Methodik und kritische Interpretation psychopharmakologischer Schizophreniestudien

Stefan Leucht und Katja Komossa, München

Die Methodik psychopharmakologischer Studien wird immer komplexer. Gleichzeitig erschweren Sponsor- und Publikationsbias deren Interpretation. In diesem Artikel versuchen wir daher, einige wichtige Aspekte zu Studiendesigns, Randomisierung, Verblindung, Fallzahlen, Ein- und Ausschlusskriterien, psychiatrischen Skalen, Vergleichssubstanzen und Dosierungen, statistischen Methoden und Maßzahlen, Publikationsbias und Darstellung von Ergebnissen zu erklären. Ziel ist es, klinisch tätigen Psychiatern das Lesen randomisierter Schizophreniestudien zu erleichtern.

Schlüsselwörter: **Randomisiert-kontrollierte Studien, Verblindung, Methodik**

Psychopharmakotherapie 2006;13:231–40.

Die Übertragung von Forschungsergebnissen psychopharmakologischer Studien in die Praxis wird durch eine Reihe von Faktoren erschwert. So wird die Methodik solcher Studien immer komplexer. Gleichzeitig steigt in der Ära von Computer und Medline die Zahl der Publikationen so sehr, dass sie für den Einzelnen kaum mehr überschaubar ist (**Abb. 1**). Die Beurteilung wird noch dadurch erschwert, dass psychopharmakologische Studien heute in aller Regel von pharmazeutischen Unternehmen durchgeführt werden. Diese Untersuchungen sind zwar unter vielen Aspekten exzellent und oft hochwertiger als von Universitäten durchgeführte Studien, sie unterliegen jedoch einem hohen wirtschaftlichen Druck. Dies führt oftmals zu Mängeln im Design und in der Darstellung der Ergebnisse. So fanden beispielsweise Heres et al. [10], dass bei Direktvergleichen atypischer Antipsychotika in 90 % der Fälle immer das Präparat des Sponsors überlegen war. Dieser Sponsorbias ist ein universelles Phänomen, das in keiner Weise psychiatriespezifisch ist [2, 17, 21, 29].

Vor diesem Hintergrund möchten wir in diesem Artikel eine Auswahl an Begriffen und Methoden erklären, die bei randomisierten Schizophreniestudien häufig angewandt werden. Ziel ist es, klinisch tätigen Psychiatern eine Hilfestellung beim Lesen und Interpretieren solcher Publikationen zu geben.

Die Auswahl erhebt keineswegs einen Anspruch auf Vollständigkeit. Die dargestellten Erläuterungen beziehen sich auf Probleme, auf die die Autoren in den letzten Jahren intensiver Beschäftigung mit solchen Studien immer wieder gestoßen sind. Dem Leser wird empfohlen, sich bei Interesse für detailliertere statistische Erklärungen in gängigen basisepidemiologischen und statistischen Lehrbüchern zu belesen [9].

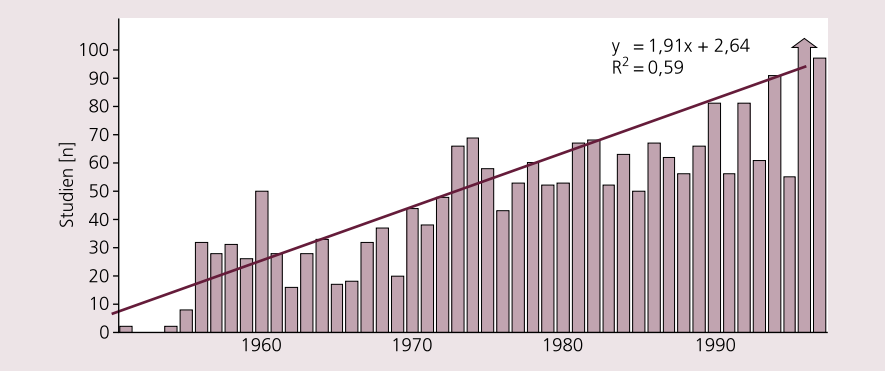

Abb. 1. Anzahl der seit 1950 jährlich publizierten randomisierten Schizophreniestudien [nach 38]

Studiendesign

Die Wahl des geeignetsten Studiendesigns hängt von verschiedenen Faktoren ab, wie beispielsweise der Prävalenz und Inzidenz einer Erkrankung, der Hypothese und bereits vorangegangenen Untersuchungen. Im Folgenden soll auf einige in der Psychopharmakologie wichtige Studiendesigns kurz eingegangen werden.

Priv.-Doz. Dr. med. Stefan Leucht, Katja Komossa, Klinik für Psychiatrie und Psychotherapie der TU München, Klinikum rechts der Isar, Ismaninger Str. 22, 81675 München, E-Mail: Stefan.Leucht@lrz.tum.de

Fallberichte oder Fallserien

Sie können in frühen Stadien der Entwicklung einer neuen Substanz nützlich sein, um Hypothesen zu generieren. Fallberichte oder Fallserien lassen jedoch eine Kontrollgruppe vermissen, die die neue Substanz mit bereits existierenden vergleicht, somit wird der Zeit als möglichem Einflussfaktor des Verlaufs einer Erkrankung keine Beachtung geschenkt. Um den Spontanverlauf mit zu erfassen, benötigt man eine Kontrollgruppe.

Fall-Kontroll-Studien

Die Aufgabe von Fall-Kontroll-Studien ist es, Risikofaktoren für bestimmte Erkrankungen oder Nebenwirkungen spezifischer Substanzen herauszufinden. Die Studienpopulation wird hierbei unterteilt in Personen, die von der Erkrankung oder dem Nebeneffekt betroffen sind, und diejenigen, die nicht betroffen sind. Kontrollen können hierbei abgeglichen auf Alter und Geschlecht zugeordnet werden. Potenzielle Einflussfaktoren – unter anderem die Behandlung –, die mit der Erkrankung oder Nebeneffekten assoziiert sind, werden retrospektiv ermittelt und zwischen den Gruppen verglichen, um Faktoren, die in Verbindung mit der Substanz stehen, zu identifizieren. Fall-Kontroll-Studien sind relativ kostengünstig und nützlich zum schnellen Erfassen von Trends und zur Formulierung neuer Hypothesen. Der Nachteil dieses Designs ist die potenzielle Beeinflussbarkeit der Ergebnisse durch verschiedenste Faktoren.

Randomisiert-kontrollierte Studien

Weitgehend unumstritten ist, dass das am besten geeignete Design zur Evaluation neuer Behandlungsmethoden die doppelblinde randomisierte kontrollierte Arzneimittelstudie („randomised controlled trial", RCT) ist. Diese vergleicht die Fähigkeit zweier Behandlungsmethoden (oder eine Behandlungsmethode und ein Plazebo) in Bezug auf das Erreichen eines vorher definierten klinischen Ziels. Der Hauptunterschied zu oben genannten Studiendesigns besteht in der aktiven Rolle des Untersuchers, der die Behandlung einer definierten Population zuteilt. Die wichtigsten Voraussetzungen für die Güte von RCT sind die Sicherstellung wirklicher Randomisierung, Verblindung auf verschiedenen Ebenen, Erlangen der notwendigen Fallzahl sowie die Definition der Interventions- und der Kontrollgruppe.

Randomisierung

Randomisierung ist eine Prozedur, die versucht sicherzustellen, dass Patienten *zufällig* dem einen oder anderen Studienarm zugeteilt werden. Wenn eine ausreichende Anzahl von Patienten randomisiert ist, sollten sich beide Studienarme idealerweise, abgesehen von der zu untersuchenden Behandlung, nicht voneinander unterscheiden. Unbekannte Faktoren, die einen möglichen Einfluss auf das Studienergebnis haben könnten, sollten somit gleich auf beide Gruppen verteilt sein und deren Effekte sich in der Analyse somit ausgleichen. Gleichzeitig können unbewusste oder bewusste Erwartungen bezüglich des Ergebnisses einer Behandlung seitens des Untersuchers durch korrekte Randomisierung vermieden werden.

Methodisch unzureichende Randomisierung kann die Ergebnisse einer Studie deutlich beeinflussen. Dies belegte eindrücklich eine Untersuchung, nach der es mit zunehmender Qualität der Randomisierung zu immer weniger statistisch signifikanten Ergebnissen kam [4].

Im Prinzip ist das Werfen einer Münze oft eine ausgezeichnete Art der Randomisierung, vorausgesetzt man wirft nur einmal [31]. Andere, nicht perfekte Vorgehensweisen sind die Zuordnung von Patienten je nach Eintreffen im Krankenhaus oder je nach Geburts-Wochentag. Diese Methoden bezeichnet man als „Quasi-Randomisierung", da beide Zuteilungskriterien beeinflusst sein könnten, beispielsweise durch eine unentdeckte Epidemie in einer Population. Heutzutage oft angewandte valide Randomisierungsmethoden sind Computerprogramme, die Randomisierungstabellen erstellen. Diese Tabellen werden dann von einer unabhängigen, nicht in die Studie involvierten Person verwaltet, die den Studienärzten nach dokumentiertem Einschluss eines Patienten in die Studie den zugeordneten Studienarm mitteilt.

Da die Randomisierung einen Schlüsselaspekt klinischer Studien darstellt, sollte die verwendete Methodik immer in den Publikationen dargestellt werden. Häufig wird die Randomisierungsmethode aber kaum beschrieben. Nach unserer Erfahrung mit der Erstellung von Cochrane Reviews (s. u.), in denen auf die Randomisierungsmethode recht genau eingegangen wird, findet sich in aller Regel leider nur die unzureichende Beschreibung „die Gruppenzuteilung erfolgte randomisiert".

Verblindung

Verblindung ist eine Prozedur, die versucht zu verhindern, dass Patienten, Untersucher oder Studienärzte wissen, welcher Patient welche Therapie erhält, damit die Ergebnisse einer Studie dadurch nicht beeinflusst werden. Die Verblindung beruht darauf, den Untersucher sowie den Patienten im Ungewissen über die zugeteilte Behandlungsmethode zu lassen, bis die Studie abgeschlossen ist. Wenn möglich, sollte eine Verblindung *aller* in die Studie einbezogenen Personen, also der Patienten, der betreuenden Ärzte, der Rater und der die Studien auswertenden Statistiker stattfinden. *Doppelblind* bedeutet, dass sowohl der Patient als auch der Beurteiler im Rahmen einer Studie verblindet sind. In *einfach verblindeten* Studien kennt nur eine Seite den Behandlungsarm nicht. Einfachblind ist eine Studie auch, wenn zwar sowohl der behandelnde Arzt als auch der Patient wissen, welcher Behandlungsgruppe der Patient zugeteilt ist, die Ratings aber von verblindeten Personen durchgeführt werden.

Die Verblindung erschweren können Medikamente, die *charakteristische Nebenwirkungen* haben, aus denen man dann leicht die Gruppenzuteilung erraten kann. Beispiele sind gastrointestinale Phänomene unter Antidementiva und extrapyramidal-motorische Nebenwirkungen (EPS) unter Haloperidol. Eine Möglichkeit, dies in Schi-

zophreniestudien zu verhindern, ist die prophylaktische Gabe von Antiparkinson-Medikation, um EPS in der Haloperidol-Gruppe zu vermeiden [28]. Diese Methode wird leider viel zu selten angewandt. Eine Möglichkeit abzuschätzen, wie gut die Verblindung war, ist es, die Rater am Ende angeben zu lassen, in welcher Gruppe der Patient sich ihrer Meinung nach befand.

Erforderliche Fallzahlen

Die Fallzahl einer Studie definiert die Robustheit des Ergebnisses. Je höher die Fallzahlen, desto akkurater sind die Ergebnisse und schmaler ist das Konfidenzintervall, ein Messwert für die mögliche Variabilität der Ergebnisse. Je kleiner der Unterschied der Wirksamkeit zweier Substanzen ist, desto höher ist die Fallzahl, die benötigt wird, um einen statistisch signifikanten Unterschied zwischen zwei Interventionen überhaupt darstellen zu können. Statistisch gesprochen ist dies das Problem der statistischen „Power", also der Wahrscheinlichkeit, einen signifikanten Unterschied zu messen, falls ein solcher besteht. Zum Beispiel ist die Wirkungsdifferenz zwischen einem Antipsychotikum und einem Plazebo so groß, dass nur eine relativ kleine Anzahl von Patienten benötigt wird, um in einer Studie einen statistisch signifikanten Unterschied zu finden. Als grobe Schätzung würden wir sagen, dass hier in der Regel 50 Patienten in jeder Gruppe ausreichten. Beim Vergleich zweier wirksamer Antipsychotika ist eine höhere Patientenzahl notwendig, möchte man einen Wirksamkeitsunterschied recht kleinen Ausmaßes erkennen.

Die Stichprobengröße sollte *vor* Beginn der Studie festgelegt werden. Hierfür stehen Computerprogramme zur Verfügung, in denen man den Unterschied zwischen den zu vergleichenden Interventionen in früheren Studien heranzieht oder – falls es solche Voruntersuchungen nicht gibt – eingibt, was ein klinisch relevanter Unterschied wäre. Für die viel zitierte CATIE-Studie, in der vier verschiedene atypische Antipsychotika miteinander verglichen wurden, kam man hierbei auf eine Größe von 300 Patienten pro Gruppe [18].

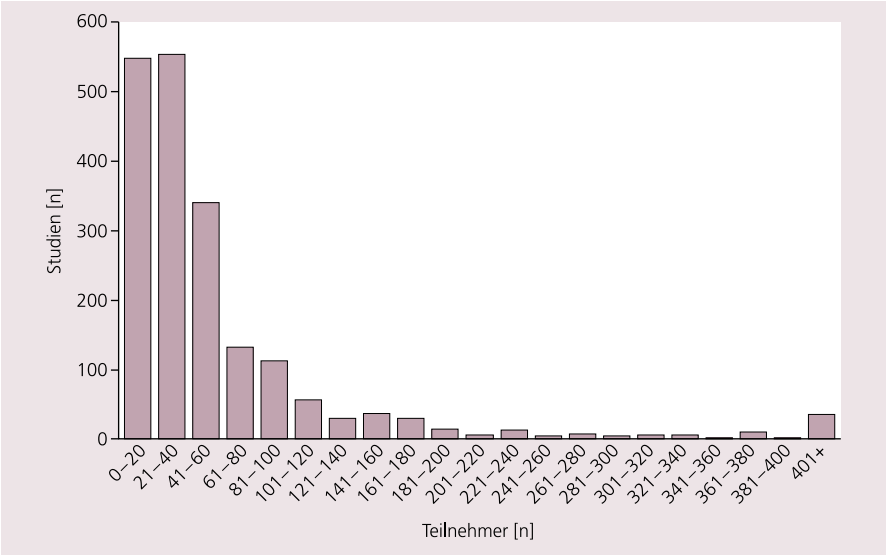

Abb. 2. Die Größe randomisierter Schizophreniestudien (n = 2 000) [38]

Leider erreichen randomisierte Schizophreniestudien die erforderlichen Fallzahlen meist nicht. So analysierten Thornley und Adams [37] im Jahr 1998 2 000 Schizophreniestudien und fanden, dass diese meist nicht mehr als 60 Teilnehmer hatten (**Abb. 2**). Es ist extrem schwierig und teuer, Studien mit hohen Fallzahlen zu organisieren. Dies ist ein Vorteil der von der pharmazeutischen Industrie durchgeführten Studien, die über die dafür notwendigen Ressourcen verfügen.

Ein weiteres Problem hinsichtlich der notwendigen Stichprobengröße zum Nachweis eines statistisch signifikanten Effekts sind die hohen *Studienabbrecherquoten*. Eine Analyse aller seit 1950 durchgeführten randomisierten Antipsychotikastudien ergab, dass sich in Kurzstudien von nur sechs bis zehn Wochen Dauer oft vorzeitige Abbruchraten von 50 % finden. Die kürzlich berichtete Abbruchrate von 74 % in der CATIE-Studie unterstreicht dieses Problem nur noch [18]. Diese hohen Abbruchraten erschweren auch die statistische Analyse.

Um die Fallzahl zu vermindern, wird oft ein *Überkreuzdesign* („Cross-over-Studie") verwendet, in dem dieselben Patienten zunächst in dem einen und dann in dem anderen Therapiearm behandelt werden und vice versa. Ein methodisches Problem bei Überkreuzdesigns sind aber *Überhangeffekte* der ersten Therapieperiode in die nächste, insbesondere wenn die medikationsfreie Auswaschphase zwischen dem Crossover nicht lang genug ist. Denn ein einmal deutlich gebesserter schizophrener Patient verschlechtert sich in einer kurzen Plazebo-Phase nicht unbedingt sofort wieder.

Ein- und Ausschlusskriterien, interne versus externe Validität, Wirksamkeit versus Effektivität

Enge Einschlusskriterien (z. B. nur Schizophrenie nach DSM-IV ohne Vorbehandlung, mit einem definierten Schweregrad, ohne gleichzeitigen Substanzabusus, nicht älter als 50 Jahre) zielen darauf ab, eine gut definierte Patientengruppe mit ähnlichen Eigenschaften untersuchen zu können. Dies vereinfacht das Auffinden von Unterschieden zwischen verschiedenen Gruppen deutlich, da die statistische Variabilität der Ergebnisse recht gering sein wird. Das Problem solcher Studien ist aber die möglicherweise *geringe Übertragbarkeit* auf die Routinearbeit (in der z. B. viele schizophrene Patienten Cannabis konsumieren). Wählt man großzügige Einschlusskriterien, mag die Analyse stärker anzuzweifeln sein und man benötigt wegen der höheren Variabilität

größere Fallzahlen, das Studienergebnis mag aber repräsentativer sein.

Methodisch gesprochen geht es hierbei um die Balance zwischen *interner* und *externer Validität* von Studien. Während die Wahl strikter Einschlusskriterien zur *internen Validität* einer Studie beiträgt (d.h., die Studie ist in der Lage, das zu messen, was man messen möchte), reduzieren derart strikte Einschlusskriterien gewöhnlich die *externe Validität* (die Übertragbarkeit der Ergebnisse auf die Routine). Das Konzept von innerer und äußerer Validität steht auch mit den aktuell viel diskutierten Begriffen *Wirksamkeit* („efficacy") und *Effektivität* („effectiveness") in Beziehung. Diese Begriffe sind nicht klar definiert, jedoch bedeutet die „efficacy" eines Medikaments seine Wirksamkeit auf die Gesundheit einer optimalen Population innerhalb einer kontrollierten Studie. Die „effectiveness" eines Medikaments beschreibt den Einfluss auf die Gesamtpopulation unter realen Bedingungen und zieht nicht nur die Wirksamkeit, sondern eine Menge anderer Aspekte, darunter Verträglichkeit, aber beispielsweise auch soziale Funktionsfähigkeit oder Kosten in Betracht.

Wirksamkeitsstudien sind bei der Entwicklung eines neuen Produktes essenziell. Die klinische Praxis jedoch kann häufig nur bedingt von solchen Ergebnissen profitieren. So wurde für Schizophreniestudien gezeigt, dass nur etwa 10 bis 15 % der Patienten, die die strengen Einschlusskriterien erfüllen, in solche Studien randomisiert werden [10, 26]. Dies schränkt die Übertragbarkeit der Ergebnisse in die Praxis erheblich ein. Deswegen werden heutzutage immer mehr so genannte *pragmatische Studien* („pragmatic trials", „large and simple trials", „effectiveness studies" – auch hier ist die Terminologie nicht eindeutig) gefordert. Diese rekrutieren große Patientenzahlen mit breitgefassten Einschlusskriterien, um die klinische Realität abzubilden. Sie untersuchen ferner einfache, klinisch intuitiv verständliche Outcome-Parameter wie beispielsweise den globalen klinischen Eindruck der Patienten oder Ärzte, Behandlungsdauer, Abbruchraten oder Hospitalisierungsraten. Diese sind leichter verständlich und besser auf den Alltag übertragbar als Mittelwerte von komplizierten psychiatrischen Skalen und daher für die Routinebehandlung nützlicher.

Ein Problem besteht aber darin, dass diese Outcomes nicht „validiert" sind. So war zum Beispiel der primäre Outcome in der CATIE-Studie (einer im Übrigen nur sehr bedingt pragmatischen Studie) der vorzeitige Therapieabbruch egal aus welchem Grund. Es ist von Arzt zu Arzt unterschiedlich, wann und aus welchen Gründen er einen Patienten aus einer Studie ausscheiden lässt.

Ein exzellentes Beispiel für pragmatische Studien sind eine Reihe von Studien über die Sedierung hoch-akuter, angespannter schizophrener Patienten, wie sie vor kurzem unter der Schirmherrschaft der Cochrane Collaboration in Rio de Janeiro und in Indien durchgeführt wurden. In der TREC Collaboration Group wurden beispielsweise 2003 innerhalb nur eines halben Jahres 300 Patienten entweder in einen Midazolam- oder einen Haloperidol-Promethazin-Arm randomisiert. Es gab so gut wie keine Ausfälle („drop-outs"), Hauptergebnis war, dass deutlich mehr Patienten in der Midazolam-Gruppe nach 20 Minuten ausreichend sediert waren [39].

Ein allgemeines Problem randomisierter Studien besteht darin, dass heutzutage jeder Teilnehmer aus ethischen Gründen in der Regel eine *schriftliche Einwilligungserklärung* unterschreiben muss. Solche Formulare sind insbesondere in von der Industrie organisierten Prüfstudien oft mehrere Seiten lang und von einem schwer erkrankten schizophrenen Patienten mit ausgeprägten formalen Denkstörungen kaum zu verstehen. Dies beeinträchtigt die Übertragbarkeit der Ergebnisse auf die klinische Praxis ganz erheblich.

Wie auch immer die Übereinkunft für die Definition von Einschlusskriterien in einer Studie sein mögen, muss der Leser darauf achten, ob die Autoren den Rekrutierungsprozess im Rahmen einer Studie genau beschrieben haben, inklusive der Anzahl und Gründe von „drop-outs". Viele Zeitschriften machen die Unterbreitung eines so genannten CONSORT-Statements (Consolidated standard of reporting trials) [22] auch zur Bedingung für eine Publikation.

Psychopathologische Messinstrumente

Ratingskalen, wie die Positive and Negative Syndrome Scale (PANSS) oder die Brief Psychiatric Rating Scale (BPRS), werden häufig dazu verwendet, die Verbesserung des psychopathologischen Zustands der Patienten zu messen. Der Vorteil dieser Skalen ist, dass ihre psychometrischen Eigenschaften sehr gut untersucht sind. Nicht auf ihre Reliabilität und Validität hin untersuchte Skalen können hingegen Studienergebnisse verfälschen [20]. Auf der anderen Seite ist aber die Interpretation solcher Skalen nicht immer einfach [15]. Was bedeutet beispielsweise die häufig verwendete Response-Definition „mindestens 20 % Reduktion des PANSS-Baseline-Scores" aus klinischer Sicht? Was bedeutet eine Differenz von zwei Punkten im PANSS-Summenscore am Ende einer Studie zwischen zwei Antipsychotika? Oder was bedeutet ein mittlerer BPRS-Summenscore von 60 Punkten?

Erstaunlicherweise wurden erst vor kurzem Untersuchungen durchgeführt, die sich mit der Beantwortung solcher Fragestellungen befassten [12, 13]. Hier wurden die PANSS- und die BPRS-Scores großer Patientenzahlen (mehrere Tausend) mit parallel erhobenen „Clinical Global Impressions" (CGI) [7] in Verbindung gesetzt. Die CGI sind zwei ganz einfache Skalen, bei denen der Arzt den *Zustand* des Patienten und seine *Veränderung* im Vergleich zum Studienbeginn auf einer Skala von 1 bis 7 (von gesund bis schwerst krank bzw. von sehr viel gebessert bis sehr viel verschlechtert) einschätzen muss. Die Ergebnisse zeigten zum Beispiel, wie in **Abbildung 3** und **4** dargestellt, dass „mindestens 20 % Reduktion des PANSS/BPRS-Baseline-Scores", ein häufig verwendeter Cut-off zur Feststellung des Ansprechens (Response), nur „minimal gebessert" nach klini-

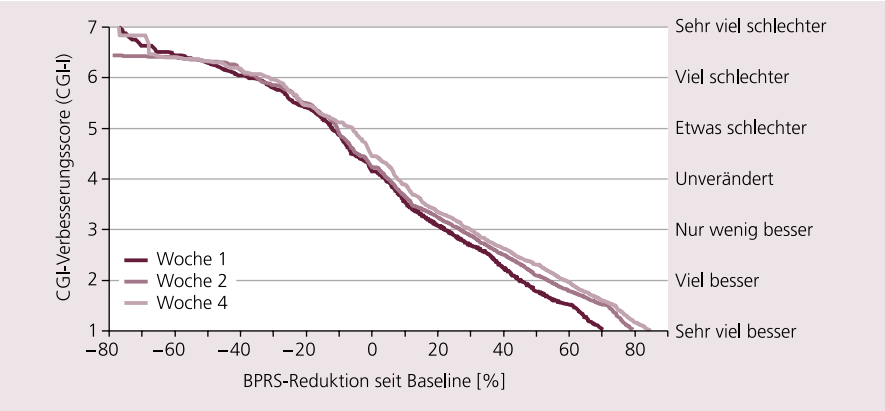

Abb. 3. Welcher CGI-Verbesserungsscore korrespondiert mit wie viel prozentualer Reduktion der BPRS vom Ausgangswert? [nach 14, 16]

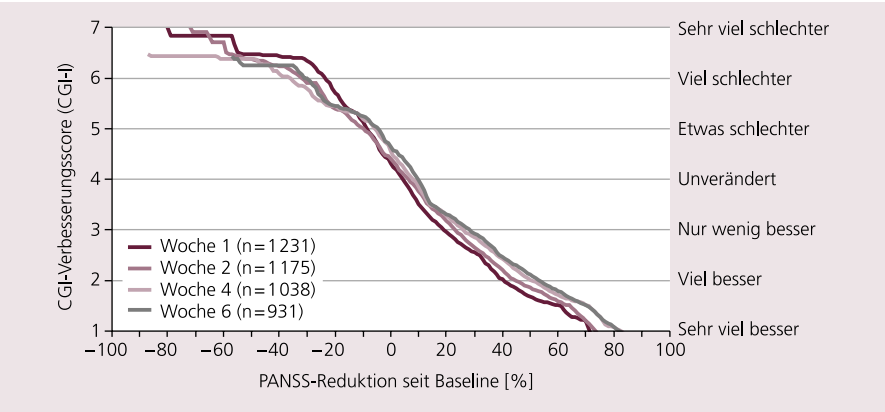

Abb. 4. Welcher CGI-Verbesserungsscore korrespondiert mit wie viel prozentualer Reduktion der PANSS vom Ausgangswert? [nach 14]

schem Eindruck bedeutet. Weitere Ergebnisse sind an anderer Stelle nachzulesen [14–16]. Daher wäre ein Cut-off „mindestens 50 % PANSS/BPRS-Reduktion" sicher sinnvoller, weil er nach klinischem Eindruck „viel besser" bedeutet. Dieser Cut-off wurde aber in letzter Zeit kaum mehr verwendet.

Angesichts dieser Schwierigkeiten bei der Interpretation des PANSS oder der BPRS wundert man sich vielleicht, warum Studien nicht einfach immer die viel intuitiver zu verstehenden Clinical Global Impressions verwenden. Das Problem ist, dass diese Skala nie wirklich auf ihre psychometrischen Eigenschaften hin untersucht worden ist. Ferner können CGI-Ratings deutlich variieren, weil verschiedene Psychiater unterschiedliche Vorstellungen davon haben, was nach CGI beispielsweise als schwer krank einzuschätzen ist. Kürzlich wurde eine verbesserte, schizophreniespezifische Version der CGI entwickelt, die für pragmatische Studien besonders geeignet sein könnte [8].

Hilfreich werden für diesen Zweck auch die Kriterien für eine „Remission" schizophrener Erkrankungen sein, die kürzlich von einer internationalen Arbeitsgruppe [1, 40] definiert wurden. Es ist zu erwarten, dass diese Definition standardmäßig in allen Publikationen verwendet werden wird und daher die Darstellung der Ergebnisse deutlich vereinheitlichen wird. Das Besondere dieser Kriterien besteht darin, dass sie die diagnostischen Kriterien für Schizophrenie nach DSM-IV mit den PANSS-Items verknüpfen, die diese messen. Wenn all diese Symptome allenfalls leicht ausgeprägt sind, spricht man von einer Remission. Von einer stabilen Remission spricht man, wenn sie mindestens sechs Monate lang vorliegt (**Tab. 1**).

Vergleichssubstanz und Dosierung

Wie oben beschrieben, sind Vergleiche mit Plazebo in der Schizophrenieforschung häufig ethisch nicht vertretbar. Bei Vergleichen mit bereits im Handel zugelassenen Medikamenten ist sowohl die Auswahl der idealen Vergleichssubstanz als auch ihrer Dosis schwierig. Ein Auswahlkriterium ist sicherlich, ein Vergleichsmedikament zu wählen, das so häufig eingesetzt wird, dass es als Standard bezeichnet werden kann. In der Schizophreniebehandlung ist *Haloperidol* so ein Medikament, so dass es in fast allen Prüfstudien über die atypischen Antipsychotika als Vergleichspräparat eingesetzt wurde. Seine hohe EPS-Rate macht es allerdings sehr wahrscheinlich, dass eine neue Substanz zumindest in Bezug auf diesen

Tab. 1. Neue, von einer amerikanischen und einer europäischen Arbeitsgruppe erstellte Remissionskriterien [angepasst nach 1 und 40]

Psychopathologie	DSM-IV-Kriterien	Positive and Negative Syndrome Scale (PANSS)	
		Symptom	Item-Nummer
Psychotizismus (Realitätsstörung)	Wahn	Wahn	P1
		Ungewöhnliche Denkinhalte	G9
	Halluzinationen	Halluzinatorisches Erleben	P3
Desorganisation	Desorganisierte Sprache	Konzeptionelle Desorganisation	P2
	Desorganisiertes oder katatones Verhalten	Manierismen	G5
Negativsymptome (psychomotorische Verlangsamung)	Negativsymptome	Verflachter Affekt	N1
		Sozialer Rückzug	N4
		Mangel an Spontaneität	N6

DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, fourth edition

wichtigen Parameter eine Überlegenheit aufweist. Es gab aber auch früher schon Medikamente, die mit weniger EPS als Haloperidol assoziiert waren, in Deutschland beispielsweise Perazin oder Sulpirid. Dies war einer der Gründe, warum in der bereits zitierten CATIE-Studie des NIMH *Perphenazin* und nicht Haloperidol als Vergleichssubstanz verwendet wurde. In der Tat fand sich nur eine geringe Überlegenheit der Atypika hinsichtlich EPS. Auf eine Reihe methodischer Probleme der CATIE-Studie kann hier aus Platzgründen nicht eingegangen werden [24].

Neben der Auswahl der geeigneten Vergleichssubstanz ist deren *Dosierung* entscheidend. Zwei Metaanalysen zeigten einen signifikanten Einfluss der in den Vergleichsgruppen verwendeten Dosierungen auf die Ergebnisse der Vergleiche zwischen Typika und Atypika [6, 13]. Bei Direktvergleichen zwischen Atypika, bei denen die Herstellerfirmen unter enormem Erfolgsdruck stehen und viel zu verlieren haben, wird manchmal nicht die optimale Dosierung des Vergleichmedikaments verwendet. Beispielsweise wurde in einer Studie Ziprasidon (80–160 mg/d) mit Olanzapin (5–15 mg/d) bei akut erkrankten schizophrenen Patienten verglichen [33]. Es fand sich kein signifikanter Wirksamkeitsunterschied, allerdings durfte die offiziell zugelassene Höchstdosis von 20 mg/d Olanzapin nicht gegeben werden. In einer anderen Studie, die vom Hersteller von Olanzapin durchgeführt wurde, fand sich unter Verwendung eines Olanzapin-Dosisbereichs von 5 bis 20 mg/d eine signifikante Überlegenheit von Olanzapin gegenüber Ziprasidon [3]. In der Literatur finden sich zahlreiche weitere Beispiele für diesen *Sponsorbias* [10]. Schon kleine Veränderungen der Dosis können möglicherweise zu größerer oder niedriger Wirksamkeit, zu mehr oder weniger Nebenwirkungen einer Substanz führen und folglich das eine oder andere Medikament begünstigen. Es ist daher für den Leser unerlässlich, zu überprüfen, ob in einer Studie adäquate Dosierungen verwendet wurden.

Statistische Verfahren

Wir können an dieser Stelle nur zwei Punkte aufgreifen, die in letzter Zeit besonders heiß diskutiert worden sind.

Zum einen ist es ein großes Problem, wie man mit den bereits angesprochenen manchmal extrem hohen Drop-out-Raten in randomisierten Schizophreniestudien umgehen soll. Ein häufig verwendetes Vorgehen ist die „last observation carried forward"-(LOCF-) Methode.

Angenommen, eine Studie dauerte sechs Wochen, aber ein Patient fiele nach zwei Wochen aus, dann würde man die Auswertung des Patienten zum Zeitpunkt der zweiten Woche für die Endpunktanalyse verwenden („last observation carried forward to endpoint"). Dass sich der Patient aber bis zum Ende der Studie nicht mehr verändert hätte, ist nur eine Annahme. Bei einem Vergleich atypisches Antipsychotikum versus Haloperidol könnten beispielsweise mehr Patienten in der Haloperidol-Gruppe aufgrund von EPS frühzeitig die Studie beenden. Dies würde bei der LOCF-Methode zu einer – beeinflusst durch die Wirkungsdauer – scheinbar besseren Wirksamkeit des Atypikums führen. Eine andere Auswertestrategie besteht darin, nur Patienten einzuschließen, die die Studie beendet haben („completers only"-Analyse). Neben der Abnahme der Stichprobengröße könnte sich auch die Zusammensetzung der Population durch die Drop-outs deutlich verändern, weil Patienten ja nicht zufällig aus Studien ausscheiden. In letzter Zeit werden daher vermehrt so genannte „mixed effects model"-Analysen in der Statistik verwendet. Vereinfacht gesagt, wird hier der bisherige Verlauf des ausgeschiedenen Patienten in die Berechnungen miteinbezogen. Aber auch bei diesen Methoden gibt es Einschränkungen. Daher ist es bei der Interpretation von RCT wichtig, Anzahl und Gründe für Drop-out-Raten zu berücksichtigen.

Der zweite Punkt, der uns für das Verständnis klinischer Studien aktuell besonders wichtig erscheint, ist das so genannte „non-inferiority"-Design. Lange Zeit wurde in Studien, in denen ein neueres antipsychotisches Medikament mit einem herkömmlichen verglichen wurde, auf statistisch signifikante *Wirksamkeitsunterschiede* zwischen den Substanzen getestet. Diese Vorgehensweise ist oft a priori falsch, da es gar keinen Wirksamkeitsunterschied gibt. Zulassungsbehörden fordern daher in letzter Zeit immer öfter so genannte „non-inferiority"-Designs, das heißt, man muss zeigen, dass ein neues Medikament nicht weniger wirksam ist als ein etabliertes Standardmedikament („non-inferiority"). „Nichtunterlegenheit" ist gegeben, wenn die Wirksamkeit beider Substanzen im selben Bereich liegt. Oft gibt es aber ein Problem mit der Definition, was ein vergleichbarer Wirksamkeitsbereich ist. Wenn a priori ein sehr großer Bereich akzeptiert wird, kann ein Nichtunterlegenheitsnachweis sehr einfach geführt werden.

Systematische Reviews und Metaanalysen

Systematische Reviews und Metaanalysen werden heutzutage international als sinnvolle Methoden der Zusammenfassung randomisierter Studien angesehen, (http://www.cebm.net/levels_of_evidence.asp). Auf eine Reihe methodischer Probleme dieser Methoden können wir aus Platzgründen nicht eingehen und müssen daher beispielsweise auf Maier et al. (2006) [19] verweisen.

Die Cochrane Collaboration (www.cochrane.org) ist eine internationale Non-Profit-Organisation, die es sich zur Aufgabe gemacht hat, systematische Reviews zu medizinischen Therapien zu erstellen, zu verbreiten und auf dem neuesten Stand zu halten. Zunächst ist eine Begriffserklärung notwendig, weil die Ausdrücke „systematischer Review" und „Metaanalyse" oft synonym verwendet werden, aber nicht genau dasselbe meinen. Ein *Review* wird dadurch *systematisch*, dass im Gegensatz zu konventionellen Reviews eine strikte Methodik angehalten wird. Vor Beginn des Reviews wird ein Protokoll geschrieben, in dem die Suchstrategie, die Ein- und Ausschlusskriterien, die zu un-

tersuchenden Outcome-Parameter, und die Methode der Zusammenfassung der Studien genau festgelegt werden. Solch ein systematisches Vorgehen hilft, Fehlerquellen konventioneller Reviews zu vermeiden. Bei letzteren stellt der Reviewer im schlimmsten Fall lediglich einige Studien zusammen, die er kennt, und kommt zu einer mehr oder weniger subjektiven Schlussfolgerung.

Um die Ergebnisse der einzelnen Studien zusammenzufassen, werden in systematischen Reviews häufig *Metaanalysen* durchgeführt. Die Metaanalyse ist ein statistisches Verfahren, mit dem man die Ergebnisse einzelner Studien zur selben Fragestellung zusammenfassen und somit einen mittleren Effekt berechnen kann (s. u.). Unter anderem kann man durch Metaanalysen die Fallzahl und damit die statistische Power erhöhen, um signifikante Unterschiede zwischen zwei Interventionen darzustellen. Eines der methodischen Hauptprobleme von Metaanalysen ist die Frage, wie ähnlich Studien sein müssen, dass ihre Kombination überhaupt gerechtfertigt ist. Dies muss im Einzelfall, am besten durch bereits vorher in einem Protokoll festgelegte Kriterien erfolgen. Gleichwohl liefern Metaanalysen im Gegensatz zu konventionellen Review-Methoden *quantitative Maße* wie p-Wert und Effektstärken (s. u.), die eine im Gegensatz zu konventionellen Reviews objektivere Bewertung der Studienlage erlauben. Dies erklärt oft auch, warum systematische Reviews und Metaanalysen zu konservativeren Schlussfolgerungen als konventionelle Reviews kommen.

Wahrheit oder Zufall

Zwei Maßzahlen helfen die berichteten Ergebnisse zu validieren: der p-Wert und das Konfidenzintervall. Der *p-Wert* beschreibt die Wahrscheinlichkeit, dass ein gemessener Effekt allein durch Zufall entsteht. Wenn die Wahrscheinlichkeit des Zufalls geringer als 5 % (das berühmte $p<0{,}05$) ist, hält man das Ergebnis generell für nicht zufallsverursacht.

Das *Konfidenzintervall* zeigt den Bereich an, innerhalb dessen sich ein erzieltes Ergebnis bewegen kann. Wenn

- 0 keinen Unterschied zwischen zwei Gruppen bedeutet und
- sowohl die untere als auch die obere Grenze des 95%-Konfidenzintervall über 0 liegt oder das komplette 95% Konfidenzintervall unter 0 liegt,

ist der Unterschied zwischen den Gruppen signifikant mit einer Fehlerwahrscheinlichkeit von 5 % ($p=0{,}05$) [9]. Diesen Effekt macht man sich auch bei der Darstellung von Metaanalysen in so genannten Forest-Plots zu Nutze, die auf diese Weise einen raschen grafischen Überblick über die Ergebnisse der einzelnen Studien und ihres mittleren Effekts erlauben (**Abb. 5**). Ein schmales Konfidenzintervall zeigt, dass der berichtete Effekt nahe beim Absolut-Wert liegt und nicht dazu neigt, stark zu variieren. Studien mit einer großen Fallzahl, haben ein schmales Konfidenzintervall und sind daher sehr präzise, während große Konfidenzintervalle erwartet werden können, wenn die Fallzahl klein ist. Bei kleinen Studien kann man nämlich nicht sicher sein, den wahren Unterschied ermittelt zu haben. Der häufig angewendete Cut-off-Score von 5 % (95%-Konfidenzintervall bzw. $p<0{,}05$) ist rein willkürlich, gilt jedoch in der Forschung als Standard, um einen signifikanten Zusammenhang abzubilden. Man könnte ihn ebenso mit 1 %, also einem 99%-Konfidenzintervall bzw. $p<0{,}01$, festlegen.

Allgemein gesprochen zeigt der p-Wert also, ob ein statistisch signifikanter Effekt vorliegt, während das Konfidenzintervall zusätzlich über die Präzision der Ergebnisse informiert (minimaler und maximaler Unterschied, den man erwarten kann, z. B. 95 % aller Ergebnisse liegen innerhalb dieses Bereichs). Ein p-Wert unter 5 % ist Voraussetzung, um Zufall als Ursache der Ergebnisse auszuschließen. Der p-Wert sagt aber nichts

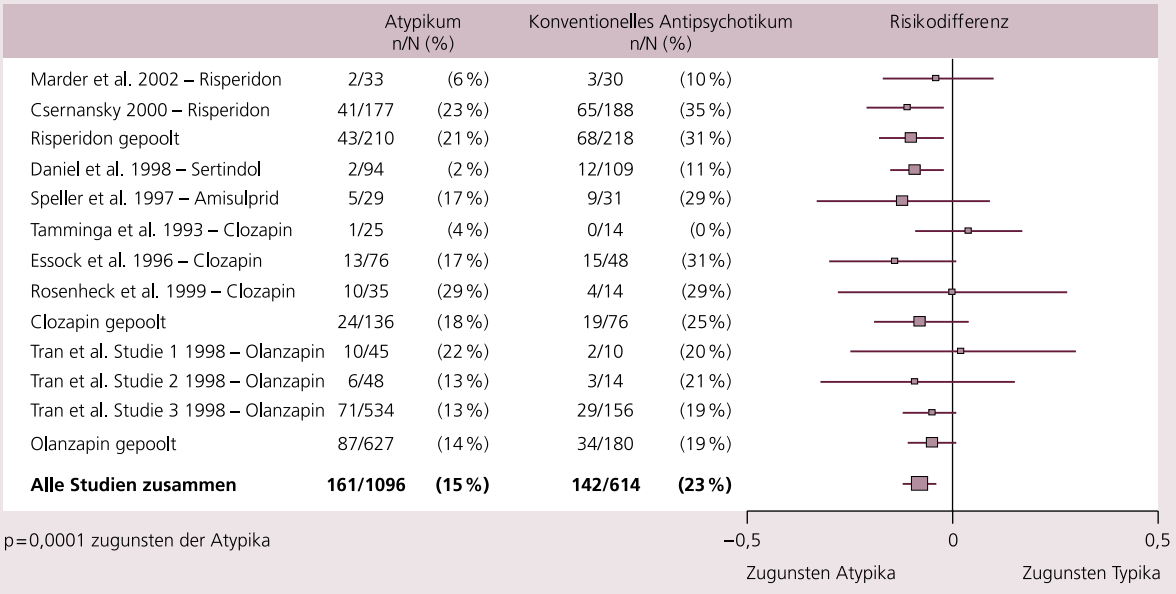

Abb. 5. Rückfallraten – konventionelle Antipsychotika (Haloperidol) vs. Atypika [nach 13]

Tab. 2. Berechnung von Risikomaßen

Risikomaße	Formel	Erklärung
Risiko	a/(a+b)	Anzahl von Patienten mit einem Ereignis geteilt durch die Gesamtzahl
Relatives Risiko	[a/(a+b)]/[c/(c+d)]	Risiko in der Interventionsgruppe geteilt durch das Risiko in der Kontrollgruppe
Risikodifferenz (RD)	[c/(c+d)] – [a/(a+b)]	Risiko in der Kontrollgruppe minus Risiko in der Interventionsgruppe
Odds-Ratio	(a/b)/(c/d)	Odds in der Interventionsgruppe geteilt durch Odds in der Kontrollgruppe
Standardisierter Unterschied	Z.B. Cohen's d oder Hedge's g	Ein Effektstärkenmaß für kontinuierliche Variablen; Ausdruck für den Unterschied zweier Interventionen in Form von Standardabweichung

		Erkrankt		
		+	–	
Exponiert	+	a	b	a+b
	–	c	d	c+d
		a+c	b+c	a+b+c+d

über die *Größe* eines Unterschieds aus. Um diese zu beurteilen, müssen Effektstärken herangezogen werden, auf die wir im Folgenden eingehen.

Effektstärke

Im Gegensatz zum p-Wert, der beschreibt, ob ein Ergebnis zufällig zustande gekommen ist oder nicht, ist die Effektstärke ein Maß für die Größe des Unterschieds zwischen zwei Interventionen. Effektstärken können sowohl für einzelne Studien als auch als Gesamtergebnis von Metaanalysen berechnet werden.

Man kann Effektstärken sowohl für dichotome Maßzahlen (erlitt der Patient eine Rückfall, ja oder nein, Schlaganfall ja oder nein) als auch für kontinuierliche Parameter (z. B. Blutdruck oder Summenscore einer Skala zur Erfassung von Psychopathologie) berechnen. Effektstärkenmaß für *kontinuierliche Parameter* ist in der Regel der *standardisierte mittlere Unterschied* (standardised mean difference, SMD), der mit sich grundsätzlich nur leicht unterscheidenden Formeln wie „Cohen's d" oder „Hedge's g" berechnet wird. Ein standardisierter mittlerer Unterschied im PANSS-Summenscore am Studienende von beispielsweise 0,3 bedeutet, dass eine Gruppe sich um 0,3 Standardabweichungen stärker besserte als die andere. Hier wird schon deutlich, dass die Interpretation der Bedeutung von SMD für die klinische Praxis schwierig ist (was bedeutet ein 0,3 Standardabweichungen großer Unterschied?).

Aus diesem Grund werden heutzutage aus *dichotomen Variablen* berechnete Effektstärken, wenn immer solche vorhanden sind, bevorzugt. Effektstärken für dichotome Variablen sind insbesondere die Risikomaße absolute Risikodifferenz (RD), relatives Risiko (RR), und Odds-Ratio (OR) (**Tab. 2**).

Risiko ist definiert als die Wahrscheinlichkeit, dass das Ergebnis sich einstellt. Haben beispielsweise 23 % der mit Haloperidol behandelten Patienten innerhalb eines Jahres einen Rückfall, so ist das Rückfallrisiko 23 %.

Die *absolute Risikodifferenz* (RD) erhält man durch einfache Subtraktion des Risikos der Interventionsgruppe von dem der Kontrollgruppe. Als Beispiel nehmen wir das Ergebnis einer Metaanalyse über Rückfallraten unter atypischen Antipsychotika im Vergleich zu Haloperidol (**Abb. 5**): Rückfallrisiko Haloperidol (23 %) minus das Rückfallrisiko unter atypischen Antipsychotika (15 %), also 23 % – 15 % = RD 8 %.

Das *relative Risiko* ist nicht die Differenz, sondern das *Verhältnis* der Risiken der beiden Gruppen, also 15 % geteilt durch 23 % = RR = 65 %. Das Risiko eines Rückfalls wurde durch die Verwendung von atypischen Antipsychotika also um 35 % (1 minus 0,65) reduziert.

Das *Odds-Ratio* wurde für Fall-Kontroll-Studien entwickelt, ist aber weniger intuitiv zu verstehen als RD und RR und wird daher zumindest in Metaanalysen heutzutage seltener angewandt (zur Formel der Berechnung von Odds-Ratios siehe **Tabelle 2**).

Das an obigem Beispiel zu erläuternde Problem besteht aber darin, dass obwohl RD und RR auf denselben Zahlen basieren, die Interpretation einer Studie deutlich anders ausfallen wird, je nach dem ob man sich auf RD oder RR bezieht. Eine absolute Risikodifferenz von 8 % ist deutlich weniger beeindruckend als eine relative Risikoreduktion von 35 %. Die absolute Risikoreduktion lässt sich übrigens leicht in die heute häufig zitierte „Number needed to treat" (NNT) umrechnen. Im konkreten Beispiel: Wie viele Patienten müsste man ein Jahr lang mit einem Atypikum anstelle von Haloperidol behandeln, um einen Rückfall zu vermeiden? Diese NNT kann man als Kehrwert der absoluten Risikodifferenz berechnen, also 1 geteilt durch RD, in diesem Fall 1 : 0,08 = 13. Je nachdem, welche Effektstärke man verwendet, wird man ein Ergebnis also völlig anders interpretieren, obwohl beide, sowohl RD (bzw. NNT) als auch RR „richtig" sind. Es ist daher ein Credo unseres Artikels, dass man beim Lesen einer randomisierten Studie nicht einfach nur die Effektstärke oder gar nur den p-Wert ansehen sollte, sondern sich über die zugrunde liegenden Zahlen Gedanken machen muss.

Publikationsbias

Das größte Problem der „Evidence-based Medicine" ist der Publikationsbias, also das Phänomen, dass Studien mit negativem Ergebnis oft nicht publiziert werden, weil zum Beispiel Pharmafirmen kein großes Interesse haben,

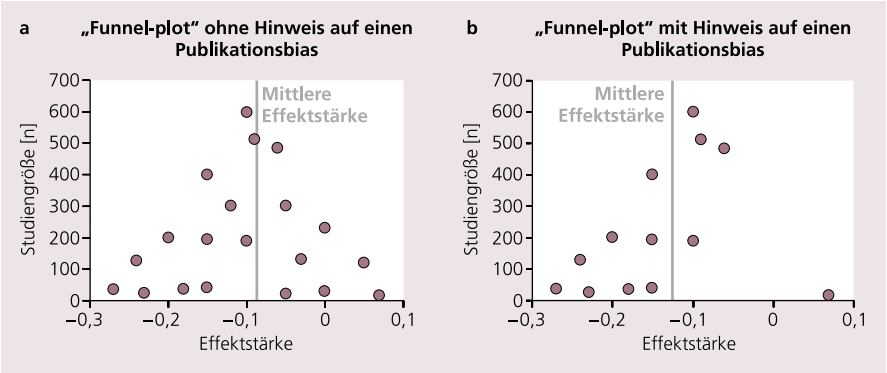

Abb. 6. Funnel-Plots. Die Effektstärken der Einzelstudien wurden auf der x-Achse, die Größe der Studien [n] auf der y-Achse aufgetragen. a) Symmetrisches Bild, d.h. kein Hinweis auf einen Publikationsbias. b) Kein symmetrisches Bild. Für ein symmetrisches Bild fehlen in der Abbildung unten rechts Studien, das wären kleinere Studien mit Ergebnissen zugunsten der Kontrollgruppe. Der Funnel-Plot gibt also einen Hinweis auf einen relevanten Publikationsbias.

Untersuchungen zu veröffentlichen, bei denen sich ihr Medikament nicht als wirksam erwies, oder weil auch Fachzeitschriften lieber positive Studien mit aufregenden neuen Ergebnissen publizieren als negative. Daher ist „Evidence-based Medicine" auch ironisch „Evidence-*biased* Medicine" genannt worden.

Immerhin bieten Metaanalysen Methoden an, um einem solchen Publikationsbias auf die Spur zu kommen. Eine solche Methode ist der „Funnel-Plot", bei dem Effektstärken versus Studiengröße der einzelnen Untersuchungen aufgetragen werden (**Abb. 6a** und **6b**). Hierbei finden sich die großen Studien oben in der Nähe der mittleren Effektstärke, während kleinere Studien um die mittlere Effektstärke streuen. Wurden alle Studien zu einer Fragestellung publiziert, so ist das Bild symmetrisch (**Abb. 6a**). Wurden kleinere Studien mit negativem Ergebnis nicht publiziert, so ist der Plot einseitig asymmetrisch (**Abb. 6b**).

Aber auch diese Methoden sind nur explorativ. Seit neuestem müssen immerhin alle Studien vor Beginn offiziell registriert werden, um später in internationalen Fachzeitschriften publiziert werden zu können. Abschließend ließe sich das Problem des Publikationsbias nur lösen, indem alle Studien publiziert werden müssen, und sei es nur in elektronischer Form (z. B. im Internet).

Zum Schluss etwas ganz Banales – die Darstellung der Ergebnisse

Eine sehr simple Möglichkeit, wie man bei der Interpretation einer Studie verwirrt werden kann, ist die Art und Weise der Präsentation der Ergebnisse. Durch Veränderungen der Skalierung der y-Achse eines Graphen kann ein Effekt größer wirken, als er wirklich ist (**Abb. 7**). Diese Art der Darstellung von Daten in der Forschung ist häufig anzutreffen.

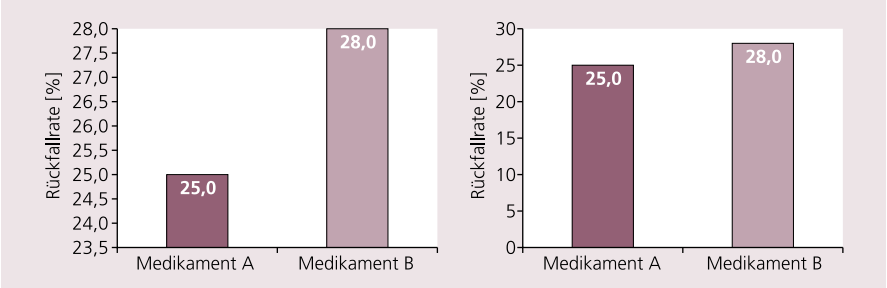

Abb. 7. Die grafische Darstellung von Ergebnissen beeinflusst deren Interpretation – hypothetisches Beispiel. Linke und rechte Abbildung stellen dieselben Ergebnisse dar. Links scheint aufgrund der Skalierung der y-Achse der Unterschied zwischen den Medikamenten aber viel deutlicher zu sein.

Wenn zwei Medikamente miteinander verglichen werden, werden häufig die Vorteile des eigenen Produkts im Abstract hervorgehoben, wie eine kürzlich veröffentlichte Untersuchung an randomisierten Antipsychotikastudien zeigte [10]. Daher ist es sicherlich nicht ausreichend, nur die Abstracts solcher Studien zu lesen. Dennoch sind Anstrengungen erforderlich, diesen unerfreulichen Sponsorbias auszuschalten. Dafür sind diese unter vielen Aspekten hochwertigen und aufwendigen Studien der Pharmaindustrie, ohne die wir weiterhin in der pharmakologischen „Steinzeit" wären, viel zu schade.

Autorenerklärung

Stefan Leucht hat von den folgenden Firmen Vortrags- und/oder Beratungshonorare erhalten: SanofiAventis, BMS, EliLilly, Janssen, Lundbeck, Pfizer. Von SanofiAventis und EliLilly hat er ferner Forschungsmittel erhalten. Katja Komossa hat keinen Interessenskonflikt anzugeben.

Der vorliegende Artikel stellt die Erweiterung eines in englischer Sprache publizierten Manuskripts dar (Leucht S. Translating research into clinical practice: critical interpretation of clinical trials in schizophrenia. Int Clin Psychopharmacol 2006;21(Suppl 2):S1–10.).

Methodology and critical interpretation of psychopharmacological studies in schizophrenia

The methodology of psychopharmacological studies is getting more and more complex. Simultaneously sponsor- and publication bias limit their interpretation. Therefore, this article tries to explain a number of aspects concerning study designs, randomisation, blinding, case numbers, in- and exclusion criteria, psychiatric rating scales, comparator compounds and doses, statistical methods and numbers, publication bias and presentation of results. The aim is to make the reading and interpretation of clinical trials in schizophrenia easier.

Keywords: Randomized controlled trial, blinding, methodology

Literatur

1. Andreasen NC, Carpenter WT, Kane JM, Lasser RA, et al. Remission in schizophrenia: proposed criteria and rationale for consensus. Am J Psychiatry 2005;162:441–9.
2. Baker C, Johnsrud M, Crismon M, Rosenheck R, et al. Quantitative analysis of sponsorship bias in economic studies of antidepressants. Br J Psychiatry 2003;183:498–506.
3. Breier A, Berg PH, Thakore JH, Naber D, et al. Olanzapine versus ziprasidone: results of a 28-week double-blind study in patients with schizophrenia. Am J Psychiatry 2005;162: 1879–87.

4. Chalmers T, Celano P, Sacks H, Smith HJ. Bias in treatment assignment in controlled clinical trials. N Engl J Med 1983;309:1358–61.
5. Eysenbach G. Tackling publication bias and selective reporting in health informatics research: register your ehealth trials in the international ehealth studies registry. J Med Internet Res 2004;6:e35.
6. Geddes J, Freemantle N, Harrison P, Bebbington P. Atypical antipsychotics in the treatment of schizophrenia: systematic overview and meta-regression analysis. BMJ 2000;321:1371–6.
7. Guy W. Clinical Global Impressions. ECDEU assessment manual for psychopharmacology, revised (DHEW Publ No ADM 76-338). Rockville, MD: National Institute of Mental Health, 1976:218–22.
8. Haro JM, Kamath SA, Ochoa S, Novick D, et al. The Clinical Global Impression-Schizophrenia scale: a simple instrument to measure the diversity of symptoms present in schizophrenia. Acta Psychiatr Scand Suppl 2003:16–23.
9. Hennekens C, Buring J. Epidemiology in medicine. Boston/Toronto: Little, Brown and Co., 1987.
10. Heres S, Davis J, Maino K, Jetzinger E, et al. Why olanzapine beats risperidone, risperidone beats quetiapine and quetiapine beats olanzapine again. An exploratory analysis of head-to-head studies on second-generation antipsychotics. Am J Psychiatry 2006;163: 185–94.
11. Hofer A, Hummer M, Huber R, Kurz M, et al. Selection bias in clinical trials with antipsychotics. J Clin Psychopharmacol 2000;20: 699–702.
12. Lambert M, Naber D. Current issues in schizophrenia: overview of patient acceptability, functioning capacity and quality of life. CNS Drugs 2004;18:5–17, 41–3.
13. Leucht S, Barnes TR, Kissling W, Engel RR, et al. Relapse prevention in schizophrenia with new-generation antipsychotics: a systematic review and exploratory meta-analysis of randomized, controlled trials. Am J Psychiatry 2003;160:1209–22.
14. Leucht S, Kane JM, Kissling W, Hamann J, et al. Clinical implications of brief psychiatric rating scale scores. Br J Psychiatry 2005;187: 366–71.
15. Leucht S, Kane JM, Kissling W, Hamann J, et al. What does the PANSS mean? Schizophr Res 2005;24:24.
16. Leucht S, Kane JM, Etschel E, Hamann J, et al. Linking the PANSS, BPRS and CGI: clinical implications. Neuropsychopharmacology. Im Druck.
17. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ 2003;326:1167–70.
18. Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, et al. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. N Engl J Med 2005;353:1209–23. Epub 2005 Sept 19.
19. Maier W, Moller HJ. Meta-analyses – highest level of empirical evidence? Eur Arch Psychiatry Clin Neurosci 2005;255:369–70.
20. Marshall M, Lockwood A, Bradley C, Adams C, et al. Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. Br J Psychiatry 2000;176:249–52.
21. Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine – selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. BMJ 2003;326: 1171–3.
22. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. BMC Med Res Methodol 2001;1:2. Epub 2001 Apr 20.
23. Mold J, Peterson K. Primary care practice-based research networks: Working at the interface between research and quality improvement. Ann Fam Med 2005;3:12–20.
24. Möller HJ. Are the new antipsychotics no better than the classical neuroleptics? The problematic answer from the CATIE study. Eur Arch Psychiatry Clin Neurosci 2005;255: 371–2.
25. Pocock SJ, Collier TJ, Dandreo KJ, de Stavola BL, et al. Issues in the reporting of epidemiological studies: a survey of recent practice. BMJ 2004;329:883.
26. Riedel M, Strassnig M, Muller N, Zwack P, et al. How representative of everyday clinical populations are schizophrenia patients enrolled in clinical trials? Eur Arch Psychiatry Clin Neurosci 2005;255:143–8.
27. Rockhill B, Spiegelman D, Byrne C, Hunter DJ, et al. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. J Natl Cancer Inst 2001;93:358–66.
28. Rosenheck R, Perlick D, Bingham S, Liu-Mares W, et al. Effectiveness and cost of olanzapine and haloperidol in the treatment of schizophrenia: a randomized controlled trial. JAMA 2003;290:2693–702.
29. Safer D. Design and reporting modifications in industry-sponsored comparative psychopharmacology trials. J Nerv Ment Dis 2002; 190:583–92.
30. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408–12.
31. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet 2002;359:515–9.
32. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005;365:1348–53.
33. Simpson GM, Glick ID, Weiden PJ, Romano SJ. Randomized, controlled, double-blind multicenter comparison of the efficacy and tolerability of ziprasidone and olanzapine in acutely ill inpatients with schizophrenia or schizoaffective disorder. Am J Psychiatry 2004;161:1837–47.
34. Sung N, Crowley WJ, Genel M, Salber P, et al. Central challenges facing the national clinical research enterprise. JAMA 2003;289:1305–6.
35. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR. Empirical assessment of effect of publication bias on meta-analyses. BMJ 2000;320:1574–7.
36. The Editors. Clinical trial registration: A statement from the international committee of medical journal editors. N Engl J Med 2004;351:1250–1.
37. Thornley B, Rathbone J, Adams C, Awad G. Chlorpromazine versus placebo for schizophrenia. Cochrane Database Syst Rev 2003;2:CD000284.
38. Thornley B, Adams C. Content and quality of 2000 controlled trials in schizophrenia over 50 years. BMJ 1998;317:1181–4.
39. TREC collaborative group. Rapid tranquillisation for agitated patients in emergency psychiatric rooms: a randomised trial of midazolam versus haloperidol plus promethazine. BMJ 2003;327:708–13.
40. Van Os J, Burns T, Cavallaro R, Leucht S, et al. Standardized remission criteria in schizophrenia. Acta Psychiatr Scand. 2006;113:91–5.
41. Wahlbeck K, Tuunainen A, Ahokas A, Leucht S. Dropout rates in randomised antipsychotic drug trials. Psychopharmacology 2001;155: 230–3.