Äquivalenzstudien und Nicht-Unterlegenheitsstudien

– Artikel Nr. 20 der Statistik-Serie in der DMW –

Equivalence and non-inferiority trials

Autoren S. Lange [1] R. Bender [1] A. Ziegler [2]

Institut
[1] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln
[2] Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck

Einführung

Zum Nachweis der Wirksamkeit neuer Therapiemaßnahmen wird heutzutage in aller Regel die Durchführung kontrollierter, randomisierter klinischer Studien (randomized controlled trials, RCTs) gefordert. Hierfür hat sich, insbesondere im Zusammenhang mit der Zulassung von Arzneimitteln, in der zweiten Hälfte des letzten Jahrhunderts eine weltweit akzeptierte, biometrische Methodik etabliert, die in den 1990er Jahren in entsprechenden Guidelines ihren Niederschlag gefunden hat [4, 12].

Üblicherweise ist es im Rahmen solcher Studien das Ziel, **Unterschiede** zwischen zwei (oder mehreren) Gruppen zu demonstrieren, das heißt zum Beispiel, die **Überlegenheit** einer neuen Therapie gegenüber Placebo (bzw. gar keiner Behandlung) oder einer anerkannten Standardtherapie im Hinblick auf ein vorab definiertes Wirksamkeitskriterium zu zeigen. Diese Studien werden als Überlegenheitsstudien bezeichnet. In den letzten Jahren mehren sich jedoch Fragestellungen, bei denen nicht mehr die Überlegenheit, sondern die **Gleichwertigkeit** zu prüfen ist.

Studien, die das letztgenannte Ziel haben, werden Studien zur **therapeutischen Äquivalenz** genannt [7, 24], im angloamerikanischen Schrifttum auch Active Control Studies oder Equivalence Studies [18, 19]. Der Begriff Äquivalenz ist allerdings nicht eindeutig definiert, da er auch für den Bereich der Bioäquivalenzprüfungen verwendet wird (und auch schon frühzeitiger verwendet wurde), bei dem es um den Vergleich der Bioverfügbarkeit von wirkstoffidentischen Formulierungen mit einer unterschiedlichen galenischen Zubereitung geht. Der Nachweis der Bioäquivalenz wird dabei allein als hinreichend für therapeutische Äquivalenz angesehen.

Unter dem Begriff Studien zur therapeutischen Äquivalenz werden zwei Ansätze voneinander unterschieden: sog. „Nicht-Unterlegenheit“ und „echte“ Äquivalenz. Dabei heißt Nicht-Unterlegenheit, dass die Prüfbehandlung in einer Richtung **höchstens irrelevant schlechter** als die Standardbehandlung ist; das bedeutet beispielsweise, dass unter einer neuen Therapie höchstens irrelevant mehr unerwünschte Ereignisse auftreten als unter der Standardtherapie (einseitige Fragestellung). Bei „echter“ Äquivalenz zeigt sich die höchstens irrelevante Unterlegenheit in beide Richtungen, z. B. weder eine zu schwache noch eine zu starke Blutdrucksenkung (zweiseitige Fragestellung). In Situationen, bei denen für eine bestimmte Fragestellung keine Standardtherapie allgemein definiert ist, kann es darüber hinaus von Interesse sein, zu prüfen, ob sich zwei Therapien nicht mehr als irrelevant voneinander unterscheiden. Auch hier handelt es sich dann um eine zweiseitige Fragestellung im Sinne von „echter“ Äquivalenz. Da diese drei Szenarien ähnliche Implikationen haben, werden Sie im Folgenden der Einfachheit halber unter dem Begriff der Äquivalenz(-studien) subsummiert, es sei denn, es werden spezifische Aspekte angesprochen [12, 17].

Für das Ziel, nicht mehr die Überlegenheit (bzw. allgemeiner: Unterschiedlichkeit), sondern die Äquivalenz von zwei (oder mehreren) Therapien zu prüfen, lassen sich vielfältige Begründungen liefern: Es gilt in vielen Bereichen als unethisch, im Rahmen von Therapieprüfungen eine Placebogruppe mitzuführen, wenn sich eine oder mehrere Behandlungen für das entsprechende Indikationsgebiet als wirksam erwiesen haben. In der zuletzt im Oktober 2000 revidierten Deklaration von Helsinki heißt es dazu unter Abschnitt 29: „The benefits, risks, burdens, and effectiveness of a new method should be tested against those of the best current prophylactic, diagnostic, and therapeutic methods.“ [25]. Neben dem ethischen Problem ist darüber hinaus der

Schlüsselwörter
- Äquivalenz
- Nicht-Unterlegenheit
- Klinisch relevante Differenz
- Konfidenzintervall

Key words
- Equivalence
- Non-inferiority
- Clinically relevant difference
- Confidence interval

Bibliografie
DOI 10.1055/s-2007-959043
Dtsch Med Wochenschr 2007; 132: e53–e56 · © Georg Thieme Verlag Stuttgart · New York · ISSN 0012-0472

Korrespondenz
Privatdozent Dr. rer. biol. hum. Ralf Bender
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln
eMail ralf.bender@iqwig.de

direkte Vergleich einer neuen Behandlung mit einer existierenden Standardtherapie Voraussetzung für differentialtherapeutische Überlegungen und somit häufig relevanter als der alleinige Vergleich mit Placebo oder keiner Therapie [11].

Natürlich kann es auch in einer kontrollierten klinischen Studie mit zwei wirksamen Therapien das Ziel sein, die Überlegenheit der einen über die andere Behandlung zu zeigen. Dann gibt es aus statistischer und methodischer Sicht keinen prinzipiellen Unterschied zu einer placebokontrollierten Studie; nur werden im Allgemeinen deutlich höhere Fallzahlen vonnöten sein, da mit kleineren Unterschieden zu rechnen ist. Diese Fallzahlen können mitunter enorme Ausmaße annehmen, beispielsweise in der Kardiologie, wo so genannte Mega-Trials, z. B. Studien zur Erprobung neuer Thrombolytika, teilweise mehrere zig-tausend Patienten umfassen. Diese Studien stoßen an finanzielle und logistische Grenzen, und deren Erfordernisse behindern möglicherweise sogar die Entwicklung innovativer Therapien und somit den therapeutischen Fortschritt [9]. Einen Ausweg aus diesem Dilemma könnten dann ggf. Studien bieten, die das ehrgeizige Ziel des Überlegenheitsnachweises aufgeben – zugunsten der vergleichbaren Wirksamkeit [8].

Abgesehen von derartigen Überlegungen, wie Studien bei begrenzten Ressourcen und gleichzeitig nur geringfügig besseren Innovationen Ziel führend durchgeführt werden können, wird beim Vergleich von zwei oder mehr aktiven Therapien möglicherweise auch gar kein oder nur ein vernachlässigbar kleiner Unterschied im Hinblick auf das Wirksamkeitskriterium zu erwarten sein. Als therapeutischer Fortschritt wird in solchen Situationen dann ein günstigeres Nebenwirkungsprofil, eine einfachere Handhabung oder ein geringerer Kostenaufwand postuliert. Auch kann es in bestimmten Fällen wünschenswert sein, mehrere gleichwertige Therapieoptionen zur Verfügung zu haben, falls es z. B. zu Resistenzentwicklungen oder allergischen oder anderen Unverträglichkeitsreaktionen kommt. Schließlich kann es bei der Betrachtung von Nebenwirkungen bzw. unerwünschten Ereignissen von Interesse sein zu zeigen, dass keine bzw. nur unwesentliche Unterschiede zum Beispiel in der Häufigkeit des Auftretens dieser Nebenwirkungen bzw. unerwünschten Ereignisse existieren.

Statistische Methodik bei Überlegenheit

Aus statistischer Sicht wird für den Nachweis der **Unterschiedlichkeit** anhand geeigneter Parameter eine **Nullhypothese** der Gleichheit der Gruppen formuliert, in dem Bestreben, diese Nullhypothese aufgrund der beobachteten Daten abzulehnen. Ziel ist es somit, die **Alternative** – das Gegenteil, nämlich die Unterschiedlichkeit – annehmen zu können [16]. Das Signifikanzniveau α (üblicherweise 5%) begrenzt hierbei die Wahrscheinlichkeit, den sog. **Fehler 1. Art** zu begehen, nämlich einen Unterschied anzunehmen, obwohl in **Wirklichkeit** gar kein Unterschied vorliegt [2]. Die andere Fehlermöglichkeit, das ist der **Fehler 2. Art,** das heißt, Gleichheit anzunehmen, obwohl sich die Gruppen in Wirklichkeit voneinander unterscheiden, wird durch die Irrtumswahrscheinlichkeit β begrenzt. Dabei gibt es unendlich viele Alternativen zur Nullhypothese, weil der Wirksamkeitsunterschied in der Regel einen beliebigen unbekannten Wert (im jeweils gültigen Wertebereich) annehmen kann. Daher kann der Fehler 2. Art nicht für alle tatsächlichen Wirksamkeitsunterschiede auf gleichem Niveau eingehalten werden, und es kann somit keine Entscheidung ***für*** die Nullhypothese mit vorgegebener Irrtumsmöglichkeit getroffen werden [17]. Auf der Basis nicht-signifikanter Ergebnisse darf demzufolge nicht auf Gleichheit geschlossen werden, weil die Möglichkeit dieser Fehlentscheidung nicht kontrolliert werden kann. Entsprechend kann die oben beschriebene klassische Hypothesenformulierung für die Untersuchung von Äquivalenzfragestellungen nicht herangezogen werden. Dieses Problem wurde in der Vergangenheit sehr prägnant umschrieben: „Absence of evidence is not evidence of absence“ [1, 10].

Statistische Methodik bei Äquivalenz

Für Äquivalenzfragestellungen ist es nun notwendig, einen Bereich gerade noch akzeptabler Ungleichheit zu definieren und für die statistische Hypothesenformulierung zu verwenden [15]. Dieser Bereich wird **Äquivalenzbereich** oder **Irrelevanzbereich** genannt [22]. Die Grenze, die den Äquivalenzbereich definiert, wird häufig mit dem griechischen Symbol Δ bezeichnet. Abweichungen vom Äquivalenzbereich können grundsätzlich in beide Richtungen auftreten, im Beispiel der Therapieprüfungen zuungunsten der neuen Therapie oder der Standardbehandlung. Allerdings ist klinisch zumeist nur der Ausschluss einer relevanten Unterlegenheit der neuen gegenüber der Standardbehandlung (in einer Richtung) bedeutsam, also die Nicht-Unterlegenheit. Zur praktischen Lösung des Nachweises von Äquivalenz wurde eine Vielzahl spezieller statistischer Verfahren entwickelt, die z. B. in Wellek [23] beschrieben werden. Dabei lässt sich dieses Problem entweder anhand eines geeigneten Signifikanztests oder über die Konstruktion von Konfidenzintervallen lösen. Letzteres wird in den meisten Empfehlungen und Richtlinien auf diesem Gebiet favorisiert [12] und auch im nachfolgenden Beispiel zur Illustration verwendet.

Wie bei Studien, die zum Ziel haben, einen Unterschied zu demonstrieren, wird allgemein auch bei echten Äquivalenzstudien sowie Nicht-Unterlegenheitsstudien empfohlen, ein ***2-seitiges*** Konfidenzintervall zur Sicherheit 1-α, in der Regel also 95%, als Entscheidungsgrundlage zu verwenden, obwohl eigentlich eine ***1-seitige*** Fragestellung vorliegt. De facto führt dieses in den meisten Studien zu einer Halbierung der Irrtumswahrscheinlichkeit für diese 1-seitige Fragestellung.

Beispiel

Als Illustration dient eine Nicht-Unterlegenheitsstudie zum Vergleich des zum damaligen Zeitpunkt neuen Thrombolytikums Reteplase mit der Standardtherapie in diesem Bereich (Streptokinase) bei Patienten mit akutem Myokardinfarkt (INJECT-Studie [14]). Hauptzielkriterium in großen Thrombolysestudien ist häufig die 30- bzw. 35-Tage-Mortalität. Aus zwei Gründen wurde eine Nicht-Unterlegenheitsstudie durchgeführt: Zum einen wurde angenommen, dass Reteplase sogar einen leichten Vorteil gegenüber Streptokinase im Hinblick auf die 35-Tage-Mortalität bieten würde, dass allerdings für einen statistischen Nachweis dieses Vorteils ein Mega-Trial erforderlich wäre. Zum anderen bietet Reteplase den Vorteil einer einfacheren Applikation, da sie als Doppel-Bolus gegeben werden kann und keine Infusion notwendig ist.

In der INJECT-Studie wurde ein Unterschied in der 35-Tage-Mortalität von einem Prozentpunkt als noch irrelevant angesehen (Δ = 1%). Das bedeutet, dass Reteplase der Streptokinase gegenüber dann als gleichwertig – bzw. präziser: als höchstens irrelevant unterlegen – anzusehen ist, wenn statistisch nachgewiesen werden kann, dass die 35-Tage-Mortalität unter Reteplase weni-

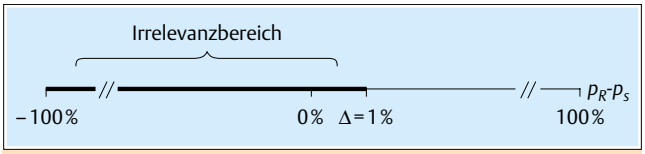

Abb. **1** Irrelevanzbereich in der INJECT-Studie [14] mit p_R und p_S als der 35-Tage-Mortalität unter Reteplase bzw. Streptokinase und Δ als der Grenze des Irrelevanzbereichs.

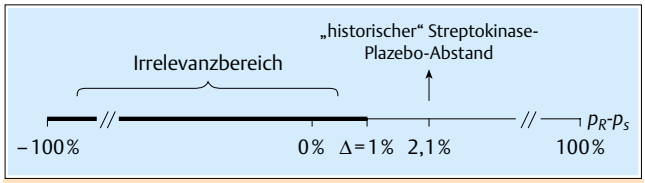

Abb. **2** Irrelevanzbereich in der INJECT-Studie [14] mit p_R und p_S als der 35-Tage-Mortalität unter Reteplase bzw. Streptokinase und Δ als der Grenze des Irrelevanzbereichs. Zusätzlich ist noch der „historische" Streptokinase-Placebo-Abstand eingezeichnet, der einer Meta-Analyse entstammt (untere Grenze des 90%-Konfidenzintervalls).

ger als ein Prozentpunkt höher ist als unter Streptokinase. Bezeichnet man nun die 35-Tage-Mortalität unter Reteplase im Folgenden mit $\boldsymbol{p}_R$ und die entsprechende Mortalität unter Streptokinase mit $\boldsymbol{p}_S$, dann lautet die Nullhypothese H_0 formell

$$H_0: \boldsymbol{p}_R - \boldsymbol{p}_S \geq 1\%$$

mit dem Ziel, diese ablehnen und die Alternativhypothese

$$H_1: \boldsymbol{p}_R - \boldsymbol{p}_S < 1\%$$

annehmen zu können. In **Abb. 1** ist das Nicht-Unterlegenheits-Problem mit dem Irrelevanzbereich noch einmal grafisch dargestellt.

In die INJECT-Studie wurden 3004 Patienten in die Reteplase- und 3006 Patienten in die Streptokinase-Gruppe eingeschlossen. Mortalitätsdaten lagen für 2993 bzw. 2992 Patienten vor, die die Auswertungspopulation bildeten. Innerhalb von 35 Tagen verstarben unter Reteplase 270 Patienten, das sind 9,02%, unter Streptokinase 285 Patienten, entsprechend 9,53%. Der Unterschied in den Mortalitätsraten in der Studie ergibt sich daher zu -0,51%, also einem tendenziellen Vorteil zugunsten der Reteplase. Das zu der Studie gehörige 95%-Konfidenzintervall ist $KI_{0,95}$ = (-1,98%, 0,96%). Das bedeutet, dass mit 95%iger Sicherheit ein knapp 2%iger Vorteil in der 35-Tage-Mortalität zugunsten von Reteplase gegenüber Streptokinase möglich ist. Es ist allerdings auch eine 0,96%ige Erhöhung der 35-Tage-Mortalität unter Reteplase gegenüber Streptokinase möglich. Ausgeschlossen ist im 95%-Konfidenzintervall hingegen eine 1%ige Erhöhung der 35-Tage-Mortalitätsrate unter Reteplase. Daher kann die Nullhypothese einer mehr als irrelevanten Unterlegenheit von Reteplase gegenüber Streptokinase verworfen und die Alternativhypothese einer höchstens irrelevanten Unterlegenheit akzeptiert werden. Damit war das Studienziel also erreicht.

Probleme

Äquivalenz- bzw. Nicht-Unterlegenheitsstudien bergen gewichtige methodische Probleme, die deren Einsetzbarkeit zum Teil stark einschränken und die Interpretation erschweren. Eines dieser Probleme betrifft die so genannte **Assay Sensitivity.** Darunter ist die Fähigkeit einer klinischen Studie zu verstehen, zwischen wirksamen (effective), weniger wirksamen (less effective) oder unwirksamen (ineffective) Therapien differenzieren zu können [13, 20]. Damit verknüpft ist die Frage nach der externen Validität der Studie: Wenn in einer klinischen Studie die Vergleichbarkeit zweier Therapien (anhand angemessener statistischer Verfahren) demonstriert wurde, so können die beiden Therapien vergleichbar **wirksam**, aber auch vergleichbar **unwirksam** sein (**Abb. 3**).

Ohne eine externe Evidenz für die Wirksamkeit der aktiven Kontrollbehandlung oder eine interne Kontrolle durch eine Placebogruppe kann zwischen diesen beiden Alternativen – vergleichbare Wirksamkeit oder Unwirksamkeit – nicht differenziert werden. Dieses markiert die große Schwäche von Äquivalenzstudien ohne Placebokontrolle: Die externe Validität kann in aller Regel – wenn überhaupt – nur aus historischen Studien abgeleitet werden. Es müssen demnach Studien vorliegen, die mit gleichem oder zumindest ähnlichem Design, d. h. vergleichbarer Studienpopulation, vergleichbaren Zielkriterien, vergleichbaren Begleitumständen usw. in der Vergangenheit regelhaft wirksame von weniger wirksamen und unwirksamen Therapien unterscheiden konnten. Falls diese fundamentale Annahme der historischen, externen Evidenz nicht getroffen werden kann, muss das Design einer Äquivalenzstudie ohne Placebokontrolle als fragwürdig angesehen werden [18, 19].

Ein weiteres Problem bei Äquivalenzstudien ohne Placebokontrolle ergibt sich schließlich aus dem Umstand, dass die Planungsinstrumente Randomisierung und Doppelblindheit nicht mehr hinreichend sicher vor Verzerrungen zugunsten des angestrebten Ziels schützen [21]. Es ist nämlich auch ohne Kenntnis der Therapie der einzelnen Patienten in der Prüfung grundsätzlich möglich, die Differenz der Behandlungsunterschiede zur Null und damit zur Alternativhypothese hin zu verschieben (z. B. im Extremfall dadurch, dass für alle Patienten der gleiche Wert eingetragen wird). Darüber hinaus besteht die Möglichkeit, in einer Studie Äquivalenz a priori dadurch zu begünstigen, dass größere Anteile von Patienten mit sehr schlechter oder sehr guter Prognose, im Extremfall nur Gesunde, in die Studie einbezogen werden.

Damit einhergehend können sich auch die in üblichen Überlegenheitsstudien eingesetzten Maßnahmen zur Umsetzung des Intention-to-treat(ITT-) Prinzips, d. h. der Auswertung aller randomisierten Patienten und gemäß dem Randomisierungscode ungünstig in dem Sinne auswirken, dass sie in der Regel zu einer Nivellierung von Unterschieden führen [11]. Damit kann dann das eigentliche Ziel, nämlich die Kontrolle der Wahrscheinlichkeit für den Fehler 1. Art nicht erreicht werden. Deshalb wird vielfach empfohlen, bei Äquivalenzstudien neben einer ITT-Auswertung auch noch eine Auswertung nur der protokollgerecht behandelten und beobachteten Patienten vorzulegen. Es werden dann also die nicht protokollgerecht behandelten und beobachteten Patienten von der Auswertung ausgeschlossen, und es wird eine so genannte **per-Protokoll-Analyse** durchgeführt, weil in dieser eher mit einer Vergrößerung von Unterschieden zu rechnen ist. Außerdem führt die Verringerung der Fallzahl zu breiteren Konfidenzintervallen, also konservativeren Ergebnissen (die Ablehnung der Nullhypothese wird dann in der Regel schwieriger). Nur falls beide Analysen zu einer Ableh-

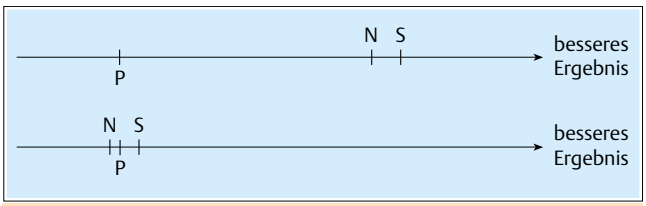

Abb. **3** Vergleichbare Wirksamkeit (oben) und Unwirksamkeit (unten) mit P für Placebo, N für neue Therapie und S für Standardtherapie.

nung der Nullhypothese führen, ist der Äquivalenz- bzw. Nicht-Unterlegenheitsnachweis erbracht [6]. Dieses Vorgehen ist allerdings auch nicht unproblematisch, weil durch die unklaren Selektionsmechanismen, die beim Ausschluss von Patienten in einer per-Protokoll-Analyse wirksam werden können, auch die Richtung einer eventuellen Verzerrung unklar bleibt.

Schließlich bleibt noch als sowohl klinisches als auch statistisches Problem bei der Planung von Äquivalenz- bzw. Nicht-Unterlegenheitsstudien die a priori Festlegung des Irrelevanzbereichs. Ein zu breit gewählter Irrelevanzbereich birgt die Gefahr, dass die zu prüfende neue Therapie tatsächlich im unwirksamen Bereich liegt. Beispielsweise wäre es im Rahmen der oben beschriebenen INJECT-Studie inadäquat gewesen, die Grenze des Irrelevanzbereichs bei drei Prozentpunkten anzusiedeln, da dies ziemlich genau dem Bereich entspricht bzw. ihn sogar ein wenig überschreitet, der in der Vergangenheit als (30 bzw. 35 Tage) Mortalitätsdifferenz zwischen der Standardtherapie Streptokinase und Placebo gefunden wurde. Idealerweise sollten daher bei der Planung von Äquivalenzstudien valide Vorinformationen über den „historischen Standard-Placebo-Abstand“ vorliegen und daran die Grenze des Irrelevanzbereichs orientiert werden [13]. Dabei muss allerdings neben dem Schätzer für diesen Abstand aus Studien auch die Variabilität, die dieser Schätzung innewohnt, berücksichtigt werden [5].

Die Planer der INJECT-Studie hatten zu diesem Zweck vor der Studie eine Meta-Analyse von Studien zum Vergleich von Streptokinase und Placebo durchgeführt. Dabei ermittelten Sie als Schätzer für die Mortalitätsdifferenz einen Wert von 2,7% mit einer unteren Grenze des zugehörigen 90%-Konfidenzintervalls von 2,1%. Die Grenze des Irrelevanzbereichs von 1% entspricht somit ziemlich genau der Hälfte dieser mindestens zu erwartenden Differenz zwischen Streptokinase und Placebo (**Abb. 2**). Fehlen solche Vorinformationen und ist dennoch das Mitführen einer Placebo- oder anderweitig inaktiven Kontrolle aus ethischen Gründen nicht vertretbar, muss eine substanzielle klinische Begründung für die Wahl des Irrelevanzbereichs erfolgen.

kurzgefasst

Aufgrund des nicht signifikanten Ergebnisses eines statistischen Tests auf Überlegenheit kann nicht auf Gleichwertigkeit geschlossen werden („absence of evidence is not evidence of absence“). Deshalb werden Äquivalenz- bzw. Nicht-Unterlegenheitsstudien konzipiert und durchgeführt. Sie haben nicht das Ziel, die Überlegenheit, sondern die Gleichwertigkeit bzw. die Nicht-Unterlegenheit einer neuen gegenüber einer anerkannten Standardtherapie zu prüfen. Solche Studien bedürfen einer besonderen Methodik, da das übliche methodische Instrumentarium nicht mehr hinreichend sicher vor falschen Schlussfolgerungen schützt.

Literatur

1 *Altman DG, Bland JM.* Absence of evidence is not evidence of absence. BMJ 1995; 311: 485
2 *Bender R, Lange S.* Was ist der *p*-Wert? Dtsch Med Wochenschr 2007; 132: e15–e16
3 *Blackwelder WC.* „Proving the null hypothesis“ in clinical trials. Control Clin Trials 1982; 3: 345–353
4 *CPMP.* Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. CPMP Working Party on Efficacy of Medicinal Products Note for Guidance III/3630/92-EN. Stat Med 1995; 14: 1659–1682
5 *CPMP.* Guideline on the choice of the non-inferiority margin. CPMP/EWP/2158/99 2005. http://www.emea.eu.int/pdfs/human/ewp/215899en.pdf (letzter Zugriff am 25.9.2005), 2005
6 *CPMP.* Points to consider on switching between superiority and non-inferiority. CPMP/EWP/482/99 2000. http://www.emea.eu.int/pdfs/human/ewp/048299en.pdf (letzter Zugriff am 25.9.2005), 2005
7 *Garbe E, Röhmel J, Gundert-Remy U.* Clinical and statistical issues in therapeutic equivalence trials. Eur J Clin Pharmacol 1993; 45: 1–7
8 *Hampton JR.* Mega-trials and equivalence trials: experience from the INJECT study. Eur Heart J 1996; 17 (Suppl E): 28–34
9 *Hampton JR.* Alternatives to mega-trials in cardiovascular disease. Cardiovasc Drugs Ther 1997; 10: 759–765
10 *Hartung J, Cottrell JE, Giffen JP.* Absence of evidence is not evidence of absence. Anesthesiology 1983; 58: 298–300
11 *Henry D, Hill S.* Comparing treatments: Comparison should be against active treatments rather than placebos. BMJ 1995; 310: 1279
12 *ICH Harmonised Tripartite Guideline.* Statistical principles for clinical trials (E9). Stat Med 1999; 18: 1905–1942
13 *ICH Harmonised Tripartite Guideline.* Choice of control group and related issues in clinical trials (E10). http://www.emea.eu.int/pdfs/human/ich/036496en.pdf (letzter Zugriff am 25.9.2005), 2000
14 *International Joint Efficacy Comparison of Thrombolytics.* Randomised, double-blind comparison of reteplase double-bolus administration with streptokinase in acute myocardial infarction (INJECT): trial to investigate equivalence. Lancet 1995; 346: 329–335
15 *Jones B, Jarvis P, Lewis JA, Ebbutt AF.* Trials to assess equivalence: the importance of rigorous methods. BMJ 1996; 313: 36–39
16 *Lange S, Bender R.* Was ist ein Signifikanztest? Dtsch Med Wochenschr 2007; 132: e19–e21
17 *Lange S, Windeler J.* Das Konzept der therapeutischen Äquivalenz. Med Klin 1997; 92: 215–220
18 *Makuch R, Johnson M.* Issues in planning and interpreting active control equivalence studies. J Clin Epidemiol 1989; 42: 503–511
19 *Makuch R, Pledger G, Hall DB, Johnson MF, Herson J, Hsu J-P.* Active Control Equivalence Studies. In: Peace KE; Hrsg. Statistical issues in drug research and development. New York: Marcel Dekker, 1990: 225–262
20 *Modell W, Houde RW.* Factors influencing clinical evaluation of drugs. With special reference to the double-blind technique. JAMA 1958; 167: 2190–2199
21 *Senn SJ.* Inherent difficulties with active control equivalence studies. Stat Med 1993; 12: 2367–2375
22 *Wellek S.* Planung und statistische Auswertung von Äquivalenzstudien im Rahmen der klassischen Theorie des Hypothesentestens. In: Michaelis J, Hommel G, Wellek S; Hrsg. Europäische Perspektiven der Medizinischen Informatik, Biometrie und Epidemiologie. München: MMV, Medizin-Verlag 1993: 143–147
23 *Wellek S.* Testing Statistical Hypotheses of Equivalence. Boca Raton: Chapman & Hall/CRC, 2003
24 *Windeler J, Trampisch HJ.* Empfehlung zur Durchführung von Studien zur therapeutischen Äquivalenz. Inform Biom Epidem Med Biol 1995; 26: 350–355
25 *World Medical Association.* Declaration of Helsinki – Ethical principles for medical research involving human subjects. Revision of the 52nd WMA General Assembly, Edinburgh, Scotland, http://www.wma.net/e/policy/b3.htm (letzter Zugriff am 25.9.2005). 2000