```
PROC MIXED DATA = meta METHOD = ML ORDER = DATA;
  CLASS study x_1i;
  MODEL theta_i = int x_1i / NOINT SOLUTION CL DDF = 10000, 10000;
  *MODEL theta_i = x_1i / SOLUTION CL;
  RANDOM int / G SUBJECT = study;
  REPEATED / GROUP = study;
  PARMS / PARMSDATA = covvars EQCONS = 2 to 12;
  LSMEANS x_1i / CL;
RUN;
```

Im `MODEL`-Statement wird der Effektschätzer durch die festen Effekte, Intercept und Kovariable, modelliert. Die Option `SOLUTION` liefert die Parameter für die festen Effekte, hier also den intercept ($\hat{\theta}$) und die Kovariable ($\hat{\beta}_1$). Bei der 0-1-Kodierung ist $\hat{\theta}$ ein Schätzer für die PP-Analyse und $\hat{\beta}_1$ ein Schätzer für den Unterschied ITT-PP. Mit dem `LSMEANS`-Statement werden Effektschätzer für die PP- *und* ITT-Analyse ausgegeben.

Die Modellierung des zufälligen Effektes `study` kann auf zwei Arten durchgeführt werden. Entweder

$$\hat{\theta}_i = \theta + \beta_1 x_{1i} + \nu_i + \varepsilon_i \sim N(\theta + \beta_1 x_{1i}, w_i^{-1} + \tau^2)$$

mit $\nu_i \sim N(0, \tau^2)$ und $\varepsilon_i \sim N(0, w_i^{-1})$. In obigem `PROC MIXED` wird ν_i in dem `RANDOM`-Statement mit einer Konstanten (`int`) und ε_i mit dem `REPEATED`-Statement modelliert. Das Setzen der Varianzkomponenten w_i^{-1} aus den Einzelstudien wird im `PARMS`-Statement durchgeführt (die erste Beobachtung in covvars ist der Startwert für die Schätzung von τ, die übrigen Varianzkomponenten werden als fest angenommen: `EQCONS = 2 to 12`). Es ist ebenso möglich ν_i und ε_i zu vertauschen, indem

$$\hat{\theta}_i = \theta + \beta_1 x_{1i} + \tilde{\nu}_i + \tilde{\varepsilon}_i \sim N(\theta + \beta_1 x_{1i}, w_i^{-1} + \tau^2)$$

mit $\tilde{\nu}_i \sim N(0, w_i^{-1})$ und $\tilde{\varepsilon}_i \sim N(0, \tau^2)$ modelliert wird. In `PROC MIXED` wird $\tilde{\nu}_i$ in dem `RANDOM`-Statement mit der Variablen `study` modelliert. $\tilde{\varepsilon}_i$ wird als normale Residualvarianz geschätzt und braucht daher nicht speziell benannt zu werden. Das Setzen der Varianzkomponenten w_i^{-1} aus den Einzelstudien wird im `RANDOM`-Statement mit der Option `GDATA` durchgeführt, dabei werden zur Zuordnung der Varianzkomponenten die Variablen `row`, `col` und `value` automatisch verwendet. Alle nicht spezifizierten Matrixeinträge werden auf Null gesetzt, so dass sich die erwünschte Diagonalmatrix $diag(w_i^{-1})$ ergibt.

```
PROC MIXED DATA = meta METHOD = ML ORDER = DATA;
  CLASS study;
  MODEL theta_i = int x_1i / NOINT SOLUTION CL DDF = 10000, 10000;
  *MODEL theta_i = int pop / NOINT SOLUTION CL DDF = 10000, 10000;
  *MODEL theta_i = x_1i / SOLUTION CL;
  RANDOM study / G GDATA = meta;
RUN;
```

Die auskommentierten `MODEL`-Statements können verwendet werden, um statt den Quantilen der Normalverteilung die der t-Verteilung zu verwenden. Wird die Kovariable `pop` (Kodierung mit -0.5 und $+0.5$) statt `x_1i` eingesetzt, so wird mit dem Intercept ein mittlerer Schätzer für beide Populationen angegeben (0.46 [-1.24, 2.15]) (REM(ADJ) in Abbildung 12 auf Seite 65).

8.5 Beispiel zur bivariaten Analyse mit SAS

Der SAS Code lehnt sich an die Arbeit von VAN HOUWELINGEN [98] an und die Ergebnisse sind in der Tabelle 9 auf Seite 69 dargestellt.

```
DATA meta2(KEEP = itt pp theta_i estimate study trial n);
  SET meta;
  itt = 0; pp = 1; theta_i = mean_pp ; estimate = var_pp ; n = n_pp ;
  OUTPUT;
  itt = 1; pp = 0; theta_i = mean_itt; estimate = var_itt; n = n_itt;
  OUTPUT;
RUN;

DATA meta2;
  SET meta2;
  arm=_n_;
RUN;
```

Der Datensatz `meta` enthält $k = 11$ Beobachtungen, `meta2` aber $2k = 22$ Beobachtungen. Die Variablen `itt` bzw. `pp` sind 1 wenn die entsprechende Beobachtung der Variablen `mean` die Ergebnisse der ITT- bzw. der PP-Analyse enhält; sonst sind die Werte von `itt` bzw. `pp` 0. Die Variable `arm` ist ein Zähler $1, 2, 3, \ldots, 2k$, für jede Zahl liegt ein anderes Analyseergebnis vor. Außerdem enthält die Variable `study` die Studiennummer: $1, 1, 2, 2, \ldots, k, k$. Zum Setzen der Varianzen w_i^{-1} aus den Einzelstudien muss wieder ein Datensatz wie schon in Abschnitt 8.4 bereitgestellt werden, der jetzt allerdings drei Startwerte enthalten muss, da noch 3 Varianzkomponenten $(\tau_0^2, \tau_1^2, \sigma_{01})$ zu schätzen sind.

```
DATA covvars2;
  estimate = 1; OUTPUT; OUTPUT; OUTPUT;
RUN;

DATA covvars2 (keep = estimate);
  SET covvars2 meta2;
RUN;
```

```
PROC MIXED DATA = meta2 METHOD = ML ORDER = DATA;
  CLASS study arm;
  MODEL theta_i = itt pp / NOINT SOLUTION CL  DDF = 10000, 10000;
  *MODEL theta_i = itt pp / NOINT SOLUTION CL;
  RANDOM itt pp / SUBJECT = study TYPE = UN;
  REPEATED / GROUP = arm;
  ESTIMATE 'PopulationsUnterschied' itt 1 pp -1 / CL DF = 10000;
  *ESTIMATE 'PopulationsUnterschied' itt 1 pp -1 / CL;
  PARMS / PARMSDATA = covvars2 EQCONS = 4 to 25;
run;
```

Das `MODEL`- und `RANDOM`-Statement definieren obiges Modell (12). Mit den Optionen des `RANDOM`-Statements (`SUBJECT = study`) werden die unabhängigen Wiederholungen festgelegt, sowie die Struktur der entstehenden 2x2-Kovarianzmatrix: hier soll eine unspezifizierte Matrix verwendet werden, also mit drei Parametern. Das `REPEATED`-Statement beschreibt nur die Kovarianzstruktur der Messfehler, hier eine Diagonalmatrix. Daher wird für jeden `arm`, also jede Analyse, ein eigener Parameter festgelegt. Diese $2k$ Parameter $(w_{01}^{-1}, w_{11}^{-1}, \ldots, w_{0k}^{-1}, w_{1k}^{-1})$ werden später durch das `PARMS`-Statement mit den Varianzschätzern aus den Einzelstudien festgelegt. Mit dem `ESTIMATE`-Statement kann der Unterschied zwischen der ITT- und der PP-Analyse geschätzt werden.

Persönliche Daten

Name	Steffen Witte, geb. Ballerstedt
Geburtsdatum	1. März 1970
Geburtsort	Hameln
Familienstand	verheiratet

Ausbildung

1976 – 1989	Schulen in Hameln
8.5.1989	allgemeine Hochschulreife
1989 – 1991	Berufsausbildung zum Bankkaufmann beim BHW in Hameln
1991 – 1997	Studium der Mathematik, Universität Göttingen
1992 – 1997	Studium der Betriebswirtschaftslehre, Universität Göttingen
13.10.1993	Vordiplom Mathematik
14.7.1994	Vordiplom Betriebswirtschaftslehre
27.6.1997	Diplom Mathematik
27.6.1997	Diplom Mathematik Thema der Diplomarbeit bei Prof. E. Brunner: „Fehlende Werte in nichtparametrischen gemischten Modellen“
25.11.2004	Promotionsprüfung

Berufliche Tätigkeiten

4/1995 – 5/1997	Studentische Hilfskraft an der Abteilung Medizinische Statistik der Georg-August-Universität Göttingen
10/1997 – 1/1998	wissenschaftlicher Mitarbeiter an der Abteilung Medizinische Statistik der Georg-August-Universität Göttingen
3/1998 – 2/2000	Programmierer und Biometriker bei Pharmaceutical Research Associates GmbH, Mannheim
seit 3/2000	wissenschaftlicher Mitarbeiter am Institut für Medizinische Biometrie und Informatik, Abteilung Medizinische Biometrie der Ruprecht-Karls-Universität Heidelberg

Danksagung

Für die Unterstützung möchte ich mich besonders bei Herrn Prof. Dr. Norbert Victor bedanken. Er stand für mich als Betreuer und Doktorvater immer gerne zur Verfügung. Unsere gemeinsamen, regelmäßigen Treffen mit anregenden Diskussionen haben mich sehr voran gebracht. Weiterer Dank gilt meiner Frau, Dr. Felicitas Witte, die mich unter anderem bestärkte, diese Arbeit zu schreiben und das Dokument redigierte. Außerdem möchte ich mich bei allen Kolleginnen und Kollegen der Abteilung für Medizinische Biometrie der Universität Heidelberg herzlich bedanken.