Zehn Anforderungen an therapeutische Äquivalenzstudien

oder Warum der fehlende Nachweis von Unterschieden und Äquivalenz nicht dasselbe bedeuten

Peter Kleist
PFC Pharma Focus AG, Volketswil

Quintessenz

- Im Vergleich zu traditionellen Studien zum Nachweis der Überlegenheit einer Behandlung gegenüber Plazebo oder einer Standardtherapie ist die Durchführung von Äquivalenz- bzw. sogenannten «Noninferioritätsstudien» mit deutlich höheren methodologischen Anforderungen verbunden.
- Von essentieller Bedeutung ist die «Assay-Sensitivität», das heisst die Frage, ob eine Äquivalenzstudie überhaupt in der Lage ist, einen bestehenden Unterschied zwischen zwei Behandlungen aufzeigen zu können.
- Die Aussagefähigkeit einer Äquivalenzstudie hängt massgeblich von einem angemessenen Studiendesign, der Wahl der Positivkontrolle, einer für die Fragestellung ausreichenden Fallzahl, dem festgelegten Äquivalenzbereich, einer adäquaten Auswertung sowie nicht zuletzt von einer korrekten, protokollgerechten Durchführung ab.
- Die meisten der bislang realisierten Studien, welche die Schlussfolgerung einer therapeutischen Äquivalenz zweier Behandlungen ziehen, erfüllen diese Anforderungen nicht.

Summary

Ten demands to therapeutic equivalence studies. Or why lack of evidence of difference and equivalence does not mean the same thing

- *Compared with traditional studies to prove the superiority of a treatment vs. placebo or standard treatment, the conduct of equivalence or noninferiority studies imposes markedly higher methodological demands.*
- *Of essential importance is assay-sensitivity, i.e. whether an equivalence study is at all capable of demonstrating an existing difference between two treatments.*
- *The validity of an equivalence study depends decisively on appropriate design, the choice of the positive control, a sufficient sample size for the question under study, the predefined equivalence margins, appropriate evaluation and, not least, its correct per-protocol conduct.*
- *Up to the present the majority of studies which draw conclusions concerning the therapeutic equivalence of two treatments have failed to meet these demands.*

Einführung

Die Mehrheit aller randomisierten kontrollierten Studien wird durchgeführt, um die Überlegenheit einer Behandlung gegenüber einer anderen nachzuweisen (sog. Überlegenheitsstudien oder «superiority trials»). Als Vergleichsbehandlung dient häufig die Verabreichung von Plazebo. Dieser Ansatz kann jedoch mit Schwierigkeiten verbunden sein:

1. Steht eine wirksame Therapie zur Verfügung, wird der Plazebogebrauch oftmals unethisch [1, 2].
2. Die in vielen therapeutischen Gebieten geforderte Untersuchung harter klinischer Endpunkte erschwert den Nachweis von Unterschieden zwischen zwei wirksamen Behandlungen, da die gegenwärtigen Therapiestandards zum Teil sehr hoch sind und harte Endpunkte – wie zum Beispiel kardiovaskuläre Ereignisse nach einem Myokardinfarkt – immer seltener auftreten [3, 4]. Die Aufdekkung eines kleinen Unterschiedes erfordert Studien mit sehr hohen Fallzahlen.
3. Bei vielen Erkrankungen sind keine therapeutischen Durchbrüche zu erwarten. Die Unterschiede zwischen einer neuen und einer bestehenden Therapie beschränken sich allenfalls auf eine bessere Verträglichkeit, eine höhere Sicherheit, eine einfachere Anwendbarkeit oder auf tiefere Kosten, während das Ausmass der Wirksamkeit vergleichbar ist [5, 6].

In diesen Situationen bietet sich die Durchführung von Äquivalenzstudien an.

Definition und Abgrenzung

Äquivalenzstudien sollen zeigen, dass sich zwei Behandlungen nicht klinisch relevant voneinander unterscheiden, das heisst, dass ihre Wirkungen in einem Bereich liegen, innerhalb welchem Äquivalenz konstatiert werden kann, da ja nicht mit einem absolut identischen Wirkungsausmass zu rechnen ist. Reine Äquivalenzstudien, bei denen zum Nachweis der Äquivalenz in beide

CME zu diesem Artikel finden Sie auf S. 809 oder im Internet unter www.smf-cme.ch.

Richtungen (d.h. «nach oben» und «nach unten») die Toleranzgrenzen für Abweichungen definiert werden, sind selten [7]. Ein Beispiel hierfür wären vergleichende Bioverfügbarkeitsstudien zwischen einem Generikum und dem Originalpräparat: Die maximale Plasmakonzentration und die Fläche unter der Konzentrations-Zeit-Kurve dürfen – zum Nachweis der Bioäquivalenz – nur innerhalb einer behördlich festgelegten Spanne (in der Regel 80–125%) voneinander abweichen.

Bei Äquivalenzstudien mit therapeutischen Fragestellungen handelt es sich überwiegend um sogenannte Nichtunterlegenheitsstudien («noninferiority trials»; nachfolgend wird nur der gebräuchlichere Begriff «Noninferiorität» verwendet). Der definierte Äquivalenzbereich erstreckt sich in der Regel nur in eine Richtung, und zwar in jene der klinisch nicht relevanten Unterlegenheit [6]. Die Fragestellung ist also asymmetrisch. Eine neue Behandlung kann empfohlen werden, wenn ihre Wirkung zumindest vergleichbar, nicht aber wenn sie schlechter als jene der Referenzbehandlung ist. Das Aufzeigen einer Überlegenheit ist als «Bonus» anzusehen [7].

Wie gross der Anteil dieses Studientyps an allen randomisierten kontrollierten Studien ist, lässt sich schwer abschätzen, da die Begriffe «Äquivalenz» bzw. «Noninferiorität» in bezug auf die Fragestellung, das Studiendesign und die nach dem Vorliegen der Ergebnisse gezogenen Schlussfolgerungen oft zu Unrecht gebraucht werden. Zwei jüngere Untersuchungen haben eindrücklich demonstriert, dass nur etwa 20% der Studien, die Äquivalenz- bzw. Noninferioritätsaussagen machen, den spezifischen Kriterien für Äquivalenzstudien gerecht werden [8, 9]. Zwei Drittel der Studien waren Überlegenheitsstudien, bei denen die angewendeten statistischen Tests keine Unterschiede zwischen den Behandlungsgruppen zeigen konnten [8]. Der fehlende Nachweis von Unterschieden, das heisst die Tatsache, dass die Nullhypothese («Es besteht *kein* Unterschied zwischen den Behandlungen.») *nicht* verworfen werden kann, impliziert jedoch nicht, dass die Behandlungen auch äquivalent sind [5]; sie sind es möglicherweise, aber eine auf den Nachweis von Unterschieden angelegte Studie ist nicht geeignet, Äquivalenz zu beweisen [10]. Oder wie es unsere englischen Kollegen formuliert haben: «Absence of evidence (of a difference) is not evidence of absence (of a difference).» [11, 12]

Im Gegensatz zu Überlegenheitsstudien unterliegen Äquivalenz- bzw. Noninferioritätsstudien nämlich besonderen methodologischen Ansprüchen, um zu aussagefähigen Ergebnissen zu gelangen:

1. Eine Überlegenheitsstudie hängt – neben ihrer internen Validität – nur von der statistischen Unsicherheit, aber nicht von externen, das heisst ausserhalb der Studie liegenden Faktoren ab [13]. Während eine grössere Fallzahl die Chance erhöht, einen Unterschied aufzuzeigen, ist eine mit einer nicht ausreichenden statistischen Power (Teststärke) versehene Studie anfällig für einen sogenannten Typ-II-Fehler, das heisst für die falsche Schlussfolgerung eines nicht bestehenden Unterschiedes zwischen zwei Behandlungen [14, 15]. Für die Aussagefähigkeit von Äquivalenz- und Noninferioritätsstudien spielt dagegen eine Vielzahl weiterer, studienunabhängiger Faktoren eine essentielle Rolle.
2. Die Validität von Äquivalenz- bzw. Noninferioritätsstudien ist entscheidend mit ihrer «Assay-Sensitivität» verbunden, das heisst ihrem Vermögen, einen Unterschied zwischen den Behandlungen aufzeigen zu können, insofern dieser besteht, also eine wirksame von einer nicht oder weniger wirksamen Behandlung abzugrenzen [3, 13]. Eine «erfolgreiche» Überlegenheitsstudie hat gleichzeitig die Assay-Sensitivität bestätigt. Eine unzureichende Assay-Sensitivität von Äquivalenzstudien hat hingegen zur Folge, dass man bei erwiesener Äquivalenz nicht wissen kann, ob eine schlechte Studiendurchführung das Aufzeigen eines bestehenden Unterschiedes verhindert hat oder ob beide Therapien gleich wirksam oder gleich unwirksam sind. Eine wichtige Voraussetzung besteht daher auch darin, dass es sich bei der Vergleichsbehandlung um eine Therapie mit zuverlässiger Wirksamkeit handelt [5].

Daraus leiten sich spezifische Anforderungen an Äquivalenz- bzw. Noninferioritätsstudien ab. Die vor kurzem veröffentlichte Erweiterung des «CONSORT-Statements» auf Äquivalenzstudien unterstreicht die Wichtigkeit ihrer Beachtung, um valide Studienergebnisse und Schlussfolgerungen zu gewährleisten [7].

Zehn Anforderungen an klinische Studien zur Untersuchung therapeutischer Äquivalenz

Ob Sie selber Äquivalenzstudien durchführen oder als Leser medizinischer Fachliteratur deren Aussagefähigkeit beurteilen möchten: Die folgenden zehn Punkte sollen Ihnen dazu konkrete Hilfestellungen geben.

1. Einer Äquivalenzstudie liegt eine klare Hypothese bezüglich der Untersuchung von Äquivalenz bzw. Noninferiorität zugrunde. Im Vergleich zu Überlegenheitsstudien sind bei Äquivalenzstudien die Null- und die Alternativhypothese vertauscht. Die Nullhypothese einer Äquivalenzstudie besagt, dass sich die untersuchten Behandlungen unterscheiden; ein Typ-I-Fehler besteht in der fälschlichen Annahme von Äquivalenz, ein Typ-II-Fehler

im Voraussetzen eines Unterschiedes, obwohl dieser in Wirklichkeit nicht besteht [5, 7]. Aus einer Studie, die darauf angelegt war, Überlegenheit zu zeigen, darf bei «negativem» Ausgang nicht die Schlussfolgerung von Äquivalenz gezogen werden (siehe Punkt 9), was in der Praxis leider häufig passiert [8]. So wurden beispielsweise nur zwei von 25 Studien, die bei der Untersuchung der kindlichen bakteriellen Meningitis Äquivalenzaussagen zu den verwendeten Antibiotika machten, tatsächlich für diese Fragestellung entworfen [16].

2. Der Äquivalenz- bzw. Noninferioritätsbereich ist im voraus zu definieren, das heisst, die Grenzwerte für einen unter klinischen Gesichtspunkten gerade noch akzeptablen Unterschied sind vor Beginn der Studie festzulegen [7]. Dieser Spanne sollte – falls es keinen allgemeinen Konsens gibt – eine unabhängige Expertenmeinung zugrunde liegen [14] – einerseits, um einen möglichen Bias durch eine subjektive Post-hoc-Festlegung zu verhindern, andererseits, um ein zu gross gewähltes Intervall zu vermeiden [17]. Insbesondere bei Studien mit dem Endpunkt Mortalität ist darauf zu achten, dass die tolerierte Abweichung nicht zu grosszügig festgelegt wird, um eine noch akzeptable Noninferiorität gegenüber einer wirksamen Therapie annehmen zu können [18].
   Ein häufig empfohlener Noninferioritätsgrenzwert liegt bei <50% des Effektes, den die Referenztherapie zuvor gegenüber Plazebo gezeigt hat. Beträgt zum Beispiel die zusätzliche blutdrucksenkende Wirkung der Referenzbehandlung gegenüber Plazebo mindestens 10 mm Hg (unteres Ende des Konfidenzintervalls), so wäre ein Äquivalenzbereich von 6 bis 14 mm Hg (entsprechend 10 mm Hg ± 40%) denkbar. Setzte man den Grenzwert bei >50% an, läge die Wirksamkeit der Testbehandlung – bei voller Ausschöpfung des Spielraums – näher bei jener von Plazebo als bei derjenigen der Vergleichstherapie, was dem Grundgedanken von Äquivalenz (bzw. Noninferiorität) widersprechen würde. Bei Mortalitätsstudien ist die Noninferioritätsspanne in der Regel enger zu definieren (siehe auch das Beispiel in Punkt 9) [14].
3. Ist die Fallzahl ausreichend? Die Fallzahl einer Äquivalenz- bzw. Noninferioritätsstudie hängt von drei Faktoren ab: der Breite des Vertrauensbereiches, der statistischen Power (Teststärke) sowie dem unter Punkt 2 beschriebenen Äquivalenzbereich. Letzterer ist in der Regel enger als der untersuchte Unterschied in einer Überlegenheitsstudie, was erklärt, weshalb die erforderlichen Fallzahlen in Äquivalenz- bzw. Noninferioritätsstudien vergleichsweise höher sind [5]. Ist die Fallzahl und somit die Power zu klein, nimmt die Assay-Sensitivität ab, das heisst die Chance, einen in Wirklichkeit bestehenden Unterschied zu zeigen. Ein extremes Beispiel stammt aus einer Vergleichsstudie zwischen Octreotid und einer Sklerotherapie bei akuten Ösophagusvarizenblutungen: Die Schlussfolgerung der therapeutischen Äquivalenz wurde auf der Basis von 100 in die Studie aufgenommenen Patienten gezogen, obwohl die in der Methodik beschriebene Fallzahlberechnung eine Anzahl von 1800 Patienten als notwendig erachtete [19]. In dieser Studie war die Wahrscheinlichkeit, überhaupt einen bestehenden Unterschied aufdecken zu können, auf 5% reduziert.
4. Bei der Auswahl der Positivkontrolle, also der aktiven Vergleichsbehandlung, muss es sich um einen etablierten Standard handeln [5, 10]. Eine behördliche Zulassung allein (insbesondere, wenn sie zeitlich lange zurückliegt) ist insofern nicht ausreichend, als sie eine gegenüber Plazebo vorhandene Wirkung nicht immer konsistent ausdrückt. Viele Arzneimittel, für die eine Wirksamkeit nachgewiesen wurde, zum Beispiel unterschiedliche Antidepressiva, Analgetika, Antianginosa oder Antihypertensiva, haben dennoch in einer nicht unbedeutenden Anzahl von Studien keine Überlegenheit gegenüber Plazebo zeigen können [13]. Das damit verbundene Problem wurde von Tramèr et al. am Beispiel von Ondansetron-Studien zur Prävention postoperativen Erbrechens veranschaulicht [20]: Wie sicher kann man die Wirksamkeit von Ondansetron – als Referenzbehandlung innerhalb einer Vergleichsstudie – voraussetzen, wenn die Substanz in 30% der zuvor durchgeführten Studien nicht signifikant besser war als Plazebo? Ohne ausreichenden Beleg für die Wirksamkeit der Kontrollbehandlung ist die Frage, ob eine nachgewiesene Äquivalenz die vergleichbare Wirksamkeit oder vergleichbare Nichtwirksamkeit der untersuchten Behandlungen ausdrückt, nicht zu beantworten.
5. Das Design einer Studie soll sich so eng als möglich an die Studie(n) anlehnen, in der oder in denen die Kontrollbehandlung zuvor ihre Wirksamkeit gezeigt hat [5, 7]. Wichtige Elemente in diesem Zusammenhang sind die Patientenauswahl, die Dosis, (un)erlaubte Begleittherapien sowie die Behandlungs- und Follow-up-Dauer [5, 6, 13]. So liess sich für die in einigen grossen Antihypertensiva-vergleichsstudien verwendeten Referenztherapien zuvor oftmals keine ausreichende Wirksamkeit in den untersuchten Patientenkollektiven belegen [21]. Auch bei Patienten mit einer schwach ausgeprägten Krankheitssymptomatik oder mit geringgradigen Abweichungen von den Normalwerten kann fälschlicherweise die Schlussfolgerung einer

therapeutischen Äquivalenz (bzw. Noninferiorität) gezogen werden, da das erforderliche Ausmass der therapeutischen Wirkung klein ist (supratherapeutischer Bereich). Inadäquate Titrationsschemata und unzureichende therapeutische Zieldosen können eine erhebliche Bedeutung für den (teilweise fragwürdigen) Ausgang von aktiv kontrollierten Vergleichsstudien aufweisen, so etwa im Bereich der Psychiatrie [22].

6. Ist der Endpunkt der Studie geeignet? Einerseits sollte der Endpunkt dem zuvor zum Nachweis der Wirksamkeit der aktiven Kontrollbehandlung verwendeten Endpunkt entsprechen; andererseits muss er bestimmte Voraussetzungen erfüllen: Er darf nicht «zu variabel» gewählt werden, denn dies erhöht per se die Wahrscheinlichkeit für eine Äquivalenz am Studienende. Die Endpunktvariabilität ist sicherlich eine der wesentlichen Ursachen für inkonsistente Studienergebnisse der unter Punkt 4 genannten Substanzen [13, 20]. Sind niedrige Ereignisraten zu erwarten (relativ selten auftretende Endpunkte), bietet sich unter Umständen die Untersuchung eines zusammengesetzten Endpunktes an – aber Vorsicht: Endpunktkomponenten, die nicht in gleicher Weise auf die Behandlung ansprechen, erhöhen wiederum die Variabilität und damit die Unsicherheit in Verbindung mit Äquivalenzstudien [23]. Aus diesem Grund ist beispielsweise «time-to-treatment-failure» als Endpunkt ungeeignet, da ein Behandlungsabbruch wegen unerwünschter Wirkungen oder eines freiwilligen Ausscheidens aus der Studie wenig Relevanz für den zugrundeliegenden Krankheitsprozess oder die pharmakologische Wirksamkeit einer Therapie aufweist [24].
7. Der Standard der Studiendurchführung muss insbesondere bei Äquivalenz- bzw. Noninferioritätsstudien sehr hoch sein [5], da Protokollabweichungen potentiell die bestehenden Unterschiede zwischen zwei Behandlungen verwischen oder aufheben können. Viele Protokollverletzungen, eine höhere Anzahl von Studienabbrechern oder von Behandlungs-Crossovers sowie nicht ausreichend standardisierte Messungen erhöhen die Wahrscheinlichkeit für eine Äquivalenz. Eine unzureichende Randomisierung kann das Studienergebnis potentiell in beide Richtungen verfälschen. Die Leser von Publikationen zu Äquivalenzstudien seien insbesondere auf die aufmerksame Betrachtung dieser Faktoren bzw. der Qualität der Studiendurchführung hingewiesen.
8. Nicht unterschätzt werden darf die (Non-)Compliance der Patienten aus der Studie, vor allem in bezug auf die zuverlässige Einnahme der Studienmedikamente [10]. Zum Beispiel erhielten am Ende von grossen Antihypertensivastudien durchschnittlich nur noch zwei Drittel der Patienten die ursprünglich randomisierte Studienmedikation [21]; wie hoch die tatsächliche Compliance war, ist darüber hinaus fraglich – schliesslich ist eine mangelnde Einnahmetreue der Hauptgrund dafür, dass der Blutdruck nur bei einem Viertel der Hypertoniker ausreichend eingestellt ist [25]. Um keine falschen Schlüsse bezüglich einer Äquivalenz zu ziehen (denn je geringer die Einnahmetreue, desto wahrscheinlicher wird es, eine Äquivalenz zu zeigen), ist der Beleg einer ausreichenden und in beiden Behandlungsgruppen vergleichbaren Compliance erforderlich; neben dem klassischen «pill count» («drug account») sollte der Einsatz von Arzneimittelbehältern, die das Öffnen bzw. die Tablettenentnahme mit Datum und Uhrzeit registrieren (sog. MEMS, «medication event monitoring systems»), grundsätzlich in Betracht gezogen werden [10].
9. Eine Äquivalenzstudie ist auf der Basis von Konfidenzintervallen (Vertrauensintervallen) auszuwerten; konventionelle statistische Testverfahren spielen für diesen Studientyp keine Rolle [5, 7]. Auch für Noninferioritätsuntersuchungen ist gegebenenfalls eine zweiseitige Fragestellung zu implementieren (zweiseitiges Konfidenzintervall von 95%), um eine (konsekutive) Auswertung im Hinblick auf eine Superiorität zu ermöglichen [18]. Denn während sich bei einer auf Superiorität angelegten Studie die Testung auf Äquivalenz in der Regel verbietet, ist es im Rahmen einer Äquivalenzstudie grundsätzlich möglich, die Analyse auf eine mögliche Überlegenheit auszuweiten [3]. Im Analyseplan sollte zuvor exakt festgehalten werden, auf welcher Basis ein Wechsel von einer Noninferioritäts- zu einer Superioritätsauswertung vorgenommen wird.

   Ein Konfidenzintervall von 95% besagt – vereinfacht ausgedrückt – dass bei 100facher Wiederholung der Studie 95 Ergebnisse in diesen Bereich fallen. Oder anders formuliert: Es zeigt mit einer Wahrscheinlichkeit von 95% die Spanne an, innerhalb der sich das tatsächliche Ergebnis bei der Untersuchung einer sehr grossen Patientenzahl in sehr vielen Studien befindet. Das Konfidenzintervall entspricht also einem Mass für die Zuverlässigkeit eines einzelnen Studienergebnisses, das zwangsläufig mit einer Unsicherheit verbunden ist. Um Äquivalenz bzw. Noninferiorität zu folgern, muss das Konfidenzintervall innerhalb des zuvor festgelegten Äquivalenz-(Noninferioritäts-)Bereichs liegen. Der in einer Studie auftretende Mortalitätsunterschied unter zwei Behandlungen A und B könnte beispielsweise 10% zu Gunsten von Behandlung A betragen (d.h. einer Senkung der Mortalität von 10% im Vergleich zu Be-

handlung B). Ein herkömmlicher statistischer Test zeigt keinen signifikanten Unterschied (p >0,05). Bei einem weiten Konfidenzintervall (z.B. von –24% bis +14%) und einem zuvor definierten Noninferioritätsbereich (–x% bis max. +10%) kann jedoch keine Schlussfolgerung der Noninferiorität von Behandlung A gezogen werden, da nicht auszuschliessen ist, dass Behandlung B klinisch signifikant besser ist als Behandlung A. Der p-Wert des statistischen Tests sagt daher in bezug auf die Äquivalenz- bzw. Noninferioritätsfragestellung wenig aus.
Ein weiteres, reelles Beispiel soll die Bedeutung von Konfidenzintervallen für die Auswertung erläutern. In der COBALT-Studie wurde die doppelte Bolusgabe mit der kontinuierlichen Infusion von t-PA bei über 7000 Patienten mit Myokardinfarkt verglichen [26]. Die Mortalität unter Bolusgabe war geringfügig höher, und das obere Ende des Vertrauensintervalls betrug 1,49%. Dieses überschritt die zuvor festgelegte Noninferioritätsgrenze von 1,4%, das heisst, dass eine Zunahme der Sterberate von maximal 0,4% noch als äquivalent betrachtet worden wäre. Die Studie konnte somit keine Äquivalenz der beiden Behandlungsschemata aufzeigen. Weitere Anschauungsbeispiele sind der Abbildung 1 zu entnehmen.

10. Bei einer auf Überlegenheit hin konzipierten Studie stellt die «intention-to-treat»-Auswertung (ITT-Auswertung) den konservativen Ansatz dar. Bei einer Äquivalenz- bzw. Noninferioritätsstudie ist eine ITT-Auswertung jedoch generell anfällig für einen Bias in Richtung Äquivalenz, da jede Abweichung von der randomisierten Behandlungszuteilung (z.B. durch Studienabbruch oder Behandlungs-Crossover) die Unterschiede zwischen den Behandlungen potentiell verkleinert und damit die Anfälligkeit für einen Typ-I-Fehler erhöht [18]. Es ist daher die zusätzliche Durchführung einer Per-Protokoll-Analyse (PP- oder «as treated»-Analyse) zu fordern, und die in einer Äquivalenzstudie gewonnenen Ergebnisse sind vertrauenswürdiger, wenn sich die ITT- und die PP-Auswertung nicht gravierend voneinander unterscheiden [5, 7, 27]. Die beispielsweise in einer Untersuchung zweier Erythropoietinverabreichungsschemata ausschliesslich vorgenommene ITT-Auswertung (bei einer Compliance von 63%) ist sicherlich als inadäquat zu betrachten [27, 28].

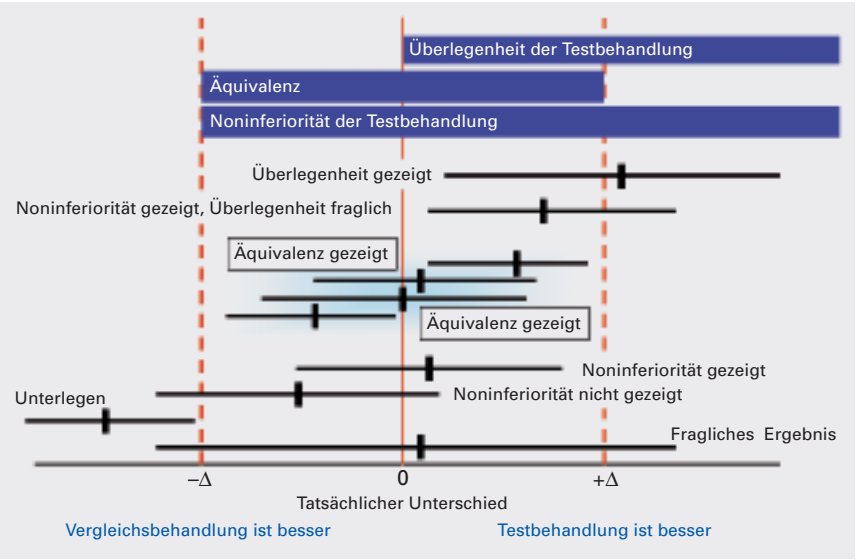

Abbildung 1
Beispiele möglicher Ergebnisse einer Äquivalenz- bzw. Noninferioritätsstudie, dargestellt durch Vertrauensintervalle. Die senkrechten gestrichelten Linien entsprechen den Grenzen des zuvor festgelegten Äquivalenzbereichs (–Δ bis +Δ)
(Darstellung adaptiert nach Jones et al. [5], Pater [10] und Gomberg-Maitland et al. [14]).

Drei weitere methodische Besonderheiten von Äquivalenzstudien

1. Im Gegensatz zu Überlegenheitsstudien spielt eine Verblindung der Behandlungen im Rahmen von Äquivalenzstudien eine untergeordnete Rolle, da diese keinen wirksamen Schutz gegen einen Bias in Richtung Angleichung der Behandlungsergebnisse bietet, etwa durch möglichst ähnliche Score-Bildungen oder Stadieneinteilungen [3].
2. Belegen Interimsanalysen die Äquivalenz bzw. Noninferiorität, besteht kein medizinischer Grund zum vorzeitigen Abbruch der Studie (insofern auch die Verträglichkeit bzw. Sicherheit vergleichbar sind) [7]. Eine Studie kann weitergeführt werden, um am Ende gegebenenfalls die Überlegenheit der Testbehandlung zu zeigen [18].
3. Wenn nur eine unzureichend etablierte Referenztherapie zur Verfügung steht, ist die Durchführung einer dreiarmigen, zusätzlich plazebokontrollierten Äquivalenzstudie in Betracht zu ziehen. Erst wenn die Referenztherapie eine Überlegenheit gegenüber Plazebo zeigt, wird in einem zweiten Schritt der Frage der möglicherweise bestehenden Äquivalenz zwischen Test- und Vergleichsbehandlung nachgegangen [6, 13, 29]. Ein solches Vorgehen entspricht der Schaffung eines internen Standards zur Bestimmung der Assay-Sensitivität.

Schlussbemerkungen

Die Durchführung von Äquivalenzstudien ist eine Alternative zu traditionellen Überlegenheitsstudien, wenn bereits eine wirksame Behandlung für die untersuchte Indikation existiert. Die methodologischen Anforderungen an Äquivalenzstudien sind jedoch ungleich höher.

Sehr viele Faktoren können dazu beitragen, eine Äquivalenz zu folgern, obwohl sich die Behandlungen in Wirklichkeit voneinander unterscheiden. Wie aussagefähig eine Äquivalenz- bzw. Noninferioritätsstudie ist, hängt massgeblich vom Studiendesign, der Wahl der Positivkontrolle, einer für die Fragestellung ausreichenden Fallzahl, dem festgelegten Äquivalenzbereich, einer adäquaten Auswertung sowie nicht zuletzt von einer korrekten, protokollgerechten Durchführung ab.

In bezug auf die Zulassung neuer Arzneimittel bzw. neuer Indikationen fordern auch die Arzneimittelbehörden explizit eine Begründung für den gewählten Äquivalenzbereich [30, 31] sowie eine ausreichende Assay-Sensitivität [32]. Hat sich die Wirkung von Referenztherapien gegenüber Plazebo als nicht konsistent erwiesen, wird eine dreiarmige, zusätzlich plazebokontrollierte Studie zum behördlich festgelegten Standard, zum Beispiel zur Untersuchung neuer Antidepressiva [33].

Korrespondenz:
Dr. med. Peter Kleist
PFC Pharma Focus AG
Chriesbaumstrasse 2
CH-8604 Volketswil
peter.kleist@pfc.ch

Empfohlene Literatur

– Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ. 1996;313:36–9.
– Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW, for the CONSORT Group. Reporting of noninferiority and equivalence randomized trials. JAMA. 2006;295:1152–60.
– Pocock SJ. The pros and cons of noninferiority trials. Fundamental Clin Pharmacol. 2003;17:483–90.
– Temple R, Ellenberg SS. Placebo-controlled trials and active-controlled trials in the evaluation of new treatments. Part I: ethical and scientific issues. Ann Intern Med. 2000;133:455–63.

Das vollständige Literaturverzeichnis [1–33] finden Sie in der Onlineausgabe dieses Artikels unter www.medicalforum.ch/pdf/pdf_d/2006/2006-37/2006-37-119.pdf.