Metaanalysen: was gilt es zu beachten?

Hans-Peter Volz, Werneck

Metaanalysen sind zu einem der wichtigsten Beurteilungsinstrumente des Effekts von Therapieinterventionen im Rahmen der Evidence-based Medicine (EbM) geworden. Diese Vorrangstellung ist nicht unproblematisch, insbesondere, wenn die den Metaanalysen inhärenten methodischen Probleme unzureichend gewürdigt werden. Demgegenüber ist die Einzelbeurteilung von Studien in den Hintergrund getreten. Allerdings sind bei der Wertung von Metaanalysen eine Reihe methodischer Gesichtspunkte zu beachten, um deren Ergebnisse nicht falsch zu interpretieren. In der vorliegenden Arbeit wird auf zwei ausgesuchte Gesichtspunkte eingegangen: Selektionseffekte und sogenannte Multi-Treatment-Comparisons. Insgesamt sollten Aussagen von Metaanalysen aufgrund vielfältiger Einflüsse methodischer Faktoren keinesfalls unkritisch übernommen werden.

Schlüsselwörter: **Metaanalysen, Selektionskriterien, Multi-Treatment-Comparisons**

Psychopharmakotherapie 2011;18:131–6.

Metaanalysen sind zu einem wichtigen Instrument der Beurteilung von Therapieeffekten geworden, vielleicht zum wichtigsten überhaupt. Sie gelten als entscheidend, wenn Evidenzgrade festgelegt werden, beispielsweise im Rahmen strukturierter Prozesse zur Erstellung von Leitlinien. Diese Vorrangstellung im Bereich der Evidence-based Medicine (EbM) ist nicht unproblematisch [9], insbesondere, wenn die in den Metaanalysen inhärenten methodischen Probleme unzureichend gewürdigt werden [7]. Unbestritten ist, dass es im Rahmen dieser Analysen möglich ist, große Datenmengen nach einheitlichen Kriterien zu sammeln, zu selektieren und auszuwählen und dann zu wenigen, in der Regel recht einfachen Aussagen zu gelangen.

Wahrscheinlich aufgrund dieser recht einfachen Aussagen werden viele Vorträge und auch Berichte in der Laienpresse – neben der herausragenden Rolle der Metaanalysen im Rahmen der evidenzbasierten Medizin – von Ergebnissen dieser Metaanalysen dominiert, insbesondere wenn diese (scheinbar) neue, nicht bekannte und damit eine Mehrheits- oder Expertenmeinung widerlegende Resultate zeigen. Die Darstellung von Einzelstudien oder die narrative Zusammenfassung und kritische Würdigung von Einzelstudien ist demgegenüber in den Hintergrund getreten.

In der folgenden Übersicht sollen einige kritische Anmerkungen zu Metaanalysen dargestellt werden; dabei geht es nicht um eine Herabwürdigung dieser Methode, vielmehr soll herausgearbeitet werden, was es unbedingt zu beachten gilt, damit Ergebnisse dieser Analysen nicht zu falschen Schlussfolgerungen führen. Das Schwergewicht der Erörterungen liegt hierbei nicht auf methodisch-statistischen Punkten (siehe hierfür z. B. [3, 6]), vielmehr sind diese auf die Selektionskriterien der in die Metaanalysen eingehenden Studien beschränkt; zudem wird (kurz) auf eine relativ neue Variante von Metaanalysen, die Multi-Treatment-Comparisons, eingegangen.

Selektionskriterien der eingeschlossenen Studien

Bei der Auswahl der in eine Metaanalyse einzuschließenden Studien werden in der Regel mehrere Filter, mehrere Selektionskriterien angewandt. So werden bei pharmakologischen Interventionen meist nur randomisierte, kontrollierte, doppelblinde Studien (RCT) eingeschlossen. (An dieser Stelle ist darauf hinzuweisen, dass bei Psychotherapiestudien Verblindungen nicht möglich sind, daher werden dort randomisierte kontrollierte, aber nicht blinde Studien in Metaanalysen eingeschlossen und so wie doppelblinde Studien gewertet; das heißt, es wird an dieser Stelle bewusst mit zweierlei Maß gemessen. Dies kann dazu führen, dass psychotherapeutische Interventionen ein ebenso hohes oder höheres Evidenzniveau als psychophar-

Prof. Dr. H.-P. Volz, Krankenhaus für Psychiatrie, Psychotherapie und Psychosomatische Medizin Schloss Werneck, Balthasar-Neumann-Platz 1, 97440 Werneck, E-Mail: hans-peter.volz@kh-schloss-werneck.de

makologische Interventionen erreichen können, obwohl sie nicht doppelblind durchgeführt worden sind.) Diese Studien werden häufig einer Qualitätskontrolle unterzogen, und es wird vorher festgelegt, welche Qualitätskriterien hierbei erfüllt werden müssen. Entweder begrenzt sich das Material auf die publizierten Studien, oder es werden auch Anstrengungen unternommen, nicht publiziertes Material einzuschließen (um keinem Publikations-Bias zu unterliegen).

Wenn nun RCTs eingeschlossen werden, so ist zu fragen, zu welchem Zweck und mit welchem Design diese durchgeführt wurden. Viele RCTs werden im Rahmen der Zulassung der entsprechenden Substanz durchgeführt. Hierbei handelt es sich häufig um Plazebo-kontrollierte Untersuchungen. Spätestens sobald Plazebo im Rahmen einer Studie verwendet wird, ändern sich die Einschlusskriterien der aufgenommenen Patienten deutlich: so sind beispielsweise im Rahmen von Depressionsstudien suizidale Patienten ausgeschlossen, zu schwer depressive Patienten sind ebenfalls ausgeschlossen. Vor allem aber ist es naiv anzunehmen, dass Patienten, die Plazebo erhalten, ausschließlich diesem Effekt ausgesetzt sind. Vielmehr erhalten auch die sogenannten Plazebo-Patienten ein hohes Maß an Zuwendung, nahezu einer supportiven Psychotherapie gleichkommend, die einen antidepressiven Effekt auch auf diese Patienten ausübt. Diese mitunter als „unspezifische Effekte" bezeichneten Einflüsse sind wahrscheinlich in der Plazebo-Gruppe ausgeprägter, da dort ja im Durchschnitt die Response geringer ausgeprägt ist, also vermehrt supportive Gespräche, Kriseninterventionen und Ähnliches notwendig sind. Zusätzliche Effekte wie Komedikation mit Benzodiazepinen, Schlafmitteln u. a. vermindern zudem den Unterschied zwischen Verum- und Plazebo-Gruppe. Daher kann es insbesondere bei Untersuchungen zur Wirksamkeit von Antidepressiva unmöglich sein, zwischen Verum- und Plazebo-Gruppe in einem RCT überhaupt noch einen Unterschied zu finden. Dies ist in etwa einem Drittel der Fälle von Plazebo-kontrollierten Antidepressiva-RCTs der Fall. Diese werden als *negative Studien* bezeichnet, wenn ein mitgeführter interner Standard (bei Antidepressiva-Studien ein Standardantidepressivum) in der Lage war, eine Plazebo überlegene Wirksamkeit zu demonstrieren (**Abb. 1a**). (Insofern kann bei einer Zwei-Arm-Studie, Verum-Plazebo, die keinen Unterschied zwischen den beiden Gruppen zeigt, nicht zwischen einer negativen und einer „failed" [s. u.] Studie unterschieden werden.) In einem Teil der Fälle der Nichtdifferenzierung von Verum und Plazebo zeigt sich auch zwischen Plazebo und dem mitgeführten internen Standard kein Unterschied, was meist an einer methodisch unzureichenden Studiendurchführung (z. B. dem Einfluss eines hohen Anteils therapierefraktärer Patienten) liegt. Eine solche Studie wird als *fehlgeschlagen („failed")* bezeichnet (**Abb. 1b**). Nun wird niemand in einer Wertung der Potenz eines Antidepressivums eine solche „failed study" einschließen, da diese ja nicht gezeigt hat, dass das zu untersuchende Antidepressivum nicht wirkt, sondern nur das Ergebnis einer methodisch falsch durchgeführten Studie berichtet, das allenfalls Rückschlüsse auf die Methodik, keinesfalls aber auf die Wirksamkeit der untersuchten Substanzen zulässt. In einer Reihe von Metaanalysen wird aber genau dies getan, was zu einer künstlichen Verminderung des Verum-Plazebo-Unterschieds führt.

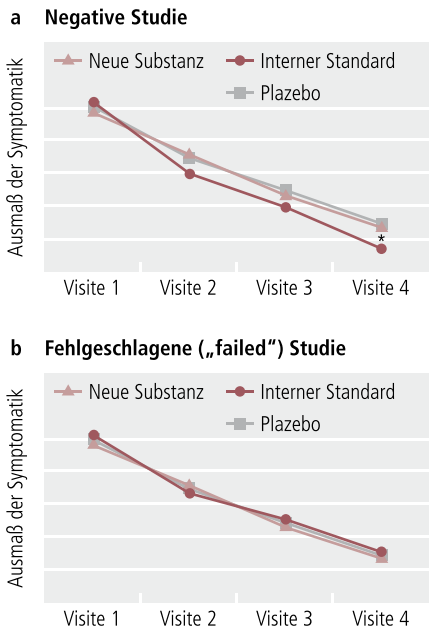

Abb. 1. „Negative" und „fehlgeschlagene" Studien.
a) Das zu untersuchende Antidepressivum zeigt keinen statistisch signifikanten Wirksamkeitsunterschied im Vergleich zu Plazebo, wohl aber der mitgeführte interne Standard (* statistisch signifikanter Unterschied zwischen internem Standard und Plazebo).
b) Weder die zu untersuchende Substanz noch ein mitgeführter interner Standard zeigen statistisch signifikanten Wirksamkeitsunterschied im Vergleich zu Plazebo.

Mitunter werden auch Ergebnisse von Dosisfindungsstudien eingeschlossen, und zwar jeder Dosisarm für sich betrachtet. Nun wird aber bei Dosisfindungsstudien die niedrigste untersuchte Dosis in der Regel so gewählt, dass sie sich vom Plazebo möglichst nur soweit differenzieren soll, dass die höheren untersuchten Dosisarme eine Chance für eine stärkere Differenzierung erhalten. Bei einem solchen Vorgehen ist es geradezu zu erwarten, dass in dieser niedrigen Dosisstufe allenfalls geringe, mitunter auch überhaupt keine Unterschiede im Vergleich zu Plazebo auftreten. (Besonders deutlich wird diese Problematik, wenn eine Gruppe ein sogenanntes Pseudo-Plazebo erhält; hierunter wird eine sehr niedrige Dosis des Verums verstanden, unter der kein oder nur ein sehr geringer Behandlungseffekt erwartet wird.) Insofern wird, wenn solche niedrigen Dosierungsarme in eine Metaanalyse eingeschlossen werden, eine insgesamt geringere Differenzierung Verum/Plazebo bewusst in Kauf genommen. Der Einschluss solcher Studien muss erneut zu einer falschen, weil zu niedrigen, Beurteilung des Plazebo-Verum-Unterschieds führen. Dabei ist es bemerkenswert, dass ergänzende Empfindlichkeits-(„sensitivity"-)Analysen im Rahmen solcher Metaanalysen diesbezügliche Dosiseffekte meist nicht erkennen; dies liegt unter anderem an der Komplexität von Dosis-Wirkungs-Beziehungen in der Psychopharmakologie (linear vs. kurvilinear [U-shaped bzw. reversed-U-shaped]).

In Metaanalysen eingeschlossene Dosisfindungsstudien können noch auf andere Weise zur Unterschätzung des

Effekts einer Substanz führen. Solche Studien müssen aus methodischen Gründen mit fixen Dosen in den einzelnen Gruppen durchgeführt werden. Die Patienten werden auf diese unterschiedlichen Gruppen randomisiert verteilt. Damit ist aber auch immanent, dass die Patienten meist nicht die für sie optimale Dosis erhalten können. (Dieser Effekt trifft auch für andere mit einer fixen Dosis durchgeführte Studien zu.) So erhält ein Patient zufällig die niedrigste Dosis, eine Anpassung ist nicht möglich; in der klinischen Praxis hätte man ihm eventuell von vorneherein eine höhere Dosis gegeben oder die Dosis im Laufe der Behandlung erhöht. Ein anderer erhält aufgrund der zufälligen Gruppenzuteilung die höchste der untersuchten Dosierungen und respondiert aufgrund von Unverträglichkeit schlechter, auch hier wäre es in der klinischen Praxis ein leichtes gewesen, die Dosis zu reduzieren und somit den Pharmakoneffekt zu erhöhen. Werden solche Studien wie Hauptwirksamkeitsstudien in einer Metaanalyse behandelt, führt dies insgesamt zu einer Unterschätzung des erzielbaren Behandlungseffekts.

Beispielgebend soll an dieser Stelle auf die letzthin so häufig zitierte Metaanalyse von Kirsch et al. [4] eingegangen werden. Diese Metaanalyse zur Wirksamkeit von Antidepressiva fand zwischen Verum und Plazebo relativ geringe Mittelwertsunterschiede in der Größenordnung von 1,8 Punkten auf der Hamilton-Depressionsskala (HAMD). Bei niedrigen Depressionsgraden war der Wert noch geringer, bei schwerer ausgeprägter Depression nahm er auf etwa vier Punkte zu. Von den Autoren wurde das Gesamtergebnis als Hinweis für eine klinisch nicht relevante Antidepressiva-Wirksamkeit gewertet (siehe unten). In diese Metaanalyse wurden Studien von vier modernen Antidepressiva (Fluoxetin, Venlafaxin, Nefazodon, Paroxetin, von den Autoren als SSRI [selektive Serotonin-Wiederaufnahmehemmer] bezeichnet) eingeschlossen, die Hauptdatenbasis entstammt der US-amerikanischen Zulassungsbehörde FDA (Food and Drug Administration), bei der die Studien aus Zulassungsgründen hinterlegt wurden. Es finden sich in dieser Publikation keine Hinweise auf eingeschlossene fehlgeschlagene oder negative Studien, auch keine auf unterschiedliche Dosierungsarme. Insofern können die oben erwähnten, zu falschen Aussagen führenden Einflussfaktoren nicht identifiziert werden und es kann nicht entschieden werden, ob der antidepressive Effekt der in diese Metaanalyse eingeschlossenen Antidepressiva nicht unterschätzt wurde.

Die Autoren kombinieren dann ihre Ergebnisse der Metaanalyse mit Betrachtungen zur klinischen Relevanz des gefundenen Unterschieds. Hier wird arbiträr nur ein Kriterium, jenes des NICE (National Institute of Clinical Excellence, dem britischen Pendant des IQWiG), verwandt (standard mean difference $\geq 0{,}5$). Dieses Kriterium ist keinesfalls so unumstritten, wie es in der Publikation dargestellt wird, bei der Zulassung von Antidepressiva gelten auf alle Fälle andere Kriterien. Da es sich bei den in diese Analyse eingeschlossenen Studien wohl größtenteils um Zulassungsstudien handelt, muss ein anderer Punkt hier noch zusätzlich Erwähnung finden: Das Ziel von Zulassungsstudien ist, unter artifiziellen Bedingungen eine Überlegenheit von Verum versus Plazebo zu finden. Diese Voraussetzung der Studie mit ihren methodischen Implikationen schließt ein, dass eine hohe Plazebo-Response ausgelöst wird, die in den letzten Jahren immer weiter angestiegen ist. Schon diese Tatsache zeigt, dass bei Zulassungsstudien mittlerweile der Verum-Plazebo-Unterschied relativ gering geworden ist, wohl nicht, weil die Antidepressiva schwächer wirksam geworden wären, sondern weil die Durchführung der Zulassungsstudien (u. a. wegen des hohen Rating-Aufwands) eine immer höhere Plazebo-Response begünstigt. Trotzdem ist es gelungen, eine Reihe neuerer Antidepressiva zuzulassen, da Einzelstudien einen Verum-Plazebo-Unterschied zeigen konnten. Werden aber dann alle Daten in der Weise, wie es bei Kirsch et al. [4] zu vermuten gilt, eingeschlossen, verwischt der Plazebo-Verum-Unterschied noch stärker. Dann zu schlussfolgern, es bestünde in der Realität kein solcher Unterschied, lässt den Zweck, zu welchem diese Studien durchgeführt wurden und wodurch auch das Ergebnis zum Teil mitbestimmt wurde, vollkommen unberücksichtigt. Im Übrigen dürfte es einen erfahrenen Kliniker oder auch Kenner von klinischen Studien kaum verwundern, dass die Verum-Plazebo-Differenz mit steigendem initialem Depressionsschweregrad deutlicher wird oder überhaupt erst entsteht; dies ist ein lange bekannter Effekt [8]. Die Bemerkung der Autoren, dass dieser Unterschied nicht auf einer erhöhten Responsivität der Patienten auf die Antidepressiva, vielmehr auf eine Abnahme der Plazebo-Response zurückzuführen ist, ist nicht, wie es wohl die Autoren verstehen, ein Zeichen für die mangelnde Wirksamkeit dieser Substanzen; vielmehr zeigt dieses Phänomen eher, dass die Antidepressiva auch bei niedrigeren Depressionsgraden wirken, sich dort aber aufgrund der höheren Plazebo-Response nicht „durchsetzen" können, was genau dann bei höheren Depressionsgraden der Fall ist [8]. Ein probates Mittel, um all diese Diskussionspunkte adäquat zu berücksichtigen, hätte darin bestanden, neben dieser allumfassenden Metaanalyse weitere Metaanalysen mit gestuften Selektionskriterien (z. B. Ausschluss aller „failed studies", Ausschluss aller Patientendaten, bei denen insuffiziente Dosen angewandt wurden, Ausschluss von Studien mit fixer Dosierung) durchzuführen.

Ein weiterer Weg wäre gewesen, Meta-Regressionsanalysen durchzuführen, die es erlauben, Einflussfaktoren auf die Ergebnisse von Metaanalysen zu erfassen. Ein verwandtes Phänomen taucht auch bei der Metaanalyse im Bereich der Antipsychotika von Leucht et al. [5] auf. Die Autoren schlussfolgern nach der Analyse von 150 RCT mit insgesamt 21 533 Teilnehmern, dass vier atypische Antipsychotika besser seien als Typika, nämlich Amisulprid, Clozapin, Olanzapin und Risperidon (**Abb. 2**). Die Autoren legen unter anderem dar, dass bei Studien mit fixen Dosen nur jene mit optimalen Dosen für die Atypika aus-

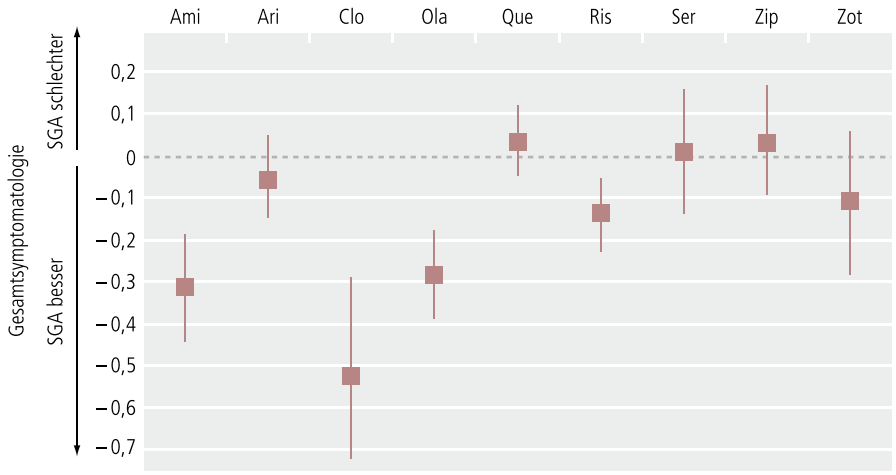

Abb. 2. Wirksamkeit der Atypika versus Typika, Ergebnisse der Metaanalyse von Leucht et al. [5]. Dargestellt sind Hedges' g (ein spezielles statistisches Maß für Effektgrößen) mit dem dazugehörigen 95%-Konfidenzintervall. Die Unterschiede sind dann statistisch signifikant, wenn dieses Konfidenzintervall 0 nicht einschließt; in der grafischen Darstellung bedeutet dies, dass die vertikale Linie die horizontale Nulllinie nicht schneidet (modifiziert nach [5]).
Ami: Amisulprid; Ari: Aripiprazol; Clo: Clozapin; Ola: Olanzapin; Que: Quetiapin; Ris: Risperidon; Ser: Sertindol; Zip: Ziprasidon; Zot: Zotepin; SGA: Second generation antipsychotic (Zweitgenerationsantipsychotikum)

gewählt wurden, und spezifizieren dies wie folgt: Amisulprid 50 bis 300 mg/Tag bei vorrangiger Negativsymptomatik und 400 bis 800 mg/Tag bei vorrangiger Positivsymptomatik, Aripiprazol 10 bis 30 mg/Tag, Olanzapin 10 bis 20 mg/Tag, Quetiapin >250 mg/Tag, Risperidon 4 bis 6 mg/Tag und Ziprasidon 120 bis 160 mg/Tag. Dieser Ansatz ist prinzipiell zu begrüßen und stellt – im Vergleich etwa zu Kirsch et al. [4] (siehe oben) – einen besseren Weg dar, zu einer klinisch relevanten Aussage zu gelangen. Über die durchschnittlichen Dosierungen bei den „flexible-dose trials" finden sich allerdings keine Informationen. Es ist unschwer zu erkennen, dass beispielsweise Olanzapin oder Risperidon relativ hoch dosiert sind, Quetiapin wohl eher niedrig; die Höchstdosis und die am häufigsten verwandte Dosis ist nicht angegeben. Wenn dies so ist, ist ein metaanalytischer Vergleich nur eingeschränkt interpretierbar, da er mehr über die Wirksamkeit einer niedrigen Dosis des einen versus einer eher höheren Dosierung des anderen Antipsychotikums aussagt als über die vergleichende Effektivität der untersuchten Antipsychotika per se. Den Autoren scheint diese Einschränkung in der Aussagekraft selbst bewusst gewesen zu sein, denn sie merken zu der niedrigen Dosisschwelle von Quetiapin an, dass die Effektivität dieser Substanz in der durchgeführten Metaanalyse bei einer höheren Schweller geringer ausgefallen wäre, da in der einzig relevanten Studie 750 mg/Tag die am wenigsten effektive Dosis gewesen sei.

Um eine solche eingeschränkte Aussagekraft einer Metaanalyse zu erkennen, ist allerdings – neben der Kenntnis der Einzelstudien – eine sorgfältige Würdigung der Metaanalyse notwendig. Ein schnelles Überfliegen des Abstracts reicht nicht, auch müssen in die Interpretation weitergehende pharmakologische Gesichtspunkte einfließen.

Die Wichtigkeit des Einflusses der Studienselektion soll an einem anderen Beispiel nochmals verdeutlicht werden, nämlich anhand zweier Metaanalysen zum Vergleich Venlafaxin – SSRI bei der Behandlung von Depressionen. Bauer et al. [1] schlossen alle RCTs, die zu dieser Frage bis zum April 2007 publiziert wurden, ein (folgende Datenbanken wurden für die Recherche benutzt: Medline, Embase, the Cochrane Library, persönliche Datenbasis der Autoren). Auch alle verfügbaren unveröffentlichten Studien wurden eingeschlossen, hierzu wurde mit der Herstellerfirma von Venlafaxin, Wyeth, zusammengearbeitet, die diese Daten nach intensiver Suche in den Firmenarchiven zur Verfügung stellte.

Weinmann et al. [12] untersuchten die gleiche Fragestellung, allerdings wurde eine Reihe von Auswahlschritten der dann eingeschlossenen Studien angewendet. So wurden nur RCTs mit einer Mindestbehandlungsdauer von sechs Wochen und einer Höchstbehandlungsdauer von sechs Monaten eingeschlossen. Studien, in denen mehr als 20 % der Patienten eine Dysthymie-Diagnose hatten, wurden ebenso ausgeschlossen wie jene, in denen mehr als 15 % der Patienten eine bipolare Depression aufwiesen. Die Ratings für das primäre Zielkriterium mussten mit der Hamilton-Depressionsskala oder der Montgomery-Åsberg-Depressionsskala erfolgt sein; Studien, die hierfür Clinical Global Impression (CGI) oder Patient Global Improvement (PGI) verwendeten, wurden nicht berücksichtigt.

Letztendlich wurden in die Metaanalyse von Bauer et al. 63 Studien eingeschlossen [1], in jene von Weinmann et al. 17 [12]. Bauer et al. fanden eine leichte Überlegenheit von Venlafaxin gegenüber den SSRI (Odds-Ratio für Response: 1,15 [95%-KI 1,02–1,29], für Remission: 1,19 [95%-KI 1,06–1,34]) [1]. Weinmann et al. fanden keine Überlegenheit von Venlafaxin über die SSRI für die Remissionsraten (Risk-Ratio 1,07 [0,99; 1,15]) und nur einen kleinen Unterschied für die Responseraten (Risk-Ratio 1,06 [1,01; 1,12]) (siehe **Abb. 3a** und **b**) [12].

Es ist zu mutmaßen, dass durch das weite Einschlusskriterium von Bauer et al. [1] eine Reihe „lebensnaher", die klinische Realität verhältnismäßig gut widerspiegelnde Studien eingeschlossen wurden, die aufgrund der rigorosen Anforderungen von Weinmann et al. keine Berücksichtigung fanden. Reine Zulassungsstudien können die Kriterien von Weinmann et al. leicht überwinden, andere Untersuchungen schwerer oder gar nicht. Insofern können die unterschiedlichen Ergebnisse dieser beiden Metaanalysen, ausgelöst durch unterschiedliche Selektionskriterien der aufgenommenen Studien, kaum überraschen.

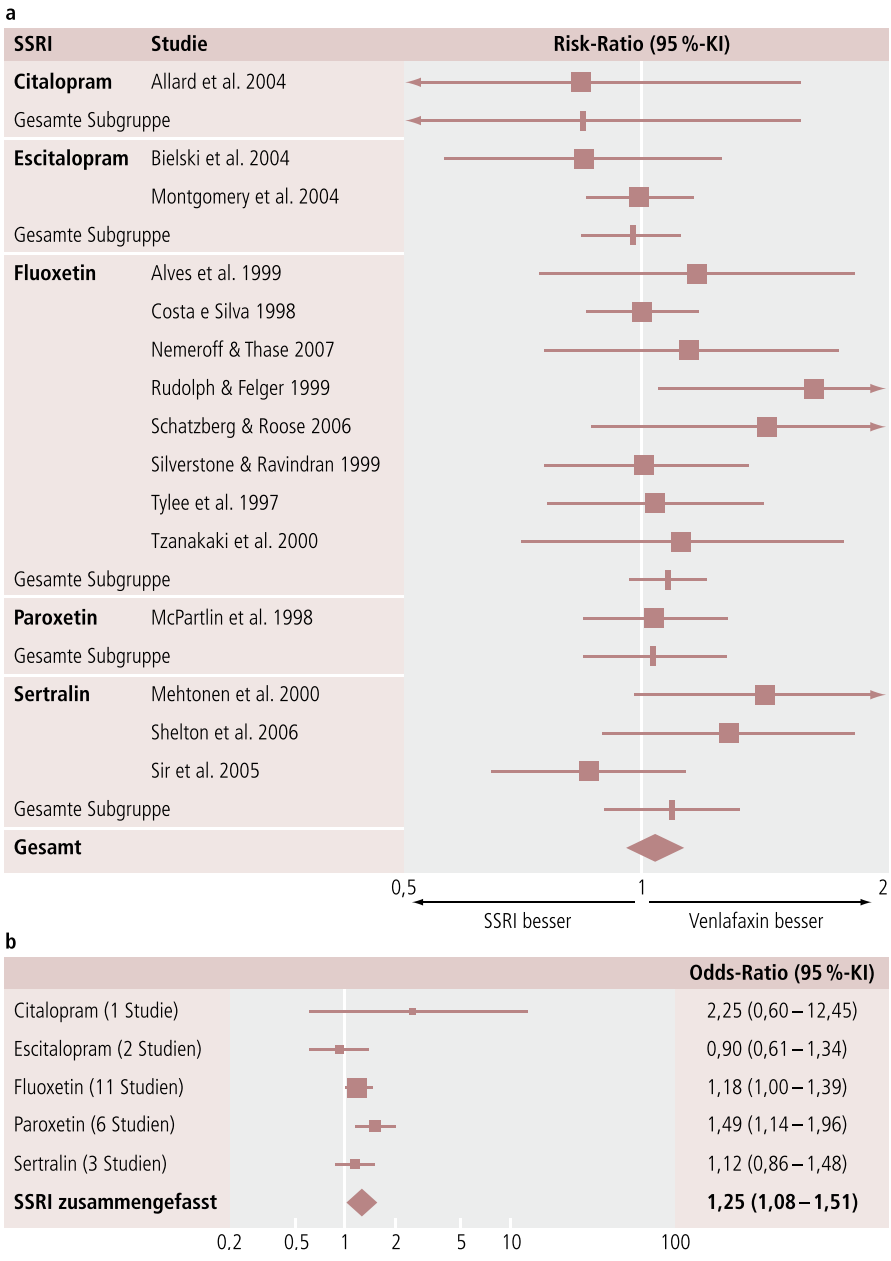

Abb. 3. Verschiedene metaanalytische Ansätze zum Vergleich des selektiven Serotonin-Noradrenalin-Wiederaufnahmehemmers Venlafaxin mit anderen Antidepressiva.
a) Mantel-Haenszel Risk-Ratio für Remission, Fixed-Effects-Model (nach [12]). Es wurde kein statistisch signifikanter Unterschied zwischen Venlafaxin und den selektiven Serotonin-Wiederaufnahmehemmern (SSRI) gefunden.
b) Venlafaxin versus aktive Komparatoren: Remission (nach [1]; nur SSRI-Ergebnisse)
95%-KI: 95%-Konfidenzintervall

Multi-Treatment-Comparisons

Eine neue Art von Metaanalysen sind sogenannte Network- oder Multi-Treatment-Comparisons (MTC). Das Prinzip dieser Art von Metaanalysen soll anhand der Metaanalyse von Cipriani et al. [2] zur Akutwirksamkeit moderner Antidepressiva kurz geschildert werden, anschließend soll ein Hauptkritikpunkt an dieser Art von Analysen umrissen werden. Cipriani et al. [2] schlossen 117 RCTs, in denen die Wirksamkeit und Verträglichkeit von acht modernen Antidepressiva untersucht wurde, mit insgesamt 25 928 Patienten ein. Das Hauptwirksamkeitskriterium war die Responserate. Nun wurde nicht jedes der hier untersuchten Antidepressiva tatsächlich gegen jedes andere Antidepressivum geprüft. Um dies an einem Beispiel klarzumachen: Zu Escitalopram gingen Vergleichsstudien versus Duloxetin, Paroxetin, Sertralin, Citalopram, Venlafaxin, Fluoxetin und Bupropion in die Untersuchung ein, die Substanz wurde nie im Rahmen eines RCTs gegen Milnacipran, Reboxetin, Mirtazapin und Fluvoxamin geprüft (siehe **Abb. 4**). Da aber Vergleichsdaten beispielsweise von Fluoxetin zu Milnacipran, Reboxetin, Mirtazapin und Fluvoxamin vorliegen, konnten die Autoren mithilfe dieser Auswertung diese „Lücken" schließen und, vereinfacht formuliert über die Verwendung von Fluoxetin als verbindendem Standard auch vergleichende Effektivitäten von Escitalopram versus Milnacipran, Reboxetin, Mirtazapin und Fluvoxamin berechnen. Das heißt, es werden in dieser Arbeit vergleichende Wirksamkeiten mit statistischen Maßen von Vergleichen, die

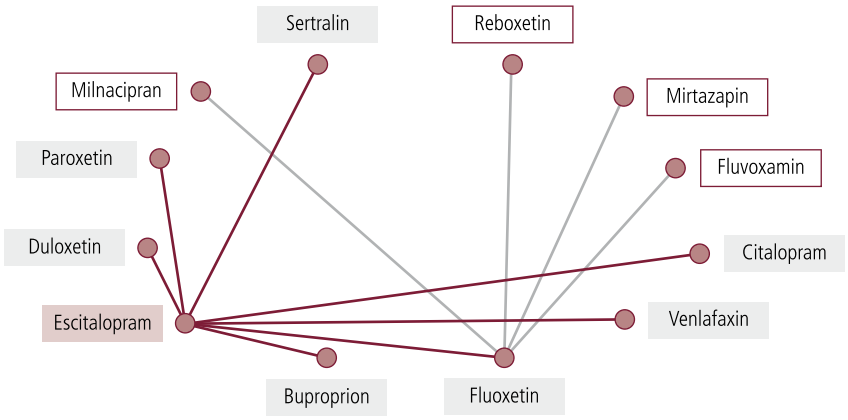

Abb. 4. Grafische Darstellung der direkten Vergleichsstudien von Escitalopram versus andere Antidepressiva. Wie zu erkennen ist, wurden zu vier Antidepressiva keine direkten Vergleiche durchgeführt (Milnacipran, Reboxetin, Mirtazapin, Fluvoxamin). In der Metaanalyse von Cipriani et al. [2] werden jedoch auch Vergleichsberechnungen zu diesen Antidepressiva durchgeführt und dargestellt, beispielsweise aufgrund von Vergleichsstudien mit Fluoxetin.

in der Realität nie stattgefunden haben, angegeben.

Ich möchte die Problematik dieser Art von Metaanalysen mit einem banalen Beispiel aus dem Fußball verdeutlichen: Argentinien spielt gegen England 1 : 2, England gegen Deutschland 0 : 2, Deutschland hat gegen Argentinien nicht direkt gespielt. In der hier vorgestellten Analyseart würde berechnet werden, dass Deutschland besser als Argentinien ist.

Diese Art der Berechnung kann zu Verzerrungen führen, wie wiederum am Beispiel Escitalopram gezeigt werden kann. So wird in der Cipriani-Analyse [2] zwischen Escitalopram und Citalopram kein statistisch signifikanter Wirksamkeitsunterschied gefunden, wobei dies sowohl in einer Reihe von Einzelstudien und auch von Metaanalysen, die die Studien mit direkten Vergleichen einschließen, hochsignifikant der Fall ist (vergleiche [11]). Das heißt, würde man nur diese beeindruckende Multiple-Comparison-Metaanalyse lesen, entginge einem diese wesentliche Information. Die Autoren gehen auf diesen und auch andere verzerrende Effekte in ihrer Diskussion allerdings nicht ein. Severinger und Kasper [10] haben diese Publikation wie folgt kritisiert: "Overall, this is a problematic publication with partly irreproducible conclusions that might lead clinicians or health authorities to false judgements."

Fazit

Metaanalysen setzen voraus, dass zu einer bestimmten Frage eine Reihe von möglichst homogenen Studien, die dann in die Metaanalyse eingeschlossen werden können, existiert. Dies ist bei der Frage, ob ein Antidepressivum oder ein Antipsychotikum in der Akut- oder auch Langzeittherapie wirkt, bereits nicht der Fall. Betrachtet man nun andere pharmakologische Strategien (z. B. wie ist zu verfahren, wenn ein Patient schon mit einem Pharmakon mit nur ungenügenden Erfolg vorbehandelt worden ist?), sind oft nur wenige RCTs vorhanden, mitunter gar keine. Insofern sind auch keine Metaanalysen erstellbar. Dies kann nun aber im Rahmen des Prozesses der Erstellung von Leitlinien dazu führen, dass für bestimmte Strategien keine Evidenz angegeben werden kann, diese Strategien dann auch nicht empfohlen werden können. Hierbei ist aber zu beachten, dass nicht jede klinische Fragestellung mit einer Reihe von RCTs, die dann metaanalysiert werden können, bearbeitet werden kann. Dieser Punkt hängt nur indirekt mit der im Zentrum dieses Artikels stehenden Problematik der Metaanalysen zusammen, er zeigt aber, wie schnell eine Schieflage in den Empfehlungen zu Interventionen entstehen kann, wenn unkritisch auf das Ergebnis von Metaanalysen fokussiert wird.

Meta-analyses – important points to consider

Meta-analyses are very important instruments to evaluate the effect of therapeutical interventions, before all regarding evidence based medicine (EbM). This priority is problematic, especially if the inherent methodological problems of meta-analyses are not sufficiently taken into account. In contrast, the evaluation of single trials seems to be nowadays relatively unimportant. However, the correct interpretation of meta-analyses is strongly dependent on methodological issues, if those issues are not properly included into the interpretation, false conclusions might be drawn. The present paper focuses on two special points: selection criteria and multi-treatment comparisons. Taken together, results from meta-analyses must be weighted very carefully.

Key words: Meta-analyses, selection criteria, multi-treatment-comparisons

Literatur

1. Bauer M, Tharmanathan P, Volz HP, Möller HJ, et al. The effect of venlafaxine compared with other antidepressants and placebo in the treatment of major depression. Eur Arch Psychiatry Clin Neurosci 2009;259:172–85.
2. Cipriani A, Furukawa TA, Geddes JR, Higgins JPT, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet January 29,2009;DOI:10.1016/S0140-6736(09)60046-5
3. Feinstein AR. Meta-analysis. Statistical alchemy for the 21st century. J Clin Epidemiol 1995;48:71–9.
4. Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, et al. Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. PLoS Medicine 2008;5:e45.
5. Leucht S, Corves C, Arbter D, et al. Second-generation versus first-generation antipsychotic drugs for schizophrenia: a meta-analysis. Lancet 2009;373:31–41.
6. Maier W, Möller HJ. Metaanalysen. Methode zur Evidenzmaximierung von Therapiestudien? Nervenarzt 2007;78:1028–36.
7. Maier W, Möller HJ. Meta-analyses: a method to maximise the evidence from clinical studies? Eur Arch Psychiatry Clin Neurosci 2010;260:17–23.
8. Möller HJ. Isn't the efficacy of antidepressants clinically relevant? A critical comment on the results of the metaanalysis by Kirsch et al. Eur Arch Psychiatry Clin Neurosci 2008;258:451–5.
9. Möller HJ, Maier W. Evidence-based medicine in psychopharmacotherapy: possibilities, problems and limitations. Eur Arch Psychiatry Clin Neurosci 2010;260:25–39.
10. Seyeringer ME, Kasper S. Ranking antidepressants. Lancet 2009;373:1760–1.
11. Volz HP. Wirksamkeitsunterschiede zwischen Escitalopram und Citalopram. Eine systematische Übersicht. Psychopharmakotherapie. Im Druck.
12. Weinmann S, Becker T, Koesters M. Re-evaluation of the efficacy and tolerability of venlafaxine vs. SSRI: meta-analysis. Psychopharmacology 2007;196:511–20.