S. Wellek[1], M. Blettner[1]

Klinische Studien zum Nachweis von Äquivalenz oder Nichtunterlegenheit – Teil 20 der Serie zur Bewertung wissenschaftlicher Publikationen*

Establishing equivalence or non-inferiority in clinical trials – part 20 of a series on evaluation of scientific publications

Hintergrund: Klinische Studien, die darauf abzielen, nachzuweisen, dass es zwischen zwei Behandlungsverfahren keine relevanten Unterschiede gibt, werden in zunehmender Anzahl durchgeführt. Für den praktizierenden Arzt vergeht kaum ein Tag, an dem er nicht zumindest indirekt von den Ergebnissen sogenannter Bio-Äquivalenzstudien Gebrauch macht. Ebenso wichtig sind aktiv kontrollierte klinische Studien, in denen die Wirksamkeit einer neuartigen Therapie durch den Nachweis der Nichtunterlegenheit gegenüber einer Standardtherapie belegt wird.
Methoden: Darstellung der Grundprinzipien und der statistischen Verfahren unter Bezugnahme auf die Originalliteratur; selektive Recherchen in der medizinischen Literatur.
Ergebnisse: Zunächst ist ein geeigneter Verteilungsparameter festzulegen, der ein sinnvolles Maß für die Unterschiedlichkeit der Behandlungswirkungen in der Grundgesamtheit darstellt. Der einfachste Ansatz für den statistischen Nachweis von Äquivalenz oder Nichtunterlegenheit beruht auf der Berechnung von Konfidenzgrenzen für diesen Parameter. Um die erforderlichen Patientenzahlen möglichst gering zu halten, empfiehlt sich auch beim Äquivalenz- und Nichtunterlegenheits-Nachweis der Einsatz von bezüglich der Trennschärfe optimierten statistischen Testverfahren.
Schlussfolgerungen: Daten aus Äquivalenz- und Nichtunterlegenheits-Studien bedürfen genauso der Signifikanzprüfung wie solche, die die Unterschiedlichkeit von Behandlungen belegen sollen. Beim Äquivalenznachweis ist es nicht zulässig, einen herkömmlichen zweiseitigen Test zu verwenden und aus einem negativen Ergebnis auf Äquivalenz zu schließen.
(Dtsch Zahnarztl Z 2014; 69: 36–42)

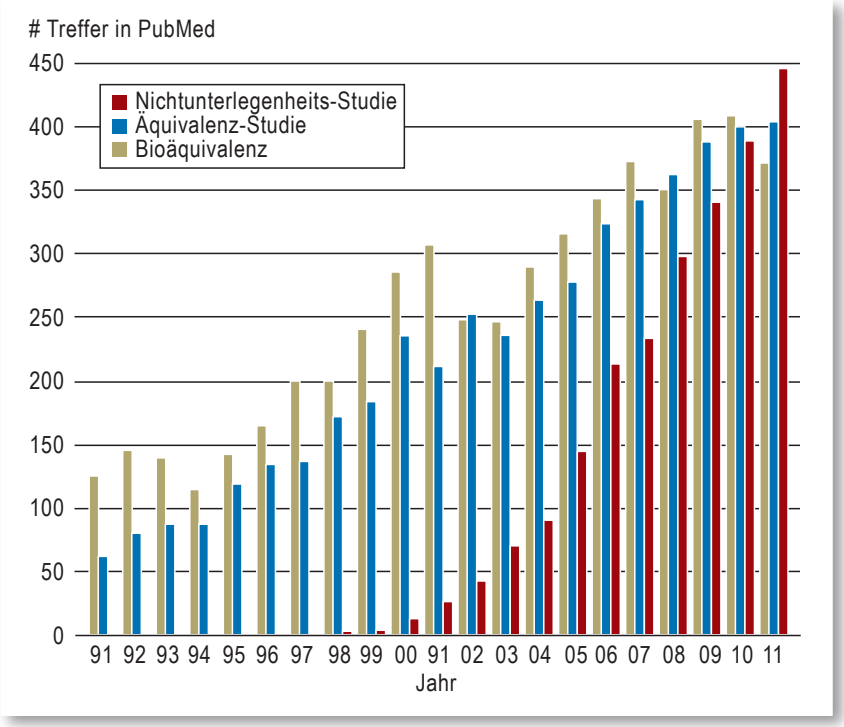
Background: An increasing number of clinical trials are being performed to show the absence of relevant differences between the effects of two treatments. The primary care physician makes use of the results of so-called equivalence studies, at least indirectly, practically every day. Equally important are active control clinical trials in which the efficacy of a new treatment has to be proven through demonstrating non-inferiority as compared to a standard treatment.
Methods: Explanation of basic principles and statistical techniques with reference to the original literature; selective searches in the medical literature.
Results: First of all, a suitable distributional parameter must be chosen that can be considered a reasonable measure of dissimilarity of the population effects of the treatments under comparison. The simplest approach to the statistical demonstration of equivalence or noninferiority is to calculate confidence intervals for that parameter. To keep the required number of subjects for equivalence and non-inferiority studies as low as possible, statistical tests should be used which are optimized with respect to power.
Conclusion: Data from equivalence and non-inferiority studies need to be assessed for statistical significance no less than data that are generated to show that two treatments have different effects. A negative result in a traditional two-sided test does not suffice for statistically proving equivalence.

* Nachdruck aus: Dtsch Arztebl Int 2012; 109(41): 674–679. DOI: 10.3238/arztebl.2012.0674 © Deutscher Ärzte-Verlag GmbH Köln
[1] Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI), Universitätsmedizin der Johannes Gutenberg-Universität
Peer reviewed article: eingereicht: 12.01.2012, revidierte Fassung angenommen: 04.07.2012
DOI: 10.3238/dzz.2014.0036–0042

Einleitung

Bei einer klassischen randomisierten kontrollierten klinischen Studie (RCT) besteht das Ziel darin, Unterschiede zwischen zwei Behandlungen zu evaluieren (oder zwischen einer Behandlung und einem Placebo) [9]. Es soll dann jeweils die Überlegenheit des neuen Behandlungsverfahrens gegenüber der Standardtherapie nachgewiesen werden. Bei Erkrankungen, für die bereits adäquate Therapien verfügbar sind, ergibt sich oft die Situation, dass ein neues Medikament entwickelt wurde, das zu geringeren Kosten erhältlich ist oder weniger Nebenwirkungen hat als existierende Präparate. In diesem Fall muss nachgewiesen werden, dass die Wirksamkeit des neuen Medikaments verglichen mit existierenden Substanzen „im wesentlichen gleich gut" (Äquivalenz) oder „nur unwesentlich schwächer" ist (Nichtunterlegenheit). Eine Fragestellung vom letzteren Typ wurde beispielsweise in der CATT-Studie (Lucentis versus Avastin, [14]) angegangen, die aufgrund der Häufigkeit des zu behandelnden Krankheitsbildes (altersbedingte Makuladegeneration) und der exorbitant hohen Kosten des als nichtunterlegen nachgewiesenen Medikaments (mindestens 1 Milliarde Euro jährlich bei flächendeckender Anwendung allein in Deutschland) auch in der Laienpresse beträchtliches Aufsehen erregt hat [10].

Eine Äquivalenzstudie ist dadurch gekennzeichnet, dass sie durchgeführt wird, um nachzuweisen, dass es zwischen zwei (oder auch mehreren) Behandlungen keine beziehungsweise keine wesentlichen Unterschiede hinsichtlich der Wirksamkeit gibt. Bei der Planung und der Bewertung solcher Studien ist daher zunächst zu definieren, was es heißt, dass zwei Therapien „gleich gut" sind, also welche Unterschiede als klinisch irrelevant toleriert werden können. Die klinisch relevanten Unterschiede sind im Studienprotokoll festzulegen. Dazu wird ein Parameter herangezogen, der diese Unterschiede charakterisiert. Dies kann zum Beispiel die Differenz oder der Quotient der Erwartungswerte der Zielvariablen sein. Außerdem wird eine untere und eine obere Grenze für die noch zu akzeptierende Abweichung von demjenigen Wert dieses Parameters festgesetzt, welcher bei identischer Wirksamkeit der Behandlungen vorliegt. Für die Werte



Grafik 1 Ergebnisse einer Literaturrecherche zur Häufigkeit von Äquivalenzsstudien.
Figure 1 Frequency of equivalence trials of a literature search.

dieser Äquivalenzgrenzen (englisch: equivalence margins) werden üblicherweise die Symbole $-\varepsilon_1$ und ε_2 verwendet, wobei ε_1 und ε_2 positive Zahlen sind. ε_1 und ε_2 werden unter Berücksichtigung der klinischen Fragestellung, des betrachteten klinischen Endpunkts und unter statistischen Aspekten (Form der zu beurteilenden Verteilungen) festgelegt. Handelt es sich zum Beispiel um eine Studie zum Nachweis der Äquivalenz zweier Antihypertensiva bezüglich der Reduktion des diastolischen Werts nach 4 Wochen Behandlungsdauer und wird die Differenz $\mu_1-\mu_2$ der mittleren Blutdrucksenkung in den Grundgesamtheiten als Zielparameter gewählt, ist $\varepsilon_1 = \varepsilon_2 = 5$ mmHg eine sinnvolle Festlegung der Äquivalenzgrenzen.

Beim Nachweis der Nichtunterlegenheit (englisch: noninferiority) soll gezeigt werden, dass die neue Therapie nicht wesentlich schlechter ist als die Referenzbehandlung. Was eine relevante Verschlechterung wäre, wird dabei festgelegt durch eine untere Schranke $-\varepsilon$ (im Falle der mittleren Blutdrucksenkung zum Beispiel −5,0), die der zur Messung des Behandlungsunterschiedes ausgewählte Parameter ungünstigstenfalls annehmen darf.

Die Bedeutung von Äquivalenz- und Nichtunterlegenheitsstudien für die klinische Forschung hat in den letzten 2 Jahrzehnten beständig zugenommen, wie sich unter anderem an den in Grafik 1 dargestellten Trefferzahlen in PubMed für die Schlüsselwörter „bioequivalence", „non(-)inferiority study (trial)" und „equivalence study (trial)" für die Jahrgänge 1991–2011 ablesen lässt. Als weiterer Indikator für diese Entwicklung kann der Anteil der auf der Basis von Äquivalenzstudien zur behördlichen Zulassung gelangten verschreibungspflichtigen Arzneimittel herangezogen werden. Nach einer in [17, § 1.4] anhand von Daten aus dem Arzneimittelreport der FDA (Food and Drug Administration der USA) vorgenommenen Hochrechnung belief sich dieser im Jahre 2008 auf nicht weniger als 78 % (Grafik 1).

Unzulässigkeit des „naiven" Ansatzes beim statistischen Testen auf Äquivalenz

Bei der Bewertung der Äquivalenz sind andere statistische Verfahren anzuwenden als in der klassischen Situati-

Kasten 1a

Durchführung des Intervallinklusions-Tests auf Äquivalenz von zwei Normalverteilungen bezüglich der Differenz der Mittelwerte

Studie: Wirksamkeitsvergleich zwischen einem neuartigen Anitdepressivum (A) und Imipramin (B) als Referenztherapie für eine Major-Depression
Zielvariable: prozentuale Reduktion des HAM-D-(Hamilton Depression Scale-)Werts nach sechswöchiger Behandlungsdauer.
Verteilungsannahme: Die Zielvariable ist unter beiden Behandlungen annähernd normalverteilt mit Mittelwerten μ_1 (←Gruppe A) und μ_2 (←Gruppe B) sowie unbekannter gemeinsamer Varianz σ^2

Auswertung: Test auf Äquivalenz dieser Verteilungen bezüglich der Mittelwerte, wobei die maximal tolerierbare Abweichung zwischen μ_1 und μ_2 sowohl nach links (←ε_1) als auch nach rechts (←ε_2) auf 5,0 [%] festgelegt wird.
Als Signifikanzniveau wird wie üblich $\alpha = 0{,}05$ gewählt.
Ergebnisse der Studie als Stichprobenmittelwerte und Standardabweichungen:

Gruppe A ($n_1 = 25$): $\overline{X} = 58{,}9$, $S_X = 5{,}82$
Gruppe B ($n_2 = 20$): $\overline{Y} = 57{,}5$, $S_Y = 4{,}94$

Konfidenzgrenzen für μ_1–μ_2 zum einseitigen Konfidenzniveau 95 %
Ausgehend von den empirischen Mittelwerten und Standardabweichungen errechnet sich die untere beziehungsweise obere Konfidenzschranke auf der Basis der zentralen t-Verteilung nach bekannten Formeln aus der elementaren Statistik [8] zu

$$C_u = -1{,}35 \text{ bzw. } C_o = 4{,}15$$

Testentscheidung:
Nach der Intervallinklusions-Regel ist zu überprüfen, ob *sowohl* $C_u > -5{,}0$
als auch $C_o < 5{,}0$ erfüllt ist.
Antwort: Da auf der Zahlenachse der Punkt –1,35 rechts von –5,0 und 4,15 links von +5,0 liegt, kann die Nullhypothese relevanter Unterschiede zwischen den Behandlungen A und B abgelehnt werden.
Also: Entscheidung zugunsten von Äquivalenz.

Alternative Darstellung der Entscheidungsregel:
Gegeben die im vorliegenden Beispiel erhaltenen Werte für die beiden Standardabweichungen führt der Äquivalenztest zu einer positiven Entscheidung, falls die beiden arithmetischen Mittel um nicht mehr als 2,25 % voneinander abweichen (Grafik 2).
Also: In den Stichproben müssen die Unterschiede noch wesentlich geringer sein, als es den unter der Hypothese zugelassenen Grenzen entspricht.

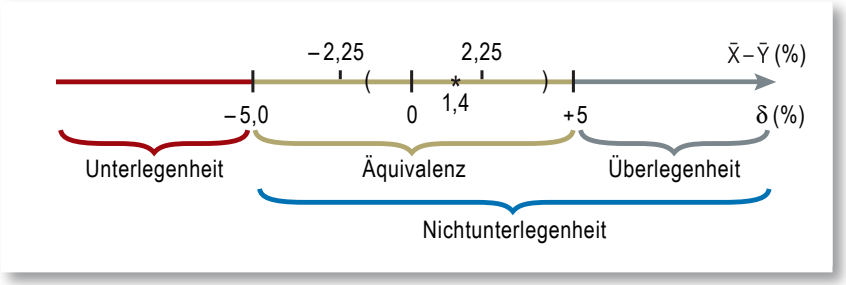
Kasten 1b

Was ändert sich, wenn in der Situation aus Kasten 1A anstatt auf Äquivalenz auf Nichtunterlegenheit getestet wird?

Hypothesenformulierung: Die Arbeits-(Alternativ-)Hypothese, die man anhand der Daten bestätigen möchte, lautet jetzt:
Der wahre Wert von μ_1 liegt oberhalb von μ_2–ε
(μ_1 [μ_2] = mittlere prozentuale Reduktion des HAM-D unter Antidepressivum A [B] in der Grundgesamtheit)
Äquivalenzgrenze: Im Nichtunterlegenheitsfall interessiert nur noch die linke Grenze –ε des Bereichs der klinisch irrelevanten Abweichungen zwischen μ_1 und μ_2.
Abweichend von den Spezifikationen in Kasten 1a soll diesmal angenommen werden, dass der Margin ε im Studienprotokoll auf 2,5 % festgesetzt wurde.
Testentscheidung: Die Entscheidung, ob Nichtunterlegenheit als statistisch gesichert angesehen werden kann oder nicht, richtet sich ausschließlich nach der unteren Konfidenzgrenze:
Der in Kasten 1a gefundene Wert von –1,35 liegt oberhalb der theoretischen Nichtunterlegenheitsgrenze von –2,5.
Also: Entscheidung zugunsten von Nichtunterlegenheit.

Hinweis: Das Beispiel zeigt, dass dieselben Daten gegebenenfalls unter dem Aspekt der Nichtunterlegenheit anders zu beurteilen sind als bei Überprüfung auf Äquivalenz. Mit auf 2,5 verringerter Toleranz ε würde der in Kasten 1a durchgeführte Test negativ ausfallen, weil die rechte Konfidenzgrenze oberhalb von +2,5 liegt.

on, in der die Überlegenheit gezeigt werden soll. Um auf Äquivalenz zu prüfen, führt ein herkömmlicher zweiseitiger Test [3] nicht weiter. Falsch ist nämlich, die Alternativhypothese der Äquivalenz der Behandlungen für gesichert zu erklären, wenn dieser Test ein negatives, das heißt nichtsignifikantes Ergebnis liefert. Der Fehler erster Art besteht hier ja darin, die Behandlungseffekte für ähnlich zu erklären, obwohl es relevante Unterschiede gibt. Wird also der herkömmliche Test durchgeführt, kann der Fehler erster Art bis zu 95 % betragen. Anders ausgedrückt: Nichtsignifikante Unterschiedlichkeit darf nicht mit signifikanter Übereinstimmung der Behandlungseffekte verwechselt werden. Eine unpräzisere, aber sehr häufig zitierte Formulierung für den gleichen Tatbestand lautet: „Absence of evidence is not evidence of absence“ [1].



Grafik 2 Visualisierung des Vorgehens in Kasten 1a: Werte unterhalb (oberhalb) der Zahlenachse beziehen sich auf den Therapieunterschied in der Grundgesamtheit (in den Stichproben); * = beobachteter Mittelwertsunterschied.

Figure 2 Visualization of the procedure in Box 1a: values above (below) the numerical axis relate to the treatment difference in the population (in the samples):; * = observed mean differenc.

Das Prinzip der Konfidenzintervall-Inklusion

Die konfirmatorische Auswertung von Äquivalenzstudien geschieht statistisch korrekt auf der Basis von Konfidenzintervallen. Die Grundidee hierzu ist bemerkenswert einfach und kam erstmals in Zusammenhang mit Bioäquivalenzprüfungen auf [18]:

Kasten 2

Test auf Nichtunterlegenheit bezüglich der Odds Ratio in Zweiarm-Studien mit dichotomer Response-Beurteilung

Ausgangssituation, Verteilungsannahme: Parallelgruppen-Design mit binären Daten (Response ja oder nein); die statistisch zu beurteilenden Parameter sind die Anteile p_1 (↔ Behandlung A) und p_2 (↔B) von Respondern in den zugehörigen Grundgesamtheiten

Nichtunterlegenheits-Hypothese: Der wahre Wert der Odds Ratio $OR = (p_1/(1-p_1))/(p_2/(1-p_2))$ liegt oberhalb von $1-\varepsilon$, mit ε als im Studienprotokoll vorgegebener Toleranz (zum Beispiel $\varepsilon = 1/3$ oder $\varepsilon = 1/2$).

Testprozedur: Verwendet als p-Wert [4] die Wahrscheinlichkeit $P_{s;\varepsilon}$ dafür, dass man in einer Situation mit denselben Stichprobenumfängen und der selben Gesamtzahl s von Behandlungserfolgen wie in der vorliegenden Studie sowie $1-\varepsilon$ als wahrem Wert der Odds-Ratio in Gruppe A mindestens so viele Responder erhält, wie tatsächlich beobachtet wurden.

Datenbeispiel: In der 2010 in Lancet publizierten Studie [5] zum Vergleich von Raltegravir (experimentelle Therapie) mit Lopinavir & Ritonavir (Positivkontrolle) bei der Behandlung von HIV-Infizierten mit stabiler viraler Suppression unter einer vorangegangenen Kombinationstherapie wurden folgende Responderhäufigkeiten beobachtet:

Medikament	Response		Σ
	+	–	
A (Raltegravir)	293 (84,4 %)	54 (15,6 %)	347 (100,0 %)
B (Lopinavir & Ritonavir)	319 (90,6 %)	33 (9,4 %)	352 (100,0 %)
Σ	612	87	699

Bei Festlegung des Noninferiority Margin ε auf 0,5 berechnet sich der p-Wert $P_{s;\varepsilon}$ unter Verwendung der SAS-Software (für Details siehe [17, § 6.6.1]) für diese Kontingenztafel zu 35,04 % und liegt somit weit oberhalb des üblichen Signifikanzniveaus von 5 %. Die Nichtunterlegenheit von Raltegravir gegenüber der Kombinationstherapie bezüglich der Odds-Ratio kann mit den vorliegenden Daten folglich nicht gesichert werden.

Kasten 3

Kriterien für die Beurteilung von einschlägigen Publikationen

(Q1) Nur Überprüfung auf „Absence of Evidence" oder Anwendung von Test auf Äquivalenz bzw. Nichtunterlegenheit?
(Q2) Äquivalenzgrenze(n) a priori (ohne Kenntnis der Daten) festgelegt?
(Q3) Nachvollziehbare Begründung für die Spezifikation der Äquivalenzgrenze(n)?
(Q4) Optimaler Test auf zweiseitige Äquivalenz oder Konfidenzintervall-Einschluss-Regel?
(Q5) Bei zweiseitiger Äquivalenzfragestellung und Anwendung des Intervallinklusionsprinzips: zweiseitiges Konfidenzniveau 90 % oder unnötig konservative Festlegung auf 95 %?

Man berechnet aus den zu analysierenden Daten eine untere Konfidenzgrenze C_u und eine obere Konfidenzgrenze C_o für den ausgewählten Parameter und vergleicht diese mit den vorgegebenen theoretischen Grenzen $-\varepsilon_1$ und ε_2. Falls das Intervall mit den Grenzen (C_u, C_o) vollständig in dem theoretischen Intervall enthalten ist, entscheidet man für die Äquivalenzhypothese. Dies trifft genau dann zu, wenn der Wert von C_u größer wird als $-\varepsilon_1$ und gleichzeitig derjenige von C_o nicht über $+\varepsilon_2$ hinausgeht. Andernfalls ist die Nullhypothese der Nichtäquivalenz beizubehalten. Bei der Anwendung dieser Regel (Kasten 1a) ist unbedingt folgendes zu beachten: Um zu garantieren, dass der durchzuführende Test auf Äquivalenz das Signifikanzniveau $\alpha = 5\,\%$ einhält, genügt es nicht, dass das verwendete Konfidenzintervall zweiseitiges Konfidenzniveau 90 % besitzt [4]. Voraussetzung ist vielmehr, dass jede der beiden Konfidenzschranken C_u und C_o einseitiges Konfidenzniveau 95 % aufweist.

Will man anstatt auf Äquivalenz nur auf Nichtunterlegenheit testen, wird lediglich die untere Konfidenzgrenze benötigt. Der zugehörige Test nach dem Intervallinklusions-Prinzip läuft dann so ab, dass Nichtunterlegenheit für statistisch gesichert erklärt wird, wenn man findet, dass C_u die unter der Hypothese spezifizierte untere Äquivalenzgrenze übersteigt (Kasten 1b, Grafik 2).

Optimale Tests auf Äquivalenz und Nichtunterlegenheit

Tests, die nach dem Intervalleinschlussprinzip arbeiten, kontrollieren zwar das Fehlerrisiko 1. Art, sind aber hinsichtlich der Power [13] suboptimal und benötigen daher größere Stichprobenumfänge als günstigstenfalls erforderlich.

In der statistischen Originalliteratur findet man für eine Vielzahl von Situationen, die sich nach dem Studiendesign und der Art der zu analysierenden Zielvariablen unterscheiden, optimale Tests für Äquivalenz- und Nichtunterlegenheitshypothesen [17]. Die praktische Umsetzung solcher Tests ist erheblich komplizierter, als man es von herkömmlichen ein- oder zweiseitigen Signfikanztests gewohnt ist, und erfordert spezielle Berechnungsverfahren, für die aber problemlos handhabbare Computerprogramme verfügbar sind.

In Kasten 2 wird für die in klinischen Studien sehr häufige Situation des Vergleichs zweier Binomialverteilungen das optimale Verfahren der Prüfung auf Nichtunterlegenheit vorgestellt und durch ein Beispiel illustriert.

Kriterien für die Beurteilung von Publikationen über Äquivalenzstudien

In Kasten 3 sind einige Grundkriterien für die Beurteilung von Publikationen über Äquivalenz- und Nichtunterlegenheitsstudien zusammengestellt. Die Tabelle enthält die Resultate einer Überprüfung dieser Kriterien in einschlägigen Publikationen der Jahrgänge 2000–2011 in den fünf wichtigsten fachübergreifenden medizinischen Zeitschriften. Danach tritt der Fehler des Schließens von nichtsignifikanten Unterschieden auf statistisch gesicherte Äquivalenz in den hochrangigen Zeitschriften nicht mehr allzu häufig auf. Weitaus weniger günstig ist das Bild bezüglich Studien mit zweiseitiger Äquivalenzfragestellung: Hier kommen anstatt optimaler Verfahren ausschließlich Konfidenzintervall-Inklusionstests zur Anwendung, und dies noch dazu in der

Verteilung der Pro's und Con's gemäß (Q1)–(Q5) aus Kasten 3*[1]

	Q1	Q2	Q3	Q4	Q5
+	180	176	46	0	2
–	10	4	131	23	21
na*[2]	0	10	13	167	167

Tabelle *[1] in durch PubMed-Suche unter den Stichworten „equivalence" und „non(-) inferiority" identifizierten Publikationen in NJEM, LANCET, JAMA, ANN INTERN MED und BMJ (Jahrgänge 2000–2012, Trefferzahl N = 190);
*[2] nicht anwendbar.

Table *[1] publications in NJEM, Lancet, JAMA, Ann Intern Med, and BMJ in the years 2000–2011 found by a PubMed search using the terms "equivalence" and "non(-)inferiority";
*[2] not applicable.

Kernaussagen

- Beim Äquivalenznachweis ist es nicht zulässig, einen herkömmlichen zweiseitigen Test zu verwenden und aus einem negativen Ergebnis auf Äquivalenz zu schließen.
- Der erste Schritt einer korrekten konfirmatorischen Analyse einer Äquivalenz- oder Nichtunterlegenheits-Studie besteht in der Festlegung eines geeigneten Verteilungsparameters, der ein sinnvolles Maß für die Unterschiedlichkeit der Behandlungswirkungen in der Grundgesamtheit darstellt.
- Der einfachste Ansatz für den statistischen Nachweis von Äquivalenz oder Nichtunterlegenheit beruht dann auf der Berechnung von Konfidenzgrenzen für diesen Parameter.
- Die Vorzüge von auf Konfidenzgrenzen basierenden Verfahren liegen hauptsächlich in der einfachen Durchführbarkeit. Dieser Vorteil wird erkauft um den Preis einer unnötig niedrigen Power der Tests.
- In Hinblick auf den möglichst ökonomischen Umgang mit Patienten- oder Probandenzahlen empfiehlt sich auch beim Gleichwertigkeitsnachweis der Einsatz von bezüglich der Trennschärfe optimierten statistischen Testverfahren.

unnötig konservativen, durch Anhebung des zweiseitigen Konfidenzniveaus auf 95 % resultierenden Version (Kasten 3, Tabelle).

Diskussion

Tests für die konfirmatorische statistische Auswertung von Äquivalenz- und Nichtunterlegenheits-Studien gehören heute zum Standardrepertoire der medizinischen Biometrie. Ein wichtiger Anwendungsbereich für diese Verfahren ist der Nachweis der Bioäquivalenz verschiedener Formulierungen des gleichen Arzneimittels. Auf die methodischen Besonderheiten dieses Studientyps, der die Grundlage für die behördliche Zulassung von Generika bildet, konnte im Rahmen dieser kurzen Übersicht nicht näher eingegangen werden (umfassende Darstellungen findet man in Kap. 10 von [17] sowie in [2, 7, 11, 15]). Der Äquivalenztest, der hierbei entsprechend den Guidelines der Zulassungsbehörden (vergleiche [6]) routinemäßig zur Anwendung gelangt, ist der in Kasten 1A dargestellte Test auf Äquivalenz zweier Normalverteilungen bezüglich der Differenz der Mittelwerte. Dieser ist durchzuführen mit den (logarithmisch transformierten) Quotienten der Messergebnisse aus den beiden Perioden eines Cross-over-Versuchs [16].

Auch klinische Studien höherer Phasen werden in zunehmender Zahl mit dem Ziel des Äquivalenz- oder Nichtunterlegenheits-Nachweises durchgeführt. In der Mehrzahl der Fälle handelt es sich dabei um randomisierte Therapiestudien [9] mit Aktiv-(Positiv-)Kontrolle. In der Kontrollgruppe wird dann anstatt Placebo eines der als wirksam bekannten etablierten Behandlungsverfahren angewandt. Inhaltlich gesehen liegt der Hauptunterschied zu Bioäquivalenz-Studien darin, dass das Zielkriterium hier das Ansprechen von Patienten mit einschlägiger Indikation auf die Behandlung ist, nicht eine pharmakokinetische Größe, die bei gesunden Probanden gemessen wird. In statistischer Hinsicht unterscheiden sich Studien zum Nachweis therapeutischer Äquivalenz von Bioäquivalenz-Studien vor allem dadurch, dass sich der Äquivalenztest sehr oft auf Zielvariablen zu beziehen hat, die keine Verteilung vom stetigen Typ besitzen (und damit insbesondere nicht normalverteilt sind) oder teilweise zensiert sind. Besonders häufig sind in aktiv kontrollierten Therapiestudien Situationen, in denen man Vergleiche durchzuführen hat zwischen Responderraten (d.h. binomialen Proportionen) oder Kaplan-Meier-Überlebensfunktionen. Für alle diese Fragestellungen sind in der Originalliteratur geeignete Äquivalenz- und Nichtunterlegenheits-Tests verfügbar. Nach gegenwärtig vorherrschender Praxis [12] werden aktiv kontrollierte klinische Studien zumeist auf der Basis von Nichtunterlegenheits-Tests geplant und ausgewertet. Dies ist jedoch von der statistischen Logik her keineswegs zwingend, sondern begründet sich in erster Linie durch die Tatsache, dass bei gleicher Festlegung der unteren Äquivalenzgrenze und bei gegebener Power für den Nachweis von Äquivalenz im strikten Sinne erheblich größere Fallzahlen benötigt werden als für den Nichtunterlegenheits-Nachweis. In Übereinstimmung damit ist die Aussage, die bei einer positiven Testentscheidung möglich ist, beim Äquivalenznachweis sehr viel präziser als beim Nachweis von Nichtunterlegenheit.

Allgemein sollte bei der Bewertung von klinischen Studien streng darauf geachtet werden, ob es sich um eine klassische Situation oder eine Studie handelt, die zum Zwecke des Nachweises von Äquivalenz oder Nichtunterlegenheit durchgeführt worden ist. Je nach Studientyp werden andere statistische Verfahren benötigt. Tests auf Äquivalenz und Nichtunterlegenheit sind zwar mittlerweile gut entwickelt und auch bekannt, werden aber bei der Interpretation der Ergebnisse und der Begründung der Annahmen, von denen man dabei ausgeht, nicht immer in der angemessenen Weise gehandhabt. Mindestanforderungen, die Publikationen, in denen über die Ergebnisse von Äquivalenz- oder Nichtunterlegenheits-Studien berichtet wird, auf dieser Ebene erfüllen sollten, sind vor einigen Jahren in einem Addendum zum sogenannten CONSORT STATEMENT zusammengestellt worden [12]. DZZ

Interessenkonflikt: Prof. *Blettner* erhielt Honorare für Beratertätigkeit von Astellas und AstraZeneca. Prof. *Wellek* erklärt, dass kein Interessenkonflikt besteht.

Korrespondenzadresse

Prof. Dr. rer. nat. Maria Blettner
Institut für Medizinische Biometrie
Epidemiologie u. Informatik der
Johannes Gutenberg-Universität
Obere Zahlbacher Straße 69, 55131 Mainz
blettner-sekretariat@imbei.uni-mainz.de

Literatur

1. Altman DG, Bland JM: Absence of evidence is not evidence of absence. BMJ 1995;311:485
2. Chow SC, Liu JP: Design and analysis of bioavailability and bioequivalence studies, 3rdEdition. Chapman & Hall/CRC, Boca Raton 2008
3. du Prel J, Röhrig B, Hommel G et al.: Choosing statistical tests: part 12 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010;107: 343–348
4. du Prel JB, Hommel G, Röhrig B: Confidence interval or p-value? Part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009;106: 335–339
5. Eron JJ, Young B, Cooper DA et al.: SWITCHMRK 1 and 2 investigators: Switch to a raltegravir-based regimen versus continuation of a lopinavir-ritonavir-based regimen in stable HIV-infected patients with suppressed viraemia (SWITCHMRK 1 and 2): two multicentre, double-blind, randomised controlled trials. Lancet 2010;375: 396–407. Epub 2010 Jan 12. PubMed PMID: 20074791
6. Food and Drug Administration (FDA): Guidance for industry: Statistical approaches to establishing bioequivalence. MD: Center for Drug Evaluation and Research (CDER), Rockville 2001
7. Hauschke D, Steinijans VW, Pigeot I: Bioequivalence studies in drug development: Methods and applications. John Wiley & Sons, Chichester 2007
8. Hilgers RD, Bauer P, Schreiber V et al.: Einführung in die Medizinische Statistik. 2 nd edition. Springer-Verlag, Berlin 2007
9. Kabisch M, Ruckes C, Seibert-Grafe M et al.: Randomized controlled trials: part 17 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2011;108:663–668
10. Kuhrt N: Gleiche Wirkung. Bei Altersblindheit helfen zwei Medikamente. Weiter verbreitet ist das teure. Warum? ZEIT ONLINE 2011. www.zeit.de/2011/20/Pharmaindustrie-Medikamente
11. Patterson S, Jones B: Bioequivalence and statistics in clinical pharmacology. Chapman & Hall/CRC, Boca Raton 2005
12. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, for the CONSORT Group: Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006;295:1152–1160
13. Röhrig B, du Prel JB, Wachtlin D et al.: Sample size calculation in clinical trials: part 13 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010;107:552–556
14. The CATT Research Group: Ranibizumab and Bevacizumab for neovascular age-related macular degeneration. NEJM 2011;364:1897–1908
15. Vollmar J (Ed.): Bioäquivalenz sofort freisetzender Arzneiformen. Gustav Fischer Verlag, Stuttgart 1991
16. Wellek S, Blettner M: λOn the proper use of the crossover design in clinical trials: part 18 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2012;109:276–281
17. Wellek S: Testing statistical hypotheses of equivalence and noninferiority. 2nd edition. Chapman & Hall/CRC, Boca Raton 2010
18. Westlake WJ: Use of confidence intervals in analysis of comparative bioavailability trials. J Pharma Sci 1972;61: 1340–1341